# Técnicas de Aprendizado de Máquina Aplicadas na Previsão de Desempenho de Operadores de Centros de Teleatendimento


Evandro Lopes de Oliveira

Faculdade de Ciências e Tecnologia - Ciências da Informação - Sistemas, Tecnologias e Gestão da Informação

Universidade Fernando Pessoa

Tese submetida ao grau de

*Doutor*

2021

# Resumo

Automatizar e melhorar a gestão e o desempenho profissional dos trabalhadores é um grande desafio enfrentado por várias empresas, sobretudo as que possuem um elevado número de funcionários. Portanto, as ferramentas de previsão de desempenho podem fornecer mecanismos especialmente importantes para o planejamento e gestão de recursos humanos das instituições intensamente dependentes do trabalho dos seus colaboradores. Esta tese foca o uso de tecnologias de aprendizado de máquina integradas num *pipeline* dinâmico que contempla a manipulação e seleção dos dados de entrada e as parametrizações dos algoritmos utilizados para otimizar a previsão do desempenho dos trabalhadores no serviço de teleatendimento.

O trabalho de previsão de desempenho, definido neste trabalho pelo absenteísmo e produtividade, foi desenvolvido para uma população-alvo pertencente a uma grande empresa de prestação de serviços de teleatendimento brasileira. As variáveis foram extraídas do perfil dos agentes de teleatendimento e, em seguida, filtradas por processos de correlação e seleção de variáveis. Mais precisamente, neste trabalho, foram extraídas características pessoais, sociais e profissionais de teleatendentes para prever o desempenho de uma população de, aproximadamente, 10,5 mil funcionários.

Foram testados alguns modelos de previsão, para os quais um conjunto vasto de variáveis de entrada foram dinamicamente selecionadas e submetidas, permitindo assim comparar o desempenho obtido pelos vários algoritmos de aprendizado máquina utilizados (cf. *LR - Logistic Regression*, *LSTM - Long Short Term Memory*, *MLP - Multilayer Perceptron*, *NB - Naive Bayes*, *RF - Random Forest*, *SVM - Suport Vector Machine* e *XGBoost - Extreme Gradient Boosting*). A hiper-parametrização destes modelos de aprendizado de máquina também foi considerada na comparação dos algoritmos mais adequados para o problema de previsão. As hiperparametrizações, assim como a seleção de variáveis, foram ajustadas através do uso de um algoritmo evolutivo, permitindo melhorar os resultados de previsão que globalmente foram bastante promissores.

O conjunto das técnicas aplicadas no trabalho permitiu melhorar o entendimento sobre o problema da previsão de desempenho da empresa. Foi, assim, possível desenvolver um estudo intensivo de aplicação dos vários algoritmos de aprendizado máquina à previsão de desempenho. Este estudo foi suportado por mecanismos dinâmicos de seleção de variáveis de entrada mais significativas, bem como de técnicas de hiper-parametrização dos algoritmos, que melhor resultados de previsão produziam, de modo a selecionar os algoritmos de aprendizagem máquina mais adequados para o efeito.

# Abstract

Automating and improving the management and professional performance of workers is a major challenge faced by several companies, especially those with a high number of employees. Therefore, performance forecasting tools can provide important mechanisms for the planning and management of human resources of institutions that are intensely dependent on the work of their employees. This thesis focuses on the use of machine learning technologies integrated in a dynamic pipeline that contemplates the manipulation and selection of input features, and the parameterization of algorithms used to optimize the prediction of workers' performance in the call center service.

The forecasting of work performance, defined in this work by absenteeism and productivity, was developed for a target population belonging to a large Brazilian call center service company. The input features were extracted from the profile of the call center agents and then filtered through processes of correlation and selection of variables. More precisely, in this work, personal, social and professional characteristics of call center agents were extracted to predict the performance of a population of approximately 10,500 employees.

Some forecasting models were tested, for which a vast set of input features was dynamically selected and submitted, allowing to compare the performance obtained by the various machine learning algorithms used (cf. *LR - Logistic Regression*, *LSTM - Long Short Term Memory*, *MLP - Multilayer Perceptron*, *NB - Naive Bayes*, *RF - Random Forest*, *SVM - Suport Vector Machine* e *XGBoost - Extreme Gradient Boosting*). The hyper-parametrization of these machine learning models was also considered when comparing the most suitable algorithms for the forecasting problem. The hyper-parameterizations, as well as the selection of variables, were adjusted through the use of an evolutionary algorithm, allowing to improve the forecast results that were very promising.

The set of techniques applied in this work improved the understanding about the forecasting problem for the company workers performance. It was possible to develop an intensive study through the application of various machine learning algorithms for performance prediction. This study was also supported by dynamic mechanisms for selecting the most significant input features, as well as the best algorithms' hyper-parameterization, which produced better forecasting results, in order to select the most suited machine learning algorithms for the job.

Aos meus pais, pela compreensão, atenção, ajuda, e por superarem as dificuldades, acreditando, sempre, no crescimento profissional dos seus filhos.

# Agradecimentos

Agradeço a Deus por minha vida e meu sucesso. Agradeço, também, a Ele por me fazer aprender e poder, com os meus fracassos, celebrar minhas vitórias e jamais desistir dos meus objetivos. Obrigado ao Professor Dr. José Manuel de Castro Torres e ao coorientador Professor Dr. Rui Silva Moreira pela excelente orientação durante o desenvolvimento deste trabalho. Ao povo de Portugal, por me acolher e fazer-me sentir tão bem quanto no meu país de origem, o Brasil. Meus agradecimentos à empresa AeC/Robbyson, por participar ativamente neste trabalho fornecendo os dados para que esta tese se concretizasse. Muito obrigado à assistência dada pelos profissionais do grupo de I&D ISUS *(Intelligent Sensing and Ubiquitous Systems)* da Faculdade de Ciências e Tecnologia (FCT) da Universidade Fernando Pessoa (UFP).

# Índice

# Índice de Figuras

# Índice de Tabelas

# Acrônimos

**AG** Algoritmo Genético

**AR** *Auto-Regression*

**ARCH** *Autoregressive Conditional Heterocedasticity*

**ARMA** *Autoregressive Moving Average*

**ARIMA** *Autoregressive Integrated Moving Average*

**BR** *Beta Regression*

**DT** *Decision Tree*

**LR** *Logistic Regression*

**LSTM** *Long Short Term Memory*

**LSVM** *Linear Support Vector Machine*

**MA** *Moving Average*

**ML** *Machine Learning*

**MLP** *Multilayer Perceptron*

**MSE** *Mean Square Error*

**NB** *Naive Bayes*

**NCV** *Nested Cross-Validation*

**NLP** *Natural Language Processing*

**PA** Pesquisa Aleatória

**RBF** *Radial Basis Function*

**RF** *Random Forest*

**RNA** Redes Neurais Artificiais

**RND** *Range Normalized Difference*

**RNN** *Recurrent Neural Network*

**ROC AUC** *Area Under the ROC Curve*

**SVM** *Support Vector Machine*

**XGBoost** *Extreme Gradient Boosting*

# Capítulo 1

# Introdução

## 1.1 Contexto do trabalho e problemas de investigação

O setor de teleatendimento se destaca no mercado econômico por oferecer serviços de atendimento primários entre os clientes e as suas empresas fornecedoras. Esses serviços empregam milhares de profissionais em todo o mundo (Silva, 2007). Um dos principais problemas enfrentados pelas empresas de teleatendimento é a gestão do desempenho dos seus colaboradores.

As empresas de teleatendimento procuram melhorar a sua gestão econômica através de estudos sobre o desempenho de seus funcionários. Contudo, entender as reais causas que influenciam os seus profissionais é um trabalho complexo, visto que, dependendo do número de características disponíveis dos funcionários, é difícil traçar um perfil de desempenho (Ávila Assunção and de Oliveira Vilela, 2003). Considera-se que este é um problema com potencialidade para ser abordado através da utilização de ferramentas de aprendizado de máquina. Neste trabalho, define-se e aplica-se um *pipeline* de previsão que utiliza vários algoritmos de aprendizado combinados com o ajuste da hiperparametrização e a seleção dinâmica de variáveis de entrada dos modelos, com o objetivo de comparação para obtenção da melhor previsão para o problema de produtividade.

Entende-se por variáveis de entrada dos modelos de classificação, todas as características do contexto de trabalho desses funcionários que possam ter relação direta ou indireta com os indicadores de desempenho profissionais. Na empresa alvo deste trabalho, consideram-se várias variáveis que caracterizam os colaboradores e definem ou condicionam em certa medida a sua performance de trabalho no setor de teleatendimento. Assim, neste trabalho e de acordo com a empresa onde ele se insere, o desempenho dos funcionários é entendido por um conjunto de indicadores. Mais concretamente, nesta tese, o desempenho dos colaboradores é medido por dois indicadores específicos: o absenteísmo e a produtividade.

Pensou-se então desenvolver um *pipeline* de previsão dos indicadores de desempenho

utilizados na empresa. Para isso, seria necessário coligir um conjunto vasto de variáveis disponíveis no perfil dos funcionários da empresa (ver Anexo A). Essas variáveis seriam posteriormente transformadas através de processos de discretização e binarização, adequando-as ao problema e modelos utilizados. O pré-processamento dos dados de entrada dos modelos deveria considerar ainda a criação e organização de variáveis em função da periodicidade e do tempo, utilizando técnicas de validação cruzada (Cochrane, 2018a).

Seria necessário também identificar as principais técnicas descritas na literatura para medir a relação entre diferentes variáveis, uma vez que nem todas teriam a mesma preponderância na previsão dos diferentes indicadores de desempenho. Estas técnicas poderiam ser úteis na ordenação e seleção de variáveis, antes mesmo dessas variáveis serem submetidas aos modelos no referido *pipeline* de previsão. Por exemplo, a seleção poderia basear-se no fator de correlação de *Pearson* ou na relevância ponderada *Relief* com as saídas que definem o desempenho do funcionário (cf. produtividade e absenteísmo). Além de avaliar essa correlação, seria útil avaliar os melhores resultados de predição conforme a combinação dessas variáveis. Essa avaliação poderia contemplar os resultados de predição em função das combinações de variáveis de entrada dos modelos. Algumas das técnicas utilizadas, poderiam contemplar algoritmos de *backward*, *forward*, ou mesmo a utilização de algoritmos genéticos (AG).

Para além do estudo das melhores combinações de variáveis de entrada dos modelos de previsão, considerou-se ainda importante estudar a hiper-parametrização dos modelos de aprendizado máquina. O objetivo seria encontrar a configuração mais adequada para a previsão de absenteísmo e para a previsão de produtividade. Foi equacionada a utilização de um algoritmo genético (AG), considerando todas as variáveis de entrada, como uma estratégia a explorar para explorar o *tuning* dos hiperparâmetros dos modelos.

Os modelos de previsão a incorporar no *pipeline* para serem avaliados neste trabalho deveriam basear-se na literatura, incluindo, por isso, algoritmos mais tradicionais, baseados em árvores de decisão e redes neurais artificiais, entre outros (e.g. LR - *Logistic Regression*, MLP - *Multilayer Perceptron*, RF - *Random Forest*, SVM - *Suport Vector Machine* e XGBoost - *Extreme Gradient Boosting*), mas também soluções mais recentes baseadas em redes neurais recorrentes (e.g. LSTM - *Long Short Term Memory*). O objetivo final seria selecionar o modelo com melhor performance, quer em termos de desempenho temporal quer em termos de acurácia.

Pretende-se por isso explorar neste trabalho uma arquitetura para um *pipeline* dinâmico ajustado a resolução de dois problemas distintos que atendem ao mesmo objetivo, ou seja, prever separadamente os indicadores de absenteísmo e produtividade. Ambos os indicadores são usados na empresa hospedeira deste trabalho na medição do desempenho dos colaboradores.

O absenteísmo é um indicador importante, pois uma taxa elevada afeta negativamente

as empresas. Esse indicador pode representar a falta de comprometimento e engajamento dos funcionários, mas também destaca a necessidade de a empresa promover medidas contra esse problema (Cohen and Golan, 2007). Geralmente, as empresas que se preocupam com capital humano desejam encontrar os motivos que levam a ausências ou atrasos de seus funcionários. Tais ausências ou atrasos podem variar de acordo com problemas pessoais, questões médicas, insatisfação com as condições de trabalho ou contexto de vida. Compreender as causas reais que levam os profissionais a se tornarem ausentes é uma tarefa complexa que depende de um grande número de variáveis dos funcionários.

Do mesmo modo, encontrar o perfil de produtividade também é uma tarefa complexa que depende de um grande número de variáveis. A possibilidade de os operadores de teleatendimento terem sucesso, ou não, nas suas funções profissionais é um tipo de informação importante para empresa na tomada de decisões relativas à sua gestão e à promoção da produtividade. Assim como no absenteísmo, as motivações que definem a produtividade podem variar de acordo com problemas pessoais, questões médicas, insatisfação com as condições de trabalho ou contexto de vida.

A produtividade dos colaboradores é um indicador de performance baseado em resultados bonificados. Esses valores bônus são pontos adquiridos conforme os resultados positivos no trabalho. De acordo com a performance em outros indicadores, os colaboradores são pontuados e subdivididos em grupos de produtividade.

A avaliação da previsão dos dois indicadores de performance (cf. produtividade e absenteísmo), que definem o desempenho dos colaboradores, seria aplicada nas unidades de atendimento por telefone da empresa alvo da pesquisa, considerando-se dados recolhidos pela organização num universo de aproximadamente 10.550 colaboradores (todos colaboradores operadores de teleatendimento). Os dados seriam recolhidos na empresa, no decurso normal do seu funcionamento interno, cumprindo os preceitos legais, éticos e respectivo consentimento informado dos seus colaboradores.

## 1.2 Contribuições do trabalho

Uma das principais características dos centros de contato de teleatendimento é a competitividade entre empresas desse ramo, no qual a eficiência e a qualidade de atendimento ao cliente é que determinam a sua permanência no mercado. Procurar soluções que atendam aos requisitos de eficiência e qualidade, prevendo cenários futuros para melhorar a produtividade, é um desafio a ser alcançado através de técnicas de previsão de desempenho laboral por aprendizado de máquina. As contribuições do trabalho na superação desse desafio estão listadas a seguir:

- Analisar e entender as peculiaridades do problema relacionado à previsão de desempenho dos colaboradores da empresa de teleatendimento;

- Realizar um estudo da literatura sobre a utilização de algoritmos de aprendizado máquina na previsão de desempenho e propôr um pipeline genérico para a previsão de absenteísmo e produtividade;
- Propor, implementar e avaliar mecanismos dinâmicos e automáticos para selecionar as variáveis de entrada nos algoritmos de aprendizado máquina;
- Propor, implementar e avaliar mecanismos dinâmicos e automáticos para selecionar a hiper-parametrização dos algoritmos de aprendizado máquina;
- Estudar e selecionar um conjunto de algoritmos de aprendizado máquina aplicados aos dois problemas de previsão de desempenho propostos, nomeadamente, previsão de absenteísmo de produtividade;
- Avaliar e comparar os vários algoritmos de aprendizado máquina utilizados no *pipeline* proposto que apresentem melhores resultados na previsão de absenteísmo e de produtividade.

O projeto pretende legar um *pipeline* que utiliza a combinação de técnicas de aprendizado máquina (aka *Machine Learning* - ML) para fazer a previsão de desempenho dos funcionários, fornecendo uma ferramenta para o planejamento administrativo dos centros de teleatendimento.

Este trabalho se propõe a avaliar vários modelos de classificação e regressão baseados em técnicas de ML aplicadas à previsão do desempenho. Sabendo que esses modelos são sensíveis à correlação dos recursos de entrada com a variável de resposta, bem como das hiper-parametrizações adotadas, pretende-se ainda avaliar o impacto, tanto da seleção das variáveis de entrada como da hiper-parametrização dos referidos modelos. Pretende-se ainda estudar os resultados quando combinados estrategicamente.

Sabendo que existem outros trabalhos prevendo o desempenho dos funcionários, com resultados interessantes, pretende-se neste estudo contribuir e destacar a definição, implementação e avaliação da arquitetura dinâmica do *pipeline* de previsão de desempenho. Adicionalmente, outros aspetos importantes prendem-se com a natureza do contexto empresarial onde o trabalho se enquadra, com a natureza, o volume e as características dos dados recolhidos, bem como pelas técnicas exploradas para abordar o problema de previsão.

Consideramos ainda que outros estudos acadêmicos não apresentam uma seleção tão completa e diversificada de variáveis de entrada, nem uma quantidade tão expressiva de colaboradores de teleatendimento, como neste projeto. Este estudo combina a seleção dinâmica das melhores características dos funcionários de teleatendimento, com uma escolha cuidadosa e sistemática da hiper-parametrização dos algoritmos de previsão, para assim conseguir obter os melhores resultados comparativos.

## 1.3 Metodologia e abordagem aos problemas de previsão de desempenho

Para o presente trabalho, os dados utilizados foram selecionados conforme a similaridade com desempenho profissional, que são o absenteísmo e a produtividade dos funcionários de teleatendimento. Quando foram analisados os conjuntos de informações das semanas, foi possível considerar os pontos de sazonalidade do consolidado mensal e anual.

Quando comparadas a modelos ingênuos, algumas técnicas de aprendizado de máquina apresentam melhora no desempenho (Caetano, 2016). Sabendo disso, pretende-se avaliar algumas alternativas como a mineração de dados. Os métodos usados seguem a arquitetura da Figura 1.1.

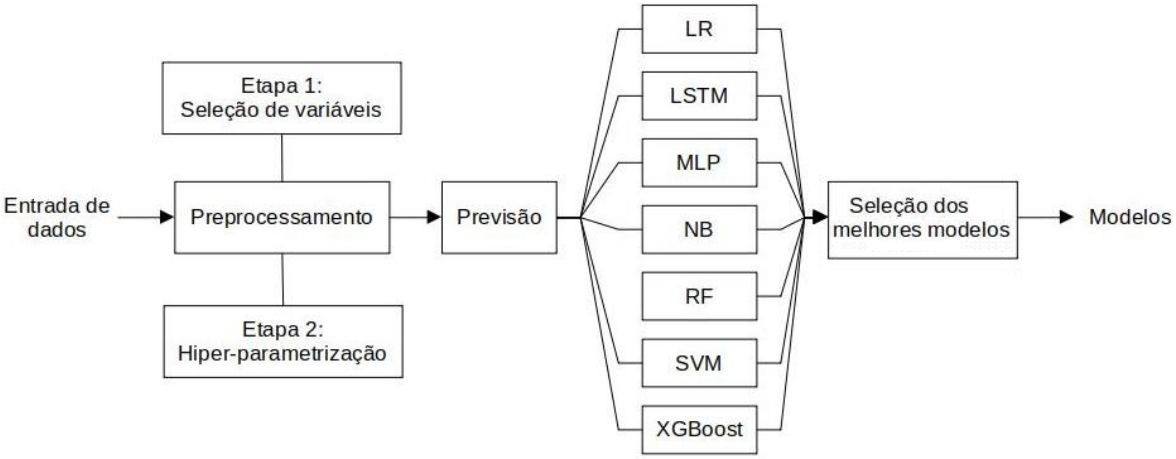


Figura 1.1: Esquema genérico do *pipeline* de previsão dos indicadores de desempenho

As entradas são representadas pela relação das variáveis de cada operador em um período de uma semana. O período considerado como entrada é de uma semana corrente para que a próxima semana seja prevista.

No pré-processamento, as variáveis dos operadores são selecionadas para cada modelo de previsão. Essas variáveis são ordenadas por correlação com a variável de saída e, depois submetidas a um algoritmo genético, com a função de avaliação sendo o próprio modelo de previsão, técnica de *forward* e a técnica de *backward*.

Ainda no pré-processamento, os hiper-parâmetros dos modelos são escolhidos usando pesquisa aleatória, em um primeiro momento e, então, submetidos a outras técnicas de hiper-parametrização como otimização por algoritmos genéticos.

Tanto para a hiper-parametrização quanto para a seleção de variáveis é usada uma técnica designada por *Nested Cross Validation (NCV)* (Cochrane, 2018b), (Varma and Simon, 2006). Essa técnica possibilita uma avaliação das quatro semanas de forma acumulativa com a média do acerto.

Com os sete modelos hiper-parametrizados e com as entradas selecionadas, inicia-se

a fase de previsões. Nessa fase são escolhidos os melhores modelos para o problema conforme os que apresentam melhores resultados.

Importa realçar que todo o processo de recolha e processamento da informação sobre os funcionários, leva em consideração as regras de legislação, bem como a obtenção de aprovação dos órgãos éticos, quer da empresa, quer da universidade e do país onde se desenvolveu. Adicionalmente, o armazenamento da informação é efetuada em repositórios de dados seguros e com anonimização dos funcionários.

## 1.4 Lista de publicações

Durante o processo de construção do projeto de pesquisa desta tese, três trabalhos científicos foram publicados:

- **Previsão de absenteísmo em centros de teleatendimento usando algoritmos de aprendizado de máquina** (Oliveira et al., 2019b). Nessa publicação acadêmica focou-se na previsão de absenteísmo dos colaboradores usando a estrutura genérica para previsão abordada no contexto desta tese.
- **Técnicas de aprendizado de máquina aplicadas na previsão de desempenho de operadores** (Oliveira et al., 2019a). Nessa publicação é abordado todo o contexto de pre-projeto que está presente com maiores detalhes e completude nesta tese.
- **Técnicas de aprendizado de máquina aplicadas na previsão de produtividade** (Oliveira et al., 2019c). Esta publicação de trabalho científico foi focada na previsão de produtividade dos colaboradores usando a estrutura genérica para previsão abordada no contexto desta tese.

## 1.5 Estrutura da tese

A secção das técnicas de aplicação de aprendizado de máquina (capítulo 2) apresenta os conceitos e a abordagem dos dados quando neles existem a dependência temporal. Aborda-se também nessa secção, técnicas estatísticas de amostragem que envolvem pequenos e grandes volumes de dados, métodos de seleção de variáveis e formas de ajustar os modelos de ML com hiper-parametrização, além da descrição do processo de construção desses modelos.

Na secção de pesquisas fundamentadas (capítulo 3), são apresentados os trabalhos relacionados ao presente projeto que ajudaram na solução do problema pesquisado. Por meio da busca do estado da arte na literatura acadêmica, chegou-se aos algoritmos e análises profundas aqui apresentados.

A secção de contexto e análise do problema de previsão de desempenho (capítulo 4) apresenta as etapas e arquitetura do trabalho. Nessa secção é apresentado o pré-processamento das informações conforme os tipos de atributos e as características do conjunto de dados, a redução de dimensionalidade, a discretização, a binarização, a transformação de variáveis e a divisão do conjunto de dados com o fracionamento da informação para prever o desempenho com os modelos de ML.

Em resultados e análise dos processos (capítulo 5), são apresentados os resultados e é feita a sua interpretação. Essa secção apresenta os resultados obtidos na previsão de absenteísmo e produtividade com objetividade e procura-se identificar os modelos de ML que obtiveram performances mais interessantes.

Na sequência final, na seção de conclusões e trabalhos futuros, são discutidas as etapas do trabalho na perspectiva do autor e são feitas sugestões de trabalhos futuros.

# Capítulo 2

# Técnicas de previsão usando ML

Uma das ideias na implementação do aprendizado de máquina é agrupar metodologias e análises de modelos algorítmicos para representar problemas de negócios. Essa abordagem é feita com objetivo de dar sentido à informação e mostrar caminhos estratégicos.

As técnicas de aprendizado de máquina são análises de dados e implementações que possibilitam a adaptação de algoritmos computacionais. Esses algoritmos reproduzem aquilo que acontece naturalmente com os humanos e com os animais: aprender com a experiência. Os algoritmos de ML usam métodos para assimilar informações diretamente dos dados e melhoram adaptativamente seu desempenho, à medida que o número de amostras disponíveis para aprendizado aumenta e se diversifica.

Neste capítulo são apresentadas estratégias de previsão usando modelos de ML, preparo das amostras, técnicas de otimização com algoritmos evolucionários, seleção de variáveis de entrada e equilíbrio dos hiperparâmetros dos modelos algorítmicos.

## 2.1 Conceitos sobre estratégias de previsão

### 2.1.1 Séries Temporais

Uma série temporal é uma sequência de dados históricos observados em um intervalo de tempo. As séries temporais têm sido estudadas com o propósito de prever o futuro baseando-se no conhecimento do passado (Chatfield, 1996). Esse processo é executado observando fenômenos ou descrevendo as principais características das séries de dados relacionados ao tempo.

A série temporal é a realização de um processo no qual as observações $x_t$ estão ordenadas em intervalos regulares de tempo (cada dia, cada mês, cada ano, etc). Esse processo se caracteriza pela função de distribuição conjunta das variáveis aleatórias $(x_1, x_2, ..x_t)$ para qualquer valor de $t$. (Morettin and Toloi, 1985) (Chatfield, 1996).

O domínio de previsão das séries temporais tem sido influenciado por modelos lineares autorregressivos há vários anos (Bontempi et al., 2013). Esses modelos são em

formato estatísticos lineares e utilizam correlações entre as observações em diversos instantes. A ideia envolve o uso de filtros lineares que identificam a própria estrutura da série automaticamente, uso desses filtros evita análises complexas. O modelo autorregressivo AR e médias móveis MA compõem o modelo ARMA, que é utilizado para séries estacionárias (Box and Pierce, 1970) (Box and Jenkins, 1970), como na Figura 2.1.

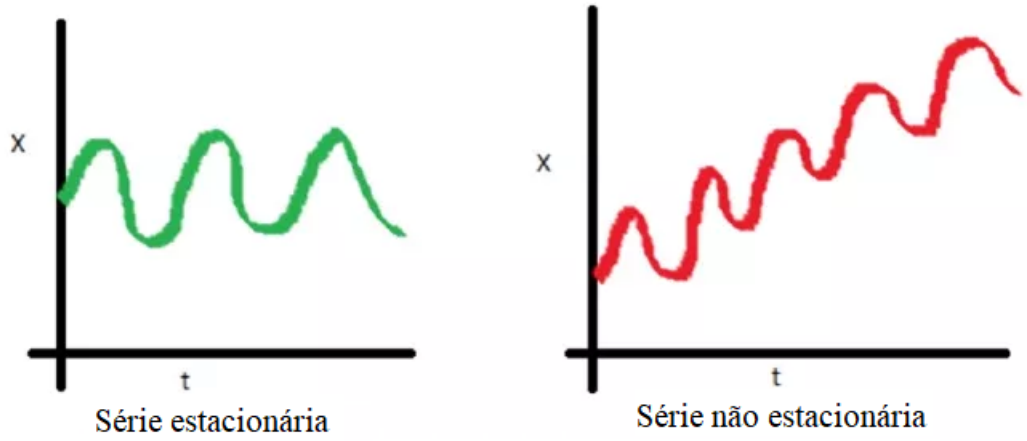


Figura 2.1: Tipos de séries temporais (Srivastava, 2015)

Quando o processo é não estacionário, a série possui tendência, visto gráfico na Figura 2.1. Uma das maneiras de analisá-lo é incorporando diferenças no modelo ARMA. O resultado é o modelo conhecido como ARIMA, modelo autorregressivo integrado de médias móveis (Diniz, 2008).

No final dos anos 1970 e início da década de 1980, observou-se que os modelos lineares não eram adaptados a aplicações reais de forma genérica (Gooijer and Hyndman, 2006) (Chatfield, 1996). No mesmo período, várias séries temporais não lineares foram propostas, como o modelo bilinear, o modelo limiar autorregressivo e o modelo heterocedástico condicional autorregressivo (ARCH). No entanto, o estudo dos modelos não lineares, na análise e previsão de séries temporais, ainda estava no início em comparação com os modelos lineares (Gooijer and Hyndman, 2006).

Quando os estudos dos modelos não lineares já estavam avançados na solução de problemas reais, alguns pesquisadores optaram por esses novos modelos, não paramétricos que usam dados históricos, para aprender a dependência estocástica entre o passado e o futuro nas séries temporais. Dentre essas opções estão os modelos de ML.

### 2.1.2 Técnicas de ML

A pesquisa sobre previsão de séries temporais tem se aperfeiçoado no ML, porque problemas de previsão que envolvem componentes de tempo possuem informações adicionais que tornam esses problemas mais complexos de serem manipulados, dependendo da situação estudada.

No ML, as representações internas são feitas através de dois métodos de treinamento: supervisionado e não supervisionado, como na Figura 2.2. No algoritmo supervisionado, a solução ótima desejada deve ser especificada a priori, com o processo de aprendizado montado a partir de valores aleatórios modificados iterativamente até a obtenção da solução ótima. O algoritmo não supervisionado não requer a especificação da solução ótima, permitindo que os dados determinem seu padrão de comportamento (Zirilli, 1996).

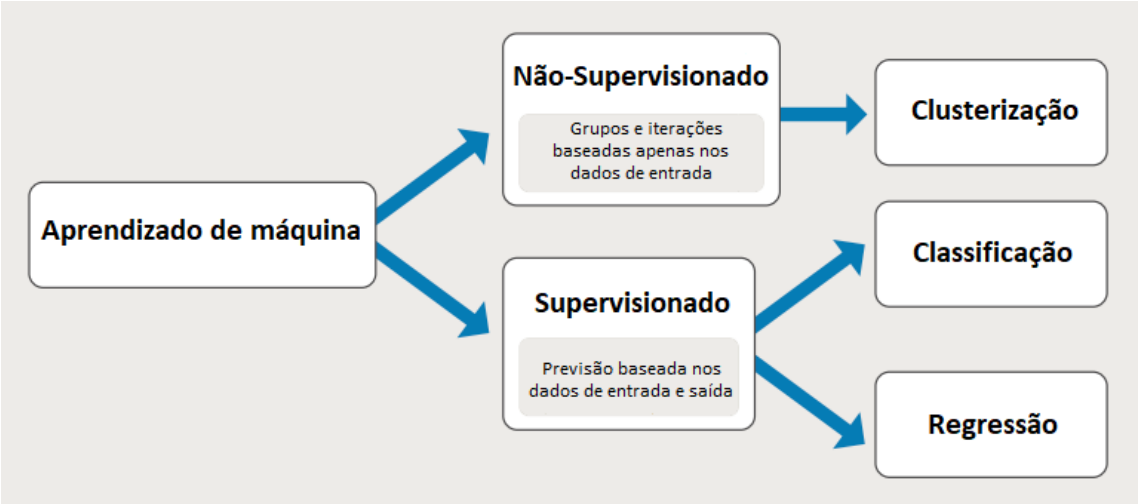


Figura 2.2: Técnicas de ML adaptado de (MathWorks, 2020)

Das técnicas de ML supervisionadas, como apresentado na Figura 2.3, um conjunto de dados pode ser definido como uma coleção de tuplas $(X_i, y_i)$ onde $X_i = (x_1, x_2, ..., x_n)$ e $y_i$ indica o rótulo ou classe correspondente a $X_i$. Quando os valores de $y_i$ são definidos por uma quantidade limitada de valores discretos, tem-se um problema de classificação. Quando tais valores de $y_i$ são contínuos, tem-se um problema de regressão (Padilha and Carvalho, 2017).

A clusterização é a classificação não supervisionada de dados, como na Figura 2.3, formando agrupamentos ou *clusters*. Ela representa uma das principais etapas de processos de análise de dados, denominada análise de *clusters* (Jain et al., 1999).

### 2.1.3 Modelando as dependências de tempo

Os problemas de previsão de séries temporais podem ser reformulados como problemas de aprendizado supervisionados. Uma abordagem para transformação desses problemas das séries em problemas de aprendizado supervisionados é feita nos dados de entrada dos modelos de formas específicas apresentadas a seguir.

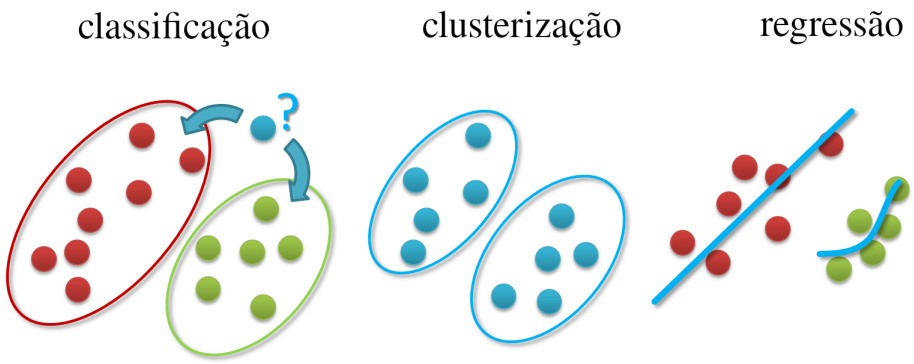


Figura 2.3: Modelos técnicos de ML (Big-Data.Tips, 2018)

### 2.1.4 Aprendizagem Supervisionada

Uma vez que se tenha um registro histórico dos dados, um problema de previsão em um passo pode ser abordado como um problema de aprendizagem supervisionada.

A aprendizagem supervisionada consiste em modelar, com base em um conjunto finito de observações, a relação entre um conjunto de variáveis de entrada e uma ou mais variáveis de saída, que são consideradas dependentes das entradas.

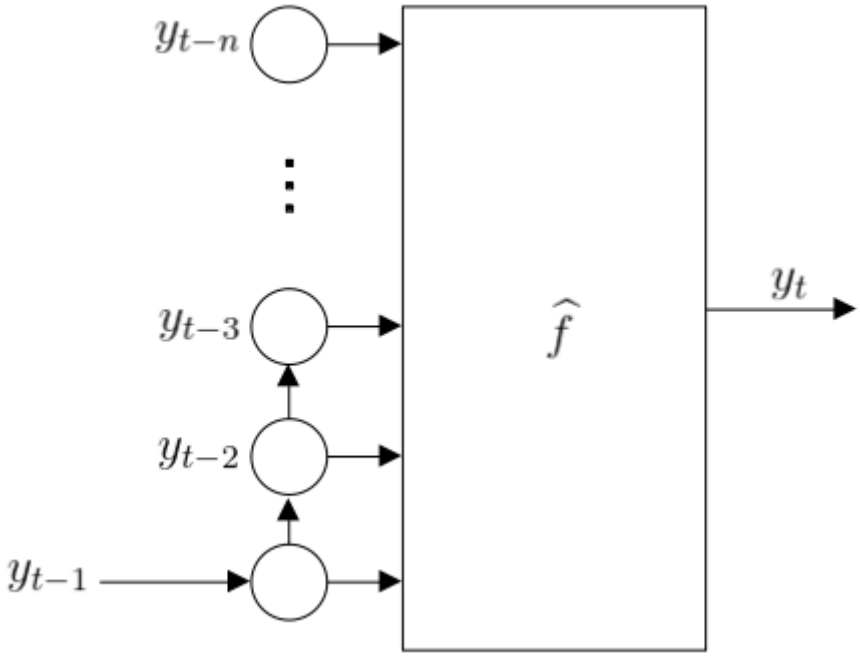


Figura 2.4: Modelo de previsão *One-step forecasting* em séries temporais

A previsão em um passo é feita com a utilização de um modelo mapeado. Nessa previsão, os $n$ valores anteriores da série estão disponíveis e o problema de previsão pode ser convertido na forma de um problema genérico de regressão (Bontempi et al., 2013), como mostrado na Figura 2.4.

Na previsão de um passo da Figura 2.4, o modelo representado por $f$ retorna à previsão

do valor das séries temporais no tempo $t$ em função dos $n$ valores anteriores (o conjunto de círculos representam os dados de um operador $y$ de atraso unitário $t-1$ a $t-n$).

## 2.2 Processo genérico para previsão baseado em ML

Os processos tecnológicos usados para a previsão que baseia-se em ML envolvem etapas dentro da estatística que são fundamentais na qualidade dos resultados. Nesse sentido, a estatística é a parte da matemática responsável por obter conclusões a partir dos dados observados. É com a estatística que se desenvolve o conjunto de técnicas que utiliza coleta de dados, classificação de dados, apresentação ou representação dos dados, análise e interpretação (Neto, 2004).

### 2.2.1 Coleta, amostragem e organização de informação

A coleta de dados pode ser classificada em relação ao tempo como contínua, periódica ou ocasional. Essa classificação só é possível se a coleta é feita de forma adequada, exigindo que o pesquisador conheça os conceitos de população e amostra (Neto, 2004).

A população é o agregado de todos os elementos sobre os quais deseja obter informações sobre algumas de suas características. Essas características são chamadas de variáveis. A amostra, por sua vez, é uma parte menor da população, mas que representa o conjunto total de elementos da população de onde foi extraída.

#### 2.2.1.1 Amostragem aleatória simples

Dentro da amostragem probabilística, a amostra aleatória simples é uma das mais populares. Nessa técnica de amostragem, todos os elementos que compõem o universo e estão descritos no marco amostral têm a mesma probabilidade de serem selecionados para a amostra, com ou sem reposição.

Considera-se que a amostragem é sem reposição quando, uma vez selecionado para pertencer à amostra, o elemento escolhido não pode voltar a participar de uma nova seleção. Seria equivalente a dizer que toda vez que sortear um elemento aleatório, esse elemento não poderá ser adicionado novamente para participar do próximo sorteio. Se, no entanto, for usada a repetição, um elemento selecionado para a amostra poderá ser selecionado uma outra vez (de Oliveira and Grácio, 2010).

Na escolha de uma amostra aleatória simples sem repetição, assegura-se a obtenção de amostras representativas de modo que a única fonte de erro que poderá afetar os resultados será o erro aleatório. Esse erro devido ao azar pode ser calculado incluindo o fator precisão.

A determinação do tamanho da amostra é um problema de grande importância, porque as amostras, desnecessariamente grandes, acarretam desperdício de tempo e amostras ex-

| Grau de Confiança | $E$ | Valor Crítico $\alpha$ |
|---|---|---|
| 90% | 0,10 | 1,645 |
| 95% | 0,05 | 1,96 |
| 99% | 0,01 | 2,575 |

Tabela 2.1: Valores críticos associados ao grau de confiança na amostra

cessivamente pequenas podem levar a resultados não confiáveis (de Oliveira and Grácio, 2010).

Em muitos casos, é possível determinar o tamanho mínimo de uma amostra para estimar um parâmetro estatístico. Uma expressão possível para cálculo do tamanho da amostra para uma estimativa confiável é dada pela equação 2.1 (de Oliveira and Grácio, 2010).

$$n = (\frac{\alpha * \sigma}{E})^2 \tag{2.1}$$

Onde na equação 2.1:

$n$ = Número de indivíduos na amostra

$\alpha$ = Valor crítico que corresponde ao grau de confiança desejado.

$\sigma$ = Desvio-padrão populacional da variável estudada.

$E$ = Margem de erro máximo da estimativa, que identifica a diferença máxima entre a média amostral e a verdadeira média populacional.

Os valores de confiança mais utilizados e os valores de $Z$ correspondentes podem ser encontrados na Tabela 2.1.

#### 2.2.1.2 Binarização de variáveis

A preparação das variáveis para os modelos de previsão pode considerar a conversão de informações. Os modelos de classificação necessitam de um padrão de entrada de dados e as entradas não podem ser aplicadas com dados no formato categórico. A variável categórica, então, pode ser convertida em variável binária assimétrica, criando uma nova variável binária para cada um dos estados nominais (Han and Kamber, 2001).

O processo de binarização impõe que, para um objeto com origem na variável categórica com um determinado valor de estado, a variável binária que representa esse estado é definida como 1, enquanto o restante das variáveis binárias criadas são definidas como 0 (Rajalaxmi and Natarajan, 2008). Isso significa que, após a conversão, o valor binário é mapeado para o valor correspondente.

Por exemplo, para codificar a variável nominal estado civil, uma variável binária pode ser criada para cada um dos três valores listados na Figura 2.5. Para um funcionário operador que tenha o estado civil casado, a variável casado é definida para 1, enquanto as duas variáveis restantes são definidas como 0.

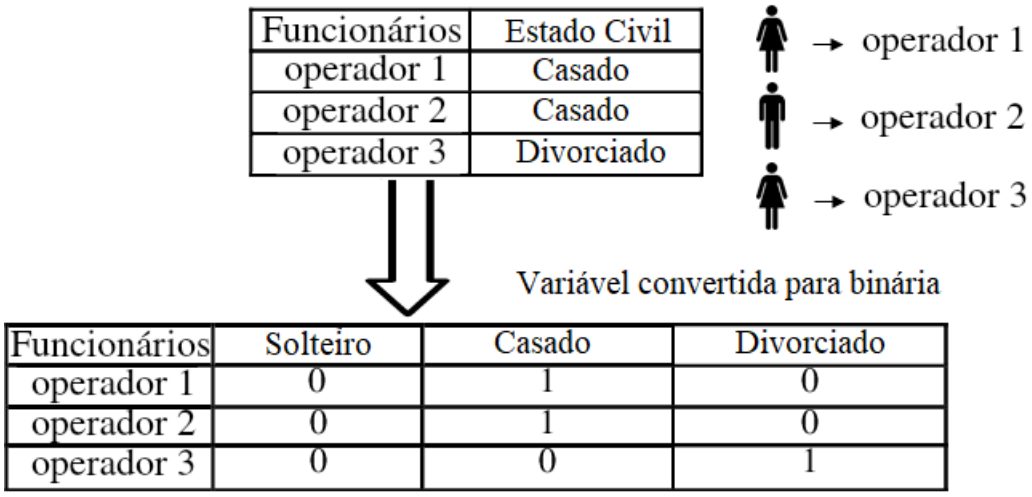


Figura 2.5: Conversão de variável categórica para binária

Uma desvantagem prática dessa técnica, que aumenta o quantidade de dados, é que o processamento das informações pelo modelo de classificação acaba se tornando mais demorado, e isso ocorre justamente pelo grande volume de dados transformados.

#### 2.2.1.3 *Nested Cross-Validation* - NCV

Para resolver problemas de erros em conjuntos destinados aos treinamentos e aos testes em modelos de previsão que têm dependência de tempo, foi criado o método *Nested Cross-Validation* (NCV) (Varma and Simon, 2006). Esse método NCV possui um *loop* para gerar a métrica ROC AUC em 4 etapas usando dados de treinamento e dados de teste, como apresentado na Figura 2.6.

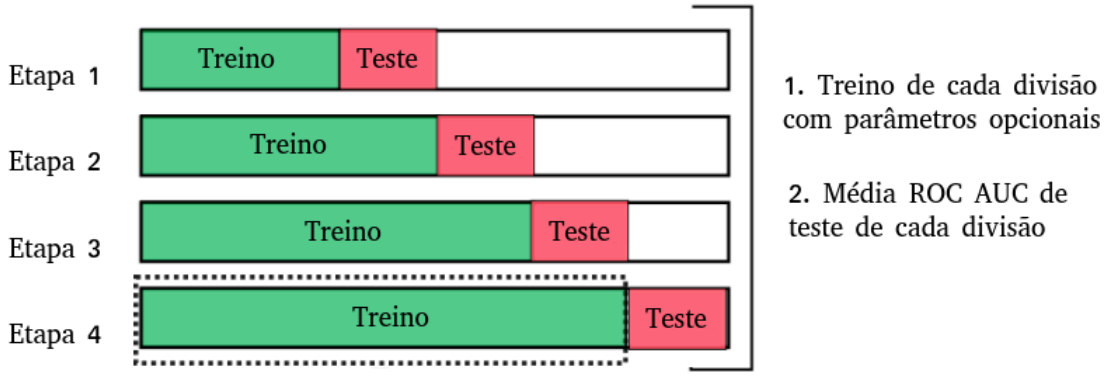


Figura 2.6: Processo de validação cruzada aninhada

O treinamento dos modelos é feito no NCV considerando que o conjunto de teste é a sequência temporal dos dados do conjunto de treinamento. O conjunto de dados de teste é o conjunto no tempo $t+1$ em relação ao conjunto de treinamento no tempo $t$. Na execução das etapas sequencialmente, o conjunto de treinamento assume os dados de teste

da etapa anterior, tornando-se o conjunto de treinamento corrente no tempo $t$ e o um novo conjunto de teste $t+1$ e utilizado na etapa em execução.

Quando uma etapa finaliza sua execução, o valor da métrica ROC AUC é registrado. Esse valor é registrado para que o valor ROC AUC final possa ser calculado pela média dos valores de todas as etapas finalizadas.

### 2.2.2 Seleção de variáveis

A seleção de variáveis inclui um conjunto de técnicas que tem se desenvolvido no decorrer dos anos com o advento da criação de grandes bancos de dados e os consequentes requisitos para boas práticas de aprendizado de máquina. À medida que surgem novos problemas, novas abordagens para a seleção de variáveis também são desenvolvidas (Dash and Liu, 1997).

A ideia da seleção de variáveis é encontrar um subconjunto de variáveis com tamanho mínimo que seja necessário e suficiente para o conceito de destino, que seria uma previsão (Kira and Rendell, 1992). A seleção resultaria em um subconjunto de variáveis $M$ com origem em um conjunto de $N$ variáveis, $M < N$, de modo que o valor de uma função seja otimizado em todos os subconjuntos de tamanho $M$ (Narendra and Fukunaga, 1977).

Para melhorar a precisão da previsão, a seleção de variáveis tem por objetivo escolher um subconjunto de variáveis para diminuir o tamanho da estrutura sem diminuir significativamente a qualidade da previsão do classificador. A intenção é justamente de melhorar a qualidade do modelo construído usando apenas as variáveis selecionadas para se aproximar da distribuição da classe original (Koller and Sahami, 1996).

#### 2.2.2.1 Seleções básicas *Backward* e *Forward*

A seleção $forward$ e *backward* é uma das opções de seleção de variáveis mais básicas e mais usadas em algoritmos disponíveis. Conceitualmente, essas técnicas são aplicáveis a muitos tipos diferentes de dados (Borboudakis and Tsamardinos, 2019).

O $forward$ é um procedimento que parte da suposição de que não há variáveis no modelo. A ideia do método é adicionar uma variável de cada vez. A primeira variável selecionada é aquela com maior correlação com a resposta, e a ordem de adição de variáveis segue a fila de correlação amostral que pode ser definida pelo algoritmo de *Pearson*, *Relief* ou outros.

No $forward$, o modelo é ajustado inicialmente com a variável com maior correlação amostral com a variável resposta. Supondo que essa variável seja $x_1$, calculamos a estatística $F$ para testar se ela realmente é significativa para o modelo. A variável entra no modelo se a estatística $F$ for maior do que o ponto crítico, chamado de $F_{in}$ ou $F$ para entrada. Nota-se que $F_{in}$ é calculado para um dado ponto crítico.

Considerando que $x_1$ foi selecionado para o modelo, o próximo passo é adicionar uma variável com maior correlação com a resposta considerando a presença da primeira variável no modelo. Supondo que a maior correlação parcial com y seja $x_2$. Se o valor da estatística é maior do que $F_{in}$, $x_2$ é selecionado para o modelo.

O processo é repetido, ou seja, a variável com maior correlação parcial com y é adicionada no modelo se sua estatística F parcial for maior que $F_{in}$, até que não seja incluída mais nenhuma variável explicativa no modelo (Marques, 2018).

Enquanto os procedimentos do $forward$ adicionam variável a variável conforme a correlação geral e avaliam a correlação parcial do modelo, o procedimento *backward* inicia adicionando todas as variáveis e depois, por etapas, cada uma pode ser ou não eliminada. A decisão de retirada da variável é tomada baseando-se em testes F parciais, que são calculados para cada variável como se ela fosse a última a entrar no modelo (BURSAC et al., 2008).

#### 2.2.2.2 Seleção por correlação de *Pearson*

A correlação de Pearson mensura a direção e o grau da relação linear entre duas variáveis. Essa correlação é uma medida de associação linear entre variáveis. Sua fórmula corresponde à equação 2.2 (Moore, 2007)(Galarça et al., 2010):

$$r = \frac{\sum_{i=1}^{n}(x_i - \overline{x})(y_i - \overline{y})}{\sqrt{\sum_{i=1}^{n}(x_i - \overline{x})^2} \cdot \sqrt{\sum_{i=1}^{n}(y_i - \overline{y})^2}} \tag{2.2}$$

onde $x1$ , $x2$ , ..., $x_n$ e $y1$ , $y2$ , ..., $y_n$ são os valores medidos de ambas as variáveis. Detalhes de $\overline{x}$ e $\overline{y}$ na Equação 2.3:

$$\overline{x} = \frac{1}{n} \cdot \sum_{i=1}^{n} x_i \quad e \quad \overline{y} = \frac{1}{n} \cdot \sum_{i=1}^{n} y_i \tag{2.3}$$

O coeficiente de correlação de *Pearson* é representado pela letra r e assume valores de -1 a 1. A relação $r = 1$ representa a correlação perfeita e positiva entre duas variáveis. Já a relação $r = -1$, representa correlação perfeita negativa entre duas variáveis, ou seja, enquanto uma aumenta a outra diminui, e à medida que se aproxima do 1 vai ficando perfeita a correlação (Galarça et al., 2010).

#### 2.2.2.3 Seleção pelo algoritmo *Relief*

O $Relief$ foi proposto como um algoritmo apropriado para estimar a relevância de atributos discretos e contínuos em dados que caracterizam duas classes (Kira and Rendell, 1992). É uma técnica que procura pelos exemplos mais próximos da mesma classe e de classes diferentes, e atribui pesos aos atributos de acordo com quão bem diferenciam esses exemplos.

O $Relief$ é o primeiro algoritmo da família $Relief-F$. Ele foi proposto para dados binários, e o seu princípio de funcionamento está em destacar atributos para vizinhos mais próximos em classe diferente, bem como penalizar atributos para vizinhos mais próximos na mesma classe. Para isso, o algoritmo busca vizinhos próximos para cada exemplo $R_i$ escolhido aleatoriamente: um de mesma classe - *nearesthit H* -, e o outro de classe diferente - *nearestmiss M*. Desse modo, o algoritmo considera $K = 1$ vizinhos mais próximos de cada classe de um problema (Demsar, 2010). O pseudocódigo que explica com mais detalhes o $Relief$ pode ser visto no Algoritmo 1.

**Algoritmo 1:** Algoritmo $Relief$ (Robnik and Kononenko, 2003)

```
Entrada: O conjunto de treinamento e o número de iterações m
Saída: o vector W de estimativas das qualidades dos atributos
inicializa o vetor de pesos W[A] := 0 : 0;
while i ≤ m do
    seleciona aleatoriamente uma instância R_i;
    encontra hit mais próximo H e o miss mais próximo M;
    while j ≤ a do
        W[A_j] := W[A_j] − diff(A_j;R_i;H)/m + diff(A_j;R_i;M)/m;
    end
end
```

$Relief$ usa a medida $diff$ para calcular a diferença entre os valores. Ela é constituída pelas medidas de $\lambda$ atributos categóricos e *Range Normalized Difference*(RND) para atributos numéricos, como nas equações do exemplo $diff(A_j;R_i;M)$, as equações 2.4, 2.5 e 2.6.

$$diff(A_j,R_i,M)\begin{cases}\lambda(A_j,R_i,M)\ se\ A_j \text{ é categórico}\\ RND(A_j,R_i,M)\ se\ A_j \text{ é numérico}\end{cases} \tag{2.4}$$

$$\lambda(A_j,R_i,M)\begin{cases}0\ se\ A_j,R_i = A_j,M\\ 1\ se\ A_j,R_i \neq A_j,M\end{cases} \tag{2.5}$$

$$RND(A_j,R_i,M) = \frac{A_j,R_i - A_j,M}{max(A_j) - min(A_j)} \tag{2.6}$$

O tratamento de domínios com mais de duas classes não poderia ser feito diretamente por este algoritmo. Foram criadas, então, outras versões (Kononenko, 1994), que são as variantes $Relief-A$, $Relief-B$, $Relief-C$, $Relief-D$, $Relief-E$ e, sendo a mais eficiente e robusta, a variante $Relief-F$.

#### 2.2.2.4 Seleção pelo algoritmo genético

Algumas técnicas tem sido usadas na ciência e na engenharia como algoritmos adaptativos para resolver problemas práticos e como modelos computacionais de sistemas evolutivos naturais (Mitchell, 1998).

Os algoritmos genéticos (AG) imitam as forças da seleção natural para encontrar valores ótimos de alguma função $F(x)$ (Mitchell, 1998). Um conjunto inicial de soluções candidatas é criado e seus valores de adequação correspondentes são calculados, sendo que os valores maiores são melhores. Esse conjunto de soluções é chamado de população inicial corrente e cada solução é considerado um indivíduo. Os indivíduos com os melhores valores de avaliação na função $F(x)$ são combinados aleatoriamente por operadores genéticos para produzir descendentes que compõem a próxima população. Para isso, os indivíduos são selecionados e submetidos a um cruzamento, imitando a reprodução genética, e também estão sujeitos a mutações aleatórias. Esse processo é repetido várias vezes e muitas gerações são produzidas criando soluções cada vez melhores (Pacheco, 1999).

Um AG simples pode ser descrito como um processo contínuo que repete ciclos de evolução controlados por um critério de parada, conforme apresentado na Figura 2.7:

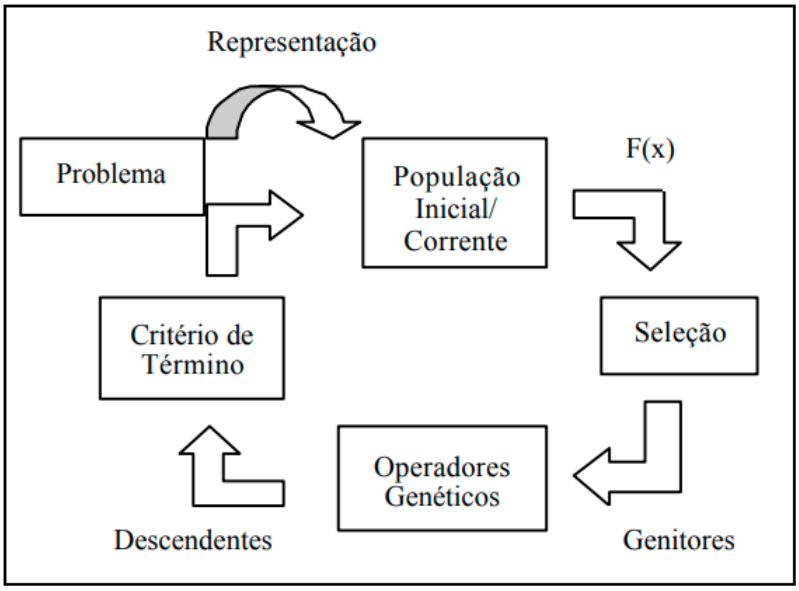


Figura 2.7: Representação de um AG Simples - (Pacheco, 1999)

O processo de seleção de variáveis tem por objetivo encontrar o menor subconjunto de atributos com a melhor acurácia de classificação. Esse processo é dividido em duas partes: o método de busca e a função de aptidão usada para medir a qualidade dos subconjuntos de atributos (Pappa et al., 2002). Os algoritmos de seleção de atributos são divididos em três grupos: exponenciais (busca exaustiva), sequenciais (*forward* e *backward*) e randômicos (algoritmo genético) (Boz, 2002).

No processo de seleção de variáveis usando AG, o gene pode ser representado por um vetor, onde cada elemento será um dígito binário (0 ou 1) representando a presença ou ausência de uma determinada variável (Vafaie and Jong, 1992), conforme a Figura 2.8. Para seleção dos indivíduos, pode ser adotado o método da roleta, e a população inicial pode ser gerada de forma aleatória. A função de avaliação poderia ser definida conforme o problema e a estratégia para resolução dele.

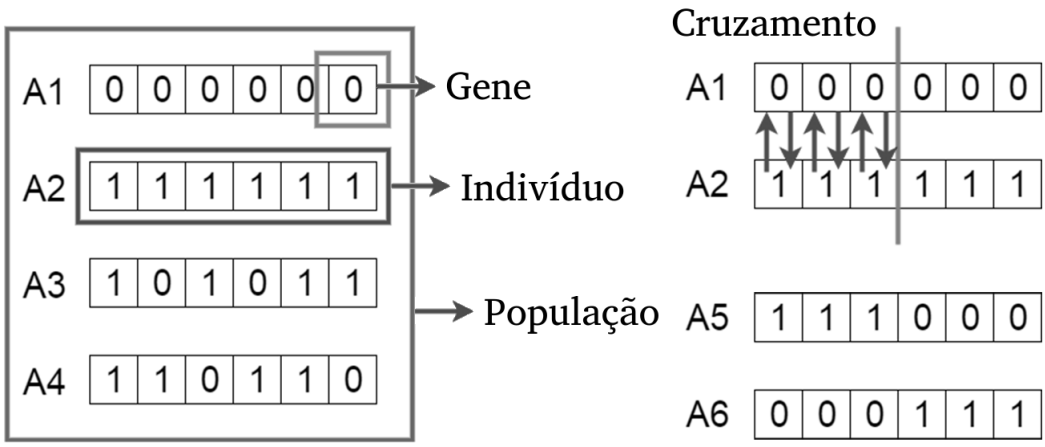


Figura 2.8: Representação de variáveis como genes no AG - (Oliveira et al., 2019b)

### 2.2.3 Modelos de ML para previsão

Os profissionais em ciência de dados e demais pesquisadores têm vários modelos de previsão à sua disposição para serem usados conforme o problema definido. Esses modelos podem ser compostos por técnicas de classificação aplicadas de acordo com a relevância para o atributo de saída de uma eventual previsão. Dentre os modelos de previsão, alguns dos mais populares são (cf. *LR - Logistic Regression, LSTM - Long Short Term Memory, MLP - Multilayer Perceptron, NB - Naive Bayes, RF - Random Forest, SVM - Suport Vector Machine e XGBoost - Extreme Gradient Boosting*) (Caruana and Niculescu-Mizil, 2006).

#### 2.2.3.1 Classificador de regressão logística - LR

*Logistic Regression* (LR) é uma técnica estatística dos métodos de ML supervisionado que podem ser dedicados às tarefas de classificação. Essa técnica tornou-se popular, principalmente no setor financeiro, por sua capacidade proeminente de detectar inadimplentes.

Como se trata de classificação, determina-se a probabilidade de uma observação fazer parte de uma determinada classe ou não. Portanto, procura-se expressar a probabilidade com um valor entre 0 e 1. Para gerar valores entre 0 e 1, utiliza-se a probabilidade,

conforme a equação 2.7.

$$p(x) = \frac{exp(\beta_0 + \beta_1 X)}{1 + exp(\beta_0 + \beta_1 X)} \tag{2.7}$$

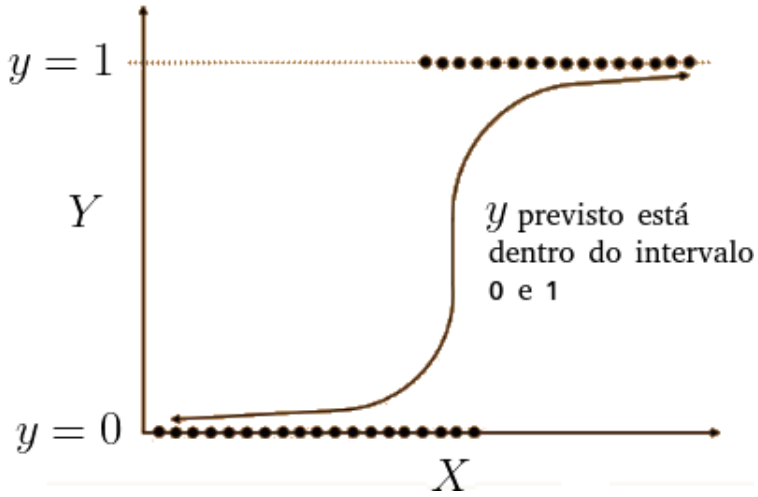


Figura 2.9: Gráfico de regressão logística

Uma variável dependente binária Y e uma variável independente $X$, sendo $p(x)$ = $P[Y = 1]$ a função usada para estimar a probabilidade de uma determinada realização da variável resposta ser um sucesso (Truett and et all, 1967).

Ao representar graficamente, como na Figura 2.9 um conjunto de dados de $Y$ e $X$, sendo que Y toma apenas os valores $Y = 1$ (sucesso) ou $Y = 0$ (insucesso), constata-se que a variação de $Y$ não tem sido linear e aditiva. Não obstante, o modelo de regressão logística pode ser ajustado recorrendo à regressão não linear. A solução tradicional consiste na transformação da seguinte forma na Equação 2.8 (Cox and Snell, 1989).

$$log(\frac{p(x)}{1 - p(x)}) = \beta_0 + \beta_1 X \tag{2.8}$$

sendo que, $\beta$ é o vetor dos $p + 1$ coeficientes de regressão logística; e $p(x)$ é o vetor de probabilidades estimadas.

Esse é um modelo de probabilidade implementado por algumas bibliotecas de ML de código aberto como *skit-learn* para a linguagem de programação *Python*. Através desse modelo, pode-se usar uma matriz de confusão para avaliar a classificação. A matriz de confusão é uma tabela usada para avaliar o desempenho de um modelo de classificação. Pode-se também visualizar o desempenho de um algoritmo por meio dela. O fundamental da matriz de confusão é o número de previsões corretas e incorretas resumidas em classe.

#### 2.2.3.2 Classificador perceptron multicamadas - MLP

O primeiro modelo de rede neural implementado foi o perceptron por Frank Rosenblatt, em 1958. Essa é uma rede neural simples constituída de uma camada de entrada, uma camada de saída (Russel and Norvig, 1995), entrada intervalar, aprendizado supervisionado e alimentação à frente. Essa rede se utiliza de um combinador adaptativo linear em que a saída de um elemento processador é a combinação linear das entradas, resultado em um vetor multiplicado por pesos (Cardon and Müller, 1994).

$$S(t) = \sum_{i=0}^{n-1} x_i(t) w_i(t) + b_i \tag{2.9}$$

Na equação perceptron 2.9, o $x_i$ representa as entradas, o $b$ representa o bias para controlar o neurônio e o $w_i$ são os pesos.

Em decorrência da identificação das limitações relativas ao perceptron de camada simples que apresentava fronteiras de decisão lineares e funções de lógica simplificadas, foi desenvolvido o *Multilayer Perceptron* (MLP), mostrado na Figura 2.10. Esse consiste em uma camada de entradas, uma ou mais camadas intermediárias, ou escondidas, e uma camada de saída.

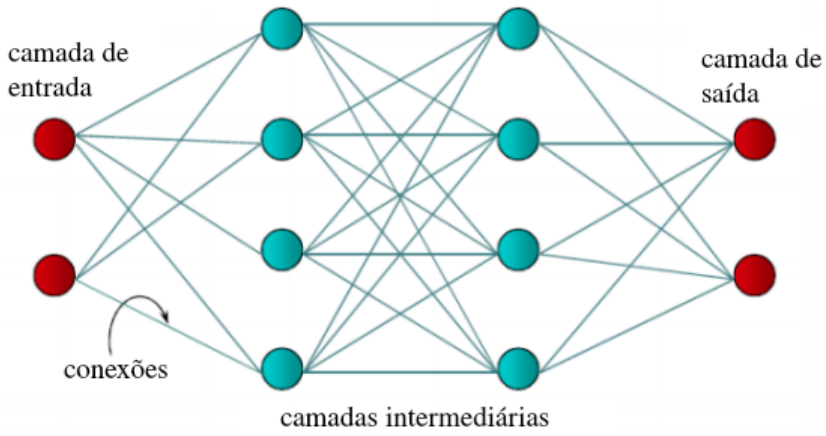


Figura 2.10: Representação de um MLP - (Soares and Silva, 2011)

O processamento da informação no MLP possui a fase de propagação, quando o sinal de entrada é propagado através da rede, camada por camada, até produzir uma saída. Há também a fase de adaptação, quando ocorrem os ajustes dos pesos da rede.

A camada de saída da rede MLP recebe os estímulos da camada intermediária e constrói o padrão que será a resposta. As camadas intermediárias funcionam como extratoras de características. Seus pesos seguem a codificação de características, pois são apresentados nos padrões de entrada e permitem que a rede crie sua própria representação (Tonsig, 2000).

As redes MLP são aplicadas a problemas através do treinamento de forma supervisio-

nada, em que a rede tem uma resposta que é comparada com a saída desejada, recebendo informações sobre o erro da resposta atual. Para minimização dos erros quadráticos, existem algoritmos conhecidos capazes de fazer o treinamento, como *Backpropagation* (Soares and Silva, 2011).

O *Backpropagation* é uma solução através da utilização de algoritmos de treino supervisionado que se baseia na heurística do aprendizado por correção do erro da camada de saída, retropropagando para as camadas intermediárias da RNA (Tonsig, 2000).

#### 2.2.3.3 Classificador de memória de longo prazo - LSTM

Um classificador *Long Short Term Memory* (LSTM) é um tipo de Rede Neural Recorrente (RNN) especialmente projetada para garantir que a saída da rede de uma determinada entrada se equilibre enquanto ela retorna resultados dentro de um ciclo. São os ciclos de retorno que permitem que as redes recorrentes sejam melhores no reconhecimento de padrões do que outras redes neurais. As redes LSTM oferecem melhor desempenho em comparação a outras arquiteturas RNN (Le et al., 2019).

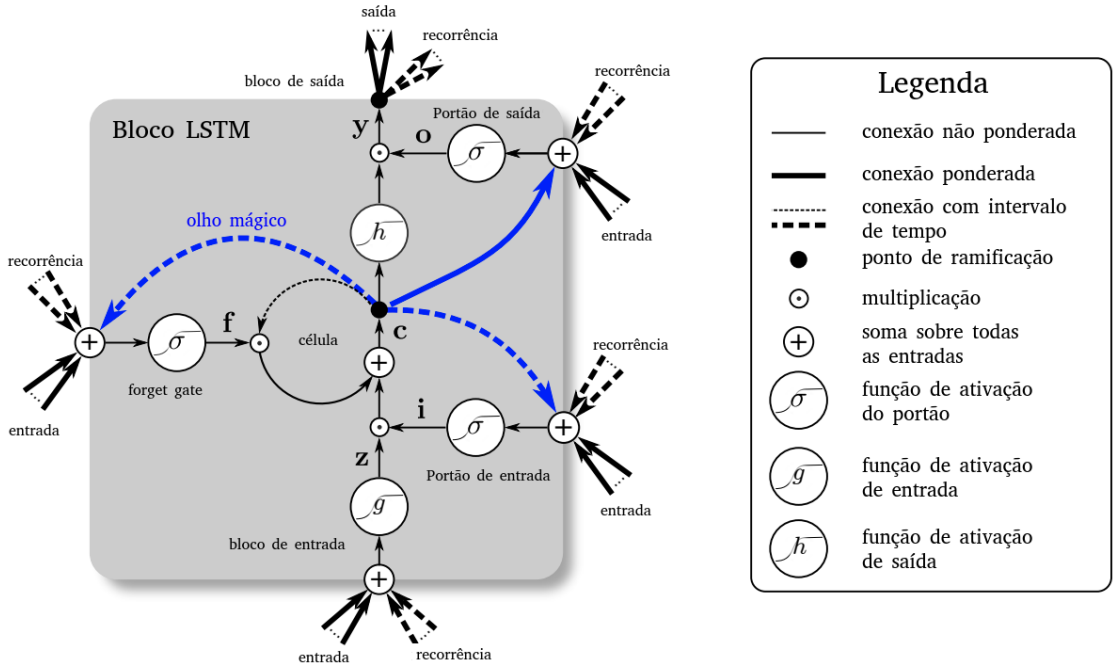


Figura 2.11: Esquema detalhado de um bloco de memória de longo prazo usado nas camadas ocultas de uma rede neural recorrente - (Greff et al., 2017)

A ideia central por trás da arquitetura LSTM é uma memória celular. Essa memória pode manter seu estado ao longo do tempo com unidades de controle não lineares que regulam o fluxo de informações dentro e fora da célula (Gers et al., 1999).

A informação é retida pelas células e as manipulações de memória são feitas por três portões (Greff et al., 2017), como apresentado na Figura 2.11.

- *Forget Gate*: Onde as informações que não são mais úteis no estado da célula são removidas. Duas entradas são alimentadas por esse portão e multiplicadas por matrizes de peso, seguidas pela adição do *bias*. O resultante é passado por uma função de ativação que fornece uma saída binária. Se para um determinado estado de célula a saída for 0, a informação é esquecida; e para a saída 1, a informação é retida para uso futuro.

- Portão de entrada: A adição de informações úteis ao estado da célula é feita por esse portão e a informação é regulada usando a função sigmoide que filtra os valores a serem lembrados de forma similar ao $forgetgate$. Um vetor é criado usando a função *tanh* que contém todos os valores possíveis. Os valores do vetor e os valores regulados são multiplicados para obter as informações úteis.

- Portão de saída: A tarefa de extrair informações úteis do estado da célula atual para serem apresentadas como uma saída é feita por esse portão e um vetor é gerado aplicando a função *tanh* na célula. A informação é regulada usando a função sigmoide que filtra os valores a serem lembrados usando as entradas. Os valores do vetor e os valores regulados são multiplicados para serem enviados como uma saída e como uma entrada para a próxima célula.

A unidade básica, na camada oculta de uma rede LSTM, é o bloco da memória, que contém uma ou mais células de memória e um par de unidades de multiplicação adaptativas que bloqueiam a entrada e a saída para todas as células no bloco. Cada célula de memória tem em seu núcleo uma autoconexão recorrente em que a ativação é chamada de estado da célula (Hochreiter and Schmidhuber, 1997).

#### 2.2.3.4 Classificador ingênuo Bayesiano - NB

O *Naive Bayes* (NB) tem origem no teorema de Bayes que é um algoritmo para aprendizado indutivo com abordagem probabilística (McCallum and Nigam, 1998). Esse modelo permite calcular probabilidades em categorias.

O classificador Bayesiano é chamado de ingênuo por assumir que os atributos são condicionalmente independentes. O método faz a classificação assumindo que a probabilidade de sua ocorrência é independente (Domingos and Pazzani, 1997). Esses tipos de classificadores atribuem para a classe mais provável um determinado exemplo descrito por seu vetor de variáveis e a aprendizagem pode ser bastante simplificada assumindo que as variáveis sejam independentes da classe em que $P(X|C) = \prod_{i=1}^{n} P(X_i|C)$ onde $X = (X_1, ..., X_n)$ é um vetor de variáveis e uma classe, como na Figura 2.12. O resultado do classificador, mesmo encarado como ingênuo, é notavelmente bem-sucedido na prática, muitas vezes competindo com uma técnica muito mais sofisticada (Rish, 2001).

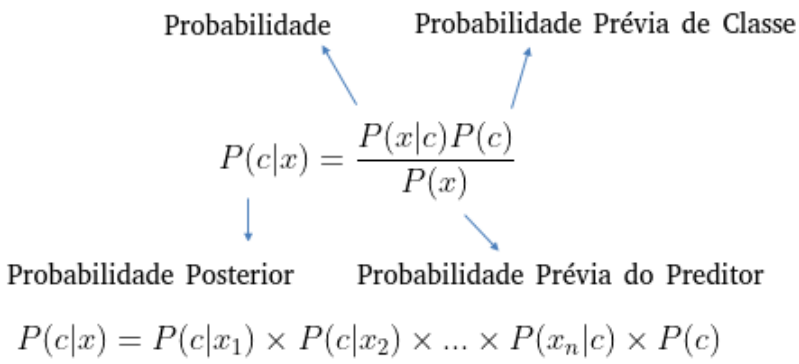


Figura 2.12: Fórmula de classificação NB

#### 2.2.3.5 Classificador de floresta aleatória - RF

O *Random Forest* (RF) consiste em uma combinação de classificadores árvore, em que cada classificador é gerado usando um vetor aleatório amostrado independentemente do vetor de entrada, e cada árvore emite um sinal de votação unitário para a classe mais popular para classificar um vetor de entrada (Breiman, 1999). Essas são melhorias significativas na precisão da classificação que resultaram no crescimento de um conjunto de árvores e possibilitam a indicação da classe mais popular. Para fazer crescer esses conjuntos, frequentemente, são gerados vetores aleatórios que gerenciam o crescimento de cada árvore em um conjunto.

Os métodos do conjunto usando classificadores do tipo árvore são definidos por $h(X, H_n)$, $n = 1, 2, 3...$, em que o $H_n$ são vetores aleatórios independentes distribuídos de forma idêntica e $X$ é um padrão de entrada (Breiman, 2001).

Para classificação, cada árvore na floresta aleatória emite um voto unitário para a classe popular na entrada $X$. A saída do classificador é determinada pelo sinal de votação majoritária das árvores, conforme Figura 2.13.

Como resultado, o algoritmo RF pode manipular dados de alta dimensão e utiliza um grande número de árvores no conjunto. Isso combinado com o fato de que a seleção de variáveis para uma divisão busca minimizar a correlação entre as árvores do conjunto (Freund and Schapire, 1996), embora consuma muito recurso computacional.

#### 2.2.3.6 Classificador máquina de vetores de suporte - SVM

*Support Vector Machine* (SVM) é um algoritmo de aprendizado de máquina supervisionado utilizado principalmente em problemas de classificação. Desde que esse algoritmo foi proposto, ele foi amplamente estudado e aplicado em muitos campos.

A ideia básica de classificadores SVM é tentar maximizar a distância entre duas classes, e a distância entre classes é tradicionalmente definida por pontos mais próxi-

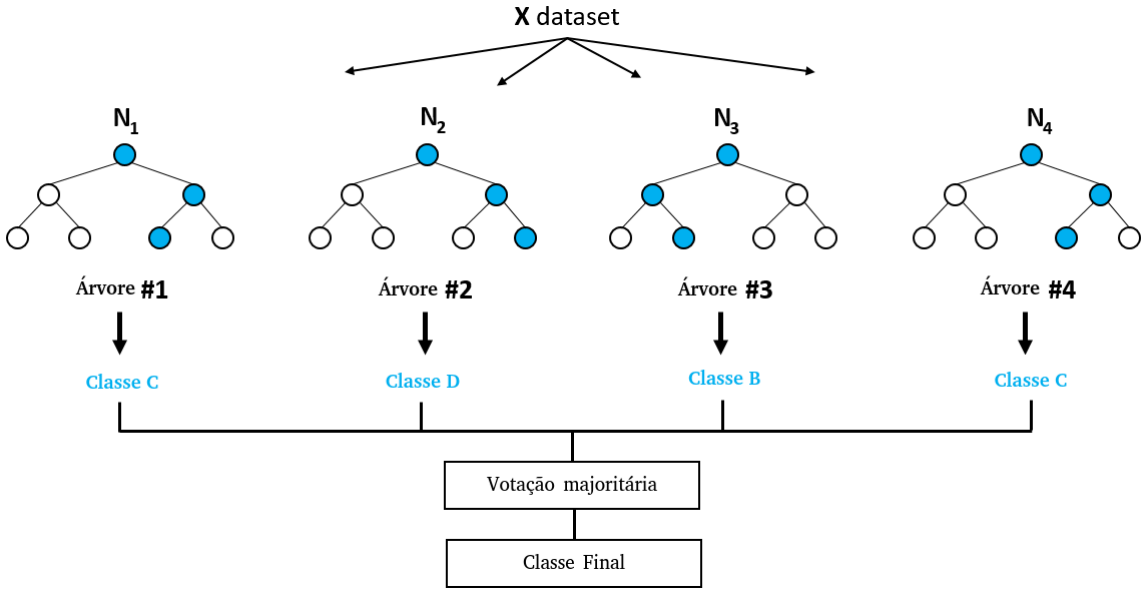


Figura 2.13: Diagrama do Classificador de Floresta Aleatória (Adalash, 2018)

mos. Os Vetores de Suporte são coordenadas da observação individual e constituem a fronteira que melhor separa as duas classes (hiperplano / linha), conforme apresentado na Figura 2.14 (Burges, 1998)(Vapnik, 1995).

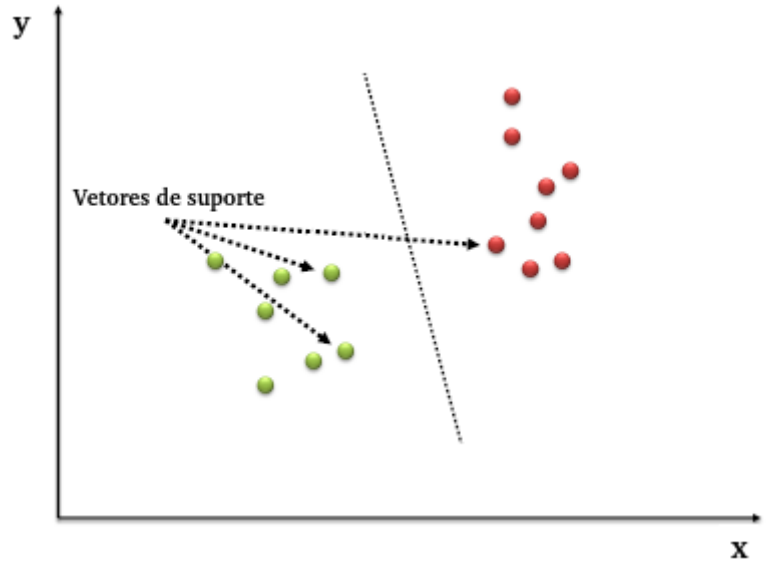


Figura 2.14: Classificação SVM (Sayad, 2020)

Considerando um problema de classificação binária, um conjunto de amostras $z = \{x_i, y_i\}_{i=1}^{m}$, onde $x_i \in \mathbb{R}^n$ e $y_i \in -1, 1$. Então, $z$ consiste em duas classes com os seguintes conjuntos de índices: $I = i|y_i = 1$ e $II = i|y_i = -1$. Seja $H$ um hiperplano dado por $w^T x + b = 0$ com $w \in \mathbb{R}^n$, $\|w\| = 1$ e $b \in \mathbb{R}$. Os valores de $I$ e $II$ são separável por $H$ se

para $i = 1, ...m$ (Vapnik, 1995)

$$\begin{aligned} w^T x_i = b > 0, \forall \in I \\ w^T x_i = b < 0, \forall \in II \end{aligned} \tag{2.10}$$

Na equação 2.10, $y_i(w^T x_i + b)$ fornece a distância entre o ponto $x_i$ e o hiperplano $H$. Em seguida, tem-se a distância de cada classe como na equação 2.11.

$$\begin{aligned} t_I(w,b) = Min_{i \in I}\ y_i(w^T x_i + b) \\ t_{II}(w,b) = Min_{i \in II}\ y_i(w^T x_i + b) \end{aligned} \tag{2.11}$$

O hiperplano de classificação correspondente é obtido conforme equação 2.12.

$$Max_{\|w\|=1,b} = t_I(w,b) + t_{II}(w,b) \tag{2.12}$$

Conforme a documentação do scikit-learn sobre o modelo SVM (scikit learn, 2020a), esse modelo possui uma complexidade alta e consome muito recurso computacional a medida que o número de vetores de treinamento aumenta. Isso pode tornar o modelo extremamente lento se o volume de dados utilizado for grande.

#### 2.2.3.7 Classificador de impulso de gradiente extremo - XGBoost

O impulsionamento de modelos baseados em árvores é um recurso altamente eficaz e amplamente utilizado nos métodos de ML. O *eXtreme Gradient Boosting* (XGBoost) é um desses impulsionamentos que é bastante utilizado pelos cientistas de dados para obter resultados em muitos desafios de aprendizado de máquina (Chen and Guestrin, 2016).

O *XGBoost* é um modelo escalável que está disponível como um pacote de código aberto. Por isso, é reconhecido na solução de vários problemas e desafios de mineração de dados.

Utilizado em problemas de aprendizado supervisionado, o *XGBoost* é uma implementação do algoritmo *Gradient Boosting* e é definido por um modelo de conjunto de árvores de decisão, o que torna o algoritmo muito semelhante ao RF (Horemuz, 2018). As árvores individuais são treinadas em sequência de florestas onde cada árvore sucessiva é treinada. Uma visão geral do treinamento algoritmo pode ser visto no Algoritmo 2.

A linha 9 do algoritmo contem os alvos de treinamento da próxima árvore, assim como os gradientes dos erros do modelo completo na iteração anterior.

A linha 10 executa o ajuste usando o algoritmo de treinamento associado à árvore de decisão $f$. As divisões da árvore de decisão são feitas usando uma combinação da função de perda e uma "pontuação da estrutura" que penaliza árvores complexas (termo de regularização).

A pontuação da estrutura penaliza a profundidade da árvore e a pontuação das folhas. Os pontos das folhas são penalizados com a regularização de $L2$, com um parâmetro

**Algoritmo 2:** Algoritmo XGBoost (Chen and Guestrin, 2016)

1 $x, y$: conjunto de dados de treinamento;
2 $\alpha \in [0,1]$: contração;
3 $M$: número de iterações;
4 $L(y, y_p)$: função de perda;
5 $f$: árvore de decisão;
6 $F_0 \leftarrow arg_\alpha \min \sum_{i=1}^{n} L(y_i, a)$ (inciar modelo)
7 $m \leftarrow 1$
8 **while** $m \leq M$ **do**
9 $\quad nr_i \leftarrow -(\frac{\partial L(y_i, F_{m-1}(x_i))}{\partial F_{m-1}(x_i)}) \forall_i$ (resíduo parcial);
10 $\quad f_m \leftarrow$ árvore de decisão para dataset $(x, r)$;
11 $\quad \gamma_m \leftarrow arg_{\gamma m} \min \sum_{i=1}^{n} L(y_i, F_{m-1}(x_i) + \gamma_m f_m(x_i))$;
12 $\quad F_m \leftarrow F_{m-1} + \alpha \gamma_m f_m$;
13 $\quad m \leftarrow m + 1$;
14 **end**
15 **return** $F_M$;

de peso de $\lambda_L 2$. Uma função de perda é usada com o *XGBoost*, porque é feita uma classificação binária.

Na linha 11 é calculado um tamanho de passo ideal $\gamma_m$ para seguir na direção dos gradientes da linha 9.

Na linha 12, atualiza-se o modelo atual com a árvore recém-aprendida. A nova árvore, treinada para prever os gradientes do erro, é multiplicada por $\gamma_m$ e pelo parâmetro de retração $\alpha$ antes de ser adicionada ao modelo. O parâmetro contração é um termo de regularização que impede que a árvore recém-criada exceda os resultados, reduzindo o tamanho da etapa $\gamma_m$.

Outro parâmetro de regularização que é considerado é a "taxa de subamostragem da coluna". A cada iteração do algoritmo de treinamento, subamostram-se as colunas (variáveis) do conjunto de treinamento. Dessa forma, cada árvore individual é treinada em um conjunto de variáveis potencialmente diferente. Isso é muito semelhante a uma técnica usada em florestas aleatórias. O modelo XGBoost se adapta muito facilmente, devido ao processo de treinamento. Dadas iterações suficientes, o modelo ajustará todo o conjunto de dados. Portanto, a regularização é muito importante.

### 2.2.4 Hiperparametrização dos modelos de ML

Existem muitos algoritmos diferentes de aprendizado de máquina, sendo que, alguns deles foram descritos nas seções anteriores. Quando são considerados os hiperparâmetros de cada um desses modelos algorítmicos, existe um número surpreendentemente grande de alternativas possíveis.

Os hiperparâmetros são parâmetros que devem ser definidos antes de treinar o modelo

para que se obtenham bons resultados na resolução de problemas. Existem algumas técnicas para encontrar os parâmetros ideais para os modelos de previsão e elas são descritas a seguir.

#### 2.2.4.1 Pesquisa em grade e manual

A pesquisa em grade e a pesquisa manual são as estratégias mais usadas para otimização de hiperparâmetros (Bergstra and Bengio, 2012). Tanto uma pesquisa como a outra são usadas para identificar regiões promissoras de possibilidades de parâmetros para os modelos, sendo que se diferenciam pela pesquisa manual a ser desenvolvida pela intuição, empiricamente, assumindo uma grande desvantagem que é a dificuldade de reproduzir resultados.

A pesquisa em grade analisa todas as combinações possíveis de hiperparâmetros e, por isso, é também conhecida como pesquisa exaustiva (Klatzer and Pock, 2015). Como já mencionado, existe um número surpreendentemente grande de alternativas possíveis de parâmetros, portanto, pode não ser uma técnica tão interessante para calibração de um modelo preditivo.

Mesmo assim, existem várias razões pelas quais a pesquisa manual e a pesquisa em grade prevalecem como o estado da arte, apesar de décadas de pesquisa em otimização global e a publicação de vários algoritmos de otimização de hiperparâmetros (Bergstra and Bengio, 2012):

- A otimização manual fornece aos pesquisadores algum grau de percepção e domínio sobre o modelo;
- Não há sobrecarga técnica ou barreira à otimização manual;
- A pesquisa em grade é simples de implementar e a paralelização é trivial;
- A pesquisa em grade (com acesso a um *cluster* de computação) normalmente encontra um conjunto de hiperparâmetros melhor que puramente a otimização manual e sequencial (na mesma quantidade de tempo);
- A pesquisa em grade é confiável em espaços de baixa dimensão.

#### 2.2.4.2 Pesquisa aleatória

A pesquisa aleatória é uma técnica algorítmica de desenho de atribuições de hiperparâmetros que usa o processo de escolha aleatória para avaliá-los. Isso faz com que nem todas as possibilidades sejam contempladas, ou seja, o espaço de busca é limitado (Bergstra and Bengio, 2012).

Usar dessa técnica para hiperparametrizar modelos preditivos pode prover o alcance de resultados melhores ou iguais, quando comparada com a técnica de pesquisa de grade padrão, tendo um custo computacional menor (Klatzer and Pock, 2015).

Na Figura 2.15, são apresentados dois gráficos comparativos das técnicas de hiperparametrização.

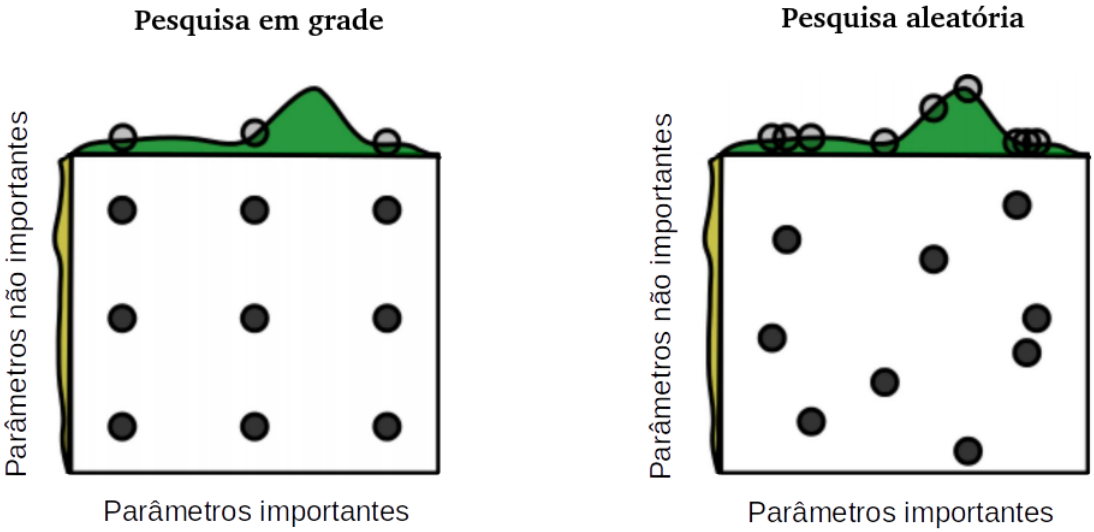


Figura 2.15: Hiperparametrização por pesquisa em grade e aleatória - (Bergstra and Bengio, 2012)

#### 2.2.4.3 Algoritmo genético para hiperparametrização

Dentre as técnicas de hiperparametrização, existem pesquisas que propõem abordagem meta-heurísticas com base em algoritmos genéticos. Os algoritmos evolutivos, como os genéticos, juntamente com técnicas de pesquisa em grade e pesquisa aleatória, são as abordagens mais usadas para resolver problemas de otimização de hiperparâmetros (Francescomarinoa et al., 2018).

Os algoritmos genéticos oferecem a oportunidade de pesquisar o espaço do hiperparâmetro de maneira aleatória e também a oportunidade de utilizar resultados anteriores para direcionar a pesquisa.

Cada hiperparâmetro a ser otimizado é codificado como um único gene para cada indivíduo. Um intervalo e uma resolução são definidos para cada gene, a fim de eliminar a pesquisa nas áreas do espaço do hiperparâmetro que não são de interesse. A população inicial é gerada por amostragem de cada gene a partir de uma distribuição aleatória uniforme, e a adequação de cada indivíduo é avaliada. A partir de então, cada geração é formada usando seleção, cruzamento e mutação com base nos indivíduos com maior aptidão da geração anterior. Esse processo pode ser visto como uma única geração de

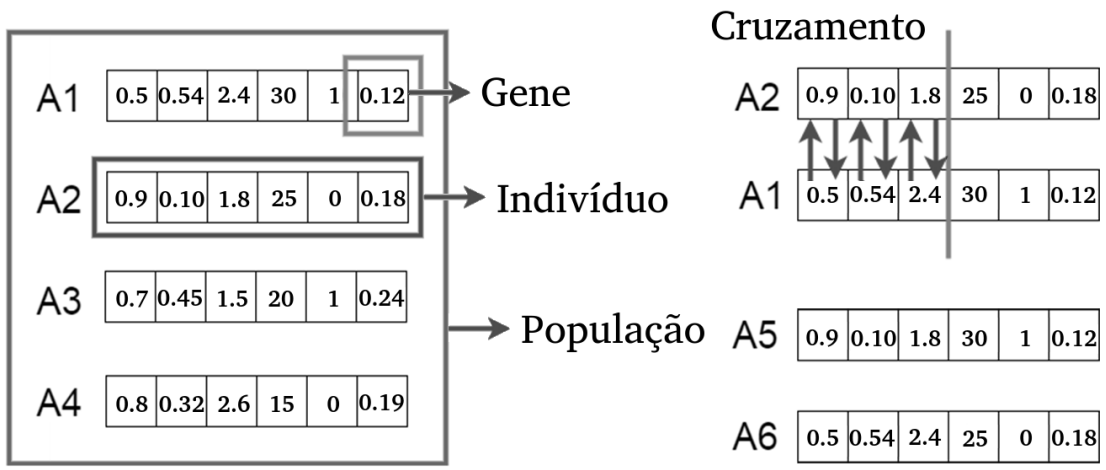


Figura 2.16: Representação do processo de cruzamento de valores de hiperparâmetros no AG

pesquisa aleatória que é seguida pela pesquisa orientada por resultados, com base nos melhores indivíduos anteriores (Young et al., 2008).

A Figura 2.16 representa o comportamento de um algoritmo genético simples combinando os valores de hiperparâmetros de cada indivíduo para encontrar a melhor combinação no processo comum, já explicado anteriormente, na Figura 2.7.

# Capítulo 3

# Trabalho relacionado na previsão de desempenho

A credibilidade de um trabalho de investigação verifica-se também pelas bases que o sustentam teoricamente. Um trabalho com qualidade deve ser criado e avaliado a partir de referências consistentes e que demonstrem resultados e contribuições relevantes para a comunidade acadêmica.

Este projeto de pesquisa faz referência a outros trabalhos acadêmicos e, ao comparar-se a esses referenciais , busca, humildemente, produzir contribuições para a previsão do desempenho de operadores de teleatendimento. Neste capítulo é feito um estudo comparativo com outros trabalhos acadêmicos que aplicam técnicas semelhantes àquelas que foram utilizadas neste projeto de pesquisa.

Na realização deste projeto de pesquisa, reafirma-se que outros trabalhos foram utilizados para embasamento científico com certos critérios; foram exploradas referências consideradas tecnicamente viáveis e, em particular, atuais, que demonstrassem resultados satisfatórios aplicados à resolução dos problemas propostos nos experimentos desta investigação.

Neste capítulo estão apresentados os trabalhos com experimentos análogos às técnicas utilizadas no presente projeto, cujos resultados foram bastante relevantes na avaliação científica.

Alguns destes trabalhos usaram modelos de classificação e outros fizeram testes com dados de absenteísmo e/ou produtividade. Uma parte destes trabalhos referenciaram também a utilização de grandes volumes de informação, tendo alguns um direcionamento específico para os dados de empresas de teleatendimento.

Tal como referido, antes neste texto (ver 1.1), este projeto de pesquisa centrou-se em dois aspeto principais:

- a fase de pré-processamento da informação;
- a etapa de obtenção de modelos de previsão de desempenho.

O foco da pesquisa bibliográfica realizada coincide, justamente, com este pipeline. No esforço de pesquisa bibliográfica procurou-se selecionar os trabalhos mais relevantes e recentes, dos últimos 10 anos, e que apresentassem similaridades com este projeto de pesquisa, tal como pode ser consultado na tabela de comparação 3.1.

## 3.1 Pré-processamento da informação

A fase de pré-processamento traz referências que tratam da preparação, organização e estruturação dos dados. Estes trabalhos, que servem como base de uma etapa tão fundamental, precedem a realização de análises de referências para as predições.

Os trabalhos que são referência para o pré-processamento estão realçados nessa secção. Esses trabalhos são apresentados conforme a sua contribuição para a construção da fase de pré-processamento do projeto final.

### 3.1.1 Volume de dados do projeto de pesquisa

O volume de dados do presente projeto de pesquisa é demasiado grande quando comparado à maioria das publicações, no meio acadêmico que tratam dos assuntos de ML. Mesmo assim, existem trabalhos que usam grandes volumes de dados e tratam as informações de diferentes formas, enfrentando desafios equivalentes.

O trabalho publicado por (Albiero et al., 2019) utiliza os dados de uma distribuidora de energia que contêm informações de consumo de cerca de 700 mil clientes entre fevereiro de 2014 e setembro de 2018. O total de variáveis utilizadas foi 64 e os dados foram amostrados em 2 conjuntos. O primeiro conjunto possuía 35 mil registros para treinamento e o segundo, conjunto de dados, sete mil registros para validação. Essa base de dados possui um volume consideravelmente grande se comparada à base de dados do presente projeto de pesquisa. O número de variáveis também chega às mesmas proporções do presente projeto de pesquisa.

A pesquisa apresentada por (Mohammed and Pang, 2011) trabalha com um volume de dados de, aproximadamente, 3 mil registros para tentar prever a distribuição de chamadas em um centro de teleatendimento. A previsão usa um conjunto de informações de chamadas de 40 dias entre as datas de 22/01/2008 a 01/03/2008; embora seja um volume de dados um pouco aquém do volume apresentado neste projeto de pesquisa que o referencia, representa relevância por estar inserido em uma problemática de negócio parecida.

O volume de dados analisado por (Valle and Ruz, 2015) foi uma amostra de 3543 registros de desempenho e atividade de vendas de operadores de teleatendimento dedicados a 91 campanhas de vendas de seguros e algumas poucas campanhas de venda de planos de celular entre maio de 2010 e novembro 2011. Assim como os dados de (Mohammed and Pang, 2011), é um volume um pouco aquém do volume apresentado neste projeto

de pesquisa que o referencia, mas é relevante por estar inserido em uma problemática de negócio parecida.

A referência de volume apresentada por (Punnoose and Ajit, 2016) é de uma população de 73.115 registros que representam equipes de liderança varejista por um período de 18 meses. O trabalho tenta prever demissões dentro de empresas usando ML e a população escolhida está distribuída em vários locais dos Estados Unidos. Essa metodologia se assemelha à distribuição de dados do projeto de pesquisa que o referencia, cujos dados foram coletados de várias unidades espalhadas pelas regiões do Brasil. Quando comparadas as duas bases de dados, o trabalho referenciado tem informações que vão além do utilizado no projeto que o referencia.

Os estudos de (Oliveira et al., 2019b), (Oliveira et al., 2019a) e (Oliveira et al., 2019c) são as principais referências do presente projeto de pesquisa e fazem uso do mesmo repositório de dados. São os trabalhos que deram origem à presente pesquisa que os referencia. A diferença de volume de informação é pequena entre os trabalhos por ocasião do período em que se faz a predição de cada um deles. Os trabalhos referenciados fazem previsões usando dados de 2017, refletindo também na quantidade de variáveis.

O trabalho de (Wang and Ni, 2019) possui um conjunto de dados fornecido por uma agência de crédito que contém mais de 10 milhões de informações comerciais de empresas nos Estados Unidos de 2006 a 2014. São utilizadas 305 variáveis independentes que fornecem informações das atividades das empresas em contas não financeiras, contas de telecomunicações e indústria. A proporção da base de dados é gigantesca quando comparada ao presente projeto de pesquisa que faz a referência. Por ser uma grande quantidade de dados, uma amostragem foi feita e os dados observados foram 8 mil.

No trabalho realizado por (Li et al., 2018) são simulados 2 cenários. O primeiro cenário tem uma geração aleatória de 10 mil registros e o segundo uma geração aleatória de 20 mil registros. Esses registros são atribuídos a um grupo de 23 operadores de telemarketing dos quais se pretende prever a performance. É um volume de dados interessante de ser referenciado, apesar de o número de atendentes ser muito menor do que os números apresentados neste presente projeto de pesquisa.

O volume de dados demonstrado no trabalho de (Wainer and Cawley, 2018) usa 9 conjuntos de dados com mais de 10 mil registros. Por ser uma grande quantidade de dados para aplicar os procedimentos do NCV, uma amostragem foi feita e os dados observados foram 5 mil registros. No projeto de pesquisa que faz referência ao trabalho de (Wainer and Cawley, 2018) também foram usadas as técnicas de NCV em um universo amostral.

A publicação de (Gamarra and Quintero, 2013) não tem um volume de registros expressivo, mas sim um volume de variáveis consideravelmente grande quando comparado ao volume de variáveis do projeto que faz essa referência. A seleção dessas variáveis é algo necessário para melhorar os resultados do trabalho e isso também se reflete no projeto de pesquisa.

### 3.1.2 Técnicas de seleção de variáveis

Selecionar variáveis para encontrar aquelas que dão resultados progressivamente bons aos projetos é algo aparentemente comum dentre os temas que abordam ML. Porém, existem técnicas que, quando combinadas, tornam-se específicas, fazendo com que as pesquisas acadêmicas muitas vezes tragam contribuições apenas na construção da solução. Isso significa que o resultado final do trabalho seria exclusivo, entretanto a forma como as variáveis foram selecionadas teria partes muito relevantes para outros projetos de pesquisa.

O estudo realizado por (Araujo et al., 2019) fez uso de uma seleção de variáveis aplicando o método Relief para verificar as características do banco de dados. A forma como essa técnica é usada se assemelha ao presente projeto por utilizar todas as variáveis do banco de dados, em um primeiro momento, e fazer a mesma avaliação, com os critérios selecionados pelo algoritmo Relief, em um segundo momento. Esse trabalho de referência usa uma rede *neuro-fuzzy* e a compara com outros modelos de regressão.

O trabalho de (Gamarra and Quintero, 2013) faz uma demonstração da utilização da seleção de variáveis com AG. O presente projeto de pesquisa faz uso da técnica de seleção de variáveis de modo similar quanto à avaliação de indivíduos. A função de avaliação passa pela execução dos modelos separadamente por indivíduo da população da mesma maneira que ele. A diferença entre os dois está na modelagem matemática do trabalho de referência: esse procura minimizar uma função, considerando a percentagem de erro, enquanto o presente projeto procura maximizar os valores encontrados pela métrica ROC AUC. As pesquisas defendidas por (Oliveira et al., 2019b), (Oliveira et al., 2019a) e (Oliveira et al., 2019c) usam as mesmas técnicas de seleção de variáveis usadas neste projeto de pesquisa. Esses trabalhos são a origem do projeto final apresentado como próprio projeto de pesquisa, logo, a forma como foram utilizadas as técnicas de seleção de variáveis são parecidas. A diferença entre os trabalhos está na quantidade de variáveis e nas variações dos parâmetros do AG.

O trabalho publicado por (Pandey and S.Taruna, 2014) usa uma técnica de seleção de variáveis com modelo DT. As variáveis ganham pontuações e algumas delas são rotuladas como nó raiz do **DF!** e os ramos aumentam para cada valor possível da variável em um processo recursivo. Essa é uma forma de dar valor às variáveis, comparando-as entre si para decidir quais delas serão selecionadas. Este projeto de pesquisa faz uso de técnicas semelhantes como *backward* e $forward$ que usam do ranqueamento das variáveis para selecioná-las da mais valorosa para a menos valorosa e vice-versa.

A análise feita por (Wang and Ni, 2019) utiliza cinco métodos de seleção de variáveis aplicados e avaliados: peso por índice gini (pontuação usada por (Pandey and S.Taruna, 2014)), peso por qui-quadrado, agrupamento de variáveis hierárquicas, peso por correlação e peso por proporção de ganho de informação. A comparação entre técnicas variadas segue a mesma ideia do projeto presente projeto de pesquisa, porém essas técnicas

são diferentes de modo geral. Apenas uma das técnicas é equivalente entre os dois trabalhos: o peso por correlação.

### 3.1.3 Técnicas de hiperparametrização

Projetos de pesquisa que incluem modelos de classificação são dependentes dos hiperparâmetros desses modelos. Essa dependência, para alcançar bons resultados, é o motivo dos projetos tentarem encontrar a melhor combinação dos hiperparâmetros por meio de diversas técnicas. Para encontrar a combinação, os trabalhos tratam dos seus problemas especificamente, mas também contribuem para construção de outras pesquisas.

Na pesquisa conduzida por (Araujo et al., 2019) houve modelos de rede neural hiperparametrizados nas camadas ocultas aleatoriamente. Os testes feitos utilizaram a técnica de pesquisa aleatória assim como no presente projeto de pesquisa. O SVM e o NB foram modelos usados na pesquisa de referência e na pesquisa referenciada, mas a evidência se relacionou ao modelo MLP que fez uso direto da técnica.

O trabalho publicado por (Kalantar et al., 2018) usou amostras selecionadas por um processo aleatório para selecionar valores para os hiperparâmetros do modelo SVM com o objetivo de melhorar o desempenho do modelo obtido. A abordagem usada por ele foi a validação cruzada que faz alusão às técnicas do projeto de pesquisa que aqui o referencia. O presente projeto de pesquisa se difere no uso da validação cruzada de modo mais complexo; é usado o NCV para avaliar a qualidade dos modelos, enquanto se alteram os seus parâmetros aleatoriamente entre as iterações.

A utilização de AG,técnica de hiperparametrização que serve como referência de (Mantovani et al., 2019), não é usada, especificamente, como no problema tratado no presente trabalho de pesquisa, uma vez que os hiperparâmetros são tratados em variáveis;todavia o autor,da mesma forma que o referenciado, combina técnicas com algoritmos evolucionários. O trabalho é específico para busca de parâmetros ótimos e propõe um sistema de recomendação de meta-aprendizagem para ajuste de hiperparâmetros do modelo SVM.

As demonstrações de técnicas de hiperparametrização de (Oliveira et al., 2019b), (Oliveira et al., 2019a) e (Oliveira et al., 2019c) são exatamente as técnicas utilizadas no presente projeto de pesquisa. Os trabalhos aqui referenciados, que usam as técnicas de AG e pesquisa aleatória para otimizar os hiperparâmetros dos modelos de classificação, foram validados e aceitos academicamente para valorizar neste projeto final.

Os fundamentos do trabalho de (Punnoose and Ajit, 2016) são semelhantes aos de (Kalantar et al., 2018). O autor utiliza um método de validação cruzada que consiste em dividir o conjunto total de dados em subconjuntos iguais. Enquanto um desses subconjuntos é usado como teste, os demais servem para estimativa de hiperparâmetros. Como já mencionado, o presente trabalho de pesquisa faz uso da técnica de validação cruzada

de uma forma mais complexa por meio do NCV.

As técnicas sugeridas por (Shah et al., 2020) são avaliadas como comuns para escolha de hiperparâmetros. Essas técnicas comuns são a pesquisa em grade e pesquisa aleatória. Para tal problema, que usa rede neural profunda, o tempo de execução é um fator de grande importância, e a pesquisa aleatória foi escolhida por economizar muito mais tempo na seleção dos hiperparâmetros otimizados.

A abordagem feita por (Wainer and Cawley, 2018) usa explicitamente o NCV, porém os procedimentos não são necessários, ao selecionar qualquer conjunto dos classificadores, desde que tenham um número limitado de hiperparâmetros que devem ser ajustados. A hiperparametrização usa a técnica de NCV do mesmo modo que o projeto de pesquisa, no entanto o resultado é baseado no erro e não na métrica de classificação.

O estudo realizado por (Wang and Ni, 2019) utilizou duas formas de hiperparametrização: as técnicas de estrutura de árvore Bayesiana e a pesquisa aleatória. A pesquisa aleatória já é uma técnica conhecida pelo presente trabalho de pesquisa, mas a otimização Bayesiana não foi tratada nesse trabalho e é uma abordagem baseada em probabilidade que usa o modelo de probabilidade para selecionar os hiperparâmetros mais promissores. Os resultado para a otimização Bayesiana foram superiores aos resultados da pesquisa aleatória.

## 3.2 Modelos de previsão de desempenho

As referências da fase de previsão de desempenho, abordadas nesse projeto acadêmico, como previsão de absenteísmo e produtividade, são acerca dos trabalhos baseados no processamento da informação, os quais expressam o momento de realização de análises e execução das predições.

O processamento de dados busca embasamento em várias fontes; algumas das citadas, nessa seção, são essenciais para identificar oportunidades de negócio e tratar problemas de previsão de desempenho.

### 3.2.1 Problema do absenteísmo

Existem referências de trabalhos que tratam o absenteísmo, muitas vezes como algo inerente a fatores externos e não fazem uma interdependência com características das próprias empresas empregadoras que não retêm, cordialmente, seus funcionários.

Em (Araujo et al., 2019), o autor infere que o absenteísmo não tem uma causa específica e sim um conjunto de causas que podem desencadear esse comportamento. O trabalho dele evidencia o problema de absenteísmo em uma empresa brasileira voltada para o setor de saúde e nutrição. Interessa destacar a similaridade, em termos de nacionalidade, com o problema tratado por este trabalho de pesquisa.

A pesquisa apresentada por (Martiniano and Sassi, 2012), abordagem da mesma nacionalidade do problema da pesquisa tratada nessa tese, trata da previsão do absenteísmo dos funcionários das agências dos Correios do Brasil e foi feita como a de (Araujo et al., 2019). O trabalho foca em especificidades da saúde física dos funcionários, mesmo entendendo que existem outros fatores que podem influenciar a falta ao trabalho, além de doenças.

Os trabalhos de (Oliveira et al., 2019b) e (Oliveira et al., 2019c) dão enfoque ao problema do absenteísmo na mesma empresa alvo da pesquisa e têm o mesmo ambiente para aplicação das técnicas de ML. Esses dois trabalhos possuem menos detalhes sobre o cenário de aplicação, quando comparados à pesquisa tratada nessa tese, mas expõem o tema de forma clara e objetiva para avaliação acadêmica.

O trabalho publicado por (Shah et al., 2020) aborda a previsão do absenteísmo dos funcionários das agências dos Correios no Brasil assim como (Martiniano and Sassi, 2012), no entanto, vai além do absenteísmo causado por doenças físicas, visto que é mais abrangente e se preocupa com fatores secionais, culturais e características socioeducacionais ligadas aos funcionários. O presente trabalho de pesquisa tem essa mesma ideia de abrangência de pesquisa e define o absenteísmo final de acordo com as horas de trabalho que funcionário esteve ausente, que também é feito no trabalho referenciado.

### 3.2.2 Problema da produtividade

A manutenção da produtividade dos colaboradores de uma empresa é um dos aspectos abordados pelos trabalhos acadêmicos referenciados nessa subsecção. A abordagem por esses trabalhos ocorre, porque a produtividade depende fortemente de fatores externos e acaba sendo complexa de ser desenvolvida, havendo a necessidade do estudo de técnicas de ML como solução.

O estudo de (Ahmed et al., 2016) descreve, detalhadamente, as decisões tomadas na construção de um sistema árabe de caracteres para avaliação e análise de sentimentos de funcionários para medir a sua produtividade. É uma metodologia que difere da previsão da produtividade do presente trabalho de pesquisa, embora busque, em um mesmo contexto, o mesmo resultado ao final. A medição da produtividade nos trabalhos não são semelhantes na prática, mas o conceito de unir fatores produtivos para compor a produtividade e quantificá-la está presente tanto no trabalho referenciado quanto no presente projeto de pesquisa.

Os trabalhos de (Oliveira et al., 2019a) e (Oliveira et al., 2019c) fazem uma análise do problema de previsão de produtividade na mesma empresa alvo da pesquisa e tem o mesmo ambiente para aplicação das técnicas de ML. Esses dois trabalhos possuem menos detalhes sobre o cenário de aplicação quando comparados à pesquisa tratada nessa tese, mas expõem o tema de forma clara e objetiva para avaliação acadêmica.

(Pandey and S.Taruna, 2014) realizou um trabalho de previsão da performance de estudantes acadêmicos para a qual considerou-se uma variedade de informações dos estudantes e, algumas delas, pessoais, como idade e região; características comuns às dos funcionários usados pelo presente trabalho de pesquisa. A intenção ao prever a produtividade seria fazer uma análise de desempenho dos alunos para que a orientação acadêmica correta pudesse ser dada em relação ao progresso, no curso, e os instrutores pudessem avaliar seu processo de ensino-aprendizagem. São intenções iguais para os dois trabalhos, este projeto de pesquisa e o trabalho referenciado; porém, cenários diferentes e com atores, também diferentes.

A abordagem feita por (Li et al., 2018) é voltada para a previsão de performance em centros de teleatendimento assim como o presente trabalho que o referencia. É um problema que revela resultados diferentes dos resultados do presente trabalho de pesquisa, porque faz uma previsão dentro de um processo de regressão. Este artigo apresentou uma abordagem de ML para avaliar programações de equipe de teleatendimento com base em uma equipe integral e tratou-se, também, de escala de turno de trabalho.

No trabalho realizado por (Valle et al., 2012) propôs-se a previsão da performance em centros de teleatendimento usando um modelo de classificação avaliado como parâmetro de comparação de outros modelos pelo presente projeto de pesquisa, o modelo de classificação NB. O trabalho não atuou sobre agentes de teleatendimento, de modo geral, mas sim naqueles que exercem funções de atendimento promocional. As características dos funcionários, que foram consideradas no estudo, foram comuns às do trabalho tratado nesta tese, porém em menor quantidade. As variáveis de indicadores de performance principais estão presentes em ambos os trabalhos, este projeto de pesquisa e o trabalho referenciado, e observou-se que é um trabalho menos complexo e bastante semelhante ao trabalho que o referencia, trazendo resultados interessantes quanto à classificação.

## 3.3 Análise comparativa

As referências analisadas para este projeto de pesquisa foram selecionadas considerando o contexto atual de ML. Um fator que merece destaque, por ocasião dessa especificidade de seleção, é o uso de classificadores similares aos usados no projeto de pesquisa. Em todas as referências apresentadas, neste capítulo, é feito o uso de algum dos modelos de classificação. De modo geral, as características dos trabalhos referenciados podem ser vistas na Tabela 3.1.

A Tabela 3.1 apresenta a comparação entre todos os trabalhos de referência e o que eles trouxeram de contribuição para este projeto de pesquisa defendido sobre previsão de desempenho. Apesar da maior similaridade entre os trabalhos ser relacionada aos modelos de classificação, existe um certo equilíbrio em relação aos trabalhos enquadrados nos outros itens técnicos, conforme contribuição.

| Trabalhos de referência | Pré-processamento | | | Previsão | | Contexto em tele-atendimento |
|---|---|---|---|---|---|---|
| | Volume de dados | Seleção de variáveis | hiper-parametrização | Absenteísmo | Produtividade | |
| (Ahmed et al., 2016) | | | | | ✓ | ✓ |
| (Albiero et al., 2019) | ✓ | | | | | |
| (Araujo et al., 2019) | | ✓ | ✓ | ✓ | | |
| (Valle and Ruz, 2015) | ✓ | | | | | ✓ |
| (Gamarra and Quintero, 2013) | ✓ | ✓ | | | | |
| (Kalantar et al., 2018) | | | ✓ | | | |
| (Mantovani et al., 2019) | | | ✓ | | | |
| (Martiniano and Sassi, 2012) | | | | ✓ | | |
| (Mohammed and Pang, 2011) | ✓ | | | | | ✓ |
| (Oliveira et al., 2019b) | ✓ | ✓ | ✓ | ✓ | | ✓ |
| (Oliveira et al., 2019a) | ✓ | ✓ | ✓ | | ✓ | ✓ |
| (Oliveira et al., 2019c) | ✓ | ✓ | ✓ | ✓ | ✓ | ✓ |
| (Pandey and S.Taruna, 2014) | | ✓ | | | ✓ | |
| (Punnoose and Ajit, 2016) | ✓ | | ✓ | | | |
| (Shah et al., 2020) | | | ✓ | ✓ | | |
| (Li et al., 2018) | ✓ | | | | ✓ | ✓ |
| (Valle et al., 2012) | | | | | ✓ | ✓ |
| (Wainer and Cawley, 2018) | ✓ | | ✓ | | | |
| (Wang and Ni, 2019) | ✓ | ✓ | ✓ | | | |

Tabela 3.1: Conteúdo com o qual cada trabalho contribuiu para o presente projeto

Pesquisas que trabalham com grandes volumes de dados se destacaram juntamente com técnicas de hiperparametrização. Isso era esperado como consequência da inclusão de trabalhos com modelos de classificação dependentes de hiperparâmetros e a abordagem de problemas com alta disponibilidade de informação.

Observa-se que não são tantos os trabalhos selecionados que contribuíram com o tema absenteísmo e produtividade, mas a união de todas as outras partes contribuintes, junto à parte do tema, compuseram o todo deste projeto de pesquisa.

Todos os projetos de referência contribuíram para que a interseção das suas técnicas chegassem ao projeto de pesquisa final aqui apresentado. A Figura 3.1 simboliza a interseção entre os 3 grandes temas que ajudam a compor o projeto final de pesquisa.

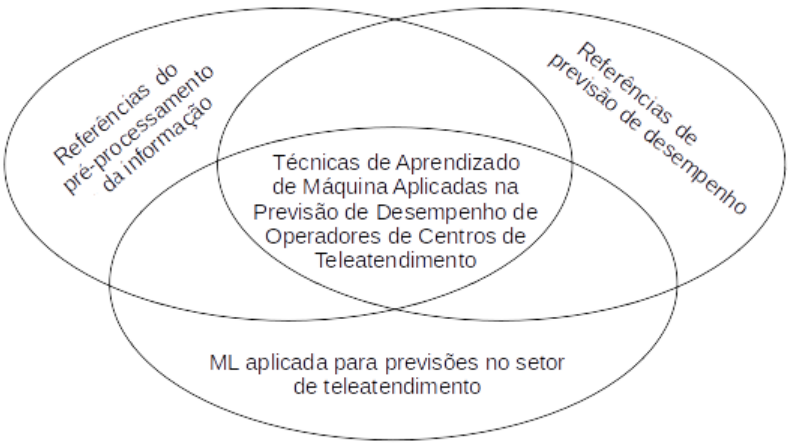


Figura 3.1: Interseção entre as áreas dos trabalhos relacionados referenciados

Algumas técnicas foram adaptadas no pré-processamento: a previsão de desempenho

foi desenvolvida voltada para cada um dos modelos de classificação e a aplicação da previsão foi específica para as características do setor de teleatendimento. Mesmo assim, foi possível, a partir das referências aqui apresentadas, compor um trabalho que tem potencial de uso para diversas áreas de negócio.

# Capítulo 4

# Processo de previsão de desempenho baseada em ML

## 4.1 Estrutura da informação

Nesta secção é apresentada a estrutura do conjunto de dados utilizado na previsão. Aqui, serão descritas as variáveis utilizadas como entrada e saída da previsão, a estrutura temporal do conjunto de dados e a representatividade desses registros dentro do conjunto.

### 4.1.1 Variáveis do modelo de produtividade

Nos serviços de atendimento, existem regras e medidores de produtividade que são particulares de cada empresa nesse seguimento do mercado. Sabendo disso, as informações de rendimento profissional dos funcionários da empresa foram medidas a partir de indicadores de produtividade. Esses indicadores básicos são pontuadores de performance do operador durante sua jornada de trabalho diária, semanal e mensal.

A seguir, são descritos os indicadores de maior relevância no estudo:

- Rechamada (variável $x_{16}$ listada no anexo A): são as ligações de clientes que retornam para o serviço de atendimento em busca de resposta para problemas apresentados, anteriormente, e não solucionados.

- Tempo médio de atendimento (variável $x_{18}$ listada no anexo A): são os minutos falados, no atendimento, divididos pelo número de ligações atendidas. Esse indicador está diretamente relacionado ao grau de complexidade dos processos que envolvem o serviço ou produto abordado no atendimento, à capacitação dos atendentes e à sua curva de aprendizado.

- Pausa nas atividades de trabalho (variável $x_{15}$ listada no anexo A): podem ser pausas produtivas ou pausas improdutivas. As pausas produtivas são geradas em função de

processo de aperfeiçoamento operacional; em geral, estas pausas correspondem a treinamentos e reuniões. As pausas improdutivas são pausas para lanche, descanso, banheiro, ambulatório ou atividades operacionais extra atendimento.

- Absenteísmo (variável $x_2$ listada no anexo A): é o percentual de não comparecimento e não cumprimento de escala expurgando-se as pausas produtivas. Na empresa alvo do estudo, o percentual utilizado é de 50% do total de horas trabalhadas em 1 dia. Fatores que influenciam diretamente esse indicador são: greve nos transportes públicos; problemas pertinentes à função, como ler e falar; problemas de saúde; ausências validadas por trabalho eleitoral, casamento do funcionário, nascimento de filhos; comprometimento do colaborador com a empresa; acompanhamento dos chefes gestores.

- Tempo logado (variável $x_{17}$ listada no anexo A): é o percentual de tempo conectado em relação ao tempo de disponibilidade esperado do agente de atendimento.

- Resultados consolidados de produtividade(variável $x_1$ listada no anexo A): o indicador de resultados consolidados representa a quantificação do trabalho dos operadores dentro do sistema de grupos de produtividade. Esse indicador representa valores obtidos, a partir de desafios criados pelos gestores, que são bonificados financeiramente com moedas virtuais de acordo com o desempenho. Para esse indicador, a empresa possui um sistema baseado em redes sociais no qual o funcionário pode visualizar os resultados dos seus indicadores e a consolidação desses resultados como moedas virtuais. Essas moedas são utilizadas para compras em uma loja virtual da própria plataforma e, também, são utilizadas para classificar o operador quanto à produtividade.

A plataforma virtual é utilizada pelos funcionários para interagir com os demais operadores de teleatendimento, melhorar o próprio desempenho e contribuir com o desempenho de seus colegas de trabalho através de sugestões e dicas de procedimentos no setor de atendimento. Esse sistema disponibiliza aos chefes gestores gráficos de rendimento e mensagens que apontam os funcionários que estão evoluindo ou regredindo nos indicadores. O sistema também disponibiliza uma secção para marcação do estado de humor, motivo para um mau humor e relatórios aos gestores sobre esse tema.

Os dados de relatórios e demais informações presentes, na plataforma social, referentes a humor ou desempenho, ficam disponíveis para os funcionários por um período de 12 semanas. Isso acontece por questões legislativas e, coincidentemente, esse é o perfil de troca de, aproximadamente, 25% do quadro de funcionários da empresa periodicamente. Por esse motivo, a largura da janela temporal deslizante, usada para alimentar os modelos criados, é de 12 semanas para treinamento e outras 4 semanas para teste conforme a técnica NCV. Os modelos usados têm uma periodicidade semanal. As 4 semanas de

testes aplicados na estrutura participaram ativamente dos processos NCV na composição de dados acumulados das etapas que serão explicadas na secção 4.2 e na secção 4.3. As 12 semanas de treinamento completam o conjunto acumulado de dados das 4 semanas de testes.

A empresa alvo do estudo possui aproximadamente 18 mil profissionais em seu quadro. Desses 18 mil, aproximadamente, 10,5 mil são do primeiro nível da escala hierárquica da empresa, os agentes de teleatendimento. Para organizar essa quantidade de funcionários, conforme seu desempenho, a empresa criou grupos de 1 a 4, sendo que o grupo 1 é composto pelos funcionários com melhores resultados e o grupo 4 é o grupo dos funcionários com baixa performance.

A divisão dos funcionários da empresa em grupos foi estabelecida de acordo com os resultado nos indicadores. O atingimento de uma meta estabelecida pelo gestor nesses indicadores determina se o funcionário é ou não produtivo.

### 4.1.2 Variáveis do perfil dos funcionários

Para compor o perfil dos funcionários, foi necessário realizar uma pesquisa nos dados disponíveis na plataforma social da empresa e nos registros do setor de recursos humanos. Essa pesquisa possibilitou a realização da previsão de indicadores de desempenho a partir de um histórico de dados.

As variáveis encontradas têm dois tipos de comportamento na escala temporal: variáveis com informações que não se alteram em curtos períodos semanais e variáveis que se alteram à medida que esses períodos avançam. A lista das variáveis utilizadas no trabalho e suas respectivas descrições podem ser consultadas no Anexo A e na Tabela 4.1.

De acordo com (Lohelin, 1998), existem dois tipos de variáveis: as variáveis do tipo endógenas e as variáveis do tipo exógenas. Os valores das variáveis endógenas são explicados por uma ou mais variáveis exógenas do modelo. Os valores das variáveis exógenas são assumidos como dados comuns, isto é, o modelo não tenta explicá-los.

Considerando a previsão de absenteísmo ou produtividade, exemplos de variáveis exógenas utilizadas neste projeto são "idade" (variável $x_{43}$ listada no anexo A), "sexo" (variável $x_{143}$ listada no anexo A) e "estado civil" (variável $x_{23}$ listada no anexo A). Essas são variáveis que não têm uma explicação por parte do absenteísmo ou produtividade. As variáveis endógenas utilizadas são "percentual de absenteísmo" (variáveis $x_{74}$ a $x_{81}$ listadas no anexo A) e "resultado de produtividade" (variáveis $x_{128}$ a $x_{135}$ listadas no anexo A) como na Expressão 4.1, ambas para períodos de 1 a 8. Essas são variáveis explicadas pela

Tabela 4.1: Variáveis gerais dos operadores de teleatendimento

| **Variáveis Pessoais** | | **Variáveis relacionadas ao ambiente de trabalho** | | **Variáveis extraídas da plataforma social** | | **Variáveis endógenas** | | **Total** |
|---|---|---|---|---|---|---|---|---|
| **Variável** | **Qtd** | **Variável** | **Qtd** | **Variável** | **Qtd** | **Variável** | **Qtd** | |
| autoavaliação | 37 | avaliação de ambiente e sobre o gestor | 18 | absenteísmo | 2 | grupo de produtividade histórico | 1 | |
| cadastro de cidadão na região natal | 1 | carga horária | 2 | dias de acesso ao sistema de gestão de resultados | 1 | grupo de resultados bonificados no período | 8 | |
| código telefônico | 27 | cidade de trabalho | 9 | grupo em que ficou maior tempo | 4 | percentual de absenteísmo no período | 8 | |
| escolaridade | 8 | cidade local de trabalho | 1 | horário de marcação do humor | 1 | percentual de bonificação no período | 8 | |
| estado civil | 3 | distancia do trabalho | 3 | humor | 10 | resultado de produtividade no período | 8 | |
| estado de nascimento | 27 | faixa etária do gestor | 4 | humor do gestor | 2 | | | |
| faixa etária | 3 | feriado no período | 2 | maior tempo consecutivo no grupo produtivo | 1 | | | |
| idade | 1 | histórico absenteísmo | 2 | média de avaliações de dicas | 2 | | | |
| imigrante | 1 | histórico do número de dias desde a última falta | 2 | média percentual de absenteísmo dos amigos | 2 | | | |
| naturalidade capital | 1 | histórico percentual de atraso no feriado | 2 | número de compras | 4 | | | |
| número de dependentes | 1 | histórico percentual de faltas consecutivas e não consecutivas | 3 | número de dicas avaliadas e recebidas | 4 | | | |
| região registro nacional de cidadão | 27 | histórico percentual de faltas no feriado | 2 | número de mensagens | 4 | | | |
| sexo | 1 | horas trabalhado | 1 | número de produtos favoritados | 3 | | | |
| telefone fixo | 1 | idade do gestor | 2 | número de trocas de perfil | 2 | | | |
| | | número de colaboradores no departamento | 2 | produtividade | 2 | | | |
| | | número de departamentos no período | 2 | situação nos indicadores principais | 20 | | | |
| | | número de dias desde o último descanso | 2 | tamanho do login | 1 | | | |
| | | número de domingo trabalhados | 2 | total de vezes nos grupos de produtividade | 4 | | | |
| | | número de gestores | 2 | | | | | |
| | | percentual de absenteísmo dos gestores no departamento | 2 | | | | | |
| | | percentual de atingimento no indicador de bonificação dos amigos | 2 | | | | | |
| | | semana do mês | 4 | | | | | |
| | | sexo gestor | 2 | | | | | |
| | | tempo de empresa | 1 | | | | | |
| | | tempo de empresa do gestor | 2 | | | | | |
| **Total** | 139 | | 76 | | 69 | | 33 | 317 |

variável absenteísmo ou pela variável produtividade que se quer prever.

$$\begin{aligned} &x_{74} = y_{t-1}, x_{75} = y_{t-2}, \ \ldots \ , x_{81} = y_{t-8} \\ &\text{se } y = \text{ percentual de absenteísmo} \\ &x_{128} = y_{t-1}, x_{129} = y_{t-2}, \ \ldots \ , x_{135} = y_{t-8} \\ &\text{se } y = \text{ resultado de produtividade} \end{aligned} \tag{4.1}$$

Neste trabalho, as variáveis exógenas são tratadas como dados comuns e compõem o objeto de dados junto às variáveis endógenas tratadas como janelas temporais ou observações do passado. Os objetos de dados são entradas independentes que representam os dados referentes a uma semana e cada funcionário contribui para a entrada de informações de 4 semanas, ou seja, cada um dos, aproximadamente, 10,5 mil funcionários dão origem a 4 objetos de dados de entrada independentes.

As variáveis exógenas dos objetos de dados se alteram conforme o estado do funcionário na semana. Como exemplo, pode-se citar a variável "estado civil", na qual um funcionário pode estar solteiro em uma semana e casado na semana seguinte. Já as variáveis endógenas, de janela temporal, são os resultados gerados de 8 semanas anteriores à previsão. Se a variável de saída é o absenteísmo, por exemplo, observa-se o absenteísmo nas 8 semanas anteriores para se tornarem variáveis de entrada.

A abordagem adotada para o problema estudado, neste trabalho, foi o modelo de classificação. Essa estratégia foi escolhida, em caráter experimental, e adaptada, considerando-se as características do problema.

### 4.1.3 Abordagem das observações do passado

Na lista de variáveis do projeto, existem variáveis diretamente ligadas à variável classe (cf. produtividade ou absenteísmo) e que possuem uma relação temporal com o período de previsão. Essas variáveis são candidatas a serem variáveis endógenas de janelas temporais.

As informações coletadas, relacionadas à variável classe, consideram os dados dos últimos 8 períodos semanais dos funcionários. Para um exemplo específico com a variável "absenteísmo", foram coletados os resultados dos 8 últimos períodos de 7 dias e cada uma dessas informações transformadas em uma variável: "percentual de absenteísmo período 1" ao "percentual de absenteísmo período 8" (variáveis $x_{74}$ a $x_{81}$ listadas no anexo A).

A distribuição das variáveis para 3 funcionários dentre os 10,5 mil funcionários da empresa é apresentado no exemplo da Figura 4.1. Nesse exemplo, são consideradas apenas 10 variáveis, sendo 8 delas observações do passado: "percentual de absenteísmo período 1" ao "percentual de absenteísmo período 8" (variáveis $x_{74}$ a $x_{81}$ listadas no anexo A) representados por $y_{t-1}$ ao $y_{t-8}$. As outras 2 variáveis, "sexo" (variável $x_{143}$ listada no anexo A) e "casado", são representações de variáveis exógenas que podem se alterar de uma semana para outra.

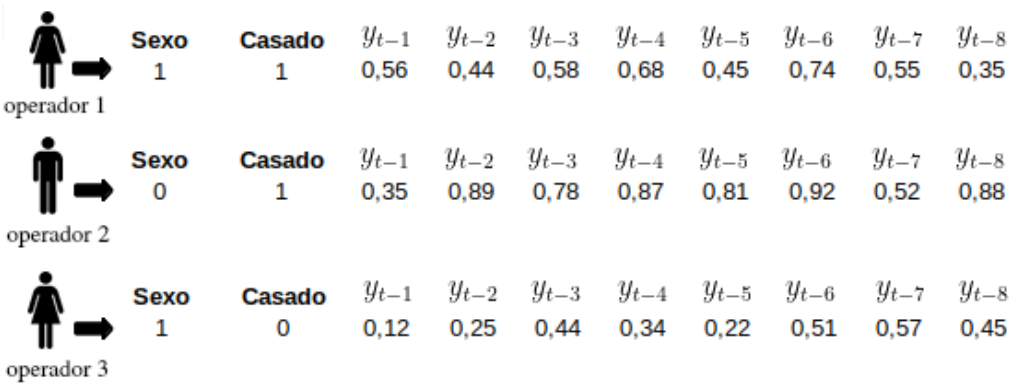


Figura 4.1: Estrutura de informações dos funcionários

A distribuição das informações das variáveis na Figura 4.1 representa 3 funcionários operadores de teleatendimento com dados do período de uma semana. Na Figura 4.2, um desses 3 funcionários, o "operador 1", foi isolado e foi apresentada a estrutura de 8 períodos *t*.

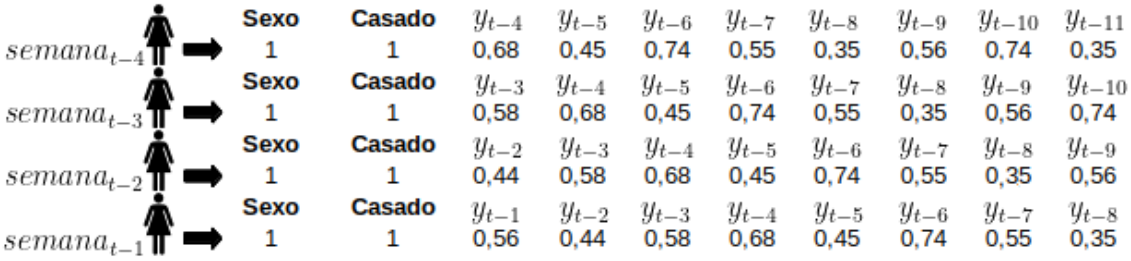


Figura 4.2: Estrutura de informações do funcionário "operador 1" da Fig. 4.1

## 4.2 Proposta de seleção de variáveis

No presente trabalho, para a seleção de variáveis aptas a resolver os problemas de previsão de desempenho dos operadores de teleatendimento, foram utilizadas algumas técnicas isoladas e, posteriormente, combinadas. A proposta inicial foi utilizar todas as variáveis como entrada dos modelos de classificação para gerar o valor da métrica ROC AUC. Esse valor ROC AUC foi a base de comparação dos modelos de classificação.

### 4.2.1 Métrica ROC AUC

As curvas ROC estão entre as métricas mais utilizadas para a avaliação de um modelo de previsão e a métrica AUC é derivada da curva ROC. A métrica ROC AUC mostra o quão bom o modelo criado pode distinguir entre valores positivos e negativos, com dois parâmetros (Bradley, 1997):

- Taxa de verdadeiro positivo, que é dado por "verdadeiros positivos / (verdadeiros positivos + falsos negativos)"
- Taxa de falso positivo, que é dado por "falsos positivos / (falsos positivos + verdadeiros negativos)"

Uma curva ROC traça a "taxa de verdadeiros positivos vs. taxa de verdadeiros negativos" em diferentes limiares de classificação. A análise da ROC AUC seria a curva ROC, em um único valor, agregando todos os limiares da ROC, calculando a "área sob a curva" (Bradley, 1997).

O valor do ROC AUC varia de 0,0 até 1,0 e o limiar entre a classe é 0,5. Acima desse limite, o algoritmo classifica em uma classe; e abaixo desse limite, na outra classe.

### 4.2.2 Aplicação de métodos de seleção de variáveis

Para avaliar os modelos LR, LSTM, MLP, NB, RF, SVM e XGBoost, cada um desses modelos foi submetido à primeira carga de informações contendo todas as variáveis.

Foi explicado, de forma sucinta na secção 4.1.1, que o conjunto de dados de treinamento é composto por 12 semanas e esses dados de treinamento são acumulados às 4 semanas usadas para teste quando executadas as etapas do processo NCV.

No trabalho foram submetidas 12 semanas de informações para treinamento seguindo cálculo amostral (ver secção 2.2.1.1) e 1 semana foi utilizada para teste dentre as 4 semanas de teste citadas. Essa semana submetida ao teste é a semana seguinte às 12 semanas na escala temporal (13ª semana de teste). Os valores de entrada, contendo a primeira parte das informações, foram submetidos ao modelo e, em seguida, os dados foram treinados. Após esse treinamento, os dados foram testados seguindo os critérios da técnica NCV e obteve-se o valor ROC AUC para a primeira etapa.

Na segunda etapa, os valores de 13 semanas de dados foram submetidos para treinamento seguindo cálculo amostral e 1 semana foi utilizada para teste (14ª semana de teste). Essas 13 semanas de treinamento são compostas pelas 12 semanas de treinamento da primeira etapa e pela semana 13 que havia sido utilizada para teste. Nessa etapa, assim como na anterior, foram seguidos os mesmos critérios da técnica NCV e obteve-se o valor ROC AUC para a segunda etapa.

Para a terceira etapa, os valores de 14 semanas de dados foram submetidos para treinamento seguindo cálculo amostral e 1 semana foi utilizada para teste (15ª semana de teste). Essas 14 semanas de treinamento são compostas pelas 13 semanas de treinamento da segunda etapa e pela semana 14 que havia sido utilizada para teste. Nessa etapa, assim como na anterior, foram seguidos os mesmo critérios presentes na técnica NCV e obteve-se o valor ROC AUC para a terceira etapa.

Na última etapa, foram submetidos os valores de 15 semanas de dados para treinamento seguindo cálculo amostral e 1 semana foi utilizada para teste (16ª semana de teste).

Essas 15 semanas de treinamento são compostas pelas 14 semanas de treinamento da terceira etapa e pela semana 15 que havia sido utilizada para teste. Nessa etapa, foram seguidos os mesmos procedimentos do NCV das etapas anteriores e obteve-se o valor ROC AUC para a quarta etapa.

As 4 semanas de teste ($13^a$, $14^a$, $15^a$ e $16^a$ semanas de teste) foram utilizadas em cada uma das etapas sequencialmente para testes e, também, compuseram os dados de treinamento. No final desse processo, foi calculada a média dos valores ROC AUC das etapas que seguiram a técnica NCV para obtenção do ROC AUC final, como apresentado na secção 2.2.1.3.

No primeiro procedimento de testes do trabalho, todas as variáveis foram submetidas e executadas conforme as etapas citadas. As demais avaliações como *Forward* e *Backward* seguem os mesmos procedimentos e padrões de repetição, mas com diferenças no número de variáveis de entrada.

#### 4.2.2.1 Aplicação de *Backward* e *Forward*

- *Forward* - Na aplicação da técnica *forward*, as variáveis são submetidas ao modelo uma a uma e, a cada vez que uma variável é adicionada, todo o processo NCV é executado para avaliação. O modelo se inicia com a primeira carga de informações da variável mais correlacionada com a variável de saída até chegar à variável menos correlacionada com a variável de saída. As variáveis foram ordenadas de acordo com os valores encontrados pelas técnicas *Pearson* ou *Relief*. A média ROC AUC obtida pelo processo NCV determina se uma variável presente no modelo torna os resultados da previsão melhores ou piores. Se uma variável submetida ao modelo tem uma métrica menor que o valor ROC AUC corrente, ela não é mantida. Se uma variável for submetida ao modelo e o valor ROC AUC for maior que o valor ROC AUC corrente, permanecerá no modelo.

- *Backward* - Na aplicação da técnica *backward*, todas as variáveis são submetidas ao modelo e, posteriormente, são retiradas uma a uma da coleção de variáveis de entrada, criando novas coleções a serem submetidas. A cada retirada de variável da coleção de entrada do modelo, todo o processo NCV é executado para avaliação. O modelo é iniciado com a primeira carga de informações contendo todas as variáveis, e são retiradas as variáveis menos relacionadas com a variável classe seguindo a ordem de retirada até chegar à variável mais relacionada com a variável classe. A ordenação é feita com base nos valores obtidos com as técnicas *Pearson* ou *Relief*. A média ROC AUC determina se uma variável presente na coleção de variáveis de entrada tornam o modelo melhor ou pior. Se uma variável é retirada e o valor da métrica é menor que a média ROC AUC corrente, ela retorna à coleção. Caso uma variável seja retirada da coleção e o valor da métrica seja maior que o

valor ROC AUC corrente, ela não permanece na coleção de variáveis de entrada do modelo.

#### 4.2.2.2 Aplicação do AG na seleção de variáveis

Uma das seleções de variáveis aplicadas no trabalho ocorre por Algoritmo Genético (AG). Na técnica utilizada, cada indivíduo da população foi composto por todas as variáveis com valores, verdadeiros ou falsos, indicando se essa variável, ou gene, será incluída ao modelo ou não. A escolha do valor, verdadeiro ou falso, para cada um dos genes foi de forma aleatória, mas garante-se que, na primeira geração, haja um indivíduo com todos os genes verdadeiros, um indivíduo com todos os genes falsos e a variável mais correlacionada com a variável de saída verdadeira. A forma de escolha dos valores de entrada e saída das variáveis, na primeira geração, também é condicionada a manter sempre a metade dos genes do indivíduo como verdadeiros e a outra metade como falsos.

No início do processo do AG, são consideradas 80 gerações e 34 indivíduos para compor a população. Essas gerações são escolhidas conforme o critério de parada do algoritmo já citado, anteriormente, na secção 2.2.2.4. A utilização dos valores de 80 gerações e 34 indivíduos foi motivada pela avaliação ROC AUC dos indivíduos que se estabiliza nesses números testados de forma empírica.

Cada indivíduo recebe uma nota de avaliação ROC AUC conforme as suas variáveis de entrada no modelo. O processo utilizado para dar notas aos indivíduos é o NCV. Em cada geração, os indivíduos que ainda não possuem uma nota são submetidos ao processo. Esses indivíduos submetem todos os seus genes a um dos modelos passando pelas etapas do NCV para obtenção da métrica.

Assim que todos os indivíduos da população passam pelo processo do NCV, eles são ordenados de forma decrescente conforme os valores obtidos. Desses indivíduos, os 50% que possuem as melhores notas são mantidos na população para a próxima geração. Nos outros 50% restantes é feita uma seleção aleatória em que cada indivíduo tem 20% de chance de se manter na população para a próxima geração. Quando a população já está composta pelos indivíduos com as melhores notas e pelos indivíduos que entraram por escolha aleatória, as vagas restantes são preenchidas pelos indivíduos resultantes do cruzamento entre os indivíduos atuais da população.

Na etapa de cruzamento do AG, são escolhidos, aleatoriamente, dois indivíduos diferentes da população, e os genes desses indivíduos são combinados para gerar um novo indivíduo. São escolhidos, aleatoriamente, metade dos genes do indivíduo mãe e metade dos genes do indivíduo pai, e o resultado é o indivíduo filho com os genes combinados. A outra metade dos genes da mãe e do pai, que não foram utilizados no cruzamento, foram combinados para gerar um outro filho. Esse procedimento possibilita que cada cruzamento dê origem a dois novos indivíduos. Se as vagas na população ainda não estiverem preenchidas para a próxima geração, uma nova escolha de pais é feita, e um novo

cruzamento é executado.

A taxa de mutação do AG foi fixada em 0,5%. Esse valor representa a possibilidade de acontecer uma mutação em um indivíduo qualquer dentro da população em cada geração. Esse indivíduo escolhido tem todos os seus genes alterados aleatoriamente.

Para utilizar as variáveis selecionadas pelas técnicas *Backward*, *Forward* e AG combinadas (Fluxo 6 da Figura 4.3), a primeira geração de uma nova execução do AG foi composta pelos seguintes indivíduos, além dos indivíduos gerados aleatoriamente:

- um indivíduo com as variáveis selecionadas pelo *Forward* e as variáveis ordenadas por *Relief* nesse processo;
- um indivíduo com as variáveis selecionadas pelo *Forward* e as variáveis ordenadas por *Pearson* nesse processo;
- um indivíduo com as variáveis selecionadas pelo *Backward* e as variáveis ordenadas por *Relief* nesse processo;
- um indivíduo com as variáveis selecionadas pelo *Backward* e as variáveis ordenadas por *Pearson* nesse processo;
- um indivíduo com melhor ROC AUC da última execução do AG;
- um indivíduo com os genes configurados para manter todas as variáveis.

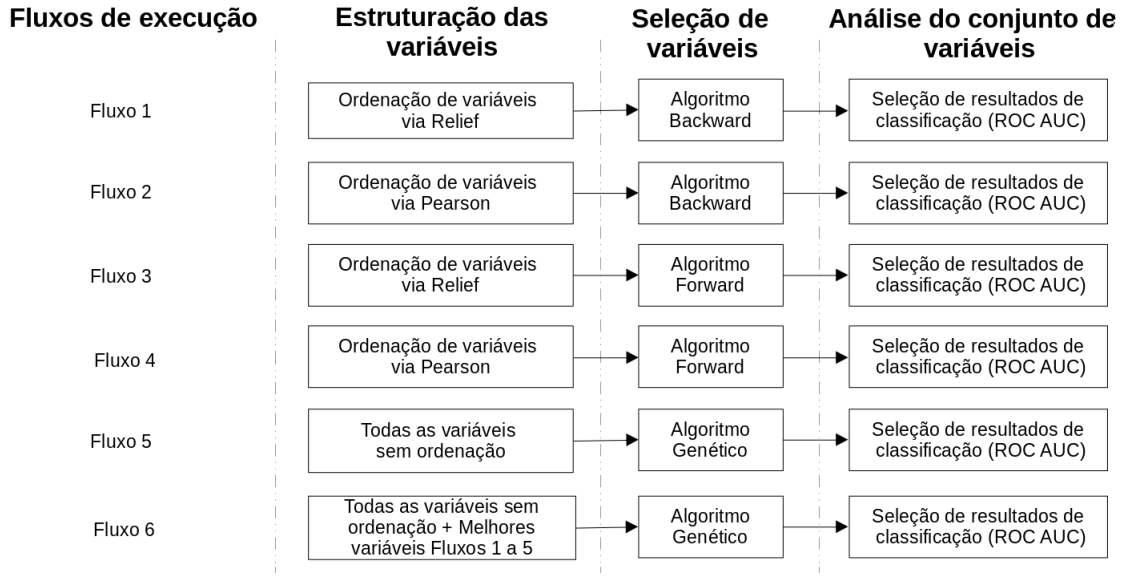


Figura 4.3: Esquema do processo de seleção de variáveis

A Figura 4.3 apresenta as possibilidades de execução da seleção de variáveis aplicadas neste trabalho. Deve-se ressaltar que existe um conjunto de melhores variáveis encontrado em cada fluxo do esquema apresentado conforme a métrica ROC AUC.

Os hiperparâmetros do AG são: o número de gerações, a quantidade de indivíduos da população, a taxa de mutação, a taxa de retenção de indivíduos na população e a quantidade de genes do cruzamento. Esses hiperparâmetros foram balanceados, manualmente, com o esforço humano, diferentemente das propostas de hiperparametrização automáticas dos modelos de classificação.

## 4.3 Proposta de hiperparametrização

Os 7 modelos de classificação utilizados neste trabalho possuem um conjunto de hiperparâmetros apresentados na Tabela 4.2 e descritos no Anexo B. A faixa de valores, nessa tabela, foi estimada conforme a documentação padrão dos modelos e nas experiências de outros cientistas de dados. Alguns hiperparâmetros não estão presentes na tabela e possuem dependência entre eles. Esses hiperparâmetros foram utilizados nos modelos com valores $default$.

Para a seleção de variáveis, foram escolhidos os valores $default$ de todos os hiperparâmetros, incluindo os parâmetros da Tabela 4.2. Para obter os melhores resultados para a métrica ROC AUC, os hiperparâmetros dessa tabela foram balanceados (hiperparametrizados) conforme a "Faixa de valores". Para a hiperparametrização dos modelos foram utilizadas duas técnicas: a técnica de pesquisa aleatória e a técnica de hiperparametrização AG.

### 4.3.1 Aplicação da pesquisa aleatória de hiperparâmetros

Na aplicação da técnica de pesquisa aleatória, foram selecionadas 50 combinações aleatórias de hiperparâmetros para os modelos de previsão (como em mais de 50% das pesquisas analisadas por (Bouthillier and Varoquaux, 2020)). Os modelos foram avaliados conforme as etapas da técnica NCV presentes na secção 2.2.1.3. As variáveis de entrada utilizadas foram as variáveis que obtiveram melhores resultados na seleção feita com os hiperparâmetros padrões, como apresentado na secção 4.2.

Para obtenção do valor ROC AUC, na primeira etapa do NCV, foram submetidas 12 semanas de treinamento aos modelos, conforme o cálculo amostral (ver secção 2.2.1.1), e 1 semana foi submetida para teste.

Na segunda etapa, os valores de 13 semanas de treinamento foram submetidos aos modelos, e foi utilizada 1 semana para teste. Essas 13 semanas eram compostas pelos valores acumulados das 12 semanas da primeira etapa e pela 13ª semana utilizada para teste. Essa etapa, assim como a anterior, seguiu os mesmos critérios da técnica NCV para obter o valor ROC AUC.

Na terceira etapa, os valores de 14 semanas de treinamento foram submetidos aos modelos, e foi utilizada 1 semana para teste. Essas 14 semanas eram compostas pelos

| Modelos | Hiperparâmetros | Faixa de valores |
|---|---|---|
| LR | Parâmetro de regularização para as classes (peso) | 1e-02 a 1e-09 |
| | Valor para critério de parada de execução do modelo | 1.0 a 2.0 |
| | | |
| LSTM | Função de ativação | 'relu', 'sigmoid', 'tanh' |
| | Número de iterações do algoritmo | 1 a 20 |
| | Representação da dimensão de saída | 20 a 300 |
| | | |
| MLP | Número de camadas | 2 a 5 |
| | Número de neurônios | <nº de variáveis> * [0.5 a 3.5] |
| | Função de ativação | 'identity', 'logistic', 'tanh', 'relu' |
| | Valor para inversão da escala da taxa de aprendizado | 0.1 a 0.9 |
| | Parâmetro para suavização (peso) | 0.0001 a 0.03 |
| | Taxa de aprendizado para atualização de pesos | 'constant', 'invscaling', 'adaptive' |
| | | |
| NB | Parâmetro para suavização (peso) | 0.1 a 1.0 |
| | | |
| RF | Pesos associados às classes do modelo | 'balanced', 'balanced_subsample', None |
| | Função para medir a qualidade de uma divisão da informação | 'entropy', 'gini' |
| | Forma de seleção do número máximo de variáveis | 'auto', 'sqrt', 'log2' |
| | Número de árvores | <nº de variáveis> * [0.5 a 4.0] |
| | | |
| SVM | Parâmetro de regularização para as classes (peso) | 0.001 a 10.0 |
| | Valor para critério de parada de execução do modelo | 1e-02 a 1e-09 |
| | | |
| XGBoost | Profundidade das árvores | 2 a 10 |
| | Parâmetro de regularização mínima para as classes (peso) | 1 a 6 |
| | Proporção de uma subamostra para treinamento | 0.5 a 1.0 |
| | Amostra de colunas para construção das árvores | 0.5 a 1.0 |
| | Taxa de aprendizado para atualização de pesos | 0.01 a 0.2 |
| | Número de árvores | <nº de variáveis> * [0.5 a 4.0] |

Tabela 4.2: Hiperparâmetros que são balanceados para cada modelo de classificação

valores acumulados das 13 semanas da segunda etapa e pela 14ª semana utilizada para teste. Essa etapa, assim como a anterior, seguiu os mesmos critérios da técnica NCV para obter o valor ROC AUC.

Na quarta etapa, os valores de 15 semanas de treinamento foram submetidos aos modelos, e foi utilizada 1 semana para teste. Essas 15 semanas eram compostas pelos valores acumulados das 14 semanas da terceira etapa e pela 15ª semana utilizada para teste. Essa etapa, assim como a anterior, seguiu os mesmos critérios da técnica NCV para obter o valor ROC AUC.

A média dos valores ROC AUC das quatro etapas foi a métrica de qualidade do modelo hiperparametrizado. Como são 50 variações de hiperparâmetros, o processo se repete 50 vezes para todos os modelos, e o valor mais alto de ROC AUC, entre essas avaliações, é o melhor conjunto de hiperparâmetros para cada um dos modelos. Nas execuções da técnica de hiperparametrização por AG, a escolha de hiperparâmetros foi otimizada.

### 4.3.2 Aplicação do AG para hiperparametrização

A hiperparametrização por AG foi outra técnica utilizada neste trabalho. Nessa técnica, cada indivíduo da população foi composto pelos parâmetros citados na Tabela 4.2 que assumem valores numéricos ou categóricos dependendo do tipo do parâmetro do modelo, como apresentado na secção 2.2.4.3. A escolha do valor para cada um dos genes foi de forma aleatória, mas garante-se que, na primeira geração, haja um indivíduo com os valores padrões comuns a cada modelo. Os modelos como o NB, que possuem apenas um gene, não foram submetidos a essa técnica por não ser possível fazer combinações.

No início do processo do AG, foram consideradas 80 gerações e 34 indivíduos compondo a população, assim como na técnica AG da seleção de variáveis. Essas gerações foram escolhidas conforme o critério de parada do algoritmo já citado, anteriormente, na secção 2.2.2.4. O motivo para que os valores fossem de 80 gerações e 34 indivíduos é baseado nos valores ROC AUC dos indivíduos que se estabilizam nessa faixa.

Cada indivíduo recebe um valor de avaliação ROC AUC conforme os valores de seus hiperparâmetros escolhidos, aleatoriamente, na primeira geração. Esse processo é o mesmo NCV utilizado pela técnica de pesquisa aleatória explicada na secção 4.3.1. Os indivíduos, que não possuíam uma nota a cada nova geração, também receberam um valor de avaliação ROC AUC utilizando os mesmos critérios.

Assim que todos os indivíduos da população passam pelo processo do NCV, eles são ordenados de forma decrescente conforme os valores obtidos. Desses indivíduos, os 50% que possuem as melhores notas são mantidos na população para a próxima geração. Nos outros 50% restantes, é feita uma seleção aleatória em que cada indivíduo tem 20% de chance de se manter na população para a próxima geração. Quando a população já está composta pelos indivíduos com as melhores notas e pelos indivíduos que entraram

por escolha aleatória, as vagas restantes são preenchidas pelos indivíduos resultantes do cruzamento entre os indivíduos atuais da população.

Na etapa de cruzamento do AG, são escolhidos, aleatoriamente, dois indivíduos diferentes da população, e os genes desses indivíduos são combinados para gerar um novo indivíduo. São escolhidos, aleatoriamente, metade dos genes do indivíduo mãe e metade dos genes do indivíduo pai, e o resultado é o indivíduo filho com os genes combinados. A outra metade dos genes da mãe e do pai, que não foram utilizados no cruzamento, foram combinados para gerar um outro filho. Esse procedimento possibilita que cada cruzamento dê origem a dois novos indivíduos. Se as vagas na população ainda não estiverem preenchidas para a próxima geração, uma nova escolha de pais é feita, e um novo cruzamento é executado.

A taxa de mutação do AG foi fixada em 0,5%. Esse valor representa a possibilidade de acontecer uma mutação em um indivíduo qualquer dentro da população em cada geração. Esse indivíduo escolhido tem todos os seus genes alterados aleatoriamente.

Para utilizar as técnicas de hiperparametrização por pesquisa aleatória e AG combinadas (Fluxo 3 da Figura 4.4), a primeira geração de uma nova execução do AG foi composta pelos seguintes indivíduos, além dos indivíduos gerados aleatoriamente:

- um indivíduo com melhor ROC AUC da última execução do AG;
- um indivíduo com os genes configurados com melhor conjunto de hiperparâmetros encontrados na pesquisa aleatória.

A Figura 4.4 apresenta todo o processo com as possibilidades de execução da hiperparametrização aplicadas neste trabalho. Deve-se ressaltar que existe um conjunto de valores de hiperparâmetros encontrados para cada fluxo de hiperparametrização.

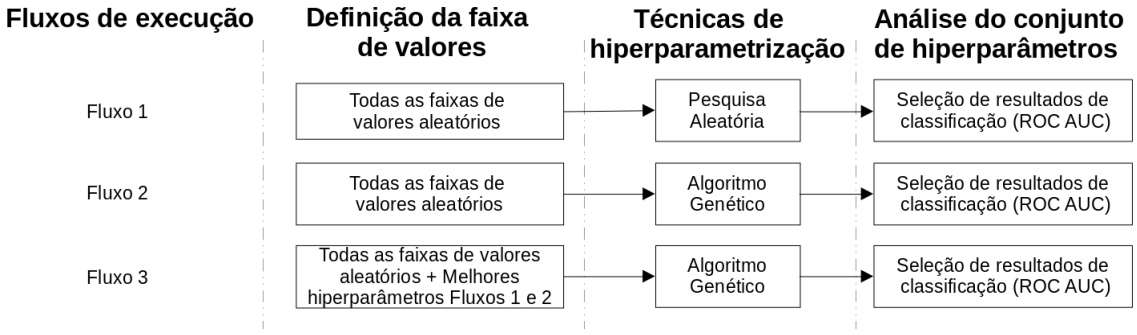


Figura 4.4: Esquema do processo de hiperparametrização

Os hiperparâmetros do AG são: o número de gerações, a quantidade de indivíduos da população, a taxa de mutação, a taxa de retenção de indivíduos na população e a quantidade de genes do cruzamento. Esses hiperparâmetros foram balanceados, manualmente, com o esforço humano, diferentemente das propostas de hiperparametrização automáticas dos modelos de classificação.

## 4.4 Treinamento dos melhores modelos de previsão

Para a realização da previsão de absenteísmo ou de produtividade com dados integrais, era esperado que as etapas de hiperparametrização e de seleção de variáveis estivessem concluídas com os dados amostrais, e os modelos pudessem ser treinados com os melhores hiperparâmetros e variáveis conforme as métricas ROC AUC.

No momento da hiperparametrização com as melhores variáveis, já definia-se quais dos modelos alcançariam as melhores métricas de previsão.

Os passos para treinamento seguem o que já foi citado nas seções anteriores (variáveis selecionadas e hiperparâmetros definidos). Foram submetidas 16 semanas de treinamento (com dados completos) ao modelo escolhido, e 1 semana foi submetida para teste.

# Capítulo 5

# Resultados e análise dos processos de previsão de desempenho

Neste trabalho, foram utilizadas técnicas isoladas e combinadas para encontrar os melhores resultados para os modelos de classificação. Foram encontrados os modelos que oferecem melhores resultados avaliados quanto à métrica e ao tempo de execução para solução dos problemas de previsão de desempenho.

No capítulo, são apresentados os parâmetros balanceados de forma semiautomática para o AG, o tempo de execução de cada modelo de classificação e os resultados da previsão de desempenho dos operadores de teleatendimento.

## 5.1 Estabilização de resultados no AG

Foi necessário realizar a hiperparametrização do AG de forma semiautomática, porque o AG já é um hiperparametrizador de modelos e os seus hiperparâmetros, quando modificados, influenciam no resultado. Para que esse processo não fosse aleatório e seguisse um padrão lógico, a tentativa de encontrar o número de gerações e a população para estabilizar o resultado seguiu a sequência de Fibonacci.

A sequência de Fibonacci é a sequência numérica proposta pelo matemático Leonardo Pisa, conhecido como Fibonacci (Zahn, 2020). A sequência é composta pela soma de dois números antecessores dando origem a um número sucessor:

1, 1, 2, 3, 5, 8, 13, 21, 34, 55, 89, ...

Foi a partir de um problema criado por Fibonacci que foi detectada a existência da regularidade matemática, e essa regularidade representada pela sequência de Fibonacci foi aplicada ao AG com um intervalo finito. O intervalo aplicado foi entre o valor 5 e o valor 55.

O conjunto de valores testados dentro do intervalo aplicado no AG foram os valores 5, 8, 13, 21, 34 e 55 para a quantidade de indivíduos na população, o número limite

de execuções foi definido em 90 gerações (para atender a um critério de convergência (Bento and Kagan, 2008)) e os valores 1, 2 e 3 da sequência de Fibonacci não foram utilizados, porque são números baixos para representar algum impacto significativo no AG relacionado ao número de indivíduos e seus genes.

Todos os resultados dessa etapa foram encontrados a partir do modelo NB, visto que este é o modelo de referência para os resultados esperados dos outros modelos de classificação. O modelo NB com hiperparâmetro padrão foi utilizado na técnica de seleção de variáveis AG para definir o limite de gerações e para definir o número de indivíduos no intervalo da sequência de Fibonacci.

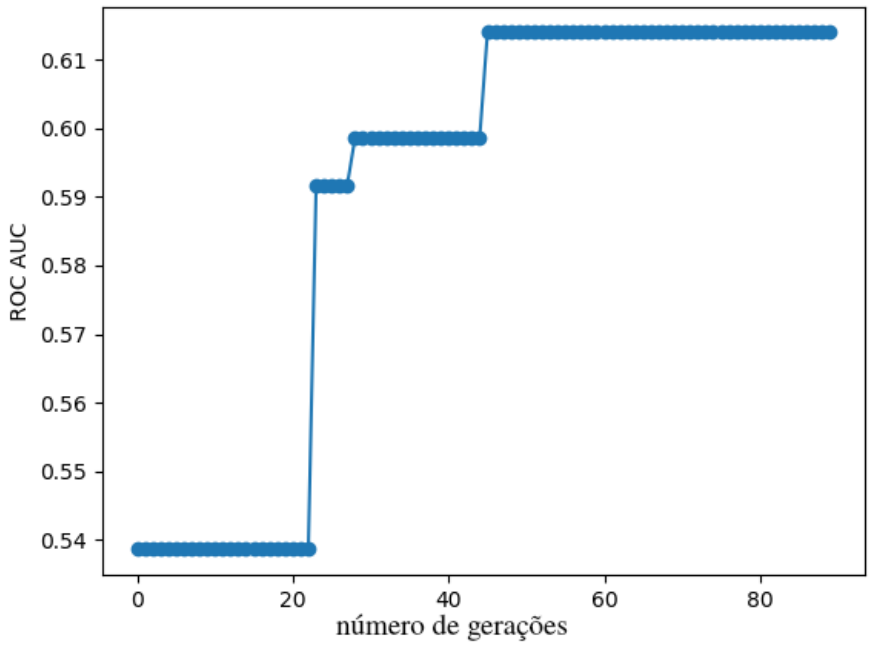


Figura 5.1: Seleção de variáveis para absenteísmo com 5 indivíduos na população utilizando NB

Na Figura 5.1, foram adicionados 5 indivíduos na população do AG para seleção de variáveis para o problema de absenteísmo e observou-se que não houve diversidade genética. Além disso, houve uma estabilização prematura dos valores próximos a 0,6. Esse gráfico apresentado expõe os valores dos indivíduos com melhores ROC AUC em cada geração com um limite de parada pré-estabelecido em 90 gerações.

O gráfico na Figura 5.2 apresenta 8 indivíduos na população do AG para a seleção de variáveis do problema de absenteísmo. Observou-se que ainda existia baixa diversidade genética quando comparado com o gráfico da população de 5 indivíduos na Figura 5.1. Comparando os gráficos da Figura 5.1 e da Figura 5.2, a pontuação melhorou à medida que cresceu o número de gerações, estabilizando os valores ROC AUC próximos a 0,7. Nesse momento do experimento, havia uma expectativa de melhora nos resultados das tentativas seguintes com os valores da sequência de Fibonacci 13, 21, 34 e 55, que correspondem aos indivíduos da população.

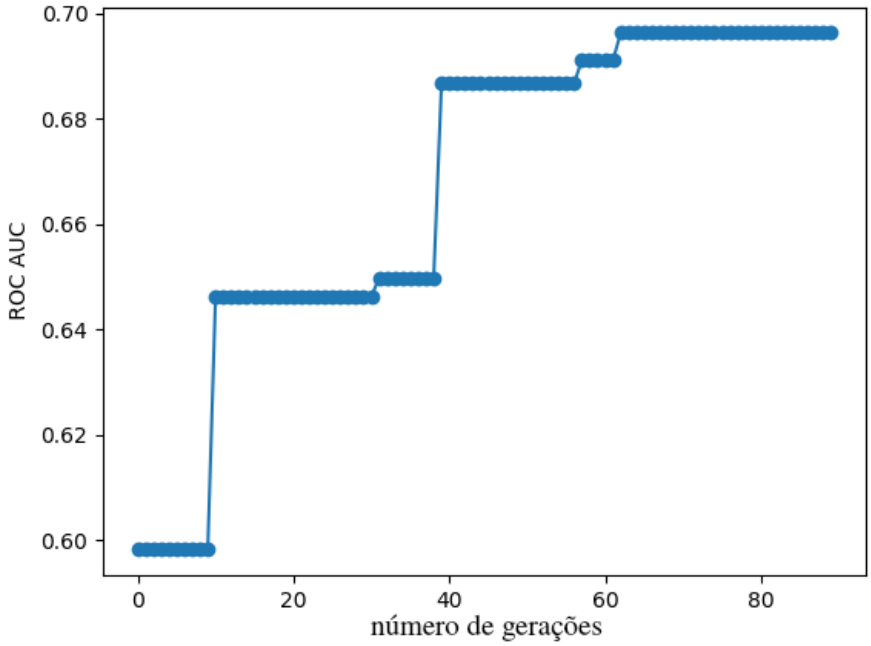


Figura 5.2: Seleção de variáveis para absenteísmo com 8 indivíduos na população utilizando NB

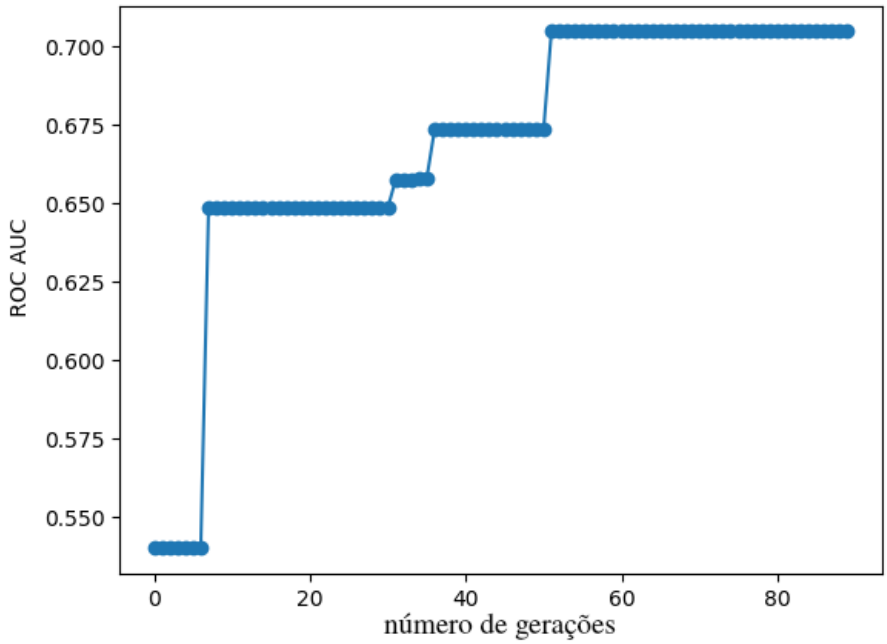


Figura 5.3: Seleção de variáveis para absenteísmo com 13 indivíduos na população utilizando NB

A Figura 5.3 apresenta 13 indivíduos na população do AG para a seleção de variáveis do problema de absenteísmo. Nesse gráfico, os valores encontrados se assemelham aos valores do gráfico apresentado na Figura 5.2 com a população de 8 indivíduos. Conclui-se que, mesmo aumentando a diversidade genética com o aumento de indivíduos na população, não houve melhoras significativas. Os valores se estabilizaram próximos a 0,7 nessa tentativa. O número de gerações em que aconteceu a estabilização dos valores ROC AUC está acima de 60 gerações e abaixo do limite 90 gerações.

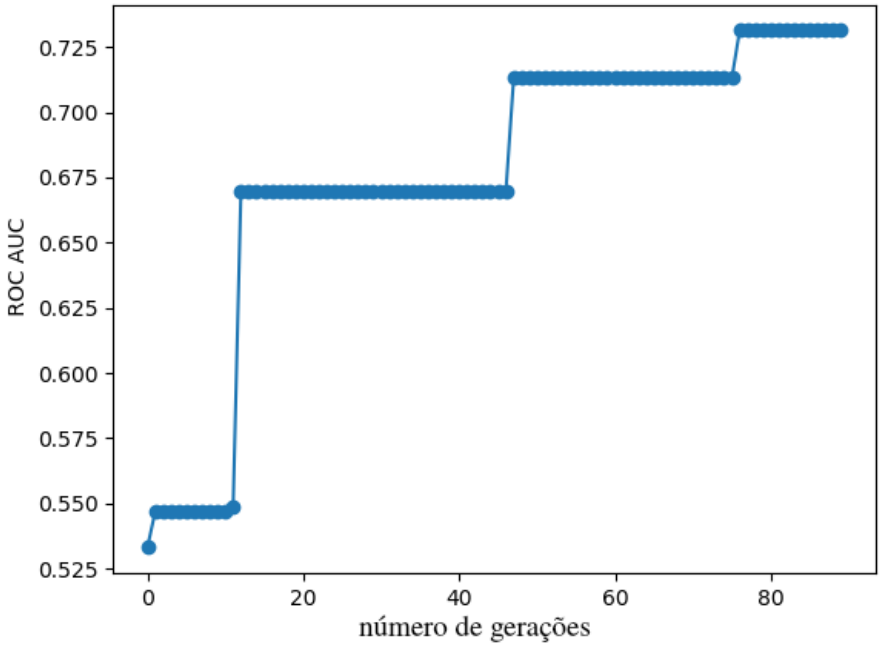


Figura 5.4: Seleção de variáveis para absenteísmo com 21 indivíduos na população utilizando NB

A expectativa para o 4ª experimento era que houvesse melhoras adicionando 21 indivíduos na população do AG para a seleção de variáveis do problema do absenteísmo. A adição dos 21 indivíduos na população é apresentada no gráfico da Figura 5.4. Observou-se que o número de gerações para se estabilizar os valores ROC AUC subiu de 60 gerações para valores próximos a 80 gerações, comparando com o experimento feito com 13 indivíduos na população. As métricas ROC AUC alcançaram 0,73, mas o número de gerações foi alto quando comparado aos testes com números inferiores de indivíduos.

O gráfico apresentado na Figura 5.5 contém o experimento com 34 indivíduos na população do AG para a seleção de variáveis do problema de absenteísmo. Observou-se que os valores ROC AUC melhoraram com a diversidade genética causada pelo aumento de indivíduos na população. Nesse experimento, os valores ROC AUC se estabilizaram próximos a 0,77, e o número de gerações da estabilização foi acima de 70 gerações e abaixo dos valores próximos a 80 gerações encontrados no experimento com 21 indivíduos.

O experimento para a seleção de variáveis do problema de absenteísmo apresentado

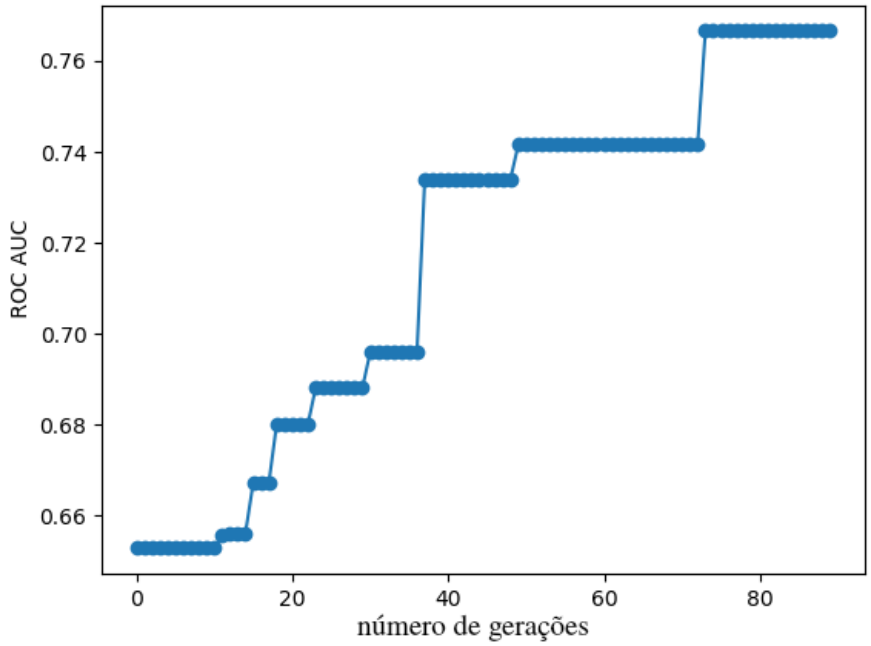


Figura 5.5: Seleção de variáveis para absenteísmo com 34 indivíduos na população utilizando NB

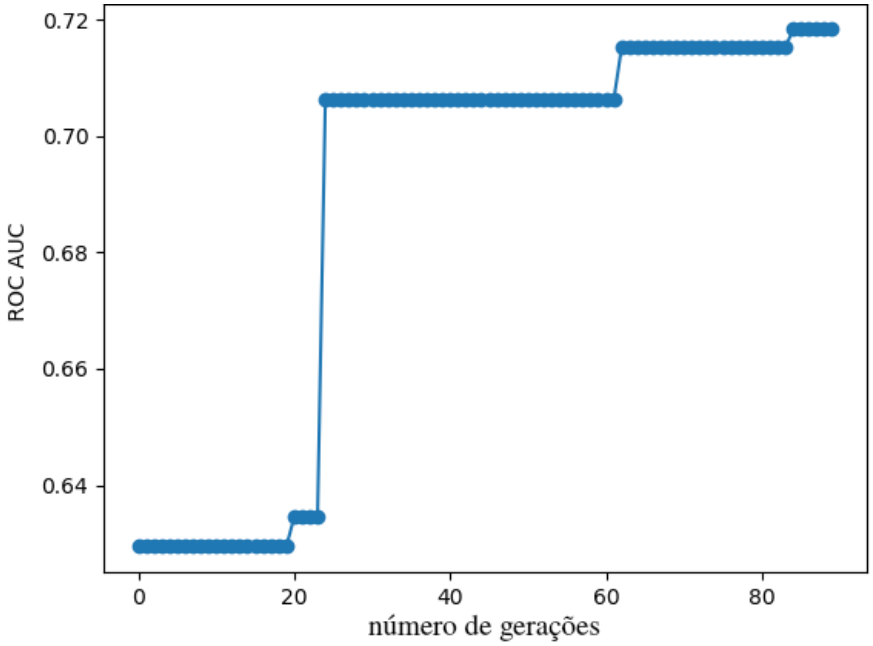


Figura 5.6: Seleção de variáveis para absenteísmo com 55 indivíduos na população utilizando NB

na Figura 5.6 teve valores ROC AUC abaixo dos valores encontrados nas tentativas anteriores com 21 e 34 indivíduos na população. Foram incluídos 55 indivíduos na população do AG e as métricas se estabilizaram próximas a 0,72. A quantidade de indivíduos era superior à dos experimentos anteriores e, por esse motivo, a faixa de valores estáveis foi pequena e bem próxima ao limite de 90 gerações.

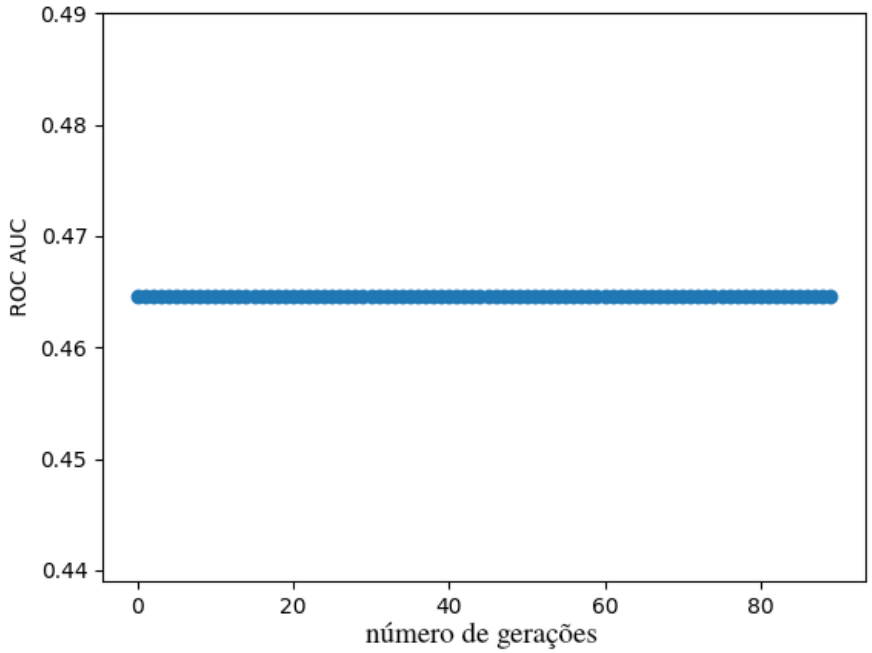


Figura 5.7: Seleção de variáveis para produtividade com 5 indivíduos na população utilizando NB

Os experimentos para a seleção de variáveis do problema de produtividade foram feitos da mesma forma que os experimentos feitos para o problema de absenteísmo. A diferença entre as duas bases de dados dos problemas está no volume de registros positivos para cada uma delas. Enquanto a base de dados de entrada para o problema de absenteísmo possui uma média aproximada de 8% de absenteístas positivos, a base de dados de produtividade é mais equilibrada entre os colaboradores produtivos positivos e negativos. A média aproximada de colaboradores com produtividade positiva na base de dados é de 60%.

Por ser uma base com classes equilibradas, o comportamento do AG foi diferente na seleção de variáveis para o problema de produtividade quando comparado com a seleção de variáveis para o problema de absenteísmo. Esse fato é mais evidente no experimento apresentado na Figura 5.7, que com 5 indivíduos na população do AG, demonstrou não alcançar uma diversidade genética que levasse à melhora da população. O melhor indivíduo surgiu na primeira geração e permaneceu até a geração 90. O valor ROC AUC foi, aproximadamente, 0,46.

Com o experimento apresentado na Figura 5.8, com 8 indivíduos na população do AG

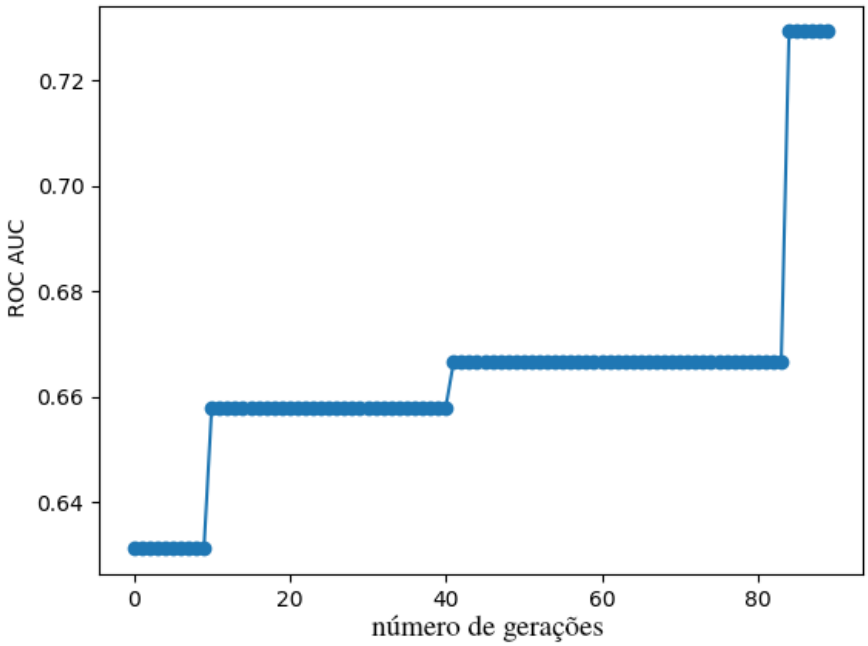


Figura 5.8: Seleção de variáveis para produtividade com 8 indivíduos na população utilizando NB

para a seleção de variáveis do problema de produtividade, observou-se que os valores se estabilizaram próximos ao limite de 90 gerações. Nessa situação, houve um aumento na diversidade genética em relação ao que foi apresentado na Figura 5.7. Para esse caso, os valores da métrica ROC AUC se estabilizaram próximos de 0,73.

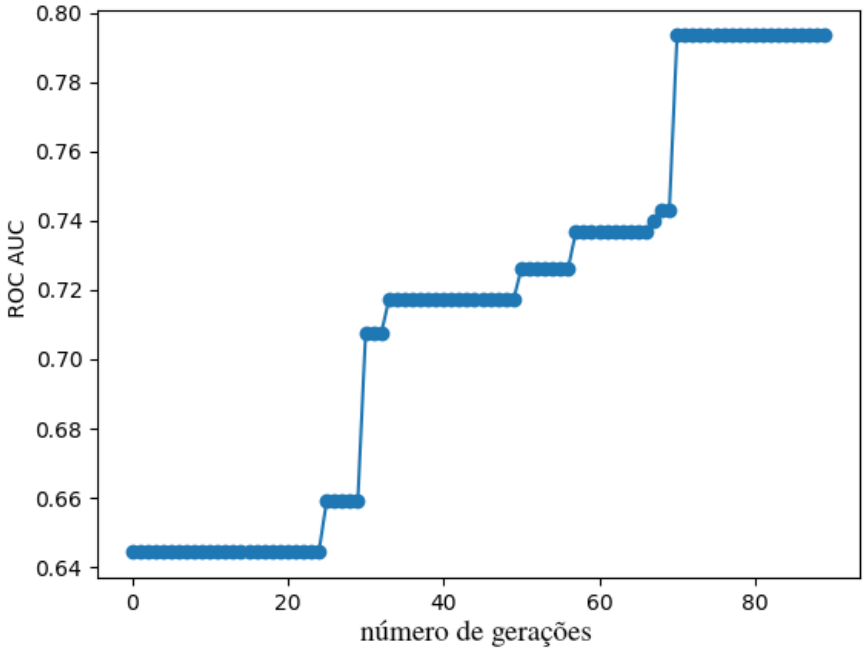


Figura 5.9: Seleção de variáveis para produtividade com 13 indivíduos na população utilizando NB

Na Figura 5.9, é apresentado o experimento com uma população de 13 indivíduos da população do AG para o problema de produtividade. Observou-se que a estabilização de valores para o melhor indivíduo da população do AG foi próxima da geração 70. Essa aproximação foi distante do limite estabelecido de 90 gerações, o que sugere que o número de gerações pode ser reduzido. O valor ROC AUC do melhor indivíduo estabilizou-se em, aproximadamente, 0,79. Esse valor representa um aumento expressivo na comparação com as tentativa anteriores que possuíam 5 e 8 indivíduos na população.

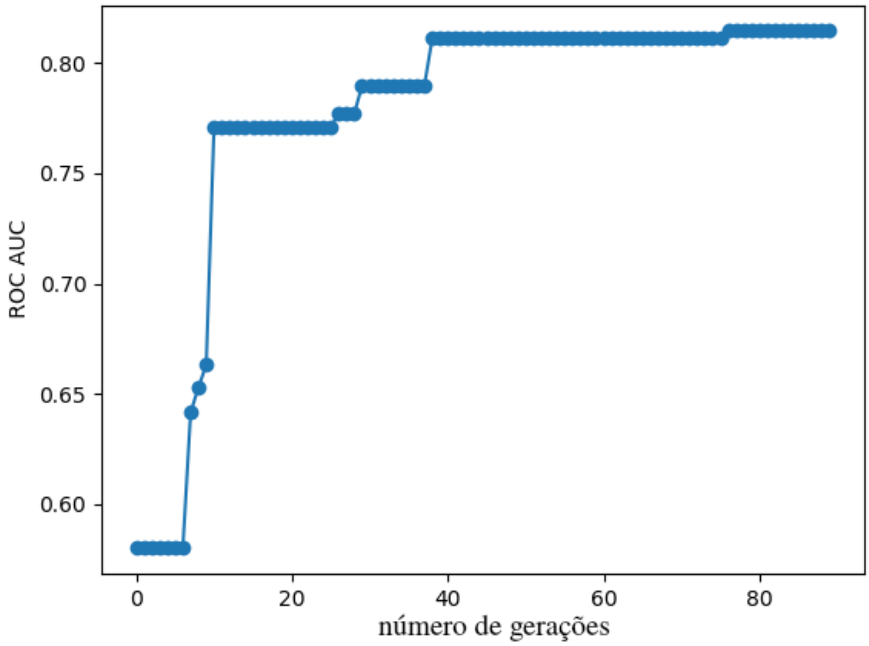


Figura 5.10: Seleção de variáveis para produtividade com 21 indivíduos na população utilizando NB

A Figura 5.10 apresenta o experimento com 21 indivíduos na população do AG para o problema de produtividade. Observou-se que o número de gerações com o valor do melhor indivíduo estabilizou-se bem próximo a 80 gerações. Os valores da métrica ROC AUC foram próximos a 0,81. As métricas alcançaram melhores resultados, entretanto o número de gerações é alto, quando comparado aos experimentos com números inferiores de indivíduos.

No experimento apresentado na Figura 5.11, com uma população de 34 indivíduos do AG para a seleção de variáveis do problema de produtividade, observou-se que os valores alcançaram uma melhora conforme a diversidade genética causada pelo aumento de indivíduos na população. Para esse experimento, os valores ROC AUC se estabilizaram próximos a 0,86 e o número de gerações ficou entre 70 e 80 gerações.

A seleção de variáveis para o problema de produtividade apresentado, na Figura 5.12, com 55 indivíduos na população do AG, alcançou uma estabilização dos valores ROC AUC próximos a 0,86 para o melhor indivíduo. Essa estabilização ocorreu próxima

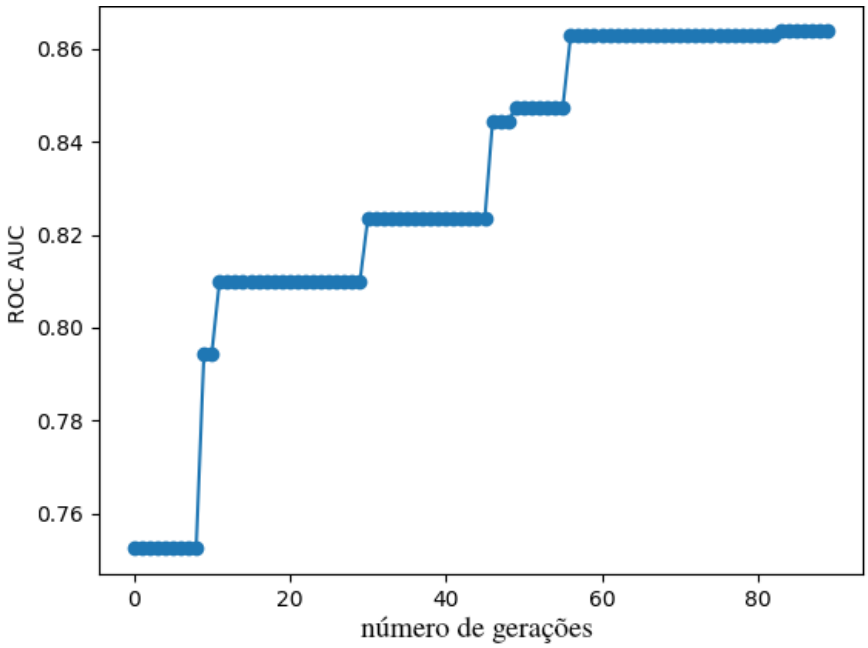


Figura 5.11: Seleção de variáveis para produtividade com 34 indivíduos na população utilizando NB

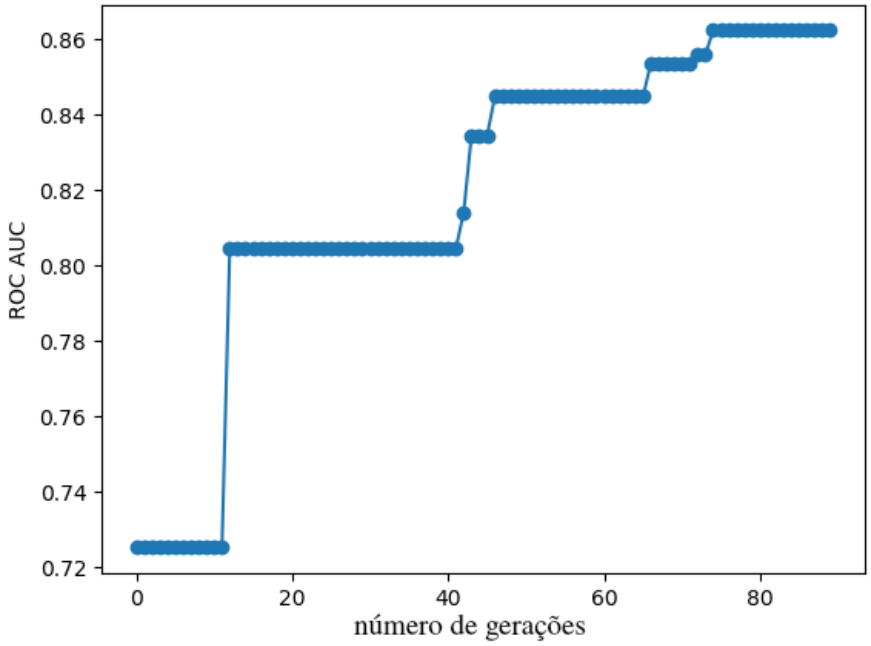


Figura 5.12: Seleção de variáveis para produtividade com 55 indivíduos na população utilizando NB

à 80ª geração. Nesse experimento, era esperada uma estabilização de resultados com um número expressivo de gerações, porque a quantidade de indivíduos era superior quando comparada aos números das populações das experiências anteriores.

A partir da comparação dos experimentos para seleção de variáveis com AG utilizando NB, tanto para o problema de absenteísmo quanto para o problema de produtividade, observou-se que a melhor probabilidade de se obter bons resultados seria com 34 indivíduos na população. Essa decisão foi tomada, porque os experimentos com essa configuração apresentaram métricas ROC AUC elevadas para um número equilibrado de gerações. Por consequência, o valor limite para as gerações foi reduzido de 90 gerações para 80 gerações.

Os experimentos apresentados possuem gráficos das métrica ROC AUC do melhor indivíduo por geração. O gráfico que demonstra que a população geral do AG evolui de forma crescente com a média de valores de todos os indivíduos é apresentado na Figura 5.13.

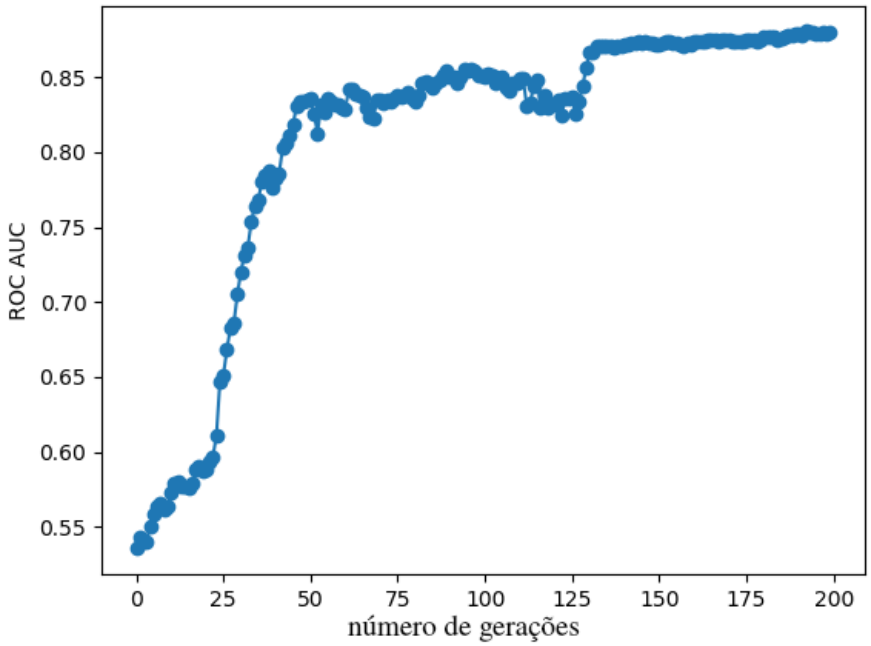


Figura 5.13: Média de valores ROC AUC para seleção de variáveis para o problema de produtividade com 34 indivíduos na população e execução de 200 gerações utilizando NB

Na Figura 5.13, é apresentado um limite de 200 gerações. Como a média considera o valor de todos os indivíduos, não houve estabilização total que coincidisse com o limite de 80 gerações para os valores ROC AUC alcançados na busca do melhor indivíduo para uma população de 34 indivíduos.

O limite de gerações foi determinado para não forçar o AG a uma busca interminável pelo melhor indivíduo. O tempo de execução de cada modelo de classificação é importante e está implícito em cada geração processada.

## 5.2 Tempo de execução dos modelos de classificação

Na execução dos modelos de classificação para obtenção dos resultados finais, foram utilizados dados amostrais, porque a base de informação era de grandes proporções e o tempo de execução dos modelos seria excessivamente longo se essas informações fossem integralmente utilizadas. Foi necessário dividir o conjunto de dados com critérios de representatividade como citadas na seção 2.2.1.1. Na presente secção serão apresentados os tempos de execução com os dados integrais e os dados amostrais.

Foram definidas 3 estimativas de execução dos modelos de classificação: "Tempo de treinamento", tempo de execução em "Processos com AG" e tempo "50 treinamentos". Para essas execuções, foram utilizadas todas as variáveis de entrada e os parâmetros padrões de cada modelo. Como o tempo de execução é algo que pode variar por fatores físicos de processamento computacional, foi calculada a média das 5 execuções para a estimativa de "Tempo de treinamento" e as outras 2 estimativas ("Processos com AG" e tempo "50 treinamentos") foram calculadas a partir desta estimativa. O limite de 5 execuções foi arbitrário, observando-se o tempo de execução do modelo mais demorado.

A Tabela 5.1 apresenta o tempo médio de execução dos modelos na resolução do problema de absenteísmo com dados amostrais. Observa-se, nessa tabela, que o tempo de execução com menor valor foi do modelo NB, e o modelo com maior tempo de execução foi o LSTM.

O tempo de execução em segundos do modelo de classificação na estimativa de "Tempo de treinamento" não é tão crítico. Porém, se forem considerados todos os "Processos com AG", são processadas as notas de 17 indivíduos para 80 gerações no pior caso. Nessas condições, a execução de uma hiperparametrização ou seleção de variáveis se torna inviável dependendo do modelo. A apresentação feita nas Tabelas dessa secção consideram sempre os melhores casos, sendo que esses valores de tempo de execução podem variar para mais.

A seleção de variáveis depende da transformação das variáveis em binárias, em que cada variável, criada a partir das 168 variáveis (número de variáveis no anexo A), será responsável por uma execução na técnica *Forward* e por uma execução na técnica *Backward*. O total de variáveis binarizadas foi 317 (variáveis geradas com a técnica da secção 2.2.1.2 e apresentadas na Tabela 4.1).

Observa-se, na Tabela 5.2, que os valores menores de tempo estimado são para o modelo LR. O modelo que alcançou a maior estimativa de tempo de execução foi o SVM, com 5,53 meses para execução na estimativa dos "Processos com AG".

A execução do modelo SVM é influenciada pela quantidade de dados de entrada como apresentado na secção 2.2.3.6. Comparando o tempo de execução dos dados amostrais da Tabela 5.3 com o tempo de execução do conjunto de dados integrais, na Tabela 5.4, a execução do modelo SVM foi consideravelmente baixa. O modelo NB continuou sendo

| Modelo | Tempo de treinamento | Treinamentos com AG | 50 treinamentos |
|---|---|---|---|
| NB | 0,92 segundos | 21,03 minutos | 46,40 segundos |
| LR | 1,40 segundos | 31,91 minutos | 1,17 minutos |
| XGBoost | 7,16 segundos | 2,70 horas | 5,95 minutos |
| MLP | 10,84 segundos | 4,09 horas | 9,02 minutos |
| RF | 19,80 segundos | 7,47 horas | 16,48 minutos |
| SVM | 24,32 segundos | 9,19 horas | 20,27 minutos |
| LSTM | 47,92 segundos | 18,10 horas | 39,93 minutos |

Tabela 5.1: Estimativas de tempo de execução – Dados amostrais do problema de previsão de absenteísmo para 4 etapas do NCV

| Modelo | Tempo de treinamento | Treinamentos com AG | 50 treinamentos |
|---|---|---|---|
| LR | 4,6 segundos | 1,74 horas | 3,83 minutos |
| NB | 6,60 segundos | 2,49 horas | 5,49 minutos |
| MLP | 40,72 segundos | 15,39 horas | 33,94 minutos |
| RF | 2,80 minutos | 2,65 dias | 2,34 horas |
| XGBoost | 2,90 minutos | 2,74 dias | 2,42 horas |
| LSTM | 5,64 minutos | 5,32 dias | 4,69 horas |
| SVM | 2,92 horas | 5,53 meses | 6,10 dias |

Tabela 5.2: Estimativas de tempo de execução – Dados completos do problema de previsão de absenteísmo para 4 etapas do NCV

| Modelo | Tempo de treinamento | Treinamentos com AG | 50 treinamentos |
|---|---|---|---|
| NB | 1,64 segundos | 37,26 minutos | 1,37 minutos |
| SVM | 2,24 segundos | 50,41 minutos | 1,85 minutos |
| LR | 2,24 segundos | 50,86 minutos | 1,87 minutos |
| XGBoost | 3,08 segundos | 1,17 horas | 2,57 minutos |
| RF | 5,76 segundos | 2,18 horas | 4,80 minutos |
| MLP | 6,76 segundos | 2,56 horas | 5,64 minutos |
| LSTM | 17,72 segundos | 6,69 horas | 14,76 minutos |

Tabela 5.3: Estimativas de tempo de execução – Dados amostrais do problema de previsão de produtividade para 4 etapas do NCV

| Modelo | Tempo de treinamento | Treinamentos com AG | 50 treinamentos |
|---|---|---|---|
| LR | 8,08 segundos | 3,05 horas | 6,72 minutos |
| NB | 19,44 segundos | 7,34 horas | 16,20 minutos |
| RF | 35,08 segundos | 13,25 horas | 29,22 minutos |
| MLP | 46,16 segundos | 17,44 horas | 38,47 minutos |
| XGBoost | 2,87 minutos | 2,71 dias | 2,40 horas |
| LSTM | 7,16 minutos | 6,76 dias | 5,96 horas |
| SVM | 8,76 horas | 1,38 anos | 18,22 dias |

Tabela 5.4: Estimativas de tempo de execução – Dados completos do problema de previsão de produtividade para 4 etapas do NCV

a execução com menor tempo, e o modelo LSTM, com maior tempo.

As estimativas de tempo apresentadas, na Tabela 5.4, comprovam que as execuções utilizando dados amostrais são mais viáveis. Os problemas de previsão trabalham com informações no intervalo de uma semana, logo um modelo que demore um tempo superior a 7 dias para ser treinado é considerado inapto para resolução do problema. Os modelos SVM e LSTM foram desconsiderados por não atenderem a esse critério. Os tempos apresentados na Tabela 5.4 foram os maiores alcançados, e o tempo estimado para os processos que usam AG para o modelo SVM foi de 1,38 anos.

## 5.3 Resultados da previsão de desempenho

Para a geração de resultados dos problemas de previsão de desempenho foram utilizadas 16 semanas. Nesse período, foram testadas as previsões para cada uma dessas semanas e verificada a média dos 16 resultados coletados. Verificou-se que a dispersão dos valores ROC AUC das semanas não apresentou uma variação expressiva. Esses valores foram arredondados para 2 casas decimais e, por esse motivo, alguns deles são apresentados iguais nas tabelas dessa secção. Os valores de desvio padrão para as semanas foram entre 1% e 4%.

Foram selecionadas as variáveis para os modelos utilizando dados amostrais das 16 semanas avaliadas e a média de resultados é observada na Tabela 5.6 e na Tabela 5.10 (ver proposta de seleção de variáveis na secção 4.2) . As opções para obter os resultados foram baseados em experimentos feitos com as técnicas de seleção de variáveis utilizando o AG, a entrada de todas as variáveis no modelo -sem o uso de técnicas de seleção (Nenhum)-, seleção de variáveis com *Pearson Forward/Backward*, seleção de variáveis com *Releaf Forward/Backward* e o conjunto de todas essas técnicas de seleção de variáveis combinadas (Completo).

A média ROC AUC apresentada, na Tabela 5.6 e na Tabela 5.10, não é o resultado final, porque os modelos precisam ser hiperparametrizados. Além disso, os menores va-

lores ROC AUC, dessas tabelas não devem ser excluídos, visto que o resultado das técnicas combinadas é alcançado utilizando-se os indivíduos com variáveis selecionadas de todos valores ROC AUC encontrados, tanto baixos quanto altos. Esses indivíduos podem cruzar-se entre si e gerar valores melhores.

Após a seleção das variáveis, passou-se à etapa seguinte: geração de resultados no processo de hiperparametrização. A Tabela 5.8 e a Tabela 5.12 mostram a média de resultados ROC AUC com os dados amostrais das 16 semanas nos modelos hiperparametrizados utilizando-se as variáveis selecionadas para os problema de previsão de desempenho (ver proposta de hiperparametrização na secção 4.3).

No processo de hiperparametrização, foram utilizadas três técnicas: AG, PA e a combinação dessas duas técnicas. Os valores ROC AUC obtidos são considerados os resultados finais da previsão, pois os valores são apresentados com as melhores variáveis selecionadas e os melhores hiperparâmetros escolhidos para os modelos na solução dos problemas.

O modelo de classificação NB foi utilizado como padrão de observação para os outros modelos. Para esse modelo não foi executada a hiperparametrização nas técnicas AG e nas técnicas combinadas representadas pela coluna "Completo" da Tabela 5.8 e da Tabela 5.12. O NB não foi submetido a essas técnicas por possuir apenas um hiperparâmetro e, consequentemente, uma população de indivíduos com um gene no AG, como citado na secção 4.3.2. Por ter apenas um hiperparâmetro, um algoritmo evolucionário não teria atuação efetiva em um eventual cruzamento de indivíduos da população.

### 5.3.1 Resultados na previsão de absenteísmo

Os resultados dos trabalhos anteriores sobre a previsão de absenteísmo estão apresentados na Tabela 5.5. Quando os resultados dessa tabela foram gerados, não havia a mesma quantidade de variáveis disponíveis que foram utilizadas na presente pesquisa e, também, não havia a implementação da combinação de todas as técnicas em um AG para seleção de variáveis e hiperparametrização.

Quando comparam-se os resultados da Tabela 5.5 com os resultados apresentados, nessa subsecção, houve uma importante melhora do ROC AUC. Os modelos LSTM e SVM já conservavam um comportamento inadequado para a resolução do problema. Como os resultados ROC AUC desses dois modelos estão menores do que os resultados dos demais modelos, confirma-se a decisão de excluí-los da fase de resultados desta pesquisa.

A seleção de variáveis com todas as técnicas combinadas obteve as melhores médias apresentadas na coluna "Completo" da Tabela 5.6. A análise dos valores médios da seleção de variáveis mostrou que os modelos tiveram progressiva melhora quando comparados ao ROC AUC de nenhuma técnica de seleção (coluna 'Nenhum' da Tabela 5.6)

| Modelo | ROC AUC |
|---|---|
| LSTM | 0,53 |
| SVM | 0,56 |
| LR | 0,60 |
| NB | 0,63 |
| MLP | 0,64 |
| RF | 0,71 |
| XGBoost | 0,73 |

Tabela 5.5: ROC AUC dos trabalhos de previsão de absenteísmo (Oliveira et al., 2019b) anteriores que usam técnicas fundamentais desta pesquisa

| | | | Forward | | Backward | | |
|---|---|---|---|---|---|---|---|
| **Modelo** | **AG** | **Nenhum** | **Pearson** | **Relief** | **Pearson** | **Relief** | **Completo** |
| XGBoost | 0,75 | 0,57 | 0,76 | 0,74 | 0,69 | 0,69 | 0,77 |
| RF | 0,75 | 0,62 | 0,77 | 0,74 | 0,72 | 0,72 | 0,78 |
| MLP | 0,65 | 0,53 | 0,79 | 0,74 | 0,62 | 0,62 | 0,79 |
| NB | 0,69 | 0,49 | 0,79 | 0,75 | 0,58 | 0,55 | 0,79 |
| LR | 0,69 | 0,51 | 0,79 | 0,78 | 0,59 | 0,59 | 0,79 |

Tabela 5.6: Média ROC AUC com dados amostrais - Seleção de variáveis dos modelos para previsão de absenteísmo

passando por todas as outras até chegar ao uso das técnicas combinadas (coluna 'Completo' da Tabela 5.6). Deve-se observar que os valores da coluna "*Pearson Forward*" são os valores mais próximos aos valores apresentados na coluna "Completo", logo tiveram grande influência nos melhores resultados.

O número médio de variáveis selecionadas, nas 16 semanas testadas, para os modelos de classificação do problema de previsão de absenteísmo, estão apresentadas na Tabela 5.7. As técnicas obtiveram números específicos, e o maior número de variáveis de entrada foi obtido pela técnica *Backward Releaf*, na qual foram utilizadas muitas variáveis e obtiveram-se resultados pouco expressivos, enquanto o número de variáveis foi equilibrado na combinação das demais técnicas apresentada na coluna "Completo" da Tabela

| | | Forward | | Backward | | | |
|---|---|---|---|---|---|---|---|
| **Modelo** | **AG** | **Pearson** | **Relief** | **Pearson** | **Relief** | **Completo** | **Maior N$^{o}$ variáveis** |
| NB | 69 | 45 | 66 | 189 | 207 | 46 | 207 |
| LR | 64 | 39 | 72 | 256 | 256 | 39 | 256 |
| RF | 68 | 22 | 65 | 260 | 261 | 49 | 261 |
| XGBoost | 67 | 18 | 57 | 260 | 261 | 44 | 261 |
| MLP | 69 | 22 | 29 | 259 | 261 | 26 | 261 |

Tabela 5.7: Número médio de variáveis selecionadas para os modelos do problema de previsão de absenteísmo com dados amostrais

| Modelo | AG | PA | Completo |
|---|---|---|---|
| NB | - | 0,79 | - |
| RF | 0,79 | 0,78 | 0,79 |
| XGBoost | 0,79 | 0,77 | 0,79 |
| LR | 0,80 | 0,79 | 0,80 |
| MLP | 0,80 | 0,79 | 0,80 |

Tabela 5.8: Média ROC AUC com dados amostrais- Hiperparametrização dos modelos para previsão de absenteísmo

| Modelo | ROC AUC |
|---|---|
| NB | 0,71 |
| LR | 0,75 |
| MLP | 0,79 |
| RF | 0,83 |
| XGBoost | 0,80 |

Tabela 5.9: ROC AUC dos trabalhos de previsão de produtividade (Oliveira et al., 2019c) anteriores que usam técnicas fundamentais desta pesquisa

5.7.

O modelo XGBoost obteve resultados abaixo do valor ROC AUC obtidos pelo modelo NB na hiperparametrização, como mostra a Tabela 5.8. Os modelos com valores acima dos valores do modelo NB foram: RF, MLP e LR. Dentre esses modelos que obtiveram melhores resultados utilizando uma quantidade equilibrada de variáveis, e em um tempo de execução menor, o modelo de classificação LR se destacou: porém, pela proximidade dos resultados, todos os modelos apresentados na tabela estão aptos para resolução do problema de previsão de absenteísmo.

Todos os modelos de classificação alcançaram melhores valores de ROC AUC, na etapa de hiperparametrização, se comparados com a etapa de seleção de variáveis.

### 5.3.2 Resultados na previsão produtividade

Quando os primeiros resultados dos trabalhos de previsão de produtividade foram apresentados, havia uma quantidade menor de variáveis disponíveis para pesquisa e não existia a implementação da combinação de todas as técnicas em um AG para seleção de variáveis e hiperparametrização. Esses resultados primordiais podem ser vistos na Tabela 5.9

A Tabela 5.9 apresenta os resultados dos trabalhos de previsão de produtividade anteriores que utilizavam técnicas fundamentais desta pesquisa. Se comparados os resultados dessa tabela com os resultados apresentados nessa subsecção, percebe-se uma importante melhora do ROC AUC para todos os modelos. Alguns deles obtiveram uma grande variação nos resultados e melhoraram a sua qualidade, principalmente, os modelos com

| | | | Forward | | Backward | | |
|---|---|---|---|---|---|---|---|
| **Modelo** | **AG** | **Nenhum** | **Pearson** | **Relief** | **Pearson** | **Relief** | **Completo** |
| LR | 0,80 | 0,48 | 0,86 | 0,85 | 0,58 | 0,56 | 0,86 |
| MLP | 0,78 | 0,50 | 0,86 | 0,84 | 0,60 | 0,60 | 0,86 |
| NB | 0,81 | 0,50 | 0,86 | 0,85 | 0,60 | 0,55 | 0,86 |
| RF | 0,83 | 0,48 | 0,85 | 0,83 | 0,70 | 0,70 | 0,86 |
| XGBoost | 0,81 | 0,46 | 0,86 | 0,84 | 0,73 | 0,71 | 0,86 |

Tabela 5.10: Média ROC AUC com dados amostrais- Seleção de variáveis dos modelos para previsão de produtividade

| | | Forward | | Backward | | | |
|---|---|---|---|---|---|---|---|
| **Modelo** | **AG** | **Pearson** | **Relief** | **Pearson** | **Relief** | **Completo** | **Maior Nº variáveis** |
| NB | 66 | 29 | 46 | 212 | 214 | 30 | 214 |
| LR | 64 | 32 | 46 | 250 | 247 | 34 | 250 |
| RF | 65 | 18 | 32 | 256 | 256 | 29 | 256 |
| XGBoost | 67 | 19 | 32 | 255 | 257 | 26 | 257 |
| MLP | 65 | 24 | 32 | 259 | 258 | 27 | 259 |

Tabela 5.11: Número médio de variáveis selecionadas para os modelos do problema de previsão de produtividade com dados amostrais

ROC AUC mais baixos. As combinações de técnicas e a entrada de novos dados podem ter influenciado na evolução deste projeto.

As características dos dados de produtividade são diferentes dos dados de absenteísmo. Foi citado, na secção 5.1, que o problema da previsão de produtividade possui uma base de dados com classes mais equilibradas de que a base de dados do problema de absenteísmo. É possível que esse seja um dos motivos para os resultados ROC AUC estarem distantes dentre os dois problemas.

Os resultados ROC AUC de produtividade, apresentados na Tabela 5.10, mostram que houve uma disparidade do modelo LR. Esse modelo foi o único capaz de obter valores superiores ao modelo base de comparação NB. No entanto, nessa etapa de seleção de variáveis os modelos não estão hiperparametrizados, e os resultados dos outros modelos ficaram próximos aos resultados do NB.

A técnica de seleção de variáveis *Backward*, normalmente, retém mais variáveis, porque inicia o processo com todas as variáveis e continua esse processo retirando, de uma a uma, da variável mais correlacionada para a menos correlacionada com a variável de saída. Essa técnica é apresentada na secção 4.2.2.1. Na técnica de seleção de variáveis *Forward*, a retenção de variáveis é menor, porque o modelo começa com a variável mais correlacionada com a variável de saída, e as demais variáveis vão sendo adicionadas em ordem decrescente conforme *Pearson* ou *Relief*. Nessas duas técnicas, o *Forward* consegue melhores resultados ROC AUC com um número menor de variáveis. Entre os

| Modelo | AG | PA | Completo |
|---|---|---|---|
| NB | - | 0,86 | - |
| LR | 0,86 | 0,86 | 0,87 |
| MLP | 0,86 | 0,86 | 0,87 |
| RF | 0,86 | 0,86 | 0,87 |
| XGBoost | 0,86 | 0,86 | 0,87 |

Tabela 5.12: Média ROC AUC com dados amostrais - Hiperparametrização dos modelos para previsão de produtividade

algoritmos de correlação *Pearson* e *Relief*, o algoritmo de *Pearson* tem se saído melhor considerando-se todos os modelos de classificação experimentados.

Os resultados dos modelos de classificação apresentados, na coluna "Completo" da Tabela 5.11, mantiveram o número de variáveis retidas equilibrado com a combinação de todas as técnicas de seleção de variáveis.

Dos quatro modelos que ficaram acima dos valores do modelo NB na 5.12 e que obtiveram melhores resultados na hiperparametrização utilizando uma quantidade equilibrada de variáveis, destaca-se o modelo de classificação LR. Os modelos RF, XGBoost e MLP obtiveram valores ROC AUC próximos ao valor do modelo LR, então, pois, todos os modelos apresentados na tabela estão aptos para resolução do problema de previsão de produtividade.

### 5.3.3 Hipóteses de desempenho

A principal métrica de qualidade dos resultados obtidos, neste projeto de pesquisa, é a métrica ROC AUC. As outras métricas apresentadas, nessa subsecção, são hipotéticas, pois dependem da decisão dos gestores perante os relatórios contendo as probabilidades de desempenho dos colaboradores da empresa de teleatendimento.

Nesse estudo, cada elemento que recebe um valor de probabilidade é um operador da empresa de teleatendimento. Desse modo, quando o problema a ser resolvido corresponde à previsão de absenteísmo, a saída do modelo se refere ao quão próximo esses funcionários estão de serem absenteístas na semana prevista. Quando a saída corresponde à solução do problema de previsão de produtividade, os valores de probabilidade se referem ao quão próximo os funcionários estão de serem produtivos na semana prevista.

A saída do modelo mostra o quão próximo os funcionários estão de se tornarem produtivos ou absenteístas, mas isso não determina se ele pode ser classificado como absenteísta, produtivo, presentista e/ou improdutivo, pois esse é um fator de proximidade com a classe que necessita de uma porcentagem limite para determinar a classificação, o *threshold*.

O *threshold*, determinante para a classificação dos funcionários, pode ser feito, em-

| | Medidas de Desempenho | | | | | |
|---|---|---|---|---|---|---|
| **Modelo** | **F-score** | **Precisão** | **Revocação** | **Acurácia** | **MSE** | **ROC AUC** |
| NB | 0,25 | 0,59 | 0,17 | 0,92 | 0,08 | 0,73 |
| XGBoost | 0,30 | 0,49 | 0,23 | 0,91 | 0,08 | 0,74 |
| MLP | 0,24 | 0,57 | 0,19 | 0,90 | 0,10 | 0,75 |
| RF | 0,23 | 0,50 | 0,17 | 0,91 | 0,08 | 0,76 |
| LR | 0,29 | 0,57 | 0,22 | 0,92 | 0,08 | 0,78 |

Tabela 5.13: Média de desempenho dos classificadores de absenteísmo com dados completos

piricamente, pelo gestor ou pode ser medido a partir do histórico de desempenho. Neste trabalho, utilizou-se um *threshold* de 40% para absenteísmo e de 80% para produtividade. Com exceção do ROC AUC (secção 4.2.1), as medidas de desempenho citadas, nesta subsecção, são dependentes desse *threshold*.

Nesta subsecção, é comparado o desempenho médio dos modelos LR, RF, XGBoost e MLP ao modelo de referência NB na previsão de desempenho de 16 semanas.

Foi apresentado na subsecção 4.4 um exemplo de previsão de desempenho de 1 semana qualquer. Na presente subsecção 5.3.3, são utilizados os dados integrais para previsão das 16 primeiras semanas de 2019 com os hiperparâmetros e variáveis selecionadas nas secções anteriores. As medidas de desempenho são as médias das previsões dessas 16 semanas.

Foram calculadas cinco medidas de desempenho da classificação: F-score, Precisão, Revocação, Acurácia, ROC AUC e o erro quadrático médio MSE, que são medidas de desempenho comuns para modelos de classificação (Japkowicz, 2013)(Chawla et al., 2007). Essas medidas de desempenho são apresentadas na Tabela 5.13 e Tabela 5.14.

A Acurácia é uma boa indicação geral de performance dos modelos, mas podem haver situações em que ela é enganosa. Uma dessas situações é a Acurácia da previsão de absenteísmo, na Tabela 5.13, que apresentou valores superiores a 0,9. Nesse exemplo, o número de casos considerados como absenteístas é pequeno em relação aos considerados não absenteístas. Ou seja, o resultado de acurácia, se analisado isoladamente, estaria validando como ótimos os modelos que falham em prever absenteístas.

Na Tabela 5.13, os dados de Revocação são mais importantes do que os dados de precisão. Os dados de Revocação têm mais importância, porque os modelos têm a função de encontrar os funcionários absenteístas, mesmo que classifiquem alguns funcionários não absenteístas nessa situação. A grande falha dos modelos pode estar em classificar uma pessoa absenteísta como não absenteísta. Outro fato que observa-se são as Revocações dos modelos de absenteísmos acima do valor do modelo de referência NB assim como o F-Score, que é uma média entre Precisão e Revocação.

O MSE dos modelos se manteve baixo, na Tabela 5.13, destacando-se o modelo MLP

| | Medidas de Desempenho | | | | | |
|---|---|---|---|---|---|---|
| **Modelo** | **F-score** | **Precisão** | **Revocação** | **Acurácia** | **MSE** | **ROC AUC** |
| MLP | 0,64 | 0,95 | 0,51 | 0,63 | 0,37 | 0,81 |
| NB | 0,53 | 0,90 | 0,39 | 0,55 | 0,45 | 0,82 |
| XGBoost | 0,58 | 0,97 | 0,42 | 0,59 | 0,41 | 0,86 |
| RF | 0,62 | 0,97 | 0,48 | 0,62 | 0,38 | 0,86 |
| LR | 0,64 | 0,97 | 0,49 | 0,63 | 0,37 | 0,87 |

Tabela 5.14: Média de desempenho dos classificadores de produtividade com dados completos

com o maior erro. Se a população absenteísta for proporcionalmente menor, qualquer alteração que gere um erro significativo pode alterar os resultados. Esse fato torna os modelos de absenteísmo muito sensíveis.

A Acurácia, na Tabela 5.14, segue a proporção dos registros de funcionários produtivos positivos na semana. Diferentemente da situação enganosa para o conjunto de funcionários absenteístas, para o problema de produtividade, essa é uma boa indicação geral de performance.

A Precisão apresentada, na Tabela 5.14, é um dado muito importante, porque quando um funcionário é classificado como produtivo, é necessário que o modelo esteja correto, mesmo que acabe classificando funcionários produtivos como não produtivos. Nessa métrica, a grande falha dos modelos pode estar na classificação de um funcionário não produtivo como produtivo.

O valor MSE mais elevado, na Tabela 5.14, é o valor do modelo NB, modelo usado de referência para os demais. Como a medida MSE é, frequentemente, usada na verificação da acurácia, é uma boa indicação para o problema de produtividade.

### 5.3.4 A previsão de desempenho e o valor de negócio

| | **Anual** | **Semanal** |
|---|---|---|
| **Faturamento da empresa** | 176.545.450€ | 3.678.030€ |
| **Faturamento por operador** | 16.100€ | 340€ |
| **Custo da empresa com salários** | 47.770.630€ | 995.220€ |
| **Custo da empresa com salários por operador** | 4.360€ | 90€ |
| **Faturamento recuperado com a previsão de absenteísmo** | 9.533.450€ | 198.610€ |

Tabela 5.15: Efeito financeiro com aplicação da arquitetura de previsão de absenteísmo para o ano de 2018

A análise financeira referente à produtividade envolve alguns pontos estratégicos que não podem ser expostos pela empresa-alvo, mas o absenteísmo, é um indicador que, financeiramente, foi disponibilizado, medido e validado com os setores de planejamento e gestão.

A solução de previsão de absenteísmo teve um efeito financeiro positivo na aplicação real para o ano de 2018, o que pode ser observado na Tabela 5.15. Nessa tabela, são apresentados os faturamentos e custos que a empresa de teleatendimento obteve no ano de 2018, por operador, e em valores gerais aproximados anual e semanal.

Para alcançar a esses valores financeiros aproximados, foram comparados os absenteístas no ano 2017, que eram em média 13,8% dos operadores de teleatendimento, e em 2018, que sofreu uma queda para a média de 8,4%. A execução das previsões de absenteísmo tiveram início no mês de outubro de 2017.

Como aconteceu uma queda do número de absenteísta de 2017 para 2018 e esse percentual se manteve baixo, em média 7,4% no início de 2019, acredita-se que a previsão, deste projeto, foi um dos fatores que influenciaram a queda no número de funcionários absenteístas da empresa. O faturamento recuperado com a previsão de absenteísmo representa a diferença de valores financeiros que a empresa não perdeu.

Não foi possível obter o retorno financeiro referente a previsão de produtividade, como mencionado inicialmente, por questões administrativas da empresa-alvo do estudo.

# Capítulo 6

# Conclusões

Importa realçar que este trabalho foi desenvolvido num contexto empresarial de uma das maiores empresas de teleatendimento do Brasil. O propósito principal do trabalho teve como enfoque a investigação das principais características dos operadores de teleatendimento dessa empresa, respeitando regras éticas e legislativas, com o intuito de fornecer às suas equipes de gestão mecanismos automáticos para determinar o desempenho desses mesmos operadores. Assim, procurou-se desenvolver um *pipeline* que permitisse utilizar e selecionar um conjunto vasto de informação sobre os operadores de teleatendimento da empresa, com vista à previsão do absenteísmo e da produtividade desses trabalhadores.

Os estudos efetuados focaram-se na exploração e comparação de vários modelos de ML, descritos na literatura, com o intuito de determinar quais desses modelos apresentavam os melhores resultados de previsão do desempenho (cf. absenteísmo e produtividade) dos teleatendentes da empresa. Para além da análise e comparação dos vários modelos de ML, na previsão do desempenho, foram ainda aplicados e combinados vários algoritmos e técnicas para selecionar dinamicamente as variáveis que mais contribuíam para os resultados de previsão de desempenho, obtidos através dos diferentes algoritmos, bem como para selecionar as melhores parametrizações desses algoritmos.

As técnicas de seleção da hiper-parametrização dos algoritmos de ML, em conjunto com as técnicas de seleção de variáveis que alimentaram esses algoritmos, permitiram obter resultados muito interessantes e promissores que forneceram às equipes de gestão as ferramentas almejadas. Desse modo, consideramos que os resultados obtidos contribuem não só com valor acadêmico para a investigação e construção do conhecimento sobre a aplicação de algoritmos de aprendizado de máquina, mas também com valor empresarial e econômico, uma vez que o *pipeline* desenvolvido passou a ser incorporado no modelo de gestão da empresa, com ganhos evidentes e quantificáveis.

Outro aspecto importante analisado neste trabalho, foi o tempo de execução dos modelos. Verificou-se que, o tempo de execução para alguns dos modelos de previsão seria muito extenso, se executados na totalidade das 16 semanas da pesquisa. Para selecionar as soluções mais viáveis, foi preciso verificar que os modelos de previsão a serem aplicados

conseguiam resultados em tempo útil e quando comparados ao modelo NB. Esse foi um dos objetivos cumpridos que permitiu desconsiderar os modelos LSTM e SVM, das as suas limitações no tempo de treinamento e utilização para previsão.

A previsão de absenteísmo, mesmo tendo uma precisão menor quando comparada com a previsão de produtividade, permitiu que a gestão da empresa de teleatendimento diminuísse as perdas relacionadas com faltas dos seus recursos humanos no ano de 2018 (quando comparadas com o ano anterior). Os valores expostos na pesquisa foram validados por coordenadores, gerentes e funcionários da área comercial da empresa. Infelizmente, não foi possível obter a mesma validação acerca da produtividade. Contudo, a aplicação das ferramentas de previsão de desempenho e sua incorporação no processo de gestão da empresa foi considerada muito positiva pela organização e, em particular, pela gestão de funcionários, com efeitos nos relatórios e alertas de gestão da empresa.

Sabemos que as empresas procuram soluções na tecnologia, para melhorar o seu desempenho produtivo. Todo conhecimento desenvolvido, neste trabalho, para a definição e afinação de um *pipeline* de previsão de desempenho, com base em algoritmos de ML, passou a ser utilizado pela empresa-alvo deste estudo, buscando melhorar ganhos econômicos e abrindo também iniciativas para novas aplicações. Este trabalho contribui, assim, com uma solução concreta que tira partido de algoritmos de ML para uma aplicação real na gestão de recursos humanos da empresa de teleatendimento. Ressalta-se no entanto que a aplicação do *pipeline* de previsão de desempenho desenvolvido não se limita, especificamente, a essa empresa; consideramos que todo o arcabouço gerado tem potencial de aplicação em outras empresas com modelos semelhantes de gestão de recursos humanos.

Este trabalho procurou avaliar vários modelos de classificação baseados em ML e mostrar que, embora sejam sensíveis à correlação dos recursos de entrada com a variável de resposta, muitos deles têm um impacto maior nos resultados quando combinados estrategicamente. O ponto de partida do estudo e a base principal para as suas etapas foi entender o problema de previsão de desempenho dos colaboradores da empresa de teleatendimento e tentar solucioná-lo. Nesse sentido, foi possível transformar o ambiente profissional, em que se desenvolveu a tese, num laboratório de pesquisa no qual todo o contexto da empresa pôde ser compreendido e vivenciado internamente. O próprio autor foi um colaborador da empresa e, também, analista dos dados presentes nela. Entender o contexto do problema e seus detalhes estando inserido na empresa-alvo facilitou bastante a análise de todas as peculiaridades pertinentes ao ambiente cultural da corporação.

Consideramos que esse é, portanto, um dos fatores diferenciadores deste trabalho, uma vez que foi desenvolvido e aplicado numa empresa real e, de forma gratificante, trouxe melhorias para a gestão e benefícios quantificáveis para a empresa. Futuramente, a empresa poderá fazer uso da mesma arquitetura descrita neste trabalho, para outros indicadores de desempenho, aproveitando as informações de profissionais com as quais já conta, bem como outras informações externas de redes de contatos que poderão melhorar

e refinar a análise.

Mesmo sabendo que existem outros trabalhos prevendo o desempenho de funcionários com resultados interessantes, o estudo apresentado nesta tese, se destaca pela arquitetura do *pipeline* desenvolvido, pelas características dos dados e pelas técnicas utilizadas para abordar o problema. Mais especificamente: outros estudos não apresentam uma seleção tão completa e diversificada de previsão, nem uma quantidade tão expressiva de colaboradores e dados recolhidos. Este estudo combina a seleção de variáveis de entrada dos algoritmos de previsão, em concomitância com uma escolha cuidadosa e sistemática da hiper-parametrização desses algoritmos, para optimizar a obtenção dos melhores resultados.

Realça-se ainda, que a estrutura da empresa de teleatendimento, seu formato de trabalho, segurança de dados e valores sociais foram analisados e foi tomado o devido cuidado de consultar as regras vigentes na legislação do Brasil para que nenhuma lei fosse infringida na manipulação dos dados dos funcionários de teleatendimento. Garantiu-se ainda que nenhum quesito ético fosse quebrado. Para isso, toda a informação recolhida para a pesquisa teve o consentimento informado dos funcionários, além da aprovação da empresa. O estudo foi, ainda, submetido e aprovado por dois comitês éticos de pesquisa, um do Brasil, conforme o programa nacional de pesquisa, e o outro de Portugal, dentro das diretrizes da universidade Fernando Pessoa, que orientou este trabalho.

Para finalizar, referimos apenas que a segurança da informação é de grande relevância e uma preocupação generalizada de instituições e governos. Neste trabalho,especificamente, a quantidade de dados pessoais trabalhada teve um valor significativo, e um eventual vazamento de informações poderia ter consequências nefastas para a empresa de teleatendimento e seus colaboradores. Assim, procurou-se tomar todos os cuidados possíveis com a informação recolhida. No que concerne à manipulação dos dados, para que os funcionários pudessem ficar protegidos nas bases de dados, foram utilizados dados sem identificação específica, ou seja, anonimizados para o seu processamento.

## 6.1 Investigação futura consequente

Neste trabalho, foi possível identificar alguns dos indicadores de desempenho com mais preponderância entre as variáveis recolhidas sobre os colaboradores. Mais concretamente, esses indicadores de desempenho são a pausa no trabalho feita pelos funcionários, o tempo que o operador fica logado no gerenciamento das ligações, o grau de retorno de chamadas dos clientes que os operadores de teleatendimento atendem, o tempo médio que o operador fica atendendo os seus clientes por telefone e o resultado geral de bonificação em moedas virtuais obtidas num sistema similar a uma rede social usada internamente na empresa. Esses indicadores de desempenho podem ser também alvo de um estudo para a sua previsão, usando a mesma arquitetura do *pipeline* proposto e implementada nesta

pesquisa. De fato, a solução proposta não se aplica apenas aos indicadores específicos utilizados nesta empresa de teleatendimento, podendo generalizar-se a problemas específicos de outras empresas com modelos de negócio e de gestão similares.

A utilização de um *ensemble* contemplando uma previsão pesada de vários algoritmos de ML em conjunto, poderá aprimorar a previsão de desempenho e tornar-se numa ferramenta mais promissora. Assim, futuramente poderá estudar-se a aplicação aos melhores modelos, de uma metodologia em que a saída de previsão de cada modelo é avaliada ou pesada numa escolha pesada ou por maioria. Por exemplo, num conjunto selecionado de três modelos de classificação, se dois deles expõem um resultado negativo para produtividade, e o terceiro um resultado positivo, uma previsão em *ensemble* poderia melhorar a resposta. Como dois dos modelos afirmam que a produtividade de um funcionário é negativa, e um dos modelos afirma que a produtividade é positiva, por maioria a saída final se tornaria negativa para produtividade. Seria interessante explorar e comparar diferentes abordagens nesta perspetiva.

Por fim, importa referir que o problema de previsão de absenteísmo teve um grau de dificuldade superior ao problema de previsão de produtividade, pois verificou-se que as classes presentes, nesta última, são menos desproporcionais quando comparadas com as classes presentes no *data-set* de absenteístas. O uso de técnicas de re-amostragem com dados artificiais é uma solução que poderá ser adotada para problemas de *data-set* que não tenham classes balanceadas. Neste trabalho, essas técnicas não foram equacionadas, uma vez que exigem mais estudos e análises para não causarem ruídos nos resultados. Os dados sobre o absenteísmo são particularmente sensíveis nesse sentido. Sabe-se, ainda, que questões técnicas sobre a ordem e o ciclo da hiper-parametrização e seleção de variáveis podem ser testadas e usadas, assim como a re-amostragem para criação de características artificiais, dependendo do problema abordado em bases de dados desbalanceadas. Embora este trabalho tenha abordado, de forma consistente e com êxito, vários pontos importantes na previsão de desempenho, consideramos que, futuramente, poderiam ser exploradas novas técnicas neste sentido.

# Anexo A - Variáveis

As variáveis listadas a seguir são tratadas como $X_1$, $X_2$, $X_3$ ... $X_{168}$ neste projeto de pesquisa.

1. produtividade: Produtividade é a variável alvo da previsão de grupo de produtividade. Essa variável é também usada como uma das variáveis na previsão de absenteísmo. Ela é apresentada como uma variável binária indicando se os colaboradores são produtivos ou não. É uma das variáveis que, junto ao absenteísmo, compõe o desempenho dos colaboradores.

2. absenteísmo: Absenteísmo é a variável alvo da previsão de absenteísmo. Essa variável é também usada como uma das variáveis na previsão de produtividade. Ela é apresentada como uma variável binária indicando se os colaboradores são absenteístas ou não. É uma das variáveis que, junto à produtividade, compõe o desempenho dos colaboradores.

3. histórico de absenteísmo no domingo: Nessa variável está o histórico de absenteísmo no domingo. Apresenta quantas vezes os colaboradores foram absenteístas, no domingo, em um intervalo de 4 meses.

4. histórico de absenteísmo na segunda: Nessa variável está o histórico de absenteísmo na segunda. Apresenta quantas vezes os colaboradores foram absenteístas, na segunda-feira, em um intervalo de 4 meses.

5. histórico de absenteísmo na terça: Nessa variável está o histórico de absenteísmo na terça. Apresenta quantas vezes os colaboradores foram absenteístas, na terça-feira, em um intervalo de 4 meses.

6. histórico de absenteísmo na quarta: Nessa variável está o histórico de absenteísmo na quarta. Apresenta quantas vezes os colaboradores foram absenteístas, na quarta-feira, em um intervalo de 4 meses.

7. histórico de absenteísmo na quinta: Nessa variável está o histórico de absenteísmo na quinta. Apresenta quantas vezes os colaboradores foram absenteístas, na quinta-feira, em um intervalo de 4 meses.

8. histórico de absenteísmo na sexta: Nessa variável está o histórico de absenteísmo na sexta. Apresenta quantas vezes os colaboradores foram absenteístas, na sexta-feira, em um intervalo de 4 meses.

9. histórico de absenteísmo no sábado: Nessa variável está o histórico de absenteísmo no sábado. Apresenta quantas vezes os colaboradores foram absenteístas, no sábado, em um intervalo de 4 meses.

10. autoavaliação: A autoavaliação é como os colaboradores se avaliam em produtividade. Um colaborador pode se achar produtivo e se autoavaliar improdutivo, ou o contrário. Essa variável indica como o colaborador se vê dentro do quadro de produtividade.

11. avaliação de ambiente: Essa variável representa a avaliação do ambiente de trabalho feita pelos colaboradores. Essa avaliação revela se os colaboradores gostam do lugar onde estão trabalhando.

12. avaliação sobre o gestor: Os colaboradores podem avaliar os seus superiores imediatos como bons ou ruins. Tal informação de avaliação feita pelos colaboradores sobre os seus gestores está presente nessa variável.

13. carga horária: A carga horária dos colaboradores não representa o número de horas que eles trabalham efetivamente, mas sim o período formal pelo qual eles devem trabalhar. O registro de horas no documento trabalhista dos colaboradores pode variar entre 4, 6 ou 8 horas de trabalho por dia.

14. cidade de trabalho: Cidade em que os aperadores de teleatendimento trabalham. As cidades onde os operadores trabalham não são, necessariamente, as cidades nas quais eles têm residência fixa.

15. situação no indicador de pausa: Essa variável define se a situação de produtividade dos operadores é baixa, média ou alta, no indicador de pausa, considerando média e desvio-padrão.

16. situação em resultados de rechamada: Essa variável define se a situação de produtividade dos operadores é baixa, média ou alta, no indicador de rechamada, considerando média e desvio-padrão.

17. situação de tempo logado: Essa variável define se a situação de produtividade dos operadores é baixa, média ou alta, no indicador de tempo logado, considerando média e desvio-padrão.

18. situação em tempo médio de atendimento: Essa variável define se a situação de produtividade dos operadores é baixa, média ou alta, no indicador de tempo médio de atendimento, considerando média e desvio-padrão.

19. situação em resultados bonificados: Essa variável define se a situação de produtividade dos operadores é baixa, média ou alta, no indicador de resultados bonificados, considerando média e desvio-padrão.

20. código telefônico: Essa variável representa o código telefônico dos operadores. Geralmente, o código telefônico representa a região em que os colaboradores residem ou trabalham, mas o advento da portabilidade numérica faz com que o registro do número possa ser feito em localidades distintas do território brasileiro.

21. distância do trabalho: Do trabalho dos colaboradores até suas residências há uma distância que pode ser longa ou curta, mas Essa variável não entra nesse mérito. É registrada uma informação contínua da distância em quilômetros.

22. escolaridade: Muitos dos colaboradores possuem nível de estudo acima do fundamental, mas essa categoria é bem dividida entre eles. nessa variável são registrados os níveis de ensino alcançados pelos colaboradores.

23. estado civil: Essa variável representa o estado de relacionamento particular, porém registrado, formalmente, pelos colaboradores. Eles podem estar casado, solteiro ou divorciado.

24. estado de nascimento: O estado brasileiro onde os colaboradores trabalham não é, necessariamente, o mesmo em que nasceram; apresenta-se, portanto, a região brasileira onde os colaboradores têm o seu registro de nascimento.

25. faixa etária: De acordo com a idade, os colaboradores podem ser classificados como jovens, adultos ou idosos. Essa variável representa a faixa etária em que os colaboradores se encontram.

26. faixa etária do gestor: De acordo com as suas idades, os superiores imediatos dos colaboradores podem ser classificados como jovens, adultos ou idosos. Essa variável representa a faixa etária em que os chefes dos colaboradores se encontram.

27. cidade local de trabalho: Essa é uma variável binária que indica se a cidade em que os colaboradores estão trabalhando é também a cidade em que eles possuem residência fixa.

28. cadastro de cidadão na região natal: Os colaboradores podem ou não estar registrados como cidadãos em suas regiões de nascimento. Essa variável identifica tal situação.

29. feriado no período: Essa é a identificação da ocorrência de feriado no período a ser previsto, pois um feriado pode ser um indicativo empírico de que pode acontecer uma queda de produtividade ou um índice alto de absenteísmo.

30. imigrante: Informação que revela se os colaboradores são imigrantes ou não. Geralmente, as maiores unidades de teleatendimento estão em grandes centros comerciais, o que incentiva muitas pessoas a migrarem do interior do país para os locais com maiores ofertas de trabalho.

31. telefone fixo: Variável binária que indica se os colaboradores possuem telefone fixo ou não. O dado sobre o usuário possuir telefone móvel acaba ficando subentendido,nessa informação,visto que,no Brasil,houve uma grande migração para tecnologia móvel e,consequentemente,abandono do telefone fixo.

32. grupo de produtividade histórico: O histórico de produtividade geral é o que apresenta quantas veze os colaboradores ficaram,no grupo maior de produtividade, em um intervalo de 4 meses.

33. grupo em que ficou maior tempo: Dentre os quatro grupos de produtividade no sistema de gestão de resultados da empresa de teleatendimento, essa variável demonstra qual desses grupos os colaboradores estiveram a maior parte do tempo de trabalho, em um intervalo de 4 meses.

34. grupo de resultados bonificados 1: Essa variável informa em qual grupo, dentre os quatro grupos de produtividade no sistema de gestão de resultados da empresa de teleatendimento, estão os colaboradores que estiveram no indicador de resultados bonificados na semana anterior tendo a semana atual como referência.

35. grupo de resultados bonificados 2: Essa variável informa em qual grupo, dentre os quatro grupos de produtividade no sistema de gestão de resultados da empresa de teleatendimento, estão os colaboradores que estiveram no indicador de resultados bonificados, há duas semanas, tendo a semana atual como referência.

36. grupo de resultados bonificados 3: Essa variável informa em qual grupo, dentre os quatro grupos de produtividade no sistema de gestão de resultados da empresa de teleatendimento, estão os colaboradores que estiveram no indicador de resultados bonificados, há três semanas, tendo a semana atual como referência.

37. grupo de resultados bonificados 4: Essa variável informa em qual grupo, dentre os quatro grupos de produtividade no sistema de gestão de resultados da empresa de teleatendimento, estão os colaboradores que estiveram no indicador de resultados bonificados, há quatro semanas, tendo a semana atual como referência.

38. grupo de resultados bonificados 5: Essa variável informa em qual grupo, dentre os quatro grupos de produtividade no sistema de gestão de resultados da empresa de teleatendimento, estão os colaboradores que estiveram no indicador de resultados bonificados, há cinco semanas, tendo a semana atual como referência.

39. grupo de resultados bonificados 6: Essa variável informa em qual grupo, dentre os quatro grupos de produtividade no sistema de gestão de resultados da empresa de teleatendimento, estão os colaboradores que estiveram no indicador de resultados bonificados, há seis semanas, tendo a semana atual como referência.

40. grupo de resultados bonificados 7: Essa variável informa em qual grupo, dentre os quatro grupos de produtividade no sistema de gestão de resultados da empresa de teleatendimento, estão os colaboradores que estiveram no indicador de resultados bonificados, há sete semanas, tendo a semana atual como referência.

41. grupo de resultados bonificados 8: Essa variável informa em qual grupo, dentre os quatro grupos de produtividade no sistema de gestão de resultados da empresa de teleatendimento, estão os colaboradores que estiveram no indicador de resultados bonificados, há oito semanas, tendo a semana atual como referência.

42. horário de marcação do humor: Para entrar no sistema de gestão de resultados da empresa, os colaboradores têm de registrar o seu humor no dia (bom, médio ou ruim). O horário do registro de humor no dia é, consequentemente, o horário de início de suas atividades profissionais. nessa variável está armazenada essa informação.

43. idade: Variável que contém a informação da idade dos colaboradores.

44. idade do gestor: Variável que contém a informação da idade dos gestores imediatos dos colaboradores.

45. maior tempo consecutivo G1: Os colaboradores podem atingir o maior grupo de produtividade e podem cair para outros grupos no decorrer dos meses de trabalho. nessa variável fica registrado o maior tempo consecutivo em que os colaboradores estiveram no melhor grupo de produtividade no indicador de resultados bonificados por um período de três meses.

46. maior tempo consecutivo G2: Os colaboradores podem atingir o segundo maior grupo de produtividade e podem cair para outros grupos, ou até mesmo podem subir para o melhor deles, no decorrer dos meses de trabalho. nessa variável fica registrado o maior tempo consecutivo em que os colaboradores estiveram no segundo melhor grupo de produtividade no indicador de resultados bonificados por um período de três meses.

47. maior tempo consecutivo G3: Os colaboradores podem atingir um grupo de produtividade e podem cair, ou podem subir para outros grupos no decorrer dos meses de trabalho. nessa variável fica registrado o maior tempo consecutivo em que os colaboradores estiveram no grupo de produtividade intermediária no indicador de resultados bonificados por um período de três meses.

48. maior tempo consecutivo G4: Os colaboradores podem atingir o pior grupo de produtividade e subir para outros grupos no decorrer dos meses de trabalho. nessa variável fica registrado o maior tempo consecutivo em que os colaboradores estiveram no pior grupo de produtividade no indicador de resultados bonificados por um período de três meses.

49. média de avaliações de dicas: Como no sistema de gestão de resultados dos colaboradores existe a possibilidade de avaliar dicas de trabalho criadas por outros colaboradores, essa variável registra a média das dicas avaliadas.

50. média percentual de absenteísmo dos amigos: Assim como é comum em redes sociais, dentro do sistema de gestão de resultados é possível agregar amigos de trabalho para auxiliar no desenvolvimento de atividades. A média de absenteísmo dos amigos adicionados é representada nessa variável.

51. naturalidade capital: Essa variável identifica os colaboradores como nascidos na capital de seus estados ou no interior.

52. número de departamentos no período: Os colaboradores podem trocar seus postos de trabalho no decorrer dos meses. Essa variável registra o número de trocas de postos de trabalho entre departamentos pelos colaboradores no intervalo de três meses.

53. número de colaboradores no departamento: O colaborador não trabalha isolado em um departamento, mas sim com vários outros colaboradores. O número de colaboradores no departamento de trabalho é representado nessa variável.

54. número de compras de avatar: No ambiente virtual do sistema de gestão de resultados dos colaboradores, é possível que ele construa um personagem digital. É possível, ainda, que ele compre itens para esse personagem usando recursos financeiros alcançados com as suas pontuações nos indicadores de desempenho.

55. número de compras de benefício: No ambiente virtual do sistema de gestão de resultados dos colaboradores, é possível que ele compre benefícios como dias de folga, entradas para cinema, entradas para teatro e visita a um restaurante, dentre outros. É possível que ele compre esses itens usando recursos financeiros alcançados com as suas pontuações nos indicadores de desempenho.

56. número compras físicas: No ambiente virtual do sistema de gestão de resultados dos colaboradores, é possível que ele compre objetos como copos, mochilas, camisetas, cadernos, etc. É possível que ele compre esses itens usando recursos financeiros alcançados com as suas pontuações nos indicadores de desempenho.

57. número de dependentes crianças e/ou idosos: Essa variável considera o número de crianças e/ou idosos que são dependentes dos colaboradores. Essa diferenciação é feita, pois estes dependentes possuem uma faixa de idade mais vulnerável e que pode requerer uma maior atenção do responsável.

58. número de dias desde o último descanso: Está presentes o número de dias que os colaboradores têm trabalhado desde a última vez que descansaram. É comum que se sigam as leis trabalhistas vigentes nessa variável para que não ultrapassem as horas normais trabalhadas na semana, mas as horas podem variar para menos.

59. número de domingos trabalhados: Uma parte dos trabalhadores tem escala para os domingos, sendo que muitos outros trabalham somente durante os outros dias da semana. nessa variável ficam registrados os números de domingos em que os colaboradores foram escalados para trabalhar.

60. número de gestores: Os colaboradores podem mudar de superior imediato durante sua carreira profissional dentro da empresa. O número de gestores, com os quais os colaboradores estiveram, também pode mudar em um curto espaço de tempo. nessa variável registra-se o número de gestores com os quais os colaboradores estiveram no intervalo de uma semana.

61. número de humores baixos seguidos: Nessa variável, são registradas,diariamente, quantas vezes os colaboradores marcaram, seguidamente, humores baixos no sistema de gestão de resultados.

62. número de humores altos seguidos: Nessa variável, são registradas, diariamente, quantas vezes os colaboradores marcaram, seguidamente, humores altos no sistema de gestão de resultados.

63. número de produtos favoritados: Assim como os colaboradores podem adquirir produtos a partir dos seus ganhos financeiros com os resultados bonificados, eles também podem favoritar aqueles produtos que não desejam comprar, naquele momento, ou não têm recurso financeiro para adquirir.

64. número de trocas de perfil: As trocas de perfil de usuário são monitoradas dentro do sistema de gestão de resultados. Registram-se, nessa variável, quantas vezes os colaboradores trocam seu perfil no sistema.

65. histórico do número de dias desde a última falta: A ausência do colaborador no ambiente de trabalho pode ter vários motivos, mas o mérito dessa variável é registrar quantos dias que os funcionários estiveram presentes, na empresa, desde sua última falta, independente da causa.

66. número de dependentes: Nessa variável fica registrado o número de dependentes dos colaboradores independente da faixa etária.

67. número de dicas avaliadas: Normalmente, os colaboradores recebem e criam dicas de trabalho dentro do próprio sistema de gestão de resultados. Essas dicas podem ter avaliações e o número de dicas avaliadas pelos colaboradores é registrado nessa variável.

68. número de dicas recebidas: Normalmente os colaboradores avaliam e criam dicas de trabalho dentro do próprio sistema de gestão de resultados. Essas dicas são recebidas por eles, e o número de dicas recebidas pelos colaboradores, é registrado nessa variável.

69. número de mensagens favoritadas: Essa variável contém o registro das mensagens das quais os colaboradores gostaram e marcaram como favoritas dentro do sistema de gerenciamento de resultados.

70. número de mensagens não automáticas: Além de mensagens automáticas que colaboradores recebem a respeito de produtos, dicas e desempenho profissional, existem mensagens vindas de vários departamentos da empresa. São anúncios e mensagens emitidas de forma direta. Esse número de mensagens recebidas, por interação humana pelos colaboradores, é registrado nessa variável.

71. número de mensagens automáticas: Assim como os usuários recebem dicas de melhoria do seu trabalho dentro do próprio sistema de gestão de resultados, eles recebem, também, mensagens automáticas, parabenizando-os, caso o resultado esteja bom; ou alertando-os com informações de como melhorar esses resultados, caso seja apresentado um resultado insatisfatório. Essa variável registra o número de mensagens, dessa natureza, recebidas pelos colaboradores.

72. número de mensagens de recomendações: Assim como os usuários recebem dicas de melhoria do seu trabalho dentro do próprio sistema de gestão de resultados, eles, também, recebem mensagens automáticas contendo informações se existe algum produto ideal para ele na loja virtual, algum produto recomendado e, também, se existem dicas de trabalho conforme o seu perfil. Essa variável registra o número de mensagens, dessa natureza, recebidas pelos colaboradores.

73. percentual de absenteísmo dos gestores no departamento: Essa variável registra a média percentual de absenteísmo dos gestores no departamento de trabalho dos colaboradores.

74. percentual de absenteísmo 1: Nessa variável está presente a informação de percentual de absenteísmo dos colaboradores, na semana anterior, tendo a semana atual como referência.

75. percentual de absenteísmo 2: Nessa variável está presente a informação de percentual de absenteísmo dos colaboradores, há duas semanas, tendo a semana atual como referência.

76. percentual de absenteísmo 3: Nessa variável está presente a informação de percentual de absenteísmo dos colaboradores, há três semanas, tendo a semana atual como referência.

77. percentual de absenteísmo 4: Nessa variável está presente a informação de percentual de absenteísmo dos colaboradores, há quatro semanas, tendo a semana atual como referência.

78. percentual de absenteísmo 5: Nessa variável está presente a informação de percentual de absenteísmo dos colaboradores, há cinco semanas, tendo a semana atual como referência.

79. percentual de absenteísmo 6: Nessa variável está presente a informação de percentual de absenteísmo dos colaboradores, há seis semanas, tendo a semana atual como referência.

80. percentual de absenteísmo 7: Nessa variável está presente a informação de percentual de absenteísmo dos colaboradores, há sete semanas, tendo a semana atual como referência.

81. percentual de absenteísmo 8: Nessa variável está presente a informação de percentual de absenteísmo dos colaboradores, há oito semanas, tendo a semana atual como referência.

82. percentual de atingimento no indicador de bonificação dos amigos: O sistema de gestão de resultados têm algumas funcionalidades de uma rede social, e em uma delas, é possível adicionar amigos. Essa variável em questão registra o percentual de atingimento no indicador que bonifica, financeiramente, os colaboradores conforme seu desempenho.

83. histórico percentual de atraso no feriado: Os colaboradores escalados para trabalhar, nos feriados, podem ter o costume de se atrasar por ser um período um pouco

informal no qual a sociedade costuma descansar do trabalho ou ir a eventos festivos. nessa variável ficam registrados os atrasos contabilizados nas últimas 12 semanas.

84. histórico percentual de faltas no feriado: Os colaboradores escalados para trabalhar nos feriados podem ter o costume de faltar por ser um período um pouco informal no qual a sociedade costuma descansar do trabalho ou ir a eventos festivos. nessa variável ficam registradas as faltas contabilizadas nas últimas 12 semanas.

85. histórico percentual de faltas consecutivas: Nessa variável ficam registradas as faltas consecutivas das últimas 12 semanas.

86. histórico percentual de faltas não consecutivas: Nessa variável ficam registradas as faltas que não foram consecutivas das últimas 12 semanas.

87. percentual de humor alto gestor: O positivismo e alegria dos superiores imediatos podem influenciar no trabalho dos colaboradores. Essa variável contém o valor percentual de marcação de humor alto do gestor no sistema de gestão de resultados na última semana.

88. percentual de humor alto: O positivismo e alegria dos colaboradores podem influenciar no trabalho. Essa variável contém o valor percentual de marcação de humor alto dos colaboradores no sistema de gestão de resultados na última semana.

89. percentual de humor baixo do gestor: O negativismo e tristeza dos superiores imediatos podem influenciar no trabalho dos colaboradores. Essa variável contém o valor percentual de marcação de humor baixo do gestor no sistema de gestão de resultados na última semana.

90. percentual de humor baixo: O negativismo e tristeza dos colaboradores podem influenciar no trabalho. Essa variável contém o valor percentual de marcação de humor baixo dos colaboradores no sistema de gestão de resultados na última semana.

91. histórico percentual motivo de humor família: No sistema de gestão de resultados, quando os colaboradores marcam 3 humores tristes seguidos, eles têm a opção de deixar uma declaração do motivo desse humor baixo. Uma dessas opções é problemas na família. Essa variável registra essa situação do humor dos usuários.

92. histórico percentual motivo de humor não obrigado: No sistema de gestão de resultados, quando os colaboradores marcam 3 humores tristes seguidos, eles têm a opção de deixar uma declaração do motivo desse humor baixo. Uma dessas opções é não declarar o que está acontecendo é, simplesmente, omitir. Essa variável registra essa situação do humor dos usuários.

93. histórico percentual motivo de humor outros: No sistema de gestão de resultados, quando os colaboradores marcam 3 humores tristes seguidos, eles têm a opção de deixar uma declaração do motivo desse humor baixo. Uma dessas opções é problemas diversos para quaisquer outras situações particulares. Essa variável registra essa situação do humor dos usuários.

94. histórico percentual motivo de humor saúde: No sistema de gestão de resultados, quando os colaboradores marcam 3 humores tristes seguidos, eles têm a opção de deixar uma declaração do motivo desse humor baixo. Uma dessas opções é problemas de saúde. Essa variável registra essa situação do humor dos usuários.

95. histórico percentual motivo de humor trabalho: No sistema de gestão de resultados, quando os colaboradores marcam 3 humores tristes seguidos, eles têm a opção de deixar uma declaração do motivo desse humor baixo. Uma dessas opções é problemas no trabalho. Essa variável registra essa situação do humor dos usuários.

96. percentual de atingimento no indicador de bonificação 1: Nessa variável está presente a informação de quanto os colaboradores ganharam dividido pelo tanto que eles poderiam ter ganhado. o valor leva em consideração a semana anterior, no indicador de bonificação financeira, tendo a semana atual como referência.

97. percentual de atingimento no indicador de bonificação 2: Nessa variável está presente a informação de quanto os colaboradores ganharam dividido pelo tanto que eles poderiam ter ganhado. o valor leva em consideração duas semanas anteriores, no indicador de bonificação financeira, tendo a semana atual como referência.

98. percentual de atingimento no indicador de bonificação 3: Nessa variável está presente a informação de quanto os colaboradores ganharam dividido pelo tanto que eles poderiam ter ganhado. o valor leva em consideração três semanas anteriores, no indicador de bonificação financeira, tendo a semana atual como referência.

99. percentual de atingimento no indicador de bonificação 4: Nessa variável está presente a informação de quanto os colaboradores ganharam dividido pelo tanto que eles poderiam ter ganhado. o valor leva em consideração quatro semanas anteriores, no indicador de bonificação financeira, tendo a semana atual como referência.

100. percentual de atingimento no indicador de bonificação 5: Nessa variável está presente a informação de quanto os colaboradores ganharam dividido pelo tanto que eles poderiam ter ganhado. o valor leva em consideração cinco semanas anteriores, no indicador de bonificação financeira, tendo a semana atual como referência.

101. percentual de atingimento no indicador de bonificação 6: Nessa variável está presente a informação de quanto os colaboradores ganharam dividido pelo tanto que

eles poderiam ter ganhado. o valor leva em consideração seis semanas anteriores, no indicador de bonificação financeira, tendo a semana atual como referência.

102. percentual de atingimento no indicador de bonificação 7: Nessa variável está presente a informação de quanto os colaboradores ganharam dividido pelo tanto que eles poderiam ter ganhado. o valor leva em consideração sete semanas anteriores, no indicador de bonificação financeira, tendo a semana atual como referência.

103. percentual de atingimento no indicador de bonificação 8: Nessa variável está presente a informação de quanto os colaboradores ganharam dividido pelo tanto que eles poderiam ter ganhado. o valor leva em consideração oito semanas anteriores, no indicador de bonificação financeira, tendo a semana atual como referência.

104. histórico percentual de absenteísmo: Nessa variável ficam registrados os percentuais de absenteísmo dos colaboradores contabilizados nas últimas 12 semanas.

105. histórico da quantidade de dias de acesso ao sistema de gestão de resultados: Nessa variável ficam registrados os dias de acesso dos colaboradores ao sistema de gerenciamento de resultados contabilizados nas últimas 12 semanas.

106. quantidade de dias de acesso ao sistema de gestão de resultados: Nessa variável ficam registrados os dias de acesso dos colaboradores ao sistema de gerenciamento de resultados.

107. região registro nacional de cidadão: A região onde os colaboradores se registraram, nacionalmente, pode ter alguma relação com seu desempenho profissional, pois eles podem ser da mesma região onde trabalham ou podem ser imigrantes. nessa variável ficam registrados essas informações.

108. média de resultados do departamento no indicador de pausa: Essa variável registra a média de resultados dos colaboradores no indicador de pausa nos seus departamentos.

109. média de resultados do departamento no indicador de rechamada: Essa variável registra a média de resultados dos colaboradores no indicador de rechamada nos seus departamentos.

110. média de resultados do departamento no indicador de tempo logado: Essa variável registra a média de resultados dos colaboradores no indicador de tempo logado nos seus departamentos.

111. média de resultados do departamento no indicador de tempo médio de atendimento: Essa variável registra a média de resultados dos colaboradores no indicador de tempo médio de atendimento nos seus departamentos.

112. média de resultados do departamento no indicador de bonificação: Essa variável registra a média de resultados dos colaboradores no indicador de bonificação nos seus departamentos.

113. resultado em pausa: O resultado do indicador de pausa, na última semana, é registrado nessa variável.

114. resultado em rechamada: O resultado do indicador de rechamada, na última semana, é registrado nessa variável.

115. resultado em tempo logado: O resultado do indicador de tempo logado, na última semana, é registrado nessa variável.

116. resultado em tempo médio de atendimento: O resultado do indicador de tempo médio de atendimento, na última semana, é registrado nessa variável.

117. resultado em bonificação: O resultado do indicador de bonificação, na última semana, é registrado nessa variável.

118. distância de resultados no departamento no indicador pausa: Existe uma diferença entre os resultados dos colaboradores dentro dos departamentos. Alguns colaboradores podem ter grandes valores de produtividade em pausa como podem ter valores pequenos: existe, ainda, a possibilidade de o departamento ter valores equilibrados. Essa média de variação de resultados é registrada por essa variável.

119. distância de resultados no departamento no indicador de rechamada: Existe uma diferença entre os resultados dos colaboradores dentro dos departamentos. Alguns colaboradores podem ter grandes valores de produtividade em rechamada como podem ter valores pequenos: existe, ainda, a possibilidade de o departamento ter valores equilibrados. Essa média de variação de resultados é registrada por essa variável.

120. distância de resultados no departamento no indicador de tempo logado: Existe uma diferença entre os resultados dos colaboradores dentro dos departamentos. Alguns colaboradores podem ter grandes valores de produtividade em tempo logado como podem ter valores pequenos: existe, ainda, a possibilidade de o departamento ter valores equilibrados. Essa média de variação de resultados é registrada por essa variável.

121. distância de resultados no departamento no indicador de tempo médio de atendimento: Existe uma diferença entre os resultados dos colaboradores dentro dos departamentos. Alguns colaboradores podem ter grandes valores de produtividade em tempo

médio de atendimento como podem ter valores pequenos: existe, ainda, a possibilidade de o departamento ter valores equilibrados. Essa média de variação de resultados é registrada por essa variável.

122. distância de resultados no departamento no indicador bonificação: Existe uma diferença entre os resultados dos colaboradores dentro dos departamentos. Alguns colaboradores podem ter grandes valores de produtividade em bonificação como podem ter valores pequenos: existe, ainda, a possibilidade de o departamento ter valores equilibrados. Essa média de variação de resultados é registrada por essa variável.

123. grupo pausa: Essa variável indica em que grupo de produtividade os colaboradores estiveram, na última semana, indicador de pausa. Dentre os 4 grupos, temos o mais produtivo, o segundo mais produtivo, o de produtividade mediana e o de baixa produtividade.

124. grupo rechamada: Essa variável indica em que grupo de produtividade os colaboradores estiveram, na última semana, indicador de rechamada. Dentre os 4 grupos, temos o mais produtivo, o segundo mais produtivo, o de produtividade mediana e o de baixa produtividade.

125. grupo tempo logado: Essa variável indica em que grupo de produtividade os colaboradores estiveram, na última semana, indicador de tempo logado. Dentre os 4 grupos, temos o mais produtivo, o segundo mais produtivo, o de produtividade mediana e o de baixa produtividade.

126. grupo tempo médio de atendimento: Essa variável indica em que grupo de produtividade os colaboradores estiveram, na última semana, indicador de tempo médio de atendimento. Dentre os 4 grupos, temos o mais produtivo, o segundo mais produtivo, o de produtividade mediana e o de baixa produtividade.

127. grupo bonificação: Essa variável indica em que grupo de produtividade os colaboradores estiveram, na última semana, indicador de bonificação. Dentre os 4 grupos, temos o mais produtivo, o segundo mais produtivo, o de produtividade mediana e o de baixa produtividade.

128. resultado de produtividade 1: Nessa variável está presente a informação de quanto os colaboradores ganharam em produtividade. o valor leva em consideração a semana anterior, no indicador de bonificação financeira, tendo a semana atual como referência.

129. resultado de produtividade 2: Nessa variável está presente a informação de quanto os colaboradores ganharam em produtividade. o valor leva em consideração duas

semanas anteriores, no indicador de bonificação financeira, tendo a semana atual como referência.

130. resultado de produtividade 3: Nessa variável está presente a informação de quanto os colaboradores ganharam em produtividade. o valor leva em consideração três semanas anteriores, no indicador de bonificação financeira, tendo a semana atual como referência.

131. resultado de produtividade 4: Nessa variável está presente a informação de quanto os colaboradores ganharam em produtividade. o valor leva em consideração quatro semanas anteriores, no indicador de bonificação financeira, tendo a semana atual como referência.

132. resultado de produtividade 5: Nessa variável está presente a informação de quanto os colaboradores ganharam em produtividade. o valor leva em consideração cinco semanas anteriores, no indicador de bonificação financeira, tendo a semana atual como referência.

133. resultado de produtividade 6: Nessa variável está presente a informação de quanto os colaboradores ganharam em produtividade. o valor leva em consideração seis semanas anteriores, no indicador de bonificação financeira, tendo a semana atual como referência.

134. resultado de produtividade 7: Nessa variável está presente a informação de quanto os colaboradores ganharam em produtividade. o valor leva em consideração sete semanas anteriores, no indicador de bonificação financeira, tendo a semana atual como referência.

135. resultado de produtividade 8: Nessa variável está presente a informação de quanto os colaboradores ganharam em produtividade. o valor leva em consideração oito semanas anteriores, no indicador de bonificação financeira, tendo a semana atual como referência.

136. pontuação de humor do departamento: Essa variável contém qual é a situação de humor geral dos colaboradores do departamento na última semana. A pontuação é calculada a partir da diferença de humores positivos e negativos divididos pelo total de humores marcados nas últimas 12 semanas.

137. humor de pontuação distância: Existe uma diferença entre os humores dos colaboradores dentro dos departamentos. Alguns colaboradores podem ter humores positivos como podem ter humores negativos. Existe, ainda, a possibilidade de o departamento ter valores equilibrados. Essa média de variação de humor é registrada por

essa variável. A pontuação é calculada a partir da diferença de humores positivos e negativos divididos pelo total de humores marcados nas últimas 12 semanas.

138. pontuação de humor gestor: Essa variável contém a situação de humor do gestor. A pontuação é calculada a partir da diferença de humores positivos e negativos divididos pelo total de humores marcados na última semana.

139. histórico de pontuação de humor: Essa variável contém a situação de humor do gestor. A pontuação é calculada a partir da diferença de humores positivos e negativos divididos pelo total de humores marcados nas últimas 12 semanas.

140. pontuação humor: Essa variável contém a situação de humor do gestor. A pontuação é calculada a partir da diferença de humores positivos e negativos divididos pelo total de humores marcados na última semana.

141. semana do mês: Nessa variável se registra qual a semana do mês atual.

142. sexo gestor: Essa variável registra se os gestores são do sexo masculino ou feminino.

143. sexo: Essa variável registra se os colaboradores são do sexo masculino ou feminino.

144. tamanho do login: O sistema de gestão de resultados é acessado pelos colaboradores a partir de um de um *login*, cujo tamanho é registrado nessa variável.

145. taxa de variação de humor do gestor: A variação de humor considera mudanças repentinas de humor. Os gestores podem apresentar humores positivos consecutivos; entretanto, de repente, podem apresentar uma queda no seu humor. Essa variação de humor, dividida pelo total de humores, compõe a taxa de humor na última semana.

146. histórico da taxa de variação de humor: A variação de humor considera mudanças repentinas de humor. Os colaboradores podem apresentar humores positivos consecutivos; entretanto, de repente, podem apresentar uma queda no seu humor. Essa variação de humor, dividida pelo total de humores, compõe a taxa de humor nas últimas 12 semanas.

147. taxa variação de humor: A variação de humor considera mudanças repentinas de humor. Os colaboradores podem apresentar humores positivos consecutivos; entretanto,de repente,podem apresentar uma queda no seu humor. Essa variação de humor, dividida pelo total de humores, compõe a taxa de humor na última semana.

148. tempo de empresa do gestor: Nessa variável é registrado o tempo de empresa dos gestores desde o início do contrato de trabalho.

149. tempo de empresa: Nessa variável é registrado o tempo de empresa dos colaboradores desde o início do contrato de trabalho.

150. tempo no departamento do gestor: A variável contém o tempo, em dias, dos gestores no departamento de trabalho.

151. tempo no departamento: A variável contém o tempo, em dias, dos colaboradores no departamento de trabalho.

152. tempo de trajeto da casa ao trabalho: Os colaboradores nem sempre possuem moradias próximas ao trabalho e, muitas vezes, precisam se deslocar, de casa para o trabalho, por meio de transporte veicular. Essa variável registra qual é o tempo de trajeto do endereço domiciliar dos colaboradores até o trabalho.

153. total G1: Número de vezes, consecutivas ou não, em que os colaboradores estiveram no grupo mais produtivo da empresa. Considera-se o indicador de bonificação por um período de três meses.

154. total G2: Número de vezes, consecutivas ou não, em que os colaboradores estiveram no segundo grupo mais produtivo da empresa. Considera-se o indicador de bonificação por um período de três meses.

155. total G3: Número de vezes, consecutivas ou não, em que os colaboradores estiveram no grupo mediano em produtividade da empresa. Considera-se o indicador de bonificação por um período de três meses.

156. total G4: Número de vezes, consecutivas ou não, em que os colaboradores estiveram no grupo menos produtivo da empresa. Considera-se o indicador de bonificação por um período de três meses.

157. total de horas trabalhado: O total de horas de trabalho dos colaboradores é registrado nessa variável.

158. trabalhou no último dia: Essa variável registra se os colaboradores faltaram no último dia escalado para trabalho, tendo a data atual como referência.

159. turno do humor: Essa variável contém o turno do último registro de humor dos colaboradores no sistema de gerenciamento de resultados.

160. último humor: Essa variável contém o último registro de humor dos colaboradores no sistema de gerenciamento de resultados.

161. último período G1: O último período semanal em que os colaboradores estiveram no grupo de maior produtividade é armazenado nessa variável. Considera-se o indicador de bonificação por um período de três meses.

162. último período G2: O último período semanal em que os colaboradores estiveram no segundo grupo de menor produtividade é armazenado nessa variável. Considera-se o indicador de bonificação por um período de três meses.

163. último período G3: O último período semanal em que os colaboradores estiveram no grupo mediano de produtividade é armazenado nessa variável. Considera-se o indicador de bonificação por um período de três meses.

164. último período G4: O último período semanal em que os colaboradores estiveram no grupo de menor produtividade é armazenado nessa variável. Considera-se o indicador de bonificação por um período de três meses.

165. última quantidade de G1 consecutivos: A variável contém a última quantidade consecutiva em que os colaboradores estiveram no grupo de maior produtividade da empresa. Considera-se o indicador de bonificação por um período de três meses.

166. última quantidade de G2 consecutivos: A variável contém a última quantidade consecutiva em que os colaboradores estiveram no segundo grupo de maior produtividade da empresa. Considera-se o indicador de bonificação por um período de três meses.

167. última quantidade de G3 consecutivos: A variável contém a última quantidade consecutiva em que os colaboradores estiveram no grupo mediano de produtividade da empresa. Considera-se o indicador de bonificação por um período de três meses.

168. última quantidade de G4 consecutivos: A variável contém a última quantidade consecutiva em que os colaboradores estiveram no grupo de menor produtividade da empresa. Considera-se o indicador de bonificação por um período de três meses.

# Anexo B - Hiperparâmetros

| | | |
|---|---|---|
| LR | Parâmetro de regularização para as classes (peso) | É um hiperparâmetro que penaliza valores muito grandes mantendo o modelo equilibrado e evitando *overfitting*. A faixa de valores desse hiperparâmetro, geralmente, é diversificada, ficando entre 0 e 100 dependendo do problema. As características observadas, neste projeto, forçaram a definição de uma faixa entre 1,0 e 2,0. O *default* para esse hiperparâmetro é 1,0. (scikit learn, 2017) |
| | Valor para critério de parada de execução do modelo | É um hiperparâmetro que representa o limite para interrupção da execução do modelo, atendendo a sua função objetivo. A faixa de valores desse hiperparâmetro, neste projeto, foi definida conforme o valor *default* 1e-04 (scikit learn, 2017) |
| LSTM | Função de ativação | É o valor que define uma função que tornará a rede neural capaz de aprender. Essa função escolhida decide se um neurônio deve ser ativado ou não. A faixa de valores desse hiperparâmetro foi definida conforme a documentação do modelo em (keras, 2020) |
| | Número de iterações do algoritmo | Os ciclos de execução em todo o conjunto de dados de treinamento são definidos por esse hiperparâmetro. Sua faixa de valores depende da quantidade de dados de entrada e, para o presente projeto, foram definidos valores entre 1 e 20. |
| | Representação da dimensão de saída | Esse hiperparâmetro determina qual será o tamanho da informação de saída do modelo. Se esse tamanho não for definido no hiperparâmetro, ele sará inferido diretamente pela entrada. A faixa de valores desse hiperparâmetro varia, neste projeto, entre 20 e 300 |
| MLP | Número de camadas | Esse hiperparâmetro representa o número dos múltiplos níveis de abstração dos dados na rede neural artificial. A faixa de valores desse hiperparâmetro é específica do problema e alguns autores como (Pellicer and Pait, 2019) defendem uma faixa de 2 a 20. Neste projeto, o espaço de busca foi de 2 a 5. |
| | Número de neurônios | Esse é o hiperparâmetro que define o número de unidades ocultas dentro de uma camada. Essas unidade aumentam a precisão dos resultados conforme as técnicas aplicadas. As faixas de valores, neste projeto, foram calculadas conforme o número de variáveis. |
| | Função de ativação | É o valor que define uma função que tornará a rede neural capaz de aprender. Essa função escolhida, decide se um neurônio deve ser ativado ou não. A faixa de valores desse hiperparâmetro foi definido conforme a documentação do modelo em (scikit learn, 2020b) |
| | Valor para inversão da escala da taxa de aprendizado | A taxa de aprendizado é uma constante de proporcionalidade no intervalo [0,1] conforme (Brownlee, 2018) |
| | Parâmetro para suavização (peso) | É um hiperparâmetro que penaliza valores muito grandes e reajusta os valores dos pesos na rede neural artificial. A faixa de valores desse hiperparâmetro esta entre 0.0001 e 0.05, conforme documentação do modelo em (scikit learn, 2020b). O presente projeto usa uma faixa de valores semelhante. |
| | Taxa de aprendizado para atualização de pesos | Esse hiperparâmetro é representado pelo conjunto de taxas de aprendizagem que fazem a atualização dos pesos da rede. Sua faixa de valores é definida na documentação do próprio modelo em (scikit learn, 2020b). |
| NB | Parâmetro para suavização (peso) | É um hiperparâmetro que resolve o problema de haver probabilidade zero em NB. É um hiperparâmetro de suavização de Laplace. Como na documentação, (scikit learn, 2020c), o valor *default* é 1,0 e o valor 0 representa a não suavização, estimou-se valores entre 0 e 1 para a faixa desse hiperparâmetro, neste projeto. |

| | | |
|---|---|---|
| RF | Pesos associados às classes do modelo | A faixa de valores desse hiperparâmetro é definida na documentação do próprio modelo (scikit learn, 2020d). Uma vez que o classificador RF tende a ser tendencioso para a classe majoritária, esse hiperparâmetro representa a penalidade quando a classificação da classe minoritária é incorreta. |
| | Função para medir a qualidade de uma divisão da informação | As medidas desse hiperparâmetro se referem à impureza de um nó da árvore. Um nó com várias classes é impuro, enquanto um nó com apenas uma classe é puro. A faixa de valores desse hiperparâmetro segue a documentação do modelo em (scikit learn, 2020d). |
| | Forma de seleção do número máximo de variáveis | É um hiperparâmetro que determina o número máximo de variáveis para construção das árvores conforme os algoritmos escolhidos na faixa de valores. Essa faixa é definida pela documentação do modelo em (scikit learn, 2020d). |
| | Número de árvores | Esse hiperparâmetro determina o número de árvores que são impulsionadas pelo modelo. Um número grande geralmente resulta em um melhor desempenho. A faixa de valores, neste projeto, foi calculada de acordo com o número de variáveis. A faixa de valores defendida por (Koehrsen, 2018) , entre 200 e 2000, mostra que essa definição é específica para o problema. |
| SVM | Parâmetro de regularização para as classes | Esse hiperparâmetro representa o desejo de encontrar um hiperplano que separa corretamente o maior número de instâncias. Sua faixa de valores com kernel $default$ RBF está entre 0 a 10 conforme (Dawson, 2019). |
| | Valor para critério de parada de execução do modelo | É um limite para interrupção da execução do modelo atendendo a sua função objetivo. A faixa de valores desse hiperparâmetro, neste projeto, foi definida conforme o valor $default$ 1e-03 na documentação do modelo em (scikit learn, 2020e). |
| XGBoost | Profundidade das árvores | Esse hiperparâmetro determina a máxima profundidade de uma árvore. É um hiperparâmetro usado para controlar o modelo permitindo que ele aprenda relações muito específicas. Conforme (Jain, 2016), a faixa de valores varia entre 3 e 10 e assim foram definidos neste projeto. |
| | Parâmetro de regularização mínima para as classes (peso) | É um hiperparâmetro que penaliza valores muito grandes, mantendo o modelo equilibrado e evitando $overfitting$. A faixa de valores desse hiperparâmetro, defendida por alguns autores como (cambridge spark, 2017), é entre 5 e 8. Neste projeto usamos uma faixa equivalente a essa. |
| | Proporção de uma subamostra para treinamento | Esse hiperparâmetro define a fração de dados a serem amostrados aleatoriamente para cada árvore. Conforme (Jain, 2016), a faixa de valores está entre 0,5 e 1. |
| | Amostra de colunas para construção das árvores | É um hiperparâmetro que determina o número máximo de variáveis para construção das árvores aleatoriamente. Os valores tipicamente usados estão entre 0,5 e 1 conforme (Jain, 2016) |
| | Taxa de aprendizado para atualização de pesos | Esse é o hiperparâmetro que define a taxa de aprendizado para tornar o modelo mais robusto, reduzindo os pesos a cada etapa. A faixa de valores típica para esse hiperparâmetro é entre 0,01 e 0,2 conforme (Jain, 2016). |
| | Número de árvores | Esse hiperparâmetro determina o número de árvores que são impulsionadas pelo modelo. Um número grande geralmente resulta em um melhor desempenho. A faixa de valores, neste projeto, foi calculada de acordo com o número de variáveis. A faixa de valores defendida por (Brownlee, 2016), entre 50 e 400, mostra que essa definição é específica para o problema. |

Tabela 1: Descrição dos hiperparâmetros

# Referências Bibliográficas


R Adalash. *Applying Random Forest (Classification) - Machine learning algorithm from scratch with real datasets*. Vooo – Insights Data Science Python Gestão. : acesso https://www.vooo.pro/insights/, 2018. ix, 25

Abdelrahman Ahmed, Sergio Toral, and Khaled Shaalan. *Agent Productivity Measurement in Call Center Using Machine Learning*. International conference on advanced intelligent systems and informatics 160–169, 2016. 37, 39

Beatriz Albiero, Estevo Uyrá, Ramon Vilarino, Juliano Andrade Silva, Tales Fonte Boa Souza, Ricardo dos Santos, Sami Yamouni, and Renato Vicente. *Employing Gradient Boosting and Anomaly Detection for Prediction of Frauds in Energy Consumption*. Anais do XVI Encontro Nacional de Inteligência Artificial e Computacional, SBC, Porto Alegre, RS, Brasil, 2019, pp. 916–925, 2019. 32, 39

Vanessa S. Araujo, Thiago S. Rezende, Augusto J. Guimarães, Vinicius J. Silva Araujo, and Paulo V. de Campos Souza. *A hybrid approach of intelligent systems to help predict absenteeism at work in companies*. SN Appl. Sci. 1, 536, 2019. 34, 35, 36, 37, 39

E. P. Bento and N. Kagan. *Algoritmos genéticos e variantes na solução de problemas de configuração de redes de distribuição*. Sba Controle e Automação vol.19 no.3, 2008. 57

James Bergstra and Yoshua Bengio. *Random Search for Hyper-Parameter Optimization*. Journal of Machine Learning Research 13, 2012. ix, 28, 29

Big-Data.Tips. *Machine Learning Methods*. Big Data Mining e Machine Learning : acesso http://www.big-data.tips/machine-learning-methods, 2018. ix, 11

Gianluca Bontempi, Souhaib Ben Taieb, and Yann-Ael Le Borgne. *Machine Learning Strategies for Time Series Forecasting*. M.-A. Aufaure and E. Zimányi (Eds.): eBISS 2012, LNBIP 138, pp. 62–77, 2013. 8, 11

Giorgos Borboudakis and Ioannis Tsamardinos. *Forward-Backward Selection with Early Dropping*. Journal of Machine Learning Research 20 1-39 Submitted 6/17; Revised 10/18; Published 1/19, 2019. 15

Xavier Bouthillier and Gaël Varoquaux. *Survey of machine-learning experimental methods at NeurIPS2019 and ICLR2020*. HAL Id: hal-02447823 Inria Saclay Ile de France, 2020. 51

G. E. P. Box and G. Jenkins. *Time Series Analysis, Forecasting and Control*. HoldenDay - San Francisco, 1970. 9

G. E. P. Box and D. A. Pierce. *Distribution of the residual autrocorrelations in autoregressive-integrated moving-average time series models*. Journal of the American Statistical Association, 65, 1509 - 1526, 1970. 9

O. Boz. *Feature Subset Selection by Using Sorted Feature Relevance.* International Conference on Machine Learning and Applications - Las Vegas, Nevada, USA, 2002. 18

Andrew P. Bradley. *The use of the area under the ROC curve in the evaluation of machine learning algorithms*. Pattern Recognit., vol. 30, no. 7, pp. 1145–115, 1997. 46, 47

L. Breiman. *Random forests - Random Features*. Technical Report 567, Statistics Department, University of California, Berkeley, 1999. 24

L. Breiman. *Random forests*. Mach. Learn. 40, 5–32, Statistics Department, University of California, Berkeley, 2001. 24

Jason Brownlee. *How to Tune the Number and Size of Decision Trees with XGBoost in Python*. Machine Learning Mastery Pty. Ltd. All Rights Reserved. - acesso: https://machinelearningmastery.com/, 2016. 100

Jason Brownlee. *How to Develop Multilayer Perceptron Models for Time Series Forecasting*. Machine Learning Mastery Pty. Ltd. All Rights Reserved. - acesso: https://machinelearningmastery.com/, 2018. 99

C. J. C. Burges. *A tutorial on support vector machines for pattern recognition*. Data Mining and Knowledge Discovery, vol. 2, no. 2, 1998. 25

Z. BURSAC, C. H. GAUSS, D. K. WILLIAMS, and D. W. HOSMER. *Purposeful selection of variables in logistic regression.* Source code for biology and medicine, v. 3, n. 1, p. 17, 2008. 16

Maitê Marques Caetano. *O Uso de Técnicas de Aprendizado de Máquina na Predição de Desempenho Acadêmico de Alunos em Cursos Superiores*. Dissertação no curso de Mestrado em Ciência da Computação da Faculdade Campo Limpo Paulista, 2016. 5

cambridge spark. *Hyperparameter tuning in XGBoost*. Cambridge Spark - acesso: https://blog.cambridgespark.com/, 2017. 100

A. Cardon and D. N. Müller. *Introdução às redes neurais artificiais*. Universidade Federal do Rio Grande do Sul - Instituto de Informática - Curso de PósGraduação em Ciência da Computação., 1994. 21

Rich Caruana and Alexandru Niculescu-Mizil. *An empirical comparison of supervised learning algorithms using different performance metrics*. Computer Science, Cornell University, Ithaca NY - Conference on Machine Learning, 2006. 19

Chris Chatfield. *The Analysis of Time Series: An introduction*. Chapman and Hall, fifth edition, NY, 1996. 8, 9

N.V. Chawla, K.W. Bowyer, L.O. Hall, and W.P. Kegelmeyer. *SMOTE: Synthetic minority over-sampling technique*. J. Artif. Intell. Res. 16, 321–357, 2007. 74

Tianqi Chen and Carlos Guestrin. *XGBoost: A Scalable Tree Boosting System*. KDD '16: Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data, 2016. 26, 27

Courtney Cochrane. *Time Series Nested Cross-Validation*. Towards Data Science/acessado 24/03/2019: https://towardsdatascience.com/time-series-nested-cross-validation-76adba623eb9, 2018a. 2

Courtney Cochrane. *Time Series Nested Cross-Validation. Towards Data Science*. https://towardsdatascience.com/time-series-nested-cross-validation-76adba623eb9 Last visited January 2018, 2018b. 5

A. Cohen and R. Golan. *Predicting absenteeism and turnover intentions by past absenteeism and work attitudes - An empirical examination of female employees in long term nursing care facilities*. Career Development International, 2007. 3

D.R. Cox and E. J. Snell. *Analysis of Binary Data. 2 ed.* London, Chapman and Hall, 1989. 20

M. Dash and H. Liu. *Feature Selection for Classification*. Intelligent Data Analysis 1 131:156. Department of Information System and Computer Science, National University of Singapore, Singapore, 1997. 15

Carl Dawson. *SVM Parameter Tuning*. Towards Data Science - acesso: https://towardsdatascience.com/, 2019. 100

Ely Francina Tannuri de Oliveira and Maria Cláudia Cabrini Grácio. *Análise a respeito do tamanho de amostras aleatórias simples: uma aplicação na área de Ciência da Informação*. Analysis regarding the size of the simple sample random: an application in the area of Information Science, 2010. 12, 13

J. Demsar. *Algorithms for subsertting attribute values with Relief*. Journal of Machine Learning Research 78:421-428, 2010. 17

Geraldo L. Diniz. *Análise harmônica do regime de precipitação em duas localidades da baixada cuiabana*. Universidade Federal de Mato Grosso - Instituto de ciências exaas tas e da terra - programa de pós graduação em física e meio ambiente, 2008. 9

P. Domingos and M. Pazzani. *On the optimality of the simple bayesian classifier under zero-one loss*. Machine Learning, 29 2/3, 103, 1997. 23

Chiara Di Francescomarinoa, Marlon Dumasb, Marco Federicic, Chiara Ghidinia, Fabrizio Maria Maggib, Williams Rizzia, and Luca Simonettoc. *Genetic Algorithms for Hyperparameter Optimization in Predictive Business Process Monitoring*. Preprint submitted to Information Systems, 2018. 29

Y. Freund and R. E. Schapire. *Experiments with a new boosting algorith*. Machine Learning. Proceedings of the Thirteenth International Conference. pp. 148–156, 1996. 24

Simone Padilha Galarça, Cláudia Simone Madruga Lima, Gustavo da Silveira, and Andreia De Rossi Rufato. *Correlação de pearson e análise de trilha identificando variáveis para caracterizar porta-enxerto de Pyrus communis L.* Ciências agrotécnicas vol.34 no.4 Lavras, 2010. 16

M. R. A. Gamarra and C. G. M. Quintero. *Using genetic algorithm feature selection in neural classification systems for image pattern recognition*. Ingeniería e Investigación vol 33 No. 1, 2013. 33, 34, 39

Felix A. Gers, Jurgen Schmidhuber, and Fred Cummins. *Learning to Forget: Continual Prediction with LSTM*. Technical Report IDSIA-01-99, 1999. 22

J.G De Gooijer and R.J Hyndman. *25 years of time series forecastin*. International Journal of Forecasting 22(3), 443–473, 2006. 9

Klaus Greff, Rupesh K. Srivastava, Jan Koutnik, Bas R. Steunebrink, and Jurgen Schmidhuber. *LSTM: A Search Space Odyssey*. Transactions on Neural Networks and Leaning Systems, 2017. ix, 22

J. Han and M. Kamber. *Data Mining: Concepts and Techniques*. Morgan Kaufmann Publishers, San Francisco, CA, 2001. 13

Sepp Hochreiter and Jiirgen Schmidhuber. *LSTM can solve hard long time lag problems*. Advances in Neural Information Processing Systems 9, NIPS'9, pages 473-479, MIT Press, Cambridge MA, 1997. 23

Michal Horemuz. *Application of Machine Learning to Financial Trading*. Degree Project in Computer Science and Engineering, Second cycle, Stockholm, Sweden, 2018. 26

A. K. Jain, M. N. Murty, and P. J. Flynn. *Data Clustering: A Review*. ACM Computing Surveys, vol. 31, no. 3, pp. 264-323, 1999. 10

Aarshay Jain. *Complete Guide to Parameter Tuning in XGBoost with codes in Python*. Analytics Vidhya - acesso: https://www.analyticsvidhya.com/, 2016. 100

N. Japkowicz. *Japkowicz.* In Imbalanced learning; John Wiley and Sons:Chichester, UK, pp. 187–206, 2013. 74

Bahareh Kalantar, Biswajeet Pradhan, Seyed Amir Naghibi, Alireza Motevalli, and Shattri Mansor. *Assessment of the effects of training data selection on the landslide susceptibility mapping: a comparison between support vector machine (SVM), logistic regression (LR) and artificial neural networks (ANN).* Geomatics, Natural Hazards and Risk VOL. 9, NO. 1, 49–69, 2018. 35, 39

keras. *LSTM layer.* desenvolvedores keras - acesso: https://keras.io/api/, 2020. 99

K. Kira and L. A. Rendell. *A practical approach to feature selection*. Proceedings of the ninth international workshop on Machine learning (pp. 249-256), 1992. 15, 16

Teresa Klatzer and Thomas Pock. *Continuous Hyper-parameter Learning for Support Vector Machines*. 20th Computer Vision Winter Workshop - Paul Wohlhart, Vincent Lepetit (eds.) - Seggau, Austria, February 9-11, 2015. 28, 29

Will Koehrsen. *Hyperparameter Tuning the Random Forest in Python*. Towards Data Science - acesso: https://towardsdatascience.com/, 2018. 100

D. Koller and M. Sahami. *Toward optimal feature selection.* Proceedings of International Conference on Machine Learning, 1996. 15

I. Kononenko. *Estimating Attributes: Analysis and Extensions of RELIEF*. University of Ljubljana, Faculty of Electrical Engineering e Computer Science, 1994. 17

Xuan-Hien Le, Hung Viet Ho, Giha Lee, and Sungho Jung. *Application of Long Short-Term Memory (LSTM) Neural Network for Flood Forecasting*. Department of Disaster Prevention and Environmental Engineering, Kyungpook National University, 2559 Gyeongsang-daero, Sangju-si 37224, Gyeongsangbuk-do, Korea, Faculty of Water Resources Engineering, Thuyloi University, 175 Tay Son, Dong Da, Hanoi, Vietnam, 2019. 22

Siqiao Li, Qingchen Wang, , and Ger Koole. *Predicting Call Center Performance with Machine Learning*. International Conference on Service Science pp.193-199, 2018. 33, 38, 39

J. C. Lohelin. *Latent variables models : an introduction to factor, path and structural analysis*. 3. ed. Mahwah, NJ : Lawrence Erlbaum, 1998. 43

Rafael G. Mantovani, André L. D. Rossi, Edesio Alcobaça, Joaquin Vanschoren, and André C. P. L. F. de Carvalho. *A meta-learning recommender system for hyperparameter tuning: predicting when tuning improves SVM classifiers*. Information Sciences, 2019. 35, 39

M. A. P. Marques. *Análise e comparação de alguns métodos alternativos de seleção de variáveis preditoras no modelo de regressão linear.* Tese de Doutorado. Universidade de São Paulo., 2018. 16

R. P. Ferreira A. Martiniano and C. Affonso R. J. Sassi. *Application of a neuro fuzzy network in prediction of absenteeism at work.* 7th Iberian Conference on Information Systems and Technologies, 2012. 37, 39

MathWorks. *Machine Learning in MATLAB*. Help Center : acesso https://se.mathworks.com/help/stats/machine-learning-in-matlab.html, 2020. ix, 10

A. K. McCallum and K. Nigam. *A Comparison of event models for naive Bayes text classification.* Proceedings of the 1º AAAI Workshop on Learning for Text Categorization, pages 41-48, Madison, USA, 1998. 23

Melanie Mitchell. *An Introduction to Genetic Algorithms*. From Complex Adaptive Systems - A Bradford Book, 1998. 18

Rafiq A. Mohammed and Paul Pang. *Agent Personalized Call Center Traffic Prediction and Call Distribution*. ICONIP'11: Proceedings of the 18th international conference on Neural Information Processing - Volume Part II, 2011. 32, 39

David Moore. *The Basic Practice of Statistics*. New York, Freeman, 2007. 16

Pedro A. Morettin and Clelia M. C. Toloi. *Previsão de séries temporais*. Atual Editora, São Paulo, 1985. 8

P.M. Narendra and K. Fukunaga. *A branch and bound algorithm for feature selection*. IEEE Transactions on Computers, C-26(9):917-922, 1977. 15

Paulo Vieira Neto. *Estatística descritiva: Conceitos básicos*. Centro Universitário Estácio de São Paulo - São Paulo, 2004. 12

Evandro Lopes Oliveira, José M. Torres, and Rui S. Moreira. *Técnicas de Aprendizado de Máquina Aplicadas na Previsão de Desempenho de Operadores de Centros de Teleatendimento*. CISTI - Conferência Ibérica de Sistemas y Tecnologías de Información, 2019a. 6, 33, 34, 35, 37, 39

Evandro Lopes Oliveira, José M. Torres, Rui S. Moreira, and Rafael Alexandre França de Lima. *Absenteeism Prediction in Call Center Using Machine Learning Algorithms*. World Conference on Information Systems and Technologies, 2019b. ix, xi, 6, 19, 33, 34, 35, 37, 39, 70

Evandro Lopes Oliveira, José M. Torres, Rui S. Moreira, and Rafael Alexandre França de Lima. *Técnicas de Aprendizado de Máquina Aplicadas na Previsão de Produtividade*. SBAI - Simpósio Brasileiro de Automação Inteligente, 2019c. xi, 6, 33, 34, 35, 37, 39, 71

Marco Aurélio Cavalcanti Pacheco. *Algoritmos genéticos: princípios e aplicações.* Laboratório de Inteligência Computacional Aplicada - ICA - PUC Rio de Janeiro, 1999. ix, 18

V. A. Padilha and A. C. P. L. F. Carvalho. *Mineração de Dados em Python*. Instituto de Ciências Matemáticas e de Computação - Universidade de São Paulo, 2017. 10

Mrinal Pandey and S.Taruna. *A multi-level classification modelpertaining to the student's academic performance prediction*. International Journal of Advances in Engineering and Technology, 2014. 34, 38, 39

G. L. Pappa, A. A. Freitas, and C. A. A. Kaestner. *A Multiobjective Genetic Algorithm for Attribute Selection.* J. Garibaldi A Lofti and R. John, editors, Proc, 4 - Int. Conf. on Recent Advances in Soft Computing (RASC-2002), pages 116-121. Nottingham Trent University, 2002. 18

Lucas Francisco Amaral Orosco Pellicer and Felipe Miguel Pait. *BarySearch: Algoritmo de tuning de Modelos de Machine Learning com o Método do Baricentro*. 8th Brazilian Conference on Intelligent Systems (BRACIS), 2019. 99

Rohit Punnoose and Pankaj Ajit. *Prediction of Employee Turnover in Organizations using Machine Learning Algorithms. A case for Extreme Gradient Boosting*. (IJARAI) International Journal of Advanced Research in Artificial Intelligence, Vol. 5, No. 9, 2016. 33, 35, 39

R. R. Rajalaxmi and A. M. Natarajan. *An Effective Data Transformation Approach for Privacy Preserving Clustering*. Article in Journal of Computer Science 4 (4): 320-326, 2008. 13

Irina Rish. *An empirical study of the naive Bayes classifier*. IJCAI workshop on empirical methods in artificial intelligence, 2001. 23

M. Robnik and I. Kononenko. *Theoretical and Empirical Analysis of ReliefF and RReliefF*. University of Ljubljana, Faculty of Electrical Engineering e Computer Science, 2003. 17

S. Russel and P. Norvig. *Artificial Intelligence - A Modern Approach*. PrenticeHall, New Jersey, 1995, p.563-597, 1995. 21

Saed Sayad. *Support Vector Machine Classification SVM*. An Introduction to Data Science : acesso http://www.saedsayad.com, 2020. ix, 25

scikit learn. *Logistic Regression*. desenvolvedores scikit-learn (Licença BSD) - acesso: https://scikit-learn.org, 2017. 99

scikit learn. *Máquinas de vetor de suporte*. desenvolvedores scikit-learn (Licença BSD) - acesso: https://scikit-learn.org, 2020a. 26

scikit learn. *MLP Classifier*. desenvolvedores scikit-learn (Licença BSD) - acesso: https://scikit-learn.org, 2020b. 99

scikit learn. *Multinomial NB*. desenvolvedores scikit-learn (Licença BSD) - acesso: https://scikit-learn.org, 2020c. 99

scikit learn. *RandomForest Classifier*. desenvolvedores scikit-learn (Licença BSD) - acesso: https://scikit-learn.org, 2020d. 100

scikit learn. *Support Vector Classification*. desenvolvedores scikit-learn (Licença BSD) - acesso: https://scikit-learn.org, 2020e. 100

Syed Atif Ali Shah, Irfan Uddin, Furqan Aziz, Shafiq Ahmad, Mahmoud Ahmad Al-Khasawneh, and Mohamed Sharaf. *An Enhanced Deep Neural Network for Predicting Workplace Absenteeism*. Hindawi Complexity, 2020. 36, 37, 39

Airton Marinho Silva. *Condições De Trabalho E Adoecimento Dos Trabalhadores Em Teleatendimento: Uma Breve Revisão*. Revista de Gestão Integrada em Saúde do Trabalho e Meio Ambiente - v.1, n.3, Artigo 7, 2007. 1

P. L. B. Soares and J. P. D. Silva. *Aplicação de Redes Neurais Artificiais em Conjunto com o Método Vetorial da Propagação de Feixes na Análise de um Acoplador Direcional Baseado em Fibra Ótica*. Revista Brasileira de Computação Aplicada (ISSN 2176-6649), Passo Fundo, v.3, n.2, p.58-72, 2011. ix, 21, 22

Tavish Srivastava. *A Complete Tutorial on Time Series Modeling in R*. Analytucs Vidhya : acesso https://www.analyticsvidhya.com/blog/2015/12/complete-tutorial-time-series-modeling/, 2015. ix, 9

S. L. Tonsig. *Redes Neurais Artificiais Multicamadas e o Algorítimo Backpropagation*. Universidade de São Carlos - Gerência de Sistemas de Informação Pontifícia Universidade Católic, 2000. 21, 22

J. Truett and et all. *A multivariate analysis of the risk of coronary heart disease in Framinghan*. Jornal of Chronic Diseases., v. 20, n. 7, pp. 511–524, 1967. 20

H. Vafaie and K. De Jong. *Genetic Algorithms as a Tool for Feature Selection in Machine Learning.* Proceedings of International Conference on Tools with Artificial Intelligence, Arlington, 1992. 19

Mauricio A. Valle and Gonzalo A. Ruz. *Turnover prediction in call center: behavioral evidence of loss aversion using randon forest and naive bayes algorithms*. Applied Artificial Intelligence 29:923-942, 2015. 32, 39

Mauricio A. Valle, Samuel Varas, and Gonzalo A. Ruz. *Job performance prediction in a call center using a naive Bayes classifier*. Elsevier - Expert Systems with Applications 39, 2012. 38, 39

V. Vapnik. *The Nature of Statistical Learning*. Springer-Verlag, 1995. 25, 26

S. Varma and R. Simon. *Bias in error estimation when using cross validation for model selection*. BMC Bioinformatics, 2006. 5, 14

Jacques Wainer and Gavin Cawley. *Nested cross-validation when selecting classifiers is overzealous for most practical applications*. arXiv, pp.1–9, 2018. 33, 36, 39

Yan Wang and Xuelei Sherry Ni. *A XGBoost risk model via feature selection and bayesian hiper-parameter optimization*. International Journal of Database Management Systems (IJDMS ) Vol.11, No.1, 2019. 33, 34, 36, 39

Steven R. Young, Derek C. Rose, Thomas P. Karnowski, Seung-Hwan Lim, and Robert M. Patton Oak. *Optimizing Deep Learning Hyper-Parameters Through an Evolutionary Algorithm*. Ridge National Laboratory PO Box, MS-6085 Oak Ridge, TN 37831, 2008. 30

Mauricio Zahn. *Sequência de Fibonacci e o Número de Ouro*. Editora Ciência Moderna, 2020. 56

Joseph S. Zirilli. *Financial Prediction Using Neural Networks*. Itp - Media, 1996. 10

Ada Ávila Assunção and Lailah Vasconcelos de Oliveira Vilela. *As condições de adoecimento em uma empresa de teleatendimento*. Editora Faculdade de Medicina UFMG, 2003. 1