Universidade do Minho
Escola de Engenharia

António Manuel Cardoso de Sousa

Predictive Analytics na Infeção Hospitalar

Dezembro de 2021

Predictive Analytics na Infeção Hospitalar

António Manuel Cardoso de Sousa

UMinho | 2021

Universidade do Minho
Escola de Engenharia

António Manuel Cardoso de Sousa

Predictive Analytics na Infeção Hospitalar

Dissertação de Mestrado

Mestrado Integrado em Engenharia e Gestão de Sistemas de Informação

Trabalho efetuado sob a orientação do(s)

Professor Doutor Manuel Filipe Vieira Torres dos Santos

Doutor Júlio Miguel Marques Duarte

Dezembro de 2021

DIREITOS DE AUTOR E CONDIÇÕES DE UTILIZAÇÃO DO TRABALHO POR TERCEIROS

Este é um trabalho académico que pode ser utilizado por terceiros desde que respeitadas as regras e boas práticas internacionalmente aceites, no que concerne aos direitos de autor e direitos conexos.

Assim, o presente trabalho pode ser utilizado nos termos previstos na licença abaixo indicada.

Caso o utilizador necessite de permissão para poder fazer um uso do trabalho em condições não previstas no licenciamento indicado, deverá contactar o autor, através do RepositóriUM da Universidade do Minho.

AGRADECIMENTOS

O desenvolvimento desta dissertação é uma acumulação de 5 anos de experiência, que contribuiu para o meu crescimento pessoal e profissional, e para isso não posso deixar de agradecer às pessoas e entidades que tiveram presentes.

Em primeiro lugar, quero agradecer ao meu orientador, Professor Manuel Filipe Santos e co-orientador, Júlio Duarte, pelo apoio, aconselhamento e disponibilidade ao longo do desenvolvimento deste projeto.

À doutora Sara Cardoso e ao Hospital da Senhora da Oliveira de Guimarães pela disponibilização dos dados e apoio prestado.

Aos meus grandes amigos de infância e da universidade, às Elites e à Lux Copa D, obrigado por todo o apoio, risos e aventuras. Um muito obrigado por todos os momentos memoráveis que me proporcionaram.

À minha namorada e melhor amiga, Catarina Castro, pelo constante apoio e paciência, não só durante o desenvolvimento desta dissertação, mas também durante estes últimos 5 anos. Obrigado pelo carinho, afeto e estares sempre lá para mim.

Por último, mas não o menos importante, à minha Família. Aos meus pais, José Sousa e Emília Sousa, obrigado pelo apoio incondicional e por me darem todas as oportunidades para seguir os meus sonhos. À minha irmã, Mafalda Sousa, obrigado por todo o apoio, preocupação e incentivo.

DECLARAÇÃO DE INTEGRIDADE

Declaro ter atuado com integridade na elaboração do presente trabalho académico e confirmo que não recorri à prática de plágio nem a qualquer forma de utilização indevida ou falsificação de informações ou resultados em nenhuma das etapas conducente à sua elaboração.

Mais declaro que conheço e que respeitei o Código de Conduta Ética da Universidade do Minho.

RESUMO

As infeções nosocomiais e a resistência antimicrobiana provocam um elevado número de morbidade e mortalidade nos pacientes hospitalizados. A Comissão de Controlo de Infeção (CCI) define medidas para combater a propagação de infeção nosocomial para outros doentes. No entanto, o controlo de infeção é ineficaz, uma vez que a sua deteção é feita de forma manual e por vezes tardia. A utilização de *Predictive Analytics* surge como uma possível solução para este problema, dado que permite a previsão automática e atempada de infeção, melhorando o tempo de resposta, e consequentemente, o controlo de infeção hospitalar.

Nesta dissertação o principal objetivo passou por desenvolver modelos preditivos com boa capacidade de previsão de infeção nosocomial, a partir de técnicas de *Data Mining* (DM) e *Machine Learning* (ML). O desenvolvimento dos modelos de previsão foi realizado em contexto local e offline, e com dados reais provenientes do Hospital da Senhora da Oliveira de Guimarães. Deste modo, foram adotadas as metodologias *Design Science Research Methodology* (DSRM) e *Cross-Industry Standard Process for Data Mining* (CRISP-DM). O DSRM foi aplicado na investigação deste projeto de dissertação e o CRISP-DM foi usado para a aplicação de técnicas de DM.

A abordagem de DM aplicada foi a Classificação e para que os modelos de DM pudessem ser criados, foram selecionadas seis técnicas baseadas em Árvores de Decisão (AD), *Random Forest* (RF), Redes Neuronais (RN), *Naive Bayes* (NB), *Support Vector Machine* (SVM) e Regressão Logística (RL). A avaliação dos modelos foi efetuada a partir da Matriz de Confusão, que permitiu a definição de sete métricas, Acuidade, Sensibilidade, Especificidade, Precisão, *F1-Score*, Índice Kappa e Curva AUC. Destas sete, a Acuidade e Sensibilidade, foram selecionadas como as mais importantes na decisão do melhor modelo.

Os modelos de previsão concebidos apresentam uma grande capacidade de previsão de infeção nosocomial, com valores de Acuidade entre 71.56% a 99.37% e valores de Sensibilidade superiores a 90%. Os resultados obtidos são positivos e podem ajudar os profissionais de saúde na tomada de decisão ao nível da gestão e controlo de infeção nosocomial.

PALAVRAS-CHAVE

Data Mining, Infeção Nosocomial, *Machine Learning* e *Predictive Analytics*.

ABSTRACT

Nosocomial infections and antimicrobial resistance cause a high number of morbidity and mortality in hospitalized patients. The Infection Control Commission (ICC) defines measures to combat the spread of nosocomial infection to other patients. However, the infection control is ineffective, since its detection is done manually and sometimes late. The use of Predictive Analytics is a possible solution to this problem, since it allows the automatic and timely prediction of infection, improving the response time, and consequently, the infection control of the hospital.

The main objective of this dissertation was to develop predictive models with good predicitve ability for nosocomial infection, based on Data Mining (DM) and Machine Learning techniques. The development of the predictive models was performed in a local and offline context, and with real data from Hospital da Senhora da Oliveira in Guimarães. Thus, the Design Science Research Methodology (DSRM) and Cross-Industry Standard Process for Data Mining (CRISP-DM) methodologies were adopted. The DSRM was applied in the research of this dissertation project and the CRISP-DM was used for the application of DM techniques.

The DM approach applied was Classification and so that the DM models could be created, six techniques were selected based on Decision Trees (DT), Random Forest (RF), Neural Networks (NN), Naive Bayes (NB), Support Vector Machine (SVM) and Logistic Regression (LR). The evaluation of the modes was performed from the Confusion Matrix, which allowed the definition of seven metrics, Accuracy, Recall, Specificity, Precision, F1-Score, Kappa Statistic and AUC Curve. Of these seven, Accuracy and Recall were selected as the most importante in deciding the best model.

The designed prediction models show a high predictive capacity for nocomial infection, with Accuracy values between 71.56% and 99.37%, and Recall values above 90%. The results obtained are positive and can help heath professionals in decision-making in nosocomial infection control and management.

KEYWORDS

Data Mining, Machine Learning, Nosocomial Infection and Predictive Analytics.

ÍNDICE

ÍNDICE DE FIGURAS

ÍNDICE DE TABELAS

LISTA DE ABREVIATURAS, SIGLAS E ACRÓNIMOS

AD – Árvore de Decisão

AUC – *Area Under ROC Curve*

BI – *Business Intelligence*

CCI – Comissão de Controlo de Infeção

CRISP-DM – *Cross-Industry Standard Process for Data Mining*

CV – *Cross-Validation*

DM – *Data Mining*

DSRM – *Design Science Research Methodology*

ICD – *International Classifcation of Diseases*

KDD – *Knowledge Discovery in Databases*

ML – *Machine Learning*

NB – *Naive Bayes*

OMS – Organização Mundial de Saúde

RF – *Random Forest*

RL – Regressão Logística

RN – Redes Neuronais

ROC - *Receiver Operating Characterisc*

SNS – Sistema Nacional de Saúde

SVM – *Support Vector Machine*

1. INTRODUÇÃO

Neste primeiro capítulo é realizado uma introdução ao projeto a ser desenvolvido, onde é descrito um enquadramento e motivação do tema, a questão de investigação, os objetivos e resultados esperados, e a estrutura do documento.

1.1 Enquadramento e Motivação

Nos dias de hoje, uns dos principais problemas nas instituições de saúde são as infeções nosocomiais e a resistência antimicrobiana, que causam um elevado número de morbidade e mortalidade nos pacientes internados nos hospitais, bem como um aumento nos custos dos cuidados de saúde associados. O custo médio estimado por doente tratado por infeção nosocomial pode ascender aos 20000 euros.

Existe a necessidade de um controlo de infeção eficaz e eficiente, de modo a garantir a melhoria de qualidade de vida e segurança dos doentes hospitalizados, a diminuição de mortalidade e custos envolvidos no tratamento. O controlo de infeção é assegurado a nível nacional a partir de programas próprios. Ao nível hospitalar, o programa de controlo de infeção é definido e implementado pela Comissão de Controlo de Infeção (CCI), que envolve outras áreas e comissões do hospital, como a Administração do Hospital, os microbiologistas, os farmacêuticos, os profissionais médicos e de enfermagem, entre outros.

As medidas preventivas estabelecidas pela CCI, apesar de estarem bem definidas, a deteção de infeções nosocomiais é realizada de forma manual e por vezes tardia, o que pode comprometer o estado de saúde dos doentes internados e de todos os que os rodeiam por força de contágio. No limite podem surgir surtos de infeções que se tornam muito difíceis de tratar, sendo necessário desencadear planos de contingência muito dispendiosos.

A aplicação de *Predictive Analytics*, com a utilização de técnicas de *Data Mining* e *Machine Learning*, é uma solução para o problema de deteção manual de infeções, uma vez que possibilitará a identificação automática das infeções nosocomiais. O uso destas técnicas permite a melhoria dos serviços de saúde prestados, possibilitando a prevenção de contágio de infeção, a salva de vidas e redução dos custos dos cuidados de saúde do hospital.

Predicitve Analytics consiste na extração de informação a partir de um conjunto de dados, com o objetivo de prever resultados futuros, utilizando técnicas de *Data Mining* e *Machine Learning* (Larose & Larose, 2015).

Este projeto de dissertação insere-se num projeto em desenvolvimento em conjunto no Laboratório de Sistemas de Dados Inteligentes do Centro de Algoritmi e no Hospital da Senhora da Oliveira de Guimarães, e tem como âmbito a utilização de técnicas de *Data Mining* e *Machine Learning* para previsão atempada e automática de infeções nosocomiais, de modo a otimizar a gestão e controlo de infeção do hospital.

A motivação na escolha deste projeto de dissertação, deve-se ao facto do principal gosto e interesse por *Data Mining* e *Machine Learning*. Ambas as áreas são fulcrais e fundamentais para as organizações e os seus projetos, ajudando na tomada de decisão e na evolução das mesmas. Além disso, este projeto está relacionado com a área da saúde, tornando-o ainda mais interessante e importante, uma vez que não só ajudará os profissionais de saúde a realizar o seu trabalho com mais eficiência e eficácia, mas também a salvar vidas humanas.

1.2 Questão de Investigação

Este projeto de dissertação tem como principal finalidade responder à seguinte questão de investigação:

É possível, através do uso de técnicas de *Data Mining* e *Machine Learning*, obter modelos que permitam perceber a probabilidade de um doente contrair uma infeção nosocomial?

1.3 Objetivos e Resultados Esperados

De modo a responder à questão científica, apresentada anteriormente, é necessário definir os objetivos principais:

- Identificar padrões na evolução do estado clínico do doente que estejam ligadas a altas probabilidades de vir a ter uma infeção nosocomial;
- Aferir previsões capazes de contribuir para a prevenção de infeção nosocomial;
- Criar modelos de previsão, utilizando técnicas de *Data Mining* e *Machine Learning*, com base nos dados dos resultados de MCDTS (parâmetros das análises laboratoriais e resultados de alguns exames) e informação clínica do paciente;
- Obter modelos com boa capacidade de previsão.

Os resultados esperados serão modelos de previsão, que indiquem se um doente possui a capacidade de ter infeção nosocomial ou não. Estes modelos de previsão vão melhorar a qualidade e serviço do Hospital, uma vez que vai otimizar o controlo de infeção, permitindo que os profissionais de saúde ajam com maior rapidez no tratamento e isolamento do paciente em questão, e também na prevenção de propagação de infeção para outros doentes.

Além disso, espera-se que os modelos de previsão de identificação futura de infeção nosocomial desenvolvidos, possam contribuir, futuramente, na construção de um protótipo Sistema de Suporte à Decisão Inteligente com base em modelos de inteligência artificial que permite, em tempo real, otimizar a gestão e controlo de infeção de doentes internados.

1.4 Estrutura do Documento

Este documento está dividido em 8 capítulos, enumerados pela seguinte ordem:

1. Introdução – tem como objetivo introduzir o tema do projeto, mostrando o ponto de situação e o que se espera obter;
2. Background – apresenta os conceitos fundamentais deste projeto de dissertação;
3. Estado de Arte – tem como propósito demonstrar o conhecimento atual e práticas existentes sobre a área de estudo deste projeto;
4. Metodologias – fornece as metodologias utilizadas na execução do projeto de dissertação;
5. Trabalho realizado – apresenta todo o trabalho prático realizado e os resultados obtidos neste projeto de dissertação;
6. Discussão de Resultados – demonstra a discussão dos resultados finais;
7. Conclusão – comentários finais do trabalho desenvolvido.

2. BACKGROUND

Neste segundo capítulo são abordados os principais conceitos deste projeto de dissertação, onde é realizado uma descrição detalhada de cada conceito, de modo a proceder à sua compreensão e, caso haja, à sua utilidade.

2.1 Infeções Nosocomiais

Infeções nosocomiais, também conhecidas como infeções associadas/adquiridas em hospitais, são infeções adquiridas por pacientes durante a estadia no hospital (Khan et al., 2015). Infeções que ocorrem dentro de 48h após o internamento, 3 dias após alta ou 30 dias após uma operação, são consideradas infeções nosocomiais (Inweregbu et al., 2005). Por outro lado, não são consideradas infeções nosocomiais, as infeções que se encontravam presentes no paciente, no momento de internamento e infeções que são adquiridas a partir da placenta após 48h do nascimento. (Khan et al., 2015)

As infeções nosocomiais podem ser endémicas ou epidémicas. As infeções endémicas são as mais regulares e são aquelas que ocorrem numa determinada zona/região. As infeções epidémicas são infeções que ocorrem durante surtos/epidemias e são aquelas onde há um aumento anormal de uma determinada infeção ou microrganismo infetante (World Health Organization, 2002).

2.1.1 Evolução

As infeções nosocomiais surgiram muitos antes da origem dos hospitais e instituições de saúde, e tornaram-se num problema grave de saúde após o aparecimento dos antibióticos. Devido a este tipo de infeções, houve um aumento crítico no uso de antibióticos e na hospitalização dos pacientes, bem como nos custos envolvidos (Khan et al., 2015). Com o crescente aumento no uso de antibióticos, verificou-se um aumento na resistência das bactérias (bactérias associadas a infeções nosocomiais) aos antibióticos (Inweregbu et al., 2005).

2.1.2 Tipos de Infeções Nosocomiais

Nas infeções nosocomiais existem vários tipos de infeções, umas mais comuns e umas mais letais. Na seguinte tabela (Tabela 1) é possível visualizar-se os critérios simplificados para vigilância de infeções nosocomiais mais comuns. Esta tabela é útil para as instituições que não possuem acesso ou a capacidade para todas as técnicas de diagnóstico de infeção nosocomial (World Health Organization, 2002).

Tabela 1 – Critérios Simplificados para a vigilância de infeções nosocomiais (adaptado de (World Health Organization, 2002))

Tipo de infeção nosocomial	Critérios Simplificados
Infeção do local cirúrgico	Qualquer exsudado purulento, abcesso ou celulite em expansão no local cirúrgico, durante o primeiro mês após a operação.
Infeção urinária	Urocultura positiva (1 ou 2 espécies) com pelo menos 105 bactérias/ml, com ou sem sintomas clínicos.
Infeção respiratória	Sintomas respiratórios com pelo menos 2 dos seguintes sinais, a surgir durante o internamento: tosse, expetoração purulenta, novo infiltrado na radiografia do tórax consistente com infeção.
Infeção do local do catéter vascular	Inflamação, linfangite ou exsudado purulento, no local de inserção do catéter.
Sepsis	Febre ou calafrios e, pelo menos, 1 hemocultura positiva.

As quatro infeções nosocomiais mais frequentes, segundo (World Health Organization, 2002), são:

- **Infeções urinárias** – são as infeções nosocomiais mais frequentes e são as que apresentam menor morbidade, no entanto, ocasionalmente levam à morte. Este tipo de infeção nosocomial é definida por critérios microbiológicos.
- **Infeções do local cirúrgico** – são infeções localizadas na ferida cirúrgica ou infeções profundas de órgãos. Este tipo de infeção nosocomial é adquirida durante a operação, podendo ser por via endógena (por exemplo a partir da flora da pele) ou exógena (por exemplo a partir de equipamento médico), e através do sangue utilizado na cirurgia.
- **Pneumonia nosocomial** – são infeções normalmente adquiridas por grupos de doentes que estão ligados a ventiladores. Este tipo de infeção nosocomial pode ser adquirida de

forma endógena (a partir do aparelho digestivo ou orofaringe) ou exógena (a partir de equipamento médico respiratório), e apresenta uma letalidade elevada.

- **Bacteriémia nosocomial** – são as infeções com maior percentagem de letalidade e que ocorrem normalmente no local de inserção da pele de dispositivos intravasculares ou no trajeto subcutâneo do catéter. Este tipo de infeção nosocomial pode ser adquirida na flora cutânea residente ou transitória.

Para além disso, existem outros tipos de infeção, apesar de serem menos frequentes, continuam a ser perigosos (World Health Organization, 2002). Esses tipos de infeção podem ser, por exemplo, infeções de pele e tecidos moles, gastrenterite, sinusite, endometrite, entre outras.

2.1.3 Comissão de Controlo da Infeção

A Comissão de Controlo da Infeção (CCI) tem como objetivo proporcionar um fórum para a cooperação e participação multidisciplinar, e para a partilha de informação entre os vários profissionais de saúde e comissões do hospital (World Health Organization, 2002). Esta comissão deve ser criada pelo conselho de administração do hospital, o qual deve dar condições e disponibilizar meios para o seu bom funcionamento (Pina et al., 2010). A comissão deve incluir uma ampla representação de outras áreas do hospital, como o conselho de administração, microbiologistas e outros profissionais de saúde, e deve reportar diretamente ao conselho de administração, de modo a assegurar a eficácia do programa (World Health Organization, 2002). A CCI deve realizar as seguintes funções:

- Revisão e aprovação de um programa anual de atividades para prevenção e vigilância epidemiológica;
- Revisão dos dados da vigilância epidemiológica e identificação das áreas de intervenção;
- Avaliação e promoção da melhoria de práticas de prestação de cuidados de saúde;
- Asseguração da formação apropriada dos profissionais responsáveis pelo controlo de infeção e segurança;
- Revisão dos riscos associados a novas tecnologias e monitorização do risco de infeção de novos dispositivos e produtos, antes da aprovação do seu uso;
- Revisão e fornecimento de dados para investigação de surtos;
- Comunicação e colaboração com outras comissões do hospital com objetivos em comum, como por exemplo, a Comissão de Farmácia e Terapêutica, Comissão de Antibióticos e Comissão de Higiene e Segurança.

2.1.4 Prevenção da Infeção Nosocomial

A prevenção e controlo de infeções nosocomiais é a responsabilidade de todos os profissionais de saúde, sejam eles enfermeiros, médicos, farmacêuticos, entre outos (World Health Organization, 2002). De modo a ter um controlo e prevenção de infeção nosocomial eficaz é necessário a criação de um programa que integre os seguintes parâmetros:

- Limitação da transmissão de microrganismos entre pacientes durante os cuidados que lhes são administrados, como a lavagem das mãos, utilização de luvas, estratégias de isolamento, de práticas de desinfeção e esterilização, e tratamento de roupas;
- Controlo dos riscos ambientais de infeção;
- Proteção dos doentes pela utilização de profilaxia antibiótica, vacinação e nutrição;
- Limitação do risco de infeção endógena, através da minimização dos procedimentos invasivos e promovendo a correta utilização dos antibióticos;
- Realização de vigilância epidemiológica das infeções, através da identificação e controlo dos surtos;
- Prevenção de infeções nos profissionais de saúde;
- Intensificação na execução e melhoria (formação contínua dos profissionais de saúde) de boas práticas de cuidados aos pacientes.

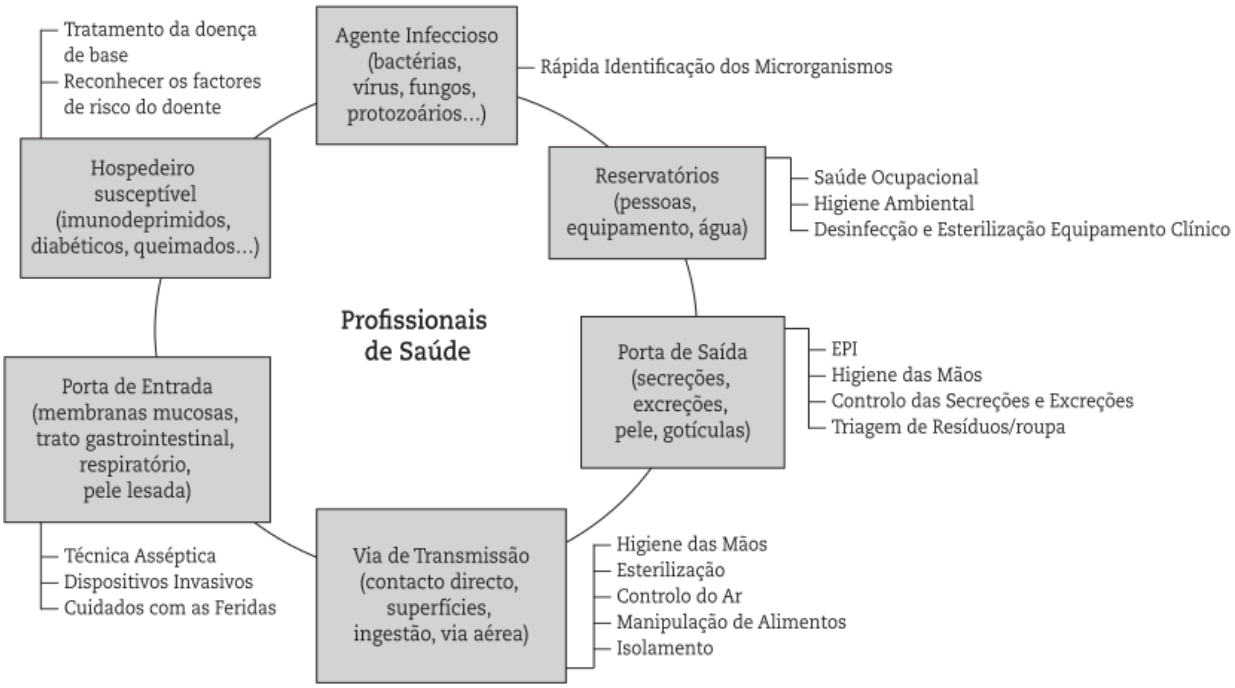

Figura 1 – Cadeia de transmissão de infeção, precauções básicas e isolamento (retirado de (Pina et al., 2010))

Na seguinte tabela (Tabela 2) é possível observar informação relativa ao risco diferencial de infeção nosocomial por doente e por intervenção.

Tabela 2 - Risco diferencial de infeção nosocomial por doente e por intervenção (adaptado de (World Healh Organization, 2002))

Risco de infeção	Tipo de doente	Tipo de procedimento
1 Mínimo	Sem imunodeficiência; sem doença subjacente significativa.	Não invasivo. Sem exposição a fluidos biológicos.
2 Médio	Doentes infetados ou com alguns fatores de risco (idade, neoplasia).	Exposição a fluidos biológicos ou procedimento invasivo não cirúrgico (p.ex., cateterização de veia periférica, introdução de algália).
3 Elevado	Doentes com imunodeficiência grave, (<500 leucócitos/ml); múltiplos traumatismos, queimaduras graves, transplante de órgãos.	Cirurgia ou procedimentos invasivos de alto risco (p.ex., cateterização venosa central, entubação endotraqueal).

2.1.5 Uso de Antimicrobianos e Resistência Antimicrobiana

Até aos dias de hoje, com o aumento no uso de antibióticos, muitos microrganismos ganharam resistência a vários agentes antimicrobianos e em alguns casos, a quase todos. A resistência das bactérias levam a um aumento de morbidade e mortalidade, tendo mais impacto nos pacientes com doenças graves ou imunodeprimidos. Esta resistência é um problema grave para as instituições de saúde, uma vez que é nos hospitais que a transmissão é elevada, devido ao facto de os indivíduos estarem mais enfraquecidos e suscetíveis (World Health Organization, 2002).

A resistência antimicrobiana e a sua disseminação entre as bactérias é devido à pressão seletiva dos antibióticos. Estas bactérias resistentes são transmitidas entre os doentes dos hospitais, e os fatores de resistência são transferidos entre as bactérias. Com o uso constante de antimicrobianos há um aumento da pressão seletiva, o que favorece a multiplicação e disseminação de estirpes resistentes. Este aumento de pressão faz com que haja um uso inapropriado e não controlado dos agentes antimicrobianos, como por exemplo, a prescrição excessiva, administração de doses sub-terapêuticas, duração insuficiente de tratamento e erros de diagnóstico, levando à escolha errada de fármacos. Além

disso, o não cumprimento das precauções básicas como a lavagem de mãos, descontaminação, entre outros, facilita a disseminação das estirpes resistentes (World Health Organization, 2002).

Segundo (World Health Organization, 2002), de modo a assegurar que exista uma prescrição eficaz e económica de fármacos, com o objetivo de minimizar as estirpes resistentes, cada instituição de saúde deve possuir um regulamento de utilização de antimicrobianos:

- A prescrição de antimicrobianos deve ser justificada com base num diagnóstico clínico e microrganismos infetantes conhecidos ou suspeitos;
- Uma colheita de produtos adequados para o estudo microbiológico deve preceder ao início do tratamento com antibiótico, a fim de confirmar a sua adequação;
- A seleção do antimicrobiano deve ser realizada com base na natureza da doença, dos agentes patogénicos, e no padrão de sensibilidade, tolerância do doente e nos custos;
- O médico envolvido deve possuir informação atempada e relevante sobre a prevalência de resistências na instituição;
- Deve ser realizado a seleção de um agente com um espectro o mais estreito possível;
- As combinações de antibióticos devem ser evitadas sempre que possível;
- Podem ser estabelecidas restrições ao uso de antibióticos selecionados;
- A dose utilizada tem de ser correta.

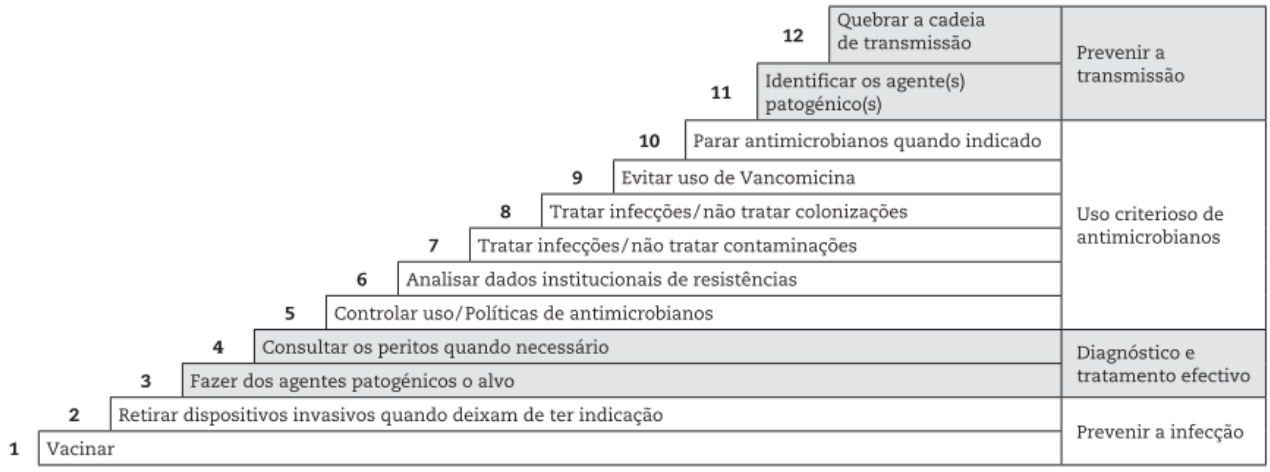

Figura 2 – 12 passos para evitar as resistências aos antimicrobianos (retirado de (Pina et al., 2010))

2.1.6 Fatores de Risco

As infeções nosocomiais são abordadas consoante o risco, de acordo com a frequência, gravidade, custos e mortalidade (Pina et al., 2010). Relativamente à sua avaliação consoante o risco, segundo (Pina et al., 2010), existem 3 tipos de fatores de risco:

- **Presença de dispositivos invasivos** – o uso de dispositivos invasivos, sejam eles para monitorização ou intervenção, nem sempre é inócuo e quem os utiliza necessita de ter um conhecimento aprofundado. Estes dispositivos, apesar de serem inofensivos, favorecem as infeções nosocomiais, uma vez que abre portas de entradas artificiais para microrganismos, como por exemplo, infeções na corrente sanguínea, nos pulmões e nas vias urinárias;
- **Procedimentos invasivos (cirurgias)** – os procedimentos invasivos podem causar infeções e são infeções que ocorrem no local da intervenção cirúrgica, e dependem do número de microrganismos e da suscetibilidade do hospedeiro. Estas contaminações podem ocorrer por forma exógena, através da equipa cirúrgica ou material utilizado, ou endógena, através da flora da pele do paciente;
- **Microrganismos multirresistentes (MMR)** – os MMR devem-se ao facto da evolução da resistência bacteriana aos antibióticos. Estes microrganismos apresentam tal resistência devido à pressão seletiva na prescrição de antibióticos e pela sua disseminação entre as bactérias, e consequentemente, pela alta transmissão entre pacientes. Os MMR podem causar infeções na corrente sanguínea, respiratórias, urinárias ou no local cirúrgico.

2.1.7 Custos das Infeções Nosocomiais

As infeções nosocomiais, para além de agravar a incapacidade funcional e o stress emocional do doente, e em alguns casos, levar a situações que diminuem a qualidade de vida ou até mesmo à morte, apresentam elevados custos económicos não só para os pacientes, mas também para as instituições de saúde. Estas infeções favorecem o desequilíbrio entre os recursos dos cuidados primários e secundários de saúde, dado que existe a necessidade de desvio de fundos para a gestão de problemas (World Health Organization, 2002).

O aumento do tempo de hospitalização dos pacientes é o fator que mais contribui para o aumento dos custos económicos, uma vez que não só vai aumentar diretamente os custos dos doentes, como

também os custos indiretos devidos à perda de produtividade. Além disso, o aumento na utilização de antibióticos, a necessidade de isolar o paciente, a utilização de recursos para estudos laboratoriais e diagnósticos de tipo de infeção, contribuem para o aumento dos custos (World Health Organization, 2002).

2.2 Business Intelligence

Business Intelligence (BI) tem como principal objetivo fornecer informação de interesse, a partir dos dados armazenados, a gestores encarregues pela tomada de decisão, de modo a melhorar ou modificar o processo de negócio de uma determinada organização, alcançando assim objetivos definidos e tornando a organização mais competitiva no mercado em que se encontra inserida.

(Negash, 2004), define BI como um processo de recolha e armazenamento de dados, e de gestão de conhecimento, com ferramentas analíticas para apresentar informações complexas e competitivas para os gestores responsáveis pela tomada de decisão.

De um modo geral, BI é visto como uma ferramenta fundamental para melhorar a qualidade e quantidade de informação disponível para a tomada de decisão (M. Santos & Ramos, 2009).

Segundo (Negash, 2004), as tarefas normalmente associadas a BI são:

- A criação de previsões baseadas em dados históricos, desempenho passado e atual da organização em causa, e estimativas para a direção que o futuro tomará;
- A criação de cenários que mostrem o impacto da alteração de diversas variáveis;
- Permissão de acesso ad-hoc (ligação temporária entre vários computadores e dispositivos) aos dados para responder a questões que não estão predefinidas;
- Visualização de toda a informação recolhida.

Os sistemas de BI têm adquirido funcionalidades de escalabilidade e segurança nos sistemas de gestão de bases de dados para construir *Data Warehouses* (DW), de forma a que no futuro sejam aplicadas técnicas de *Data Mining* e *On-line Analytical Processing* (OLAP).

Em suma, *Business Intelligence* é um recurso essencial para o desenvolvimento e evolução do negócio de uma determinada organização. BI apresenta tal capacidade, uma vez que extrai informação de outros sistemas, como é possível visualizar na figura abaixo (Figura 3).

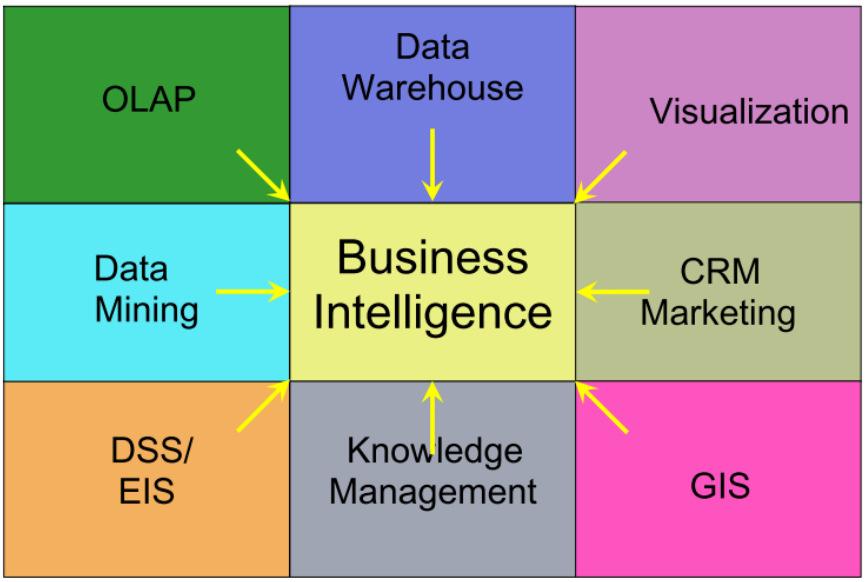

Figura 3 – Sistemas que afetam Business Intelligence (retirado de (Negash, 2004))

2.3 Knowledge Discovery in Databases

Segundo (Fayyad et al., 1996), *Knowledge Discovery in Databases* (KDD) é a extração de conhecimento útil a partir de dados. KDD é a extração não trivial de informação implícita, anteriormente desconhecida e potencialmente útil dos dados (Frawley et al., 1992).

O processo KDD consiste em utilizar métodos de *Data Mining* para extrair conhecimento útil, de acordo com as medidas e restrições especificadas, a partir de uma base de dados que necessite de qualquer pré-processamento e transformação (Azevedo & Santos, 2008).

KDD é um processo iterativo e interativo, uma vez que pode haver retrocesso para fases anteriores e necessita da participação do utilizador sempre que for necessário a tomada de decisão (Azevedo & Santos, 2008).

2.3.1 Fases da KDD

Knowledge Discovery in Datasabes possui 5 fases, segundo (Fayyad et al., 1996), sendo elas:

1. ***Selection*** – Esta primeira fase tem como objetivo a seleção dos dados, ou seja, criar um *target dataset* (conjunto de dados alvo), ou um *subset* (subconjunto) de variáveis ou uma amostra de dados, onde a extração de conhecimento deve ser realizada. Esta fase é fundamental para obter uma boa solução final, uma vez que a base do processo de construção dos modelos é definida nesta etapa;

2. ***Preprocessing*** – A segunda fase tem como propósito a limpeza e pré-processamento dos dados selecionados, ou seja, os dados serão sujeitos a processos de limpeza, filtragem, alteração e remoção, de modo a obter dados mais consistentes;
3. ***Transformation*** – Nesta terceira fase é realizado a transformação dos dados, com a utilização de métodos de redução ou transformação da dimensão do *dataset*, de modo a obter os melhores modelos de dados;
4. ***Data Mining*** – Esta quarta fase consiste na procura de padrões de interesse numa determinada forma representacional, dependendo do objetivo de *Data Mining*. Nesta etapa são selecionados métodos específicos, segundo os objetivos definidos anteriormente, com o intuito de alcançar os melhores resultados possíveis;
5. ***Interpretation/Evaluation*** – Esta quinta e última fase tem como objetivo intrepretar e avaliar os padrões de interesse obtidos, tendo em conta as metas definidas nas primeiras fases. Se os padrões de interesse forem ótimos o processo de KDD termina. Por outro lado, caso seja necessário um reajuste dos padrões, o processo é repetido, de modo a tentar obter os melhores resultados.

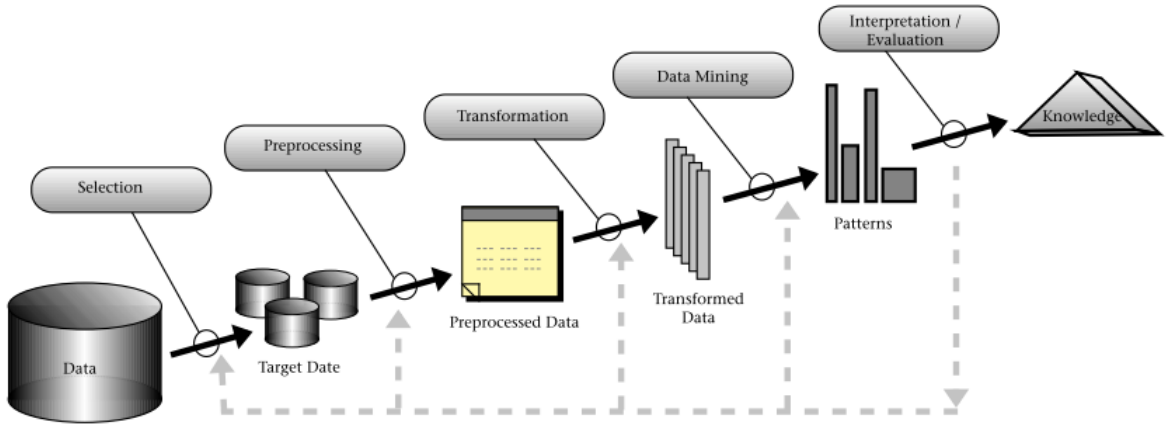

Figura 4 – Fases da KDD (retirado de (Fayyad et al., 1996))

2.4 Machine Learning

Para definirmos *Machine Learning* (ML), primeiro temos de saber o que significa *"learning"* neste contexto de tecnologia e computadores. *Learning* é o processo de construção de um modelo científico após descobrir o conhecimento a partir de um *dataset* ou *datasets*. Desse modo, podemos definir, em geral, que ML é um processo de aplicação de recursos de computação para implementar algoritmos de aprendizagem. Por outro lado, podemos definir, cientificamente, que ML é um processo de computação muito complexo de reconhecimento automático de padrões e de tomadas de decisões inteligentes baseadas em amostras de dados treinadas (Dua & Du, 2016).

ML investiga, com base nos dados, como é que os computadores podem aprender ou melhorar o seu desempenho (Han et al., 2012).

Segundo (Dua & Du, 2016), os métodos de ML podem ser divididos em 4 grupos de atividades de aprendizagem, sendo elas:

- ***Symbol-based*** – Esta atividade de ML afirma que todo o conhecimento pode ser representado por símbolos e que ML tem a capacidade de criar novos símbolos e conhecimentos, a partir dos símbolos conhecidos. Nesta atividade as decisões são entendidas através de procedimentos de inferência de lógica;
- ***Connectionist-based*** – Esta atividade de ML é construída com base na imitação de sistemas de conexão de redes neuronais do cérebro. Nesta atividade as decisões são realizadas depois do reconhecimento dos padrões e do treino dos sistemas;
- ***Behavior-based*** – Esta atividade de ML afirma que existem soluções para a identificação do comportamento e que é projetada para encontrar a melhor solução para a resolução do problema;
- ***Immune system-based*** – Esta atividade de ML aprende através dos seus encontros com objetos que não se encontram no seu ambiente e desenvolve a capacidade de encontrar padrões nos dados.

Dos quatro métodos de ML definidos anteriormente, nenhum deles possui vantagens sobre os outros, desse modo, não existe a necessidade de os selecionar com base nas distinções que possuem (Dua & Du, 2016).

A avaliação de ML deve ser empírica, dado que o seu desempenho depende fundamentalmente do tipo de treino sofrido, das métricas de avaliação de desempenho e da definição do problema. Os métodos de ML são avaliados tendo em atenção o tempo e a viabilidade do método de ML. Além disso,

são avaliados com base na comparação de resultados de aprendizagem de métodos aplicados sobre o mesmo *dataset* ou na quantificação de resultados de aprendizagem dos mesmos métodos aplicados sobre amostras do *dataset* (Dua & Du, 2016).

Os métodos de ML utilizam padrões de treino para aprender a forma de um modelo de classificação, sendo que os modelos podem ser paramétricos ou não. A vantagem de usar os algoritmos de ML é a redução do erro de classificação na amostra de dados treinada (Dua & Du, 2016).

2.4.1 Aprendizagem Supervisionada

Aprendizagem supervisionada é praticamente um sinónimo de classificação, em que a supervisão da aprendizagem vem dos exemplos rotulados no *dataset* de treino (Han et al., 2012).

Na aprendizagem supervisionada, são fornecidos pares de *inputs* e um *target* de *output* para um modelo de aprendizagem ser treinado, de modo a que o *target* da função possa ser previsto a um custo mínimo (Dua & Du, 2016).

Os métodos de aprendizagem supervisionada são categorizados com base nas estruturas e funções dos algoritmos de aprendizagem, sendo que podem incluir categorizações como Redes Neuronais, *Support Vector Machine* e Árvores de Decisão (Dua & Du, 2016).

2.4.2 Aprendizagem Não Supervisionada

Segundo (Han et al., 2012), aprendizagem não supervisionada é basicamente um sinónimo de *Clustering* (agrupamento).

Na aprendizagem não supervisionada, não é fornecido nenhum *target* ou rótulo à amostra de dados. Este método é utilizado para resumir as principais características dos dados e para formar *clusters* naturais de padrões de *inputs* para uma função de custo específico (Dua & Du, 2016).

Os principais métodos de aprendizagem não supervisionada são *Clustering K-Means*, *Hierarchial Clustering* e *Self-Organization-Map* (Dua & Du, 2016).

2.4.3 Aprendizagem Semi-supervisionada

Aprendizagem semi-supervisionada é utilizada para resolver problemas semelhantes aos da aprendizagem supervisionada e tem como objetivo otimizar a forma como se prevê classes de dados futuros comparativamente aos modelos em que apenas se usam dados rotulados (Mohammed et al., 2016).

A aprendizagem semi-supervisionada é uma técnica que utiliza dados rotulados e não rotulados, isto é, para alguns dados são fornecidos as respostas e para outros não. Os dados rotulados são utilizados para aprender modelos de classe e os dados não rotulados são utilizados para aperfeiçoar os limites entre as classes (Han et al., 2012). Esta combinação permite gerar um modelo apropriado de classificação de dados (Mohammed et al., 2016).

2.5 Data Mining

Para (Han et al., 2012), *Data Mining* (DM) é um termo complexo, podendo ser definido de várias formas diferentes, e até diretamente confundido com *Knowledge Discovery in Databases* (KDD). Afirma também que DM é apenas uma etapa de todo o processo de KDD, no entanto, decidiu definir DM como um processo de descoberta de padrões e conhecimento de interesse, a partir de grandes quantidades de dados, que podem estar armazenados em bases de dados, *Data Warehouses* (DW), na *Web* ou em outros tipos de repositórios de dados.

DM envolve algoritmos de inferência para realizar exploração nos dados, desenvolvendo modelos matemáticos e descobrindo padrões importantes, sendo eles implícitos ou explícitos, que anteriormente eram desconhecidos (Maimon & Rokach, 2011) . Esses modelos são utilizados para compreender fenómenos existentes nos dados, análise e previsão.

De um modo geral, DM é um conjunto de algoritmos que têm como objetivo encontrar conhecimento novo, útil e interessante nas bases de dados (Zaitseva et al., 2015). Os algoritmos são baseados em campos aplicados da matemática e informática, como por exemplo, a estatística, probabilidades e redes neuronais, e são utilizados para encontrar relações entre os dados, de modo a criar modelos que possam prever algum tipo de comportamento ou propriedades em comum. (Zaitseva et al., 2015).

2.5.1 Taxonomia dos Métodos de Data Mining

Data Mining possui vários métodos, segundo (Maimon & Rokach, 2011), os quais são utilizados para diferentes finalidades e objetivos. A taxonomia é fundamental para ajudar na compreensão da variedade de métodos, nas suas inter-relações e agrupamentos. Para (Maimon & Rokach, 2011), existem dois principais tipos de DM, nomeadamente, orientado à Verificação, onde o sistema verifica a hipótese do utilizador, e orientado à Descoberta, onde o sistema descobre novos padrões e regras de forma autônoma. A figura a seguir representa a taxonomia de DM (Figura 5).

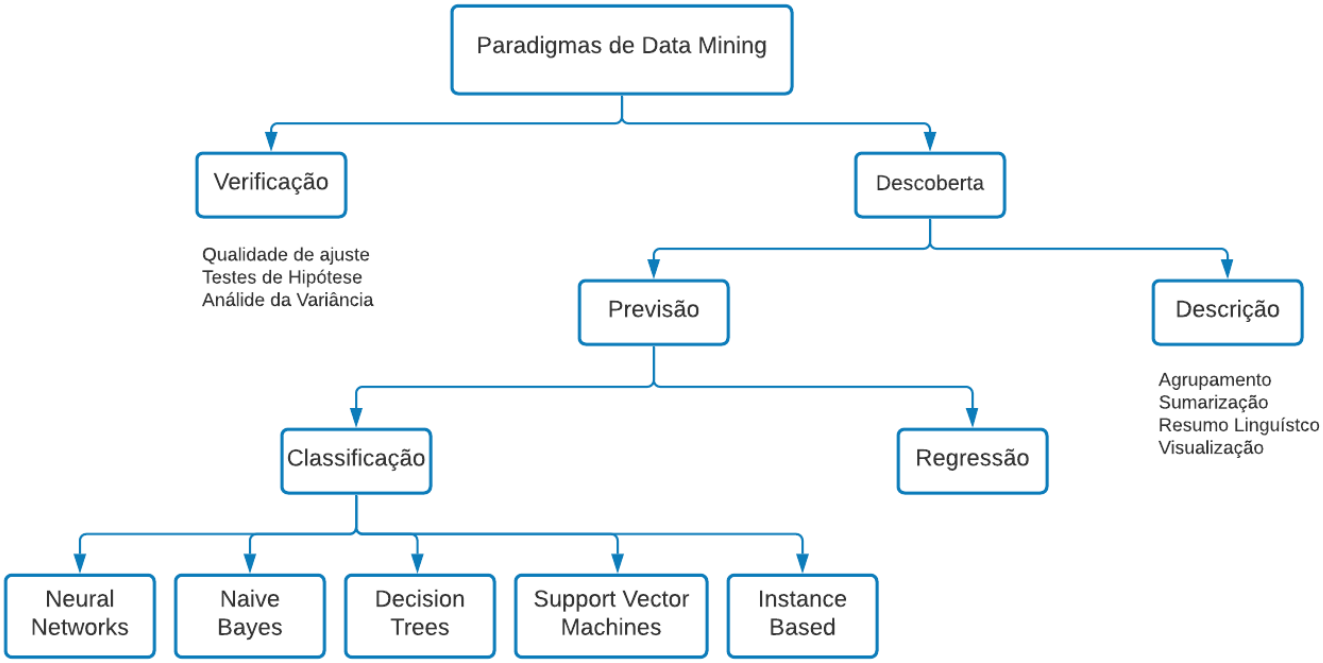

Figura 5 – Taxonomia de Data Mining (adaptado de (Maimon & Rokach, 2011))

Os métodos de Verificação lidam com a avaliação de uma hipótese proposta por uma fonte externa, como por exemplo, um especialista ou analista. Estes métodos envolvem os métodos mais comuns da estatística tradicional, como Qualidade de Ajuste, Testes de Hipótese e Análise da Variância. Além disso, os métodos de Verificação são menos associados ao DM do que os métodos de Descoberta, uma vez que a maior parte dos problemas de DM passa pela descoberta de uma nova hipótese, em vez de testar uma hipótese já conhecida, e os métodos tradicionais de estatística focam-se maioritariamente na estimativa do modelo, enquanto que um dos objetivos principais de DM é a identificação e construção de modelos baseado em evidências (Maimon & Rokach, 2011).

Os métodos de Descoberta descobrem e identificam padrões nos dados automaticamente. O método de Descoberta consiste em dois métodos, sendo um orientado à Descrição e um orientado à Previsão. Os métodos orientados à Descrição têm como objetivo a interpretação dos dados, ou seja, foca-se na compreensão de como é que os dados subjacentes se relacionam com as suas partes, a partir de Visualização, Agrupamento, Sumarização ou Resumo Linguístico. Os métodos orientados à Previsão têm como propósito a construção automática de modelos comportamentais (modelos de Classificação ou Regressão), obtendo amostras novas e anteriormente desconhecidas, e prevendo valores de uma ou mais variáveis associadas à amostra de dados. Além disso, métodos de Previsão podem auxiliar na compreensão de dados (Maimon & Rokach, 2011).

A maior parte das técnicas de DM que utilizam métodos orientados à Descoberta baseiam-se em aprendizagem indutiva, onde é construído um modelo, a partir da generalização de um número suficiente de exemplos de treino. A abordagem da aprendizagem indutiva, possui um pressuposto, que consiste na aplicação do modelo de treino em exemplos futuros nunca vistos (Maimon & Rokach, 2011).

Os modelos de Classificação consistem na identificação de uma função que associe um caso a uma classe dentro de diversas classes discretas de classificação, isto é, tem como objetivo procurar uma função que mapeie (classifique) um item dos dados a várias classes predefinidas (Fayyad et al., 1996). Os modelos de Regressão têm como propósito procurar uma função para a previsão de uma determinada variável (Fayyad et al., 1996).

2.5.2 Modelos Preditivos

Árvore de Decisão (AD) – Árvore de Decisão tem origem na àrea de *Machine Learning*, uma vez que foi desenvolvida por um investigador em *Machine Learning*, o J. Ross Quinlan. Uma AD é um fluxograma com uma estrutura semelhante a uma árvore, onde o nó superior é a raiz da árvore, cada nó interno especifica a utilização de um teste para cada atributo da variável, cada ramo corresponde a um valor possível desse atributo e cada folha contém um rótulo da classe, isto é, a classificação ou decisão (Han et al., 2012). Este tipo de modelo preditivo tem um conjunto de regras que segue uma hierarquia de classes ou valores e pode ser utilizado em problemas de classificação e regressão.

***Random Forest* (RF) –** *Random Forest* consiste em muitas árvores de decisão e é o algoritmo mais popular de *"bagging ensemble"* (Dua & Du, 2016). RD baseia-se no conceito de *"ensemble learning"*, que é uma técnica de combinação de vários classificadores para resolver problemas e melhorar o desempenho do modelo (Han et al., 2012). O resultado final de um modelo de RD é decidido pelos votos dados por todas as árvores de decisão individuais. Cada árvore de decisão é construída a partir da classificação de amostras de *bootstrap* dos dados de entrada, utilizando um algoritmo de árvore, isto é, utilizando o método *"bagging"*. De seguida, cada árvore de decisão será usada para classificar os dados teste e cada árvore tem a decisão de classificar quaisquer dados de teste. Esta classificação é chamada de votação. Por fim, o resultado da classificação dos dados de testes é decidido depois de recolher o maior número de votos entre as árvores (Dua & Du, 2016). Este tipo de modelo preditivo pode ser utlizado em problemas de classificação e regressão.

Redes Neuronais (RN) – Redes Neuronais são modelos que transformam os *inputs* em *outputs* que correspondem aos *target*, através do processamento de informações não lineares dentro de um grupo ligado a neurónios artificiais (Dua & Du, 2016). Do ponto de vista biológico, as RN são modelos simples que imitam o sistema nervoso central do ser humano. Por outro lado, do ponto de vista da computação, as RN são métodos que representam funções, utilizando redes de elementos de computação simples, e aprendem essa representação através de exemplos predefinidos. Este tipo de modelo preditivo pode ser utilizado em problemas de classificação, segmentação e associação.

***Naive Bayes* (NB) –** *Naive Bayes* é baseado no Teorema de Bayes de Thomas Bayes, um inglês que trabalhou inicialmente na teoria da probabilidade e da decisão durante o século XVIII (Han et al., 2012). As NB utilizam a teoria da probabilidade para a construção de um modelo que trabalha sobre variáveis incertas (Dua & Du, 2016).

Teorema de Bayes:

$$P(A|B) = \frac{P(B|A)P(A)}{P(B)} \quad (1)$$

- P(A) é a probabilidade de a hipótese (A) ser verdadeira;
- P(A | B) é a probabilidade de que a hipótese (A) é verdadeira segundo os dados (B);
- P(B | A) é a probabilidade dos dados (B), dado que a hipótese (A) é verdadeira;

Desse modo, podemos concluir que as NB realizam o cálculo da probabilidade de um evento ocorrer, condicionando pela ocorrência de outro evento. Este tipo de modelo preditivo pode ser utilizado em problemas de classificação, segmentação, sumarização e associação.

***Support Vector Machine* (SVM) –** *Support Vector Machine* foi desenvolvido por Vladimir Vapnik e os seus colegas Bernhard Boser e Isabelle Guyon (Han et al., 2012). SVM é um conjunto de métodos relacionados para aprendizagem supervisionada, podendo ser aplicada a problemas de classificação e regressão (Maimon & Rokach, 2011). O algoritmo de SVM utiliza um mapeamento não linear para transformar os dados originais numa dimensão superior, isto é, num grande espaço característico multidimensional, onde irá depois procurar o melhor hiperplano de separação linear, ou seja, é onde irá selecionar os vetores de suporte à previsão (Han et al., 2012). Esta transformação dos dados depende da função *kernel* k(x,y), que ao ser utilizada, promove a dimensionalidade dos dados.

Regressão Logística (RL) – Regressão Logística são modelos que usam funções logísticas para modelar uma variável binária (variável dependente) (O'Connell, 2011). No ramo da matemática, os modelos logísticos binários modelam a probabilidade de um certo evento ocorrer (variável dependente), e a partir dessas probabilidades calculam uma estimativa dos efeitos nas variáveis independentes (O'Connell, 2011). A probabilidade de um certo evento ocorrer é um quociente que compara a probabilidade do evento ocorrer com a probabilidade de ele não ocorrer, como se pode verificar na seguinte equação:

$$\frac{P(Y)}{1-P(Y)} \quad (2)$$

Este tipo de algoritmo preditivo pode ser usado em problemas de classificação.

2.5.3 Avaliação de Modelos

Após a criação de modelos de previsão é necessário de seguida proceder à sua avaliação para verificar até que ponto estão aptos para realizarem previsões corretas. Esta tarefa é fundamental, uma vez que é necessário perceber qual o modelo mais adequado para garantir os objetivos de negócio (Pete et al., 2000). Os modelos criados serão avaliados consoante um conjunto de métricas, sendo que algumas dessas métricas só se aplicam a algoritmos de classificação e outras a algoritmos de regressão.

2.5.3.1 Métricas Associadas à Classificação

Matriz Confusão – a matriz de confusão é a técnica mais utilizada para a avaliação de desempenho de modelos de classificação (Novakovic et al., 2017). Este método permite a definição de várias métricas, como taxas de erro e as curvas *Receiver Operating Characterisc* (ROC). A matriz de confusão possui uma dimensão de 2x2, em que as linhas representam os valores reais e as colunas representam os valores previstos de uma classe, como é possível verificar na seguinte tabela (Tabela 3).

Tabela 3 – Matriz de Confusão (adaptado de (Witten et al., 2011))

		Classe Prevista	
		Positivo (P)	**Negativo (N)**
Classe Atual	**Positivo (P)**	TP	FN
	Negativo (N)	FP	TN

- **TP –** True Positive (Verdadeiros Positivos) são o número de previsões corretas com saída positiva;
- **FN –** False Negative (Falsos Negativos) são o número de previsões incorretas com saída negativa;
- **FP –** False Positive (Falsos Positivos) são o número de previsões incorretas com saída positiva;
- **TN –** True Negative (Verdadeiros Falsos) são o número de previsões corretas com saída negativa.

A partir matriz de confusão é possível calcular outras medidas de desempenho, sendo elas:

Acuidade (*Accuracy*) – esta métrica é a percentagem de acerto do modelo, que corresponde à quantidade de registos corretamente classificados, tanto positivos como negativos, divididos pelo número total de registos (Novakovic et al., 2017).

$$Acuidade = \frac{TP + TN}{n} * 100(\%) \ (3)$$

Sensibilidade (*Recall* ou Taxa de Verdadeiros Positivos) – esta medida é a percentagem de verdadeiros positivos corretamente classificados como positivos. É calculada tendo em conta a quantidade de registos que o modelo identificou corretamente como positivos, divididos pelo número total de registos positivos (Novakovic et al., 2017).

$$Sensibilidade = \frac{TP}{TP + FN} * 100(\%) \ (4)$$

Especificidade (Taxa de Verdadeiros Negativos) – corresponde à percentagem de verdadeiros negativos corretamente classificados como negativos. Esta métrica é calculada tendo em conta a quantidade de registos que o modelo identificou corretamente como negativos, divididos pelo número total de registos negativos (Han et al., 2012).

$$Especificidade = \frac{TN}{TN + FP} * 100(\%) \ (5)$$

Precisão (*Precision*) – esta métrica corresponde à quantidade de registos que o modelo identificou corretamente como positivos, divididos pelo número total de registos que o modelo classificou como positivos (Novakovic et al., 2017).

$$Precisão = \frac{TP}{TP + FP} * 100(\%) \ (6)$$

***F1-Score* –** esta medida exibe o balanço entre a precisão e o *recall*. É calculada multiplicando-se pelo dobro, a divisão da precisão vezes o *recall* pela soma da precisão com o *recall* (Han et al., 2012).

$$F1 - Score = \frac{Precisão * Recall}{Precisão + Recall} * 2 \ \ (7)$$

Índice Kappa – esta medida é um método estatístico que avalia o nível de concordância entre dois conjuntos de dados (Carletta, 2008).

$$Kappa = \frac{P(A) - P(E)}{1 - P(E)} \ (8)$$

Curva ROC – a curva ROC (*Receiver Operating Characterisc*) é uma ferramenta visual útil para comparar dois modelos de classificação (Han et al., 2012). Esta curva é uma representação gráfica, em que o eixo do X representa a taxa de falsos positivos e o eixo do Y representa a taxa dos verdadeiros positivos (Novakovic et al., 2017).

Esta métrica de desempenho permite visualizar um trade-off entre a taxa de verdadeiros positivos e a taxa dos falsos positivos, isto é, entre a sensibilidade e a especificidade respetivamente (Han et al., 2012).

CURVA AUC – a curva AUC (*Area Under ROC Curve*) calcula o desempenho de discriminação, definida como a capacidade de separar os zeros dos zeros. AUC representa a probabilidade de uma presença selecionada aleatoriamente ter uma previsão contínua maior, do que uma ausência selecionada aleatoriamente em todos os *thresholds* (Sofaer et al., 2019).

2.5.3.2 Cross-Validation

Cross-Validation (CV) é uma técnica utilizada para avaliar modelos de previsão, onde a amostra de dados é dividida aleatoriamente em ***n*** subjconjuntos, partições ou folds (Witten et al., 2011).

Em *k-fold cross-validation*, os dados são divididos aleatoriamente em ***k*** subjconjuntos mutuamente exclusivos ***D1, D2, … DK*** com tamanhos aproximadamente iguais. O treino e o teste são realizados ***k*** vezes. Na iteração ***i***, a partição ***Di*** é reservada para testes e as restantes partições são utilizadas de forma coletiva para treinar o modelo. Isto é, na primeira iteração, os subconjuntos ***D2,…DK*** são utilizados coletivamente para treino, de modo a obter o primeiro modelo, que é testado no ***D1***. Na segunda iteração, o modelo é treinado utilizado os subconjuntos ***D1,… DK*** e é testado no ***D2***. Este processo é realizado até satisfazer o valor de ***k*** (Han et al., 2012).

A forma mais comum de prever a taxa de erro de uma técnica de *learning* é a partir de um conjunto de dados único e fixo, sendo esta a estratificação de *10-folds Cross-Validation*. É o método mais usado, uma vez que foram realizados testes em diversos conjuntos de dados, com recurso a diferentes técnicas de *learning*, e verificou-se que o 10 é o número correto de *folds* para obter melhores estimativas de erro. Apesar de estes argumentos não serem aceites como totalmente conclusivos, o método *10-folds Cross-Validation* tornou-se o método comum em termos práticos (Witten et al., 2011).

2.6 Predictive Analytics

Segundo (Larose & Larose, 2015), *Predictive Analytics* é um processo que consiste na extração de informação a partir de grandes *datasets* com o objetivo de realizar previsões e estimativas sobre futuros *outcomes* (resultados). Predictive Analytics é baseada nas relações entre as variáveis e tem como propósito responder a questões difíceis, como por exemplo, se os *inputs* X e Y mudarem, o que acontecerá a Z? *Predictive Analytics* ao utilizar os dados poderá responder a este tipo de perguntas, isto é, poderá prever os futuros resultados (Fitz-enz & Mattox II, 2014).

Predictive Analytics envolve a utilização de *data analytics* para fazer previsões com base nos dados. Este processo utiliza os dados, realizando técnicas de análise, estatística, *Data Mining* e *Machine Learning*, de modo a criar modelos preditivos para prever futuros eventos (Ashfaque, 2020). A utilização de *Predictive Analytics* inicia-se com um objetivo de negócio, isto é, utilizar os dados para economizar tempo, reduzir custos e desperdícios. O processo envolve a agregação de conjuntos de dados em modelos que criam resultados simples, claros e útéis para ajudar a alcançar o objetivo definido, com menos desperdício de tempo e recursos (Ashfaque, 2020).

A aplicação de *Predictive Anlytics* nas organinações é extremamente útil, uma vez que ajuda na otimização dos processos de trabalho, no melhor entendimento do comportamento do cliente, na identificação de oportunidades e na antecipação de problemas (Boonsiritomachai et al., 2016).

3. Estado de Arte

Neste terceiro capítulo é demonstrado práticas existentes relativas ao tema deste projeto de dissertação, como projetos e estudos já desenvolvidos, descrevendo o seu propósito e utilidade. Além disso, é apresentado o processo de pesquisa realizado neste projeto de dissertação.

3.1 Processo de Pesquisa

O processo de pesquisa envolveu a procura de artigos, jornais, livros, dissertações, entre outros. Para a realização da pesquisa, utilizou-se as seguintes plataformas:

- Google Scholar (https://scholar.google.pt/);
- Research Gate (https://www.researchgate.net/);
- Elsevier (https://www.elsevier.com/);
- ScienceDirect (https://www.sciencedirect.com/);
- Zlibrary (https://1lib.eu/);
- Repositório da UM (https://repositorium.sdum.uminho.pt/).

Para a pesquisa e recolha de documentos, foram utilizadas palavras-chaves nas plataformas, sendo elas, "Infeções Nosocomiais, "*Predictive Analytics*", "*Data Mining*" e "*Machine Learning*".

Os documentos foram armazenados na ferramenta *Mendeley* e a sua seleção foi realizada com base num conjunto de critérios:

- Reconhecimento do autor na área de estudo (número de citações);
- Ano do documento;
- Explicação clara do conceito a investigar.

3.2 Predictive Analytics nos Cuidados de Saúde

Atualmente, *Predictive Analytics* é maioritariamente utilizada como ferramenta de negócio nas organizações, no entanto, tem havido um aumento do seu uso nas instituições de saúde (Engelgau et al., 2019).

Na área da saúde tem-se verificado um elevado aumento dos dados armazenados, criando grandes e complexos *datasets*. Este aumento acontece, dado que é gerado novos dados ao longo de todo o processo de atendimento e tratamento do paciente (Alharthi, 2018). Os dados gerados são muito complexos, uma vez que podem ser estruturados, como lista de medicações, resultados de análises de sangue, entre outras, ou não estruturados, como notas sobre o progresso de tratamento ou outro tipo de relatório. As bases de dados relacionais ou *Data Warehouses* (DW) não possuem a capacidade de analisar e organizar esse tipo de dados, no entanto, *Predictive Analytics* é um processo que possibilita e facilita a análise desses dados, fazendo previsões com bases no dados, como por exemplo, prever se um dado indivíduo tem alto de risco de ter cancro da mama (Alharthi, 2018).

3.2.1 Diagnosing breast cancer with an improved artificial immune recognition system

Um bom exemplo de *Predictive Analytcis* na área da saúde, é um estudo realizado para a previsão de cancro da mama, "*Diagnosing breast cancer with an improved artificial immune recognition system*". Nesse estudo foi desenvolvido um sistema híbrido que utilizou o *Artificial Immune Recognition System* (algoritmo baseado em aprendizagem supervisionada) e algoritmos de *Data Mining*, para a criação de um modelo preditivo que distinguisse entre diagnósticos de cancro da mama e não cancro da mama. Este modelo foi criado com base nos dados do *dataset* do *Wisconsin Breast Cancer*, que contem 699 amostras de tecido, e foi concebido com o intuito de retornar valores de previsão de sim cancro da mama ou não cancro da mama. Neste estudo, o modelo apresentou 100% de precisão (Saybani et al., 2016).

3.3 Predictive Analytics na Previsão de Infeções Noscomiais

A utlização de *Predictive Analytics* nas infeções nosocomiais, através de técnicas de *Data Mining* e *Machine Learning*, é extremamente útil, uma vez que promove a segurança e a eficácia do instituto de saúde na gestão de controlo de infeção. Com a utilização das técnicas de *Data Mining*, é possível criar modelos preditivos que calculem a probabilidade de uma determinada infeção nosocomial segundo determinados parâmetros (Silva et al., 2015).

3.3.1 Predicting Nosocomial Infection by Using Data Mining Technologies

Um exemplo de *Predictive* Anlytics nas infeções nosocomiais é o estudo realizado por (Silva et al., 2015), "*Predicting Nosocomial Infection by Using Data Mining Technologies*", que tem como objetivo prever infeções nosocomiais para isolamento e minimzação da incidência de infeção. Neste estudo é desenvolvido modelos preditivos, através de técnicas de *Data Mining*, que consigam classificar um paciente com capacidade de ter infeção nosocomial ou não. Os modelos foram criados com base num *dataset* composto por informações de formulários de infeção nosocomial das Unidades de Medicina do Centro Hospitalar do Porto. Os atributos recolhidos dos formulários de infeção nosocomial não tinham qualidade suficiente para serem utilizados no processo de *Data Mining*, desse modo, realizaram uma seleção de atributos com o objetivo de escolher as variáveis representativas para a execução do estudo. As variáveis escolhidas foram: Infeção nosocomial, Idade, Sexo, Especialidade clínica, Dias de hospitalização, Fatores de risco, Cateter urinário, Cateter periférico, Cateter central, Intubação Nasogástrica e Intubação Nasotraqueal. Após a seleção dos dados e variáveis, realizaram certas atividades, como limpeza, correção e construção de dados, de modo a obter dados mais consistes. Além disso, aplicaram técnicas de *oversampling* (sobreamostragem) ao conjunto de dados, com o objetivo de replicar os dados associados à ocorrência de uma infeção nosocomial. Esta *oversampling* (sobreamostragem) deu origem a três conjuntos de dados: um conjunto de dados sem dados replicados (Abordagem A), um conjunto de dados com dados replicados (Abordagem B) e um conjunto de dados com dados replicados e a variável Idade agregada em classes (Abordagem C). Para a criação dos modelos foi necessário criar 4 cenários em que cada um possuía diferentes variáveis. Os cenários criados foram: Ausência de Fatores de Risco (Cenário 1), Ausência de Intubação (Cenário 2), Ausência de Cateterização (Cenário 3) e Todas as variáveis (Cenário 4). As técnicas de DM escolhidas foram o *Support Vector Machine* (SVM) e *Naive Bayes* (NB), uma vez que foram as que apresentaram melhores resultados. Estas técnicas foram depois aplicadas a todas as combinações de cenários e abordagens, com o objetivo de criar novo conhecimento e obter o melhor modelo para a resolução do problema. Com os resultados,

procederam à avaliação dos modelos através da Matriz Confusão e por fim à seleção do modelo que apresentasse os melhores resultados possíveis. Os melhores modelos foram selecionados consoante os valores de sensibilidade, dado que é importante identificar todas as não ocorrências de infeção nosocomial. A melhor combinação de cenário e abordagem para as duas técnicas de *Data Mining* escolhidas foi o Cenário 4 e Abordagem B, sendo que estes apresentam um maior valor de sensibilidade quando modelado com SVM (Silva et al., 2015).

3.3.2 Predicting Hospital-Acquired Infections by Scoring System with Simple Parameters

Outro exemplo é o estudo realizado por (Chang et al., 2011), "*Predicting Hospital-Acquired Infections by Scoring System with Simple Parameters*", onde foi desenvolvido um sistema de pontuação, com base em técnicas de modelação de Redes Neuronais Artificiais (RNA) ou Redes Neuronais (RN), e Regressão Logística (RL), para a previsão de infeções nosocomiais com parâmetros simples. Os dados utilizados na construção deste sistema são provenientes do *Eletronic Health Records* (EHR) do Hospital Wan Fang da Universidade Médica de Taipei. Os dados foram divididos em 3 grupos, sendo eles, *trainning set* (conjunto de dados de treino), *selection set* (conjunto de dados de seleção) e *test set* (conjunto de dados de teste). O *trainning set* possui cerca de 50% dos dados e foi utilizado para a construção dos modelos de RNA e RL, o *selection set* possui cerca de 25% dos dados e foi usado na modelação de RNA, e o *teste set* possui cerca de 25% dos dados e foi utilizado para avaliação interna. Para a criação dos modelos foram aplicados certos métodos, como o uso de diferentes cenários, para obter os melhores modelos possíveis. Os métodos aplicados foram ambos usados em RNA e RL. Após a criação dos modelos, procederam à sua avaliação e seleção. Por fim, conseguiram concluir que 48.2% dos pacientes eram pacientes do sexo feminino, com idade entre os 17 e 80 anos, e que os pacientes que contraíram infeções nosocomiais eram mais velhos e do sexo masculino. Para além disso, conseguiram concluir que a infeção nosocomial era mais frequente em hemodiálise, em dispositivos utilizados como cateteres arteriais, intubações endotraqueais e traqueostomia, entre outros (Chang et al., 2011).

4. METODOLOGIAS

Neste quarto capítulo são abordadas as metodologias que serão utilizadas na elaboração deste projeto de dissertação, sendo elas CRISP-DM e a *Design Science Research Methodology (DSRM)*.

4.1 Metodologia CRISP-DM

No desenvolvimento prático deste projeto de dissertação será seguida a metodologia *Cross-Industry Standard Process for Data Mining* (CRISP-DM). Esta metodologia é descrita num modelo de processo hierárquico, constituído por quatro níveis de abstração, sendo hierarquizada do nível mais geral para o mais específico (Pete et al., 2000). Estes quatro níveis são:

1. ***Phases*** – neste nível, o processo de *Data Mining* está organizado em fases e cada fase consiste em várias tarefas genéricas do segundo nível;
2. ***Generic Tasks*** – este segundo nível tem como objetivo cobrir todas as situações possíveis de *Data Mining*, ou seja, as tarefas genéricas devem ter a capacidade de cobrir todos os processos e aplicações de *Data Mining*, e devem ser válidas para novas técnicas de modelação;
3. ***Specialized Tasks*** – o tercerio nível tem como propósito definir como é que as ações das tarefas genéricas devem ser executadas em situações específicas;
4. ***Process Instances*** – o quarto e último nível tem como objetivo armazenar todas as ações, decisões e resultados do *Data Mining*. Este nível é organizado de acordo com as tarefas definidas nos níveis superiores, representando o que acontece numa situação específica.

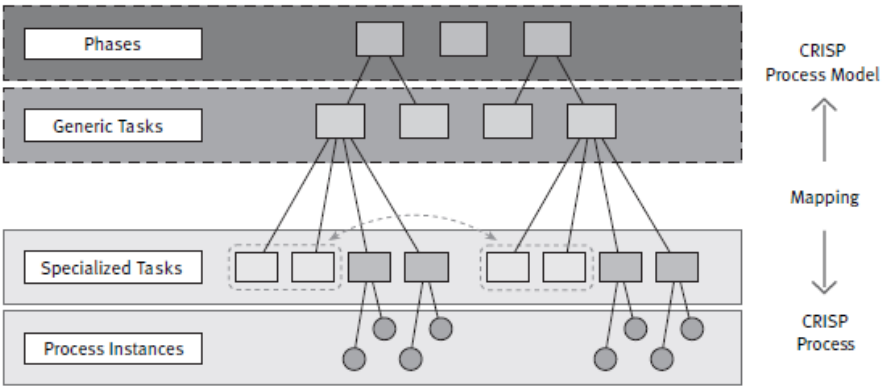

Figura 6 – Níveis da Metodologia CRISP-DM (retirado de (Pete et. al., 2000))

A metodologia possui um modelo de referência, o qual mostra uma visão geral do ciclo de vida de um projeto de *Data Mining*. Este modelo contem as fases do projeto, as respetivas tarefas e as relações entre as tarefas (Pete et al., 2000). O ciclo de vida de CRISP-DM consiste em seis fases, como é possível observar na figura abaixo (Figura 7).

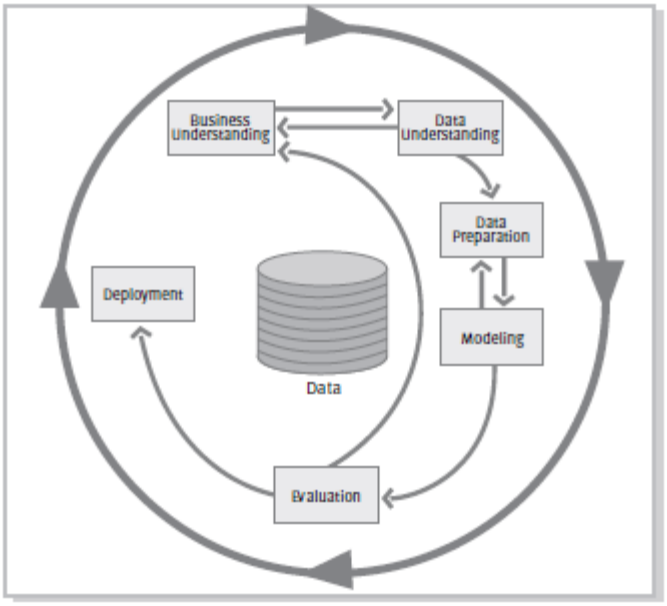

Figura 7 – Ciclo de vida do CRISP-DM (retirado de (Pete et. al., 2000))

- **Compreensão do Negócio** – Esta fase tem como propósito a compreensão e formação de uma perspetiva do negócio, bem como a definição dos objetivos de negócio e uma avaliação da situação atual. Esta informação obtida é depois convertida num problema de *Data Mining*, ou seja, é determinado os objetivos de *Data Mining* para que depois seja possível produzir um plano de projeto;
- **Compreensão dos Dados** – Nesta fase é realizado a recolha inicial dos dados, no qual é executado atividades, como a descrição, exploração e verificação de qualidade, para a compreensão e identificação de possíveis problemas nos dados;
- **Preparação dos Dados** – A fase de Preparação dos Dados cobre todas as atividades necessárias para a construção do *dataset* final a partir dos dados iniciais. Este *dataset* será alcançado a partir de atividades como, seleção de dados, limpeza de dados, correção de dados, construção de novos dados e integração de dados, e será utilizado na seguinte fase de Modelação;

- **Modelação** – Esta fase tem como objetivo a seleção e a aplicação de técnicas de modelação no *dataset* trabalhado anteriormente. Os parâmetros dos modelos são definidos e calibrados para os melhores valores possíveis;
- **Avaliação** – Nesta fase será realizado a avaliação do(s) modelo(s) que apresentarem os melhores valores. O(s) modelo(s) serão avaliados com profundidade e rigor, e as tarefas utlizadas na construção do(s) modelo(s) serão revistas, de modo a garantir que o(s) modelo(s) atinjam propriamente os objetivos do negócio;
- **Implementação** – Esta fase tem como finalidade organizar e apresentar o conhecimento adquirido nos modelos de uma forma que o cliente consiga usá-lo com facilidade. Nesta fase é trabalhado um plano de monitorização e manutenção que integra toda a estratégia, para que depois o conhecimento possa ser acedido sobre a forma de um relatório ou outro tipo de visualização de informação. A fase de Implementação pode ser simples, como pode ser complexa, dependendo dos requisitos impostos pelo cliente.

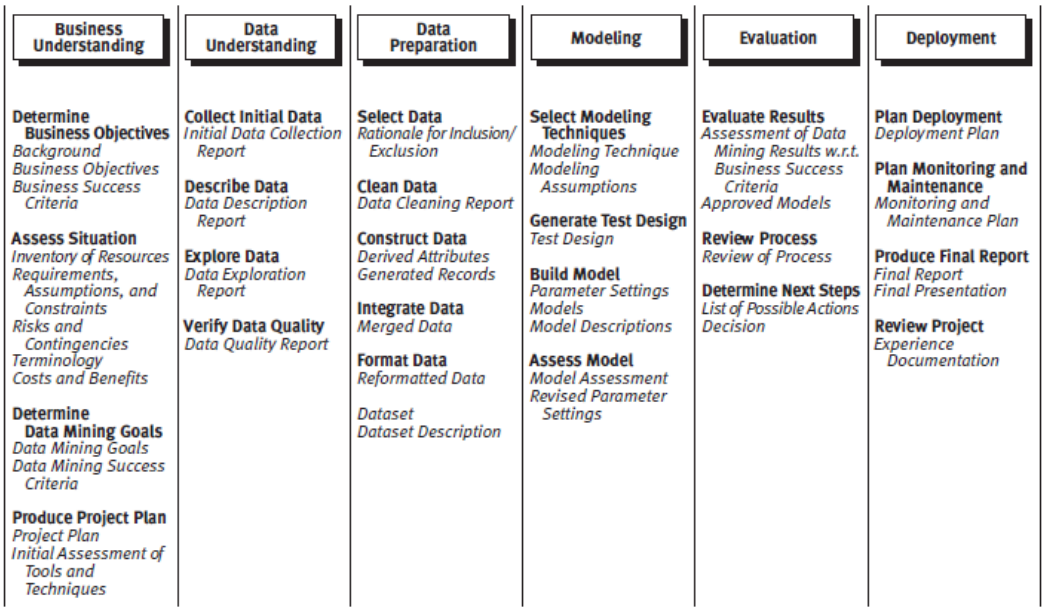

Figura 8 – Fases detalhadas do ciclo de vida de CRISP-DM (retirado de (Pete et. al., 2000))

4.2 Metodologia Design Science Research

Para a elaboração desta dissertação será utilizada a metodologia de investigação científica *Design Science Research Methodology* (DSRM). Esta metodologia consiste num processo para projetar artefactos, de modo a resolver problemas, realizar contribuições de pesquisa, avaliar projetos e comunicar os resultados a audiências apropriadas (Peffers et al., 2007).

A metodologia encontra-se dividida em seis fases, segundo (Peffers et al., 2007), como se pode verificar na figura abaixo (Figura 9).

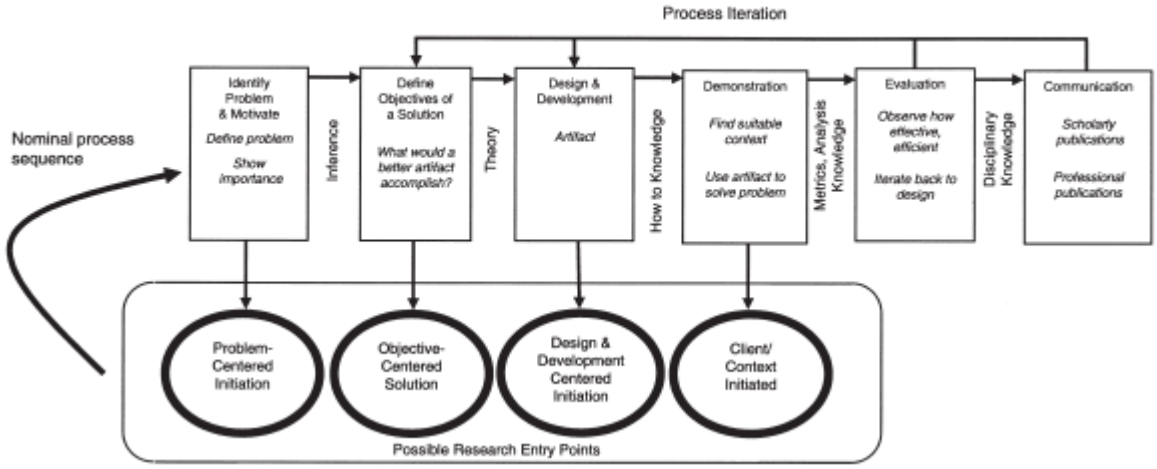

Figura 9 – Fases da Metodologia Design Science Research (retirado de (Peffers et al., 2007))

1. **Identificação e Motivação do Problema** – Nesta fase é realizado a identificação dos aspetos mais importantes na definição do problema e o valor da solução. Para que esta primeira fase seja realizada é fundamental ter conhecimento do estado do problema e da importância da solução;
2. **Definição dos Objetivos da Solução** – Esta fase tem como propósito a definição dos objetivos que ajudam na resolução do problema. Para a realização desta fase é necessário ter conhecimento do estado do problema e das soluções, caso hajam, e a sua eficácia;
3. ***Desgin* e Conceção** – Esta fase permite a criação de artefactos que podem ser construções, modelos, métodos ou instâncias. Tem como âmbito determinar a funcionalidade desejada do artefacto e a sua arquitetura, para de seguida, criar o artefacto real. Para que esta fase seja realizada é necessário ter conhecimento da teoria que será utilizada na solução;

4. **Demonstração** – A fase de demonstração consiste na validação da solução desenvolvida, isto é, na implementação e validação da arquitetura anteriormente proposta, para resolver uma ou mais instâncias do problema. Para que esta fase seja executada é necessário ter conhecimento sobre como usar o artefacto na resolução do problema;
5. **Avaliação** – Esta fase envolve a comparação dos objetivos de uma solução com os resultados reais obtidos no uso do artefacto validado anteriormente. Para que a avaliação seja concluída é necessário possuir conhecimento de métricas e técnicas de análise;
6. **Comunicação** – Nesta última fase é apresentado e divulgado os resultados obtidos, mostrando a utilidade do trabalho realizado. Para que esta fase seja concluída é necessário possuir conhecimento de cultura disciplinar.

5. TRABALHO REALIZADO

Neste sexto capítulo é exposto todo o trabalho prático realizado e os resultados obtidos neste projeto de dissertação. Para além disso, apresenta a abordagem à metodologia CRISP-DM, a qual foi utilizada para orientar todo o trabalho prático desenvolvido.

5.1 Ferramentas Utilizadas

Ao longo deste projeto de dissertação foi necessário utilizar diversas ferramentas. Assim, em seguida, estão enumeradas as ferramentas utilizadas no desenvolvimento do trabalho prático, bem como a respetiva descrição.

- ***Microsoft Excel 2019***

Esta ferramenta é fornecida pela *Microsoft*, no pacote *Office*, e permite a visualização, análise e exploração dos dados. Neste projeto este software foi principalmente utilizado para a visualização dos dados.

- ***Jupyter Notebook by Anaconda***

Esta ferramenta é fornecida pelo *software Anaconda*, sendo este último uma distribuição da linguagem *Python*, que possibilita a criação de ambientes virtuais. O *Jupyter Notebook* é uma aplicação *web open source* que permite a criação, edição e partilha de documentos que contêm *live code*, equações, visualizações e texto narrativo. Este programa possibilita a limpeza e transformação de dados, simulações numéricas, modelação estatística, ML e entre outros (Jupyter, 2016). Neste projeto este *software* foi principalmente utilizado para a visualização, análise, processamento e tratamento de dados, e na implementação de técnicas de DM e ML, através do uso da linguagem *Python*.

- ***Python***

Esta linguagem de programação foi criada em 1991 por Guldo Von Rossum e é uma linguagem *open source*, interativa e simples de usar e aprender (Nelli, 2018). *Python* foi a linguagem de programação escolhida para a realização da parte prática deste projeto e foi principalmente utilizada na visualização, análise, processamento e tratamento de dados, e na realização das técnicas de DM e ML. As técnicas de DM e ML foram concebidas com recurso à *package sklearn*.

5.2 Compreensão do Negócio

Para realizar este projeto de dissertação foi necessário determinar os objetivos de negócio que respondam aos problemas que a Comissão de Controlo de Infeção (CCI) do Hospital da Senhora da Oliveira de Guimarães possui. O CCI até à data não dispõe nenhum mecanismo capaz de detetar atempadamente infeção nosocomial, podendo neste caso comprometer o estado de saúde de doentes internados e de todos os que o rodeiam. Desse modo, o objetivo de negócio consiste na previsão automática e atempada de infeção nosocomial, com o intuito de otimizar a gestão de controlo de infeção do hospital e dos custos associados.

A nível de *Data Mining*, os objetivos consistem na construção de modelos preditivos, através da técnica de Classificação, e consequente, a obtenção de bons modelos de previsão. Após a criação dos modelos preditivos, serão identificados os que apresentarem melhores valores, para que no futuro possam ser introduzidos num protótipo Sistema de Suporte à Decisão Inteligente.

5.3 Compreensão dos Dados

Neste projeto os dados foram fornecidos pelo Hospital da Senhora da Oliveira de Guimarães e estão relacionados com os parâmetros e regras que o CCI utiliza para a deteção de infeção. São dados relativos a cirurgias realizadas, antibióticos administrados, internamentos e urgências. O espaço temporal da amostra de dados encontra-se entre os anos de 2018 a 2021.

Os dados recolhidos estão divididos em 5 *datasets* e cada *dataset* possui um número diferente de registos. Os *datasets* recolhidos são:

- **Cirurgias**

Este *dataset* é constituído por 4 campos e 20734 registos, e possui informação relativa a todos as cirurgias realizadas. Na tabela a seguir (Tabela 4) é possível visualizar as colunas, com a respetiva descrição, que constituem esta amostra de dados.

Tabela 4 – Descrição dos dados referentes ao dataset Cirurgias

Coluna	Descrição	Tipo
NUM_SEQUENCIAL	ID do paciente.	*Integer*
DTA_NASC	Data de nascimento do paciente.	*Datetime*
SEXO	Sexo do paciente. (1 – Masculino, 2 – Feminino, 3 – Indefinido)	*Categorical*
DATAMOV	Data da cirurgia.	*Datetime*

- **Antibióticos**

Esta *dataset* é constituída por 7 campos e 27407 registos, e possui informação relativa a todos os pacientes que necessitaram de ser administrados com antibióticos após cirurgia. Na tabela a seguir (Tabela 5) é possível visualizar as colunas, com a respetiva descrição, que constituem esta amostra de dados.

Tabela 5 – Descrição dos dados referentes ao dataset Antibióticos

Coluna	Descrição	Tipo
NUM_SEQUENCIAL	ID do paciente.	*Integer*
DATA_INICIO	Data de início de administração de antibiótico.	*Datetime*
DATA_FIM	Data de fim de administração de antibiótico.	*Datetime*
MODULO	Área do paciente. (COM – Consulta, INT – Internamento, URG – Urgência)	*String*
MED_DESIGNACAO	Nome do antibiótico administrado.	*Categorical*
ART_DESIGNACAO	Nome do antibiótico e a respetiva dosagem administrada.	*Categorical*
DATA_CIRURGIA	Data da cirurgia.	*Datetime*

- **Internamentos**

Este *dataset* é constituída por 6 campos e 23453 registos, e possui informação relativa a todos os pacientes que necessitaram de ser internados. Na tabela a seguir (Tabela 6) é possível visualizar as colunas, com a respetiva descrição e tipo, que constituem esta amostra de dados.

Tabela 6 – Descrição dos dados referentes ao dataset Internamentos

Coluna	Descrição	Tipo
NUM_SEQUENCIAL	ID do paciente.	*Integer*
ADMISSAO	Data de admissão no internamento.	*Datetime*
ALTA	Data de alta do internamento.	*Datetime*
DTA_NASCIMENTO	Data de nascimento do paciente.	*Datetime*
SEXO	Sexo do paciente. (1 – Masculino, 2 – Feminino, 3 – Indefinido)	*Categorical*
CIRURGIA	Número de cirurgias realizadas durante o tempo de internamento.	*Integer*

- **Urgências**

Esta *dataset* é constituída por 7 campos e 157199 registos, e possui informação relativa a todos os indivíduos que deram entrada nas urgências do hospital. Na tabela a seguir (Tabela 7) é possível visualizar as colunas, com a respetiva descrição e tipo, que constituem esta amostra de dados.

Tabela 7 – Descrição dos dados referentes ao dataset Urgências

Coluna	Descrição	Tipo
NUM_SEQUENCIAL	ID do paciente.	*Integer*
URG_EPISODIO	ID do episódio de urgência.	*Integer*
DATAHORA_ADM	Data de admissão nas urgências.	*Datetime*
DATAHORA_ALTA	Data de alta das urgências.	*Datetime*
COD_LOCAL	Código local do tipo de urgência.	*Categoriocal*
COD_DIAG_ALTA	Código do diagnóstico de alta hospitalar.	*Categorical*
DES_DIAGNOSTICO	Designação do diagnóstico.	*String*

- **Infeções**

Esta *dataset* é constituída por 3 campos e 5385 registos, e possui informação relativa a todos os indivíduos que possuíam risco de infeção. Na tabela a seguir (Tabela 8) é possível visualizar as colunas, com a respetiva descrição e tipo, que constituem esta amostra de dados.

Tabela 8 – Descrição dos dados referentes ao dataset Infeções

Coluna	Descrição	Tipo
ID Registo	ID de registo do paciente no departamento de infeção.	*Integer*
Data Admissão	Data de admissão no departamento de infeção.	*Datetime*
Tem infeção	Identificação de infeção ou não infeção no paciente.	*String*

De seguida, foi realizado uma exploração dos dados dos 5 *datasets*, de modo a compreendê-los com maior detalhe.

- **Cirurgias**

Elaborado a análise geral ao *dataset* cirurgias, podemos concluir que todas as colunas estão completamente preenchidas, com um total de 20734 linhas, como é possível verificar na figura seguinte (Figura 10).

```
RangeIndex: 20734 entries, 0 to 20733
Data columns (total 4 columns):
 #   Column          Non-Null Count  Dtype
---  ------          --------------  -----
 0   DATAMOV         20734 non-null  object
 1   NUM_SEQUENCIAL  20734 non-null  int64
 2   DTA_NASC        20734 non-null  object
 3   SEXO            20734 non-null  int64
```

Figura 10 – Análise geral das colunas do dataset Cirurgias

Para além disso, foi realizado uma análise detalhada para cada coluna do *dataset*, como é possível visualizar a seguir:

- **NUM_SEQUENCIAL –** A figura a seguir apresentada (Figura 11), mostra a contagem efetuada sobre os valores (mostrando apenas o top 10). Podemos verificar que a moda da nossa amostra se encontra no indivíduo com o ID de paciente igual a 188027, com cerca de 31 linhas. Os restantes apresentam valores entre 15 a 30 linhas.

	NUM_SEQUENCIAL	Contagem	Percentagem %
0	188027	31	0.149513
1	240667	30	0.144690
2	251269	19	0.091637
3	82651	19	0.091637
4	503706	17	0.081991
5	558620	16	0.077168
6	11558	16	0.077168
7	210780	16	0.077168
8	164811	15	0.072345
9	230325	15	0.072345

Figura 11 – Análise à frequência de valores da coluna "NUM_SEQUENCIAL"

- **DTA_NASC –** A figura a seguir apresentada (Figura 12), mostra a contagem efetuada sobre os valores (mostrando apenas o top 10). Podemos verificar que a moda da nossa amostra se encontra nos indivíduos com a data de nascimento igual a 1964.09.07 00:00:00, com cerca de 32 linhas. Os restantes apresentam valores entre 16 a 30 linhas.

	DTA_NASC	Contagem	Percentagem %
0	1964.09.07 00:00:00	32	0.154336
1	1982.11.12 00:00:00	30	0.144690
2	1969.02.09 00:00:00	21	0.101283
3	1946.02.26 00:00:00	20	0.096460
4	1950.10.30 00:00:00	19	0.091637
5	1948.03.22 00:00:00	19	0.091637
6	1961.08.14 00:00:00	18	0.086814
7	1939.03.05 00:00:00	18	0.086814
8	1944.03.29 00:00:00	17	0.081991
9	1942.07.04 00:00:00	16	0.077168

Figura 12 – Análise à frequência de valores da coluna "DTA_NASC"

- **SEXO –** A figura a seguir apresentada (Figura 13), mostra a contagem efetuada sobre os valores. Podemos verificar que a moda da nossa amostra se encontra nos indivíduos com valor igual a 2 (sexo feminino), possuindo cerca de 12193 linhas. O outro valor, valor 1 (sexo masculino), apresenta cerca de 8541 linhas.

	SEXO	Contagem	Percentagem %
0	2	12193	58.806791
1	1	8541	41.193209

Figura 13 – Análise à frequência de valores da coluna "SEXO"

- **DATAMOV –** A figura a seguir apresentada (Figura 14), mostra a contagem efetuada sobre os valores (mostrando apenas o top 10). Podemos verificar que a moda da nossa amostra se encontra nos indivíduos com data de cirurgia igual 2021.06.18, com cerca de 104 linhas. Os restantes apresentam valores entre 88 a 98 linhas.

	Contagem	Percentagem %
2021.06.18 00:00:00	104	0.501592
2021.05.18 00:00:00	98	0.472654
2021.05.04 00:00:00	96	0.463008
2021.06.01 00:00:00	96	0.463008
2021.06.15 00:00:00	93	0.448539
2021.06.12 00:00:00	92	0.443716
2021.04.20 00:00:00	91	0.438893
2021.03.23 00:00:00	89	0.429247
2021.05.08 00:00:00	89	0.429247
2021.05.22 00:00:00	88	0.424424

Figura 14 – Análise à frequência de valores da coluna "DATAMOV"

- **Antibióticos**

Elaborado a análise geral ao *dataset* antibióticos, podemos concluir que todas as colunas estão completamente preenchidas, com um total de 27407 linhas, exceto as colunas "DATA_FIM", que possui 1984 linhas vazias, e as colunas "MODULO" e "DATA_CIRURGIA", que não têm qualquer tipo de registo. Isto tudo pode verificar-se na figura seguinte (Figura 15).

```
RangeIndex: 27407 entries, 0 to 27406
Data columns (total 7 columns):
 #   Column          Non-Null Count  Dtype
---  ------          --------------  -----
 0   NUM_SEQUENCIAL  27407 non-null  int64
 1   DATA_INICIO     27407 non-null  object
 2   DATA_FIM        25423 non-null  object
 3   MODULO          0 non-null      float64
 4   MED_DESIGNACAO  27407 non-null  object
 5   ART_DESIGNACAO  27407 non-null  object
 6   DATA_CIRURGIA   0 non-null      float64
```

Figura 15 – Análise geral das colunas do dataset Antibióticos

Para além disso, foi realizado uma análise detalhada para cada coluna do *dataset*, como é possível visualizar a seguir:

- **NUM_SEQUENCIAL –** A figura a seguir apresentada (Figura 16), mostra a contagem efetuada sobre os valores (mostrando apenas o top 10). Podemos verificar que a moda da nossa amostra se encontra nos indivíduos com ID de paciente igual a 141583, com cerca de 32 linhas. Os restantes apresentam valores entre 23 a 31 linhas.

	NUM_SEQUENCIAL	Contagem	Percentagem %
0	141583	32	0.116758
1	192777	31	0.113110
2	252566	29	0.105812
3	69200	27	0.098515
4	48108	27	0.098515
5	593162	26	0.094866
6	240667	26	0.094866
7	101928	25	0.091218
8	222266	24	0.087569
9	596020	23	0.083920

Figura 16 – Análise à frequência de valores da coluna "NUM_SEQUENCIAL"

- **DATA_INICIO –** A figura a seguir apresentada (Figura 17), mostra a contagem efetuada sobre os valores (mostrando apenas o top 10). Podemos verificar que a moda da nossa amostra se encontra nos indivíduos com data de início de antibiótico igual a 2020.09.19 08:00:00, com cerca de 11 linhas. Os restantes apresentam valores entre 6 a 10 linhas.

	DATA_INICIO	Contagem	Percentagem %
0	2020.09.19 08:00:00	11	0.040136
1	2021.03.02 08:00:00	10	0.036487
2	2021.03.30 08:00:00	8	0.029190
3	2021.05.20 08:00:00	7	0.025541
4	2021.05.06 08:00:00	7	0.025541
5	2020.09.03 00:00:00	7	0.025541
6	2020.07.23 00:00:00	6	0.021892
7	2021.03.04 08:00:00	6	0.021892
8	2021.04.15 08:00:00	6	0.021892
9	2021.02.08 07:00:00	6	0.021892

Figura 17 – Análise à frequência de valores da coluna "DATA_INICIO"

- **DATA_FIM –** A figura a seguir apresentada (Figura 18), mostra a contagem efetuada sobre os valores (mostrando apenas o top 10). Podemos verificar que a moda da nossa amostra se encontra nos indivíduos com data de fim de antibiótico igual a 2021.06.03 00:00:00, com cerca de 33 linhas. Os restantes apresentam valores entre 27 a 31 linhas.

	DATA_FIM	Contagem	Percentagem %
0	2021.06.03 00:00:00	33	0.129804
1	2021.04.29 00:00:00	31	0.121937
2	2021.05.29 00:00:00	30	0.118003
3	2021.05.05 00:00:00	30	0.118003
4	2021.06.02 00:00:00	29	0.114070
5	2021.05.13 00:00:00	29	0.114070
6	2021.03.19 00:00:00	28	0.110136
7	2021.04.21 00:00:00	27	0.106203
8	2021.05.07 00:00:00	27	0.106203
9	2021.04.17 00:00:00	27	0.106203

Figura 18 – Análise à frequência de valores da coluna "DATA_FIM"

- **MED_DESIGNACAO –** A figura a seguir apresentada (Figura 19), mostra a contagem efetuada sobre os valores (mostrando apenas o top 10). Podemos verificar que a moda da nossa amostra se encontra nos indivíduos com nome de antibiótico igual a cefaZOLINA, com cerca de 4434 linhas. Os restantes apresentam valores entre 722 a 4326 linhas.

	MED_DESIGNACAO	Contagem	Percentagem %
0	cefaZOLINA	4434	16.178349
1	Amoxicilina + Ácido clavulânico	4326	15.784289
2	cefTRIAXONA	3152	11.500711
3	Azitromicina	3088	11.267195
4	Piperacilina + Tazobactam	2706	9.873390
5	metRONIDazol	945	3.448024
6	Meropenem	869	3.170723
7	cefOXITINA	781	2.849637
8	Ampicilina	763	2.783960
9	LEVOfloxacina	722	2.634363

Figura 19 – Análise à frequência de valores da coluna "MED_DESIGNACAO"

- **ART_DESIGNACAO –** A figura a seguir apresentada (Figura 20), mostra a contagem efetuada sobre os valores (mostrando apenas o top 10). Podemos verificar que a moda da nossa amostra se encontra nos indivíduos com nome de antibiótico e respetiva dosagem igual a cefAZOLINA 1G AMP IV, com cerca de 4433 linhas. Os restantes apresentam valores entre 781 a 3151 linhas.

	ART_DESIGNACAO	Contagem	Percentagem %
0	cefAZOLINA 1G AMP IV	4433	16.174700
1	cefTRIAXONA 1G AMP IV	3151	11.497063
2	AMOXICILINA+AC CLAVULÂNICO 1,2G AMP IV	2235	8.154851
3	PIPERACILINA TAZOBACTAM 4,5 G AMP IV	2130	7.771737
4	AZITROMICINA 500 MG AMP IV	2037	7.432408
5	AMOXICILINA+AC CLAVULÂNICO 625 MG CP	1037	3.783705
6	AZITROMICINA 500 MG CP	1001	3.652352
7	AMOXICILINA+AC CLAVULÂNICO 2,2G AMP IV	884	3.225453
8	MEROPENEM 1G AMP IV	789	2.878827
9	cefOXITINA 1G AMP IV	781	2.849637

Figura 20 – Análise à frequência de valores da coluna "ART_DESIGNACAO"

- **MODULO e DATA_CIRURGIA –** Não foi realizado qualquer tipo de exploração e análise de dados nestas duas colunas, uma vez que possuem todas as linhas vazias.

- **Internamentos**

Elaborado a análise geral ao *dataset* internamentos, podemos concluir que todas as colunas estão completamente preenchidas, com um total de 23453 linhas, exceto as colunas "DTA_NASCIMENTO" e "SEXO", que possuem ambas 2 linhas vazia, como se pode verificar na figura seguinte (Figura 21).

```
RangeIndex: 23453 entries, 0 to 23452
Data columns (total 6 columns):
 #   Column          Non-Null Count  Dtype
---  ------          --------------  -----
 0   NUM_SEQUENCIAL  23453 non-null  int64
 1   ADMISSAO        23453 non-null  object
 2   ALTA            23453 non-null  object
 3   DTA_NASCIMENTO  23451 non-null  object
 4   SEXO            23451 non-null  float64
 5   CIRURGIA        23453 non-null  int64
```

Figura 21 – Análise geral das colunas do dataset Internamentos

Para além disso, foi realizado uma análise detalhada para cada coluna do *dataset*, como é possível visualizar a seguir:

- **NUM_SEQUENCIAL –** A figura a seguir apresentada (Figura 22), mostra a contagem efetuada sobre os valores (mostrando apenas o top 10). Podemos verificar que a moda da nossa amostra se encontra nos indivíduos com ID de paciente igual a 192777, com cerca de 11 linhas. Os restantes apresentam valores entre 7 a 9 linhas.

	NUM_SEQUENCIAL	Contagem	Percentagem %
0	192777	11	0.046902
1	501236	9	0.038375
2	289	9	0.038375
3	79614	8	0.034111
4	50552	8	0.034111
5	177311	8	0.034111
6	574861	7	0.029847
7	188965	7	0.029847
8	441000	7	0.029847
9	178592	7	0.029847

Figura 22 – Análise à frequência de valores da coluna "NUM_SEQUENCIAL"

- **ADMISSAO –** A figura a seguir apresentada (Figura 23), mostra a contagem efetuada sobre os valores (mostrando apenas o top 10). Podemos verificar que a moda da nossa amostra se encontra nos indivíduos com data de admissão igual a 2020.01.02 13:00:00, com cerca de 10 linhas. Os restantes apresentam valores entre 4 a 8 linhas.

	ADMISSAO	Contagem	Percentagem %
0	2020.01.02 13:00:00	10	0.042638
1	2021.01.01 13:30:00	8	0.034111
2	2020.01.02 12:30:00	6	0.025583
3	2021.01.02 12:30:00	6	0.025583
4	2020.01.03 15:30:00	5	0.021319
5	2020.01.01 13:30:00	5	0.021319
6	2020.01.03 15:00:00	4	0.017055
7	2020.07.03 08:00:00	4	0.017055
8	2020.08.05 09:00:00	4	0.017055
9	2020.01.14 09:00:00	4	0.017055

Figura 23 – Análise à frequência de valores da coluna "ADMISSAO"

- **Alta –** A figura a seguir apresentada (Figura 24), mostra a contagem efetuada sobre os valores (mostrando apenas o top 10). Podemos verificar que a moda da nossa amostra se encontra nos indivíduos com data de alta igual a 2020.01.23 12:30:00, com cerca de 21 linhas. Os restantes apresentam valores entre 12 a 15 linhas.

	ALTA	Contagem	Percentagem %
0	2020.01.23 12:30:00	21	0.089541
1	2021.06.03 15:00:00	15	0.063958
2	2020.03.21 15:00:00	14	0.059694
3	2020.03.05 14:00:00	14	0.059694
4	2020.01.16 13:30:00	14	0.059694
5	2020.01.06 13:30:00	14	0.059694
6	2021.05.26 13:00:00	13	0.055430
7	2021.04.12 13:00:00	13	0.055430
8	2020.07.10 13:30:00	12	0.051166
9	2020.02.28 13:30:00	12	0.051166

Figura 24 – Análise à frequência de valores da coluna "ALTA"

- **DTA_NASCIMENTO –** A figura a seguir apresentada (Figura 25), mostra a contagem efetuada sobre os valores (mostrando apenas o top 10). Podemos verificar que a moda da nossa amostra se encontra nos indivíduos com data de nascimento igual a 2020.09.05 00:00:00 e 2020.01.03 00:00:00, com cerca de 17 linhas cada um. Os restantes apresentam valores entre 12 a 16 linhas.

	DTA_NASCIMENTO	Contagem	Percentagem %
0	2020.01.03 00:00:00	17	0.072492
1	2020.09.05 00:00:00	17	0.072492
2	2020.03.19 00:00:00	16	0.068227
3	2021.02.06 00:00:00	14	0.059699
4	2020.07.26 00:00:00	14	0.059699
5	2020.05.02 00:00:00	13	0.055435
6	2020.04.03 00:00:00	13	0.055435
7	2021.02.05 00:00:00	13	0.055435
8	2021.04.26 00:00:00	12	0.051171
9	2019.12.31 00:00:00	12	0.051171

Figura 25 – Análise à frequência de valores da coluna "DTA_NASCIMENTO"

- **SEXO –** A figura a seguir apresentada (Figura 26), mostra a contagem efetuada sobre os valores. Podemos verificar que a moda da nossa amostra se encontra nos indivíduos com valor igual a 2 (sexo feminino), possuindo cerca de 12978 linhas. O outro valor, valor 1 (sexo masculino) apresenta cerca de 10473 linhas.

	SEXO	Contagem	Percentagem %
0	2.0	12978	55.340924
1	1.0	10473	44.659076

Figura 26 – Análise à frequência de valores da coluna "SEXO"

- **CIRURGIA –** A figura a seguir apresentada (Figura 27), mostra a contagem efetuada sobre os valores (mostrando apenas o top 10). Podemos verificar que a moda da nossa amostra se encontra nos indivíduos com número de cirurgias igual a 0, com cerca de 15953 linhas. Os restantes apresentam valores entre 1 a 7066 linhas.

	CIRURGIA	Contagem	Percentagem %
0	0	15953	68.021149
1	1	7066	30.128342
2	2	356	1.517929
3	3	56	0.238775
4	4	12	0.051166
5	5	5	0.021319
6	32	1	0.004264
7	6	1	0.004264
8	7	1	0.004264
9	11	1	0.004264

Figura 27 – Análise à frequência de valores da coluna "CIRURGIA"

- **Urgências**

Elaborado a análise geral ao *dataset* urgências, podemos concluir que todas as colunas estão completamente preenchidas, com um total de 157199 linhas, exceto as colunas "COD_DIAG_ALTA" e "DES_DIAGNOSTICO", que possuem 59358 e 59399 linhas vazias, respetivamente, como se pode verificar na figura seguinte (Figura 28).

```
RangeIndex: 157199 entries, 0 to 157198
Data columns (total 7 columns):
 #   Column           Non-Null Count   Dtype
---  ------           --------------   -----
 0   NUM_SEQUENCIAL   157199 non-null  int64
 1   URG_EPISODIO     157199 non-null  int64
 2   DATAHORA_ADM     157199 non-null  datetime64[ns]
 3   DATAHORA_ALTA    157199 non-null  datetime64[ns]
 4   COD_LOCAL        157199 non-null  int64
 5   COD_DIAG_ALTA    97841 non-null   object
 6   DES_DIAGNOSTICO  97800 non-null   object
```

Figura 28 – Análise geral das colunas do dataset Urgências

Para além disso, foi realizado uma análise detalhada para cada coluna do *dataset*, como é possível visualizar a seguir:

- **NUM_SEQUENCIAL –** A figura a seguir apresentada (Figura 29), mostra a contagem efetuada sobre os valores (mostrando apenas o top 10). Podemos verificar que a moda da nossa amostra se encontra nos indivíduos com ID de paciente igual a 108446, com cerca de 99 linhas. Os restantes apresentam valores entre 34 a 82 linhas.

	NUM_SEQUENCIAL	Contagem	Percentagem %
0	108446	99	0.062977
1	2808	82	0.052163
2	25118	64	0.040713
3	9948	47	0.029898
4	336072	45	0.028626
5	317270	40	0.025445
6	74229	38	0.024173
7	501236	36	0.022901
8	53819	35	0.022265
9	179531	34	0.021629

Figura 29 – Análise à frequência de valores da coluna "NUM_SEQUENCIAL"

- **DATAHORA_ADM –** A imagem a seguir apresentada (Figura 30), mostra a contagem efetuada sobre os valores (mostrando apenas o top 10). Podemos verificar que a moda da nossa amostra se encontra nos indivíduos com data de admissão igual a 2020-09-04 18:14:18, com cerca de 6 linhas. Os restantes possuem 4 linhas cada um.

	DATAHORA_ADM	Contagem	Percentagem %
0	2020-09-04 18:14:48	6	0.003817
1	2020-01-04 19:46:16	4	0.002545
2	2020-05-31 09:30:41	4	0.002545
3	2020-02-12 08:16:59	4	0.002545
4	2020-04-25 10:56:33	4	0.002545
5	2020-09-04 23:43:50	4	0.002545
6	2020-09-29 14:36:53	4	0.002545
7	2020-01-02 10:43:40	4	0.002545
8	2020-06-28 18:02:28	4	0.002545
9	2020-01-08 09:17:17	4	0.002545

Figura 30 – Análise à frequência de valores da coluna "DATAHORA_ADM"

- **DATAHORA_ALTA –** A figura a seguir apresentada (Figura 31), mostra a contagem efetuada sobre os valores (mostrando apenas o top 10). Podemos verificar que a moda da nossa amostra se encontra nos indivíduos com data de alta igual a 2020-12-31 14:10:00 22:01:00, com cerca de 12 linhas. Os restantes apresentam valores entre 9 a 11 linhas.

	DATAHORA_ALTA	Contagem	Percentagem %
0	2020-12-31 14:10:00	12	0.007634
1	2020-12-19 01:07:00	11	0.006997
2	2020-11-12 21:01:00	10	0.006361
3	2020-12-30 00:23:00	10	0.006361
4	2021-01-01 16:51:00	10	0.006361
5	2020-12-15 00:07:00	10	0.006361
6	2021-04-26 22:01:00	10	0.006361
7	2020-12-31 20:46:00	10	0.006361
8	2020-10-28 13:14:00	10	0.006361
9	2021-06-14 20:23:00	9	0.005725

Figura 31 – Análise à frequência de valores da coluna "DATAHORA_ALTA"

- **COD_LOCAL –** A figura a seguir apresentada (Figura 32), mostra a contagem efetuada sobre os valores (mostrando apenas o top 10). Podemos verificar que a moda da nossa amostra se encontra nos indivíduos com código local igual a 1, com cerca de 120921 linhas. Os restantes apresentam valores entre 12702 a 23576 linhas.

	COD_LOCAL	Contagem	Percentagem %
0	1	120921	76.922245
1	2	23576	14.997551
2	3	12702	8.080204

Figura 32 – Análise à frequência de valores da coluna "COD_LOCAL"

- **COD_DIAG_ALTA –** A figura a seguir apresentada (Figura 33), mostra a contagem efetuada sobre os valores (mostrando apenas o top 10). Podemos verificar que a moda da nossa amostra se encontra nos indivíduos com código diagnóstico igual a 78099, com cerca de 3125 linhas. Os restantes apresentam valores entre 1048 a 2749 linhas.

	COD_DIAG_ALTA	Contagem	Percentagem %
0	78099	3125	3.193958
1	M2480	2749	2.809661
2	U072	2575	2.631821
3	R6889	2331	2.382437
4	9249	2150	2.197443
5	8299	1851	1.891845
6	U071	1468	1.500393
7	07999	1241	1.268384
8	7806	1113	1.137560
9	7242	1048	1.071126

Figura 33 – Análise à frequência de valores da coluna "COD_DIAG_ALTA"

- **DES_DIAGNOSTICO –** A figura a seguir apresentada (Figura 34), mostra a contagem efetuada sobre os valores (mostrando apenas o top 10). Podemos verificar que a moda da nossa amostra se encontra nos indivíduos com designação de diagnóstico igual a "OUTROS SINTOMAS GERAIS", com cerca de 3125 linhas. Os restantes apresentam valores entre 1048 a 2749 linhas.

	DES_DIAGNOSTICO	Contagem	Percentagem %
0	OUTROS SINTOMAS GERAIS	3125	3.195297
1	O perturb articulares específicas, artic NE, NCOP	2749	2.810838
2	COVID-19, confirmação laboratorial inconclusiv...	2575	2.632924
3	O sintomas e sinais gerais	2331	2.383436
4	CONTUSAO DE LOCAL NAO ESPECIFICADO	2150	2.198364
5	FRACTURAS DE OSSOS SOE	1851	1.892638
6	COVID-19	1424	1.456033
7	INFECCAO VIRICA NAO ESPECIFICADA	1241	1.268916
8	FEBRE	1113	1.138037
9	LUMBAGO	1048	1.071575

Figura 34 – Análise à frequência de valores da coluna "DES_DIAGOSTICO"

- **Infeções**

Elaborado a análise geral ao *dataset* infeções, podemos concluir que todas as colunas estão completamente preenchidas, com um total de 5385 linhas, como é possível verificar na figura seguinte (Figura 35).

```
RangeIndex: 5385 entries, 0 to 5384
Data columns (total 3 columns):
 #   Column         Non-Null Count  Dtype
---  ------         --------------  -----
 0   ID Registo     5385 non-null   object
 1   Data Admissão  5385 non-null   datetime64[ns]
 2   Tem infeção    5385 non-null   object
```

Figura 35 – Análise geral das colunas do dataset Infeções

Para além disso, foi realizado uma análise detalhada para cada coluna do dataset, como é possível visualizar a seguir:

- **ID Registo –** A figura a seguir apresentada (Figura 36), mostra a contagem efetuada sobre os valores (mostrando apenas o top 10). Podemos verificar que a moda da nossa amostra se encontra nos indivíduos com ID Registo igual a 25010776, com cerca de 5 linhas. Os restantes apresentam valores entre 3 a 4 linhas.

	ID Registo	Contagem	Percentagem %
0	25010776	5	0.092851
1	28014803	4	0.074280
2	29	3	0.055710
3	27	3	0.055710
4	99	3	0.055710
5	95024756	3	0.055710
6	33	3	0.055710
7	12	3	0.055710
8	14	3	0.055710
9	7	3	0.055710

Figura 36 – Análise à frequência de valores da coluna "ID Registo'

- **Data Admissão –** A figura a seguir apresentada (Figura 37), mostra a contagem efetuada sobre os valores (mostrando apenas o top 10). Podemos verificar que a moda da nossa amostra se encontra nos indivíduos com data de admissão igual a 2020-03-08,com cerca de 20 linhas. Os restantes apresentam valores entre 15 a 16 linhas.

	Data Admissão	Contagem	Percentagem %
0	2020-03-08	20	0.371402
1	2020-01-03	16	0.297122
2	2020-01-15	16	0.297122
3	2018-06-26	16	0.297122
4	2019-05-03	16	0.297122
5	2018-09-14	16	0.297122
6	2018-11-23	16	0.297122
7	2019-11-22	15	0.278552
8	2020-01-17	15	0.278552
9	2019-05-10	15	0.278552

Figura 37 – Análise à frequência de valores da coluna "Data Admissão"

- **Tem infeção –** A figura a seguir apresentada (Figura 38), mostra a contagem efetuada sobre os valores (mostrando apenas o top 10). Podemos verificar que a moda da nossa amostra se encontra nos indivíduos com infeção igual a "Não", com cerca de 5097 linhas. O outro valor, valor "Sim", apresenta cerca de 8541 linhas.

	Tem infeção	Contagem	Percentagem %
0	Não	5097	94.651811
1	Sim	288	5.348189

Figura 38 – Análise à frequência de valores da coluna "Tem infeção"

Ao fim da exploração de dados, foi necessário realizar uma verificação da qualidade dos mesmos, de modo a identificar os principais problemas que os dados possuem. Dos 5 *datasets* recolhidos, todos apresentavam alguns problemas, sendo eles:

- **Cirurgias**

Tabela 9 – Verificação da qualidade dos dados do dataset cirurgias

Coluna	Erros	Soluções
DTA_NASC	Esta coluna está identificada como tipo *String* e possui o nome incorreto.	Esta coluna será convertida para *Datetime* e renomeada para "DTA_NASCIMENTO".
SEXO	Esta coluna está identificada como tipo *Integer*.	Esta coluna será convertida para *Categorical*.
DATAMOV	Esta coluna está identificada como tipo *String*.	Esta coluna será convertida para *Datetime*.

- **Antibióticos**

Tabela 10 – Verificação da qualidade dos dados do dataset antibióticos

Coluna	Erros	Soluções
DATA_INICIO	Esta coluna está identificada como tipo *String*.	Esta coluna será convertida para *Datetime*.
DATA_FIM	Esta coluna está identificada como tipo *String* e apresenta 1984 linhas vazias.	Esta coluna será convertida para *Datetime* e as linhas vazias serão eliminadas.
MED_DESIGNACAO	Esta coluna está identificada como tipo *String*.	Esta coluna será convertida como para *Categorical*.
ART_DESGINACAO	Esta coluna está identificada como tipo *String*.	Esta coluna será convertida como para *Categorical*.
MODULO	Esta coluna apresenta todas as linhas vazias.	Esta coluna será descartada da amostra de dados.
DATA_CIRURGIA	Esta coluna apresenta todas as linhas vazias.	Esta coluna será descartada da amostra de dados.

- **Internamentos**

Tabela 11 – Verificação da qualidade dos dados do dataset internamentos

Coluna	Erros	Soluções
ADMISSAO	Esta coluna está identificada como tipo *String*.	Esta coluna será convertida para *Datetime*.
ALTA	Esta coluna está identificada como tipo *String*.	Esta coluna será convertida para *Datetime*.
DTA_NASCIMENTO	Esta coluna está identificada como tipo *String* e apresenta 2 linhas vazias.	Esta coluna será convertida para *Datetime* e as linhas vazias serão eliminadas.
SEXO	Esta coluna está identificada como tipo *Float* e apresenta 2 linhas vazias.	Esta coluna será convertida para *Categorical* e a linhas vazias serão eliminadas.

- **Urgências**

Tabela 12 – Verificação da qualidade dos dados do dataset urgências

Coluna	Erros	Soluções
COD_DIAG_ALTA	Esta coluna está identificada como tipo *String* e apresenta 59358 linhas vazias.	Esta coluna será convertida para *Categorical* e as linhas vazias serão preenchidas aleatoriamente a partir de uma lista de valores
DES_DIAGNOSTICO	Esta coluna apresenta 59399 linhas vazias.	Esta coluna será descartada da amostra de dados.

- **Infeções**

Tabela 13 – Verificação da qualidade dos dados do dataset infeções

Coluna	Erros	Soluções
ID Registo	Esta coluna está identificada como tipo *String,* com nome incorreto e possuía um "~" numa linha.	Esta coluna será convertida para *Integer*, renomeada para "NUM_PROCESSO" e o "~" será retirado.
Data Admissão	Esta coluna está identificada com nome incorreto.	Esta coluna será renomeada para "DATA_ADMISSAO".
Tem infeção	Esta coluna está identificada com o nome incorreto e apresenta valores em *String*.	Esta coluna será renomeada para "INFECAO", os valores "SIM" e "NÃO" serão convertidos para 1 e 0, respetivamente, e o seu formato sera convertido para *Integer*.

5.4 Preparação dos Dados

Para poder avançar e realizar a próxima fase de modelação de forma eficiente e eficaz, foi necessário executar ações de melhoria, como seleção, limpeza, correção e construção de novos dados.

Primeiramente foi necessário realizar alterações aos *datasets* inicialmente recolhidos, como descrito em seguida:

- **Cirurgias –** Nesta amostra de dados o formato das colunas "DATAMOV", "DTA_NASC" e "SEXO" não se encontravam com o formato correto, desse modo, foi necessário realizar a transformação do seu tipo. Para as colunas "DATAMOV" e "DTA_NASC", converteu-se de *String* para *Datetime*, colocando estas colunas com o novo formato de Ano-Mês-Dia Hora-Minuto-Segundo. Para a coluna "SEXO", converteu-se de *Integer* para *Categorical*. Além disso, a coluna "DTA_NASC" foi renomeada para "DTA_NASCIMENTO", de modo a coincidir com o nome da coluna do *dataset* internamentos.

- **Antibióticos –** Neste *dataset* o formato das colunas "DATA_INICIO", "DATA_FIM", "MED_DESIGNACAO" e "ART_DESIGNACAO" não se encontravam com o formato correto, desse modo foi necessário realizar a transformação do seu tipo. Nas colunas "DATA_INICIO", "DATA_FIM" converteu-se de *String* para *Datetime*, colocando estas colunas com o novo formato de Ano-Mês-Dia Hora-Minuto-Segundo, e nas colunas "MED_DESIGNACAO" e "ART_DESIGNACAO" converteu-se de *String* para *Categorical*. Além disso, na coluna "DATA_FIM" existiam 1984 linhas vazias e optou-se pela sua eliminação, uma vez que não havia justificação para o seu preenchimento e esse número de linhas não iam acrescentar nenhum valor ao modelo. As colunas "MODULO" e "DATA_CIRURGIA" não apresentavam qualquer tipo de registo, assim sendo, foram excluídas da amostra de dados.
- **Internamentos –** Nesta amostra de dados o formato das colunas "ADMISSAO", "ALTA", "DTA_NASCIMENTO" e "SEXO" não se encontravam com o formato correto, desse modo foi necessário realizar a transformação do seu tipo. Para as colunas "ADMISSAO", "ALTA" e "DTA_NASCIMENTO", converteu-se de *String* para *Datetime*, colocando estas colunas com o formato de Ano-Mês-Dia Hora-Minuto-Segundo. Para a coluna "SEXO", converteu-se de *Float* para *Categorical*. Além disso, as colunas "DTA_NASCIMENTO" e "SEXO" apresentavam 2 linhas vazias e optou-se pela sua eliminação em ambas as colunas.
- **Urgências –** Neste *dataset* as colunas "URG_EPISÓDIO", "COD_LOCAL" e "DES_DIAGNOSTICO" foram eliminadas, uma vez que não possuíam informação essencial para a conceção dos modelos preditivos. Para além disso, a coluna "COD_DIAG_ALTA" não se encontrava com o formato correto e possuía 59358 linhas vazias. Desse modo, converteu-se o seu formato de *String* para *Categorical* e optou-se pelo preenchimento aleatório das linhas vazias. As linhas foram preenchidas a partir de uma lista de valores. Essa lista possuía valores provenientes da própria coluna, a "COD_DIAG_ALTA".

- **Infeções –** Nesta amostra de dados a coluna "ID Registo" possuía um "~" num valor e esse "~" foi eliminado da linha. A coluna "ID Registo" não se encontrava com o formato correto, desse modo, foi necessário realizar a transformação do seu tipo, convertendo de *String* para *Integer*. A coluna "Tem infeção" possuía os seus valores em *String*, desse modo, optou-se por converter os seus valores "SIM" e "NÃO", por 1 e 0, respetivamente, e consequentemente converteu-se o seu formato, de *String* para *Integer*. As colunas "ID Registo", "Data Admissão" e "Tem infeção" foram renomeadas para "NUM_PROCESSO", "DATA_ADMISSAO" e "INFECAO", respetivamente.

Em segundo lugar, realizou-se um conjunto de *merges* entre os datasets Cirurgias, Antibióticos, Internamentos e Urgências. Estes *merges* foram efetuados, uma vez que é fundamental ter uma amostra de dados que envolva todos os parâmetros que possibilitam a identificação de infeção nosocomial. Desse modo, os *merges* foram concretizados com o objetivo de satisfazer todas as regras e procedimentos para identificação de infeção nosocomial e com o intuito de obter a melhor amostra de dados possível, isto é, uma amostra de dados com o máximo de informação relativa a doentes internados no hospital, ou seja, informação do *dataset* internamentos, e com o maior número de linhas preenchidas. O primeiro *merge* efetuado foi entre os *datasets* Internamentos e Antibióticos, e foi realizado sobre a coluna "NUM_SEQUENCIAL" e com tipo *left join*, dando assim, preferência aos dados do *dataset* Internamentos. O segundo *merge* foi efetuado entre o *dataset* resultante do primeiro *merge* e o *dataset* Cirurgias, e foi executado também sobre a coluna "NUM_SEQUENCIAL" e com tipo *left join*, continuando assim, a dar prioridade aos dados relativos a Internamentos. O terceiro e último *merge*, foi efetuado entre o *dataset* resultante do segundo *merge* e o *dataset* Urgências, e foi realizado também sobre a coluna "NUM_SEQUENCIAL" e com tipo *left join*, continuando assim, a dar prioridade aos dados relativos a Internamentos. Estes *merges* originaram um *dataset*, chamado Infeção_Nosocomial, com todas as colunas dos 4 *datasets* (Cirurgias, Antibióticos, Internamentos e Urgências), ou seja, possui o máximo de informação relativo a pacientes internados, e consequentemente, se esses doentes foram administrados com antibióticos, se foram submetidos a cirurgias e se foram às urgências. Este *dataset* é constituído por 14 campos e 277586 registos, como é possível verificar na figura seguinte (Figura 39).

```
Int64Index: 277586 entries, 0 to 277585
Data columns (total 14 columns):
 #   Column          Non-Null Count   Dtype
---  ------          --------------   -----
 0   NUM_SEQUENCIAL  277586 non-null  int64
 1   ADMISSAO        277586 non-null  datetime64[ns]
 2   ALTA            277586 non-null  datetime64[ns]
 3   DTA_NASCIMENTO  277586 non-null  datetime64[ns]
 4   SEXO            277586 non-null  category
 5   CIRURGIA        277586 non-null  int64
 6   DATA_INICIO     249013 non-null  datetime64[ns]
 7   DATA_FIM        249013 non-null  datetime64[ns]
 8   MED_DESIGNACAO  249013 non-null  category
 9   ART_DESIGNACAO  249013 non-null  category
 10  DATAMOV         168845 non-null  datetime64[ns]
 11  DATAHORA_ADM    271987 non-null  datetime64[ns]
 12  DATAHORA_ALTA   271987 non-null  datetime64[ns]
 13  COD_DIAG_ALTA   271987 non-null  category
```

Figura 39 – Dataset Infeção_Nosocomial após os 3 merges

Em terceiro lugar, foi necessário realizar a construção de novos dados para avaliar o risco de infeção, isto é, criou-se 6 novas colunas para representarem o risco de infeção do paciente. Estas colunas foram criadas a partir de regras e procedimentos, fornecidos pelo Hospital, para a identificação de infeção nosocomial. As regras utilizadas foram:

1. Se houve reinternamento no mês seguinte à cirurgia e se ultrapassa o tempo medio de internamento – 7 dias - Reinternamentos: definir 32 dias após CRG;
2. Se houve idas às urgências no mês seguinte à cirurgia;
3. Se houve prescrição de antibiótico no mês seguinte à cirurgia;
4. Se esteve internado nos últimos 6 meses;
5. Se ao fim de 72h em internamento inicia antibiótico;
6. Se houve um reinternamento no espaço de 72h.

A primeira coluna criada está relacionada com a 3º regra e foi necessário conceber uma função, com a utilização das colunas “DATA_INICIO” e “DATAMOV”, que verificasse se o paciente começou a ser administrado com antibiótico no mês seguinte à cirurgia e se começou a tomar passado 1 dia da cirurgia. Se ambas as condições se verificarem como verdadeiras, o resultado é 1 (risco de infeção), se não, o resultado é 0 (sem risco de infeção). Esta função originou a criação da coluna “INFECAO_CIRANTI”, com valores binários de “0” ou “1”.

A segunda coluna criada está relacionada com a 5º regra e foi necessário conceber uma função, com a utilização das colunas “DATA_INICIO” e “ADMISSAO”, que verificasse se o paciente começou a

ser administrado com antibiótico ao fim de 72h em internamento. Se esta condição se verificasse como verdadeira, o resultado é 1 (com risco de infeção), se não, o resultado é 0 (sem risco de infeção). Esta função originou a criação da coluna "INFECAO_INTERANTI", com valores binários de "0" ou "1".

A terceira coluna criada está relacionada com a 1º regra e foi necessário conceber uma função, com a utilização das colunas "ADMISSAO" e "DATAMOV", que verificasse se o paciente foi internado no mês seguinte à cirurgia e se ultrapassava o tempo médio de internamento (7 dias de tempo médio). Se ambas as condições se verificassem como verdadeiras, o resultado é 1 (com risco de infeção), se não, o resultado é 0 (sem risco de infeção). Esta função originou a criação da coluna "INFECAO_CIRINTER", com valores binários de "0" ou "1".

A quarta coluna criada está relacionada com a 4º regra e foi necessário conceber uma função, com a utilização das colunas "ADMISSÃO" e "NUM_SEQUENCIAL", que verificasse se o paciente esteve internado nos últimos 6 meses. Se a condição se verificasse como verdadeira, o resultado é 1 (com risco de infeção), se não, o resultado é 0 (sem risco de infeção). Esta função originou a criação da coluna "INFECAO_INTERINFEC", com valores binários de "0" ou "1".

A quinta coluna criada está relacionada com a 6º regra e foi necessário conceber uma função, com a utilização das colunas "ADMISSAO", "ALTA" e "NUM_SEQUENCIAL", que verificasse se o paciente foi reinternado num espaço de 72h. Se a condição se verificasse como verdadeira, o resultado é 1 (com risco de infeção), se não, o resultado é 0 (sem risco de infeção). Esta função originou a criação da coluna "INFECAO_REINTER", com valores binários de "0" ou "1".

A sexta e última coluna criada está relacionada com a 2º regra e foi necessário conceber uma função, com a utilização das colunas "DATAHORA_ADM" e "DATAMOV", que verificasse se o paciente foi às urgências no mês seguinte à cirurgia. Se a condição se verificasse como verdadeira, o resultado é 1 (com risco de infeção), se não, o resultado é 0 (sem risco de infeção). Esta função originou a criação da coluna "INFECAO_CIURG", com valores binários de "0" ou "1".

```
Int64Index: 277586 entries, 0 to 277585
Data columns (total 20 columns):
 #   Column              Non-Null Count   Dtype
---  ------              --------------   -----
 0   NUM_SEQUENCIAL      277586 non-null  int64
 1   ADMISSAO            277586 non-null  datetime64[ns]
 2   ALTA                277586 non-null  datetime64[ns]
 3   DTA_NASCIMENTO      277586 non-null  datetime64[ns]
 4   SEXO                277586 non-null  category
 5   CIRURGIA            277586 non-null  int64
 6   DATA_INICIO         249013 non-null  datetime64[ns]
 7   DATA_FIM            249013 non-null  datetime64[ns]
 8   MED_DESIGNACAO      249013 non-null  category
 9   ART_DESIGNACAO      249013 non-null  category
 10  DATAMOV             168845 non-null  datetime64[ns]
 11  DATAHORA_ADM        271987 non-null  datetime64[ns]
 12  DATAHORA_ALTA       271987 non-null  datetime64[ns]
 13  COD_DIAG_ALTA       271987 non-null  category
 14  INFECAO_CIRANTI     277586 non-null  int32
 15  INFECAO_INTERANTI   277586 non-null  int32
 16  INFECAO_CIRINTER    277586 non-null  int32
 17  INFECAO_INTERINFEC  277586 non-null  int64
 18  INFECAO_REINTER     277586 non-null  int64
 19  INFECAO_CIURG       277586 non-null  int32
```

Figura 40 – Dataset Infeção_Nosocomial após a criação das 6 colunas de risco de infeção

Em quarto lugar, foi também necessário a construção de novos dados, no entanto, dados que ajudassem a perceber o historial hospitalar do paciente. Assim sendo, criaram-se 2 novas colunas, uma respetiva ao número de internamentos do paciente e outra respetiva ao número de entradas nas urgências do paciente. Além disso, criou-se uma coluna referente à idade do doente.

A primeira coluna criada é referente ao número de internamentos e foi concebida contando o número de datas de admissão no internamento diferentes (coluna "ADMISSAO") por ID de paciente (coluna "NUM_SEQUENCIAL"). Esta função originou a coluna "N_INTERNAMENTOS", constituída por valores inteiros.

A segunda coluna criada é referente ao número de entradas nas urgências e foi concebida contando o número de datas de admissão nas urgências diferentes (coluna "DATAHORA_ADM") por ID de paciente (coluna "NUM_SEQUENCIAL"). Esta função originou a coluna "N_URGENCIAS", constituída por valores inteiros.

A terceira coluna criada é referente à idade do paciente e foi concebida calculando a diferença entre o ano atual e o ano presente na coluna "DTA_NASCIMENTO". Esta função originou a coluna "IDADE", constituída por valores inteiros.

```
Int64Index: 277586 entries, 0 to 277585
Data columns (total 23 columns):
 #   Column              Non-Null Count   Dtype
---  ------              --------------   -----
 0   NUM_SEQUENCIAL      277586 non-null  int64
 1   ADMISSAO            277586 non-null  datetime64[ns]
 2   ALTA                277586 non-null  datetime64[ns]
 3   DTA_NASCIMENTO      277586 non-null  datetime64[ns]
 4   SEXO                277586 non-null  category
 5   CIRURGIA            277586 non-null  int64
 6   DATA_INICIO         249013 non-null  datetime64[ns]
 7   DATA_FIM            249013 non-null  datetime64[ns]
 8   MED_DESIGNACAO      249013 non-null  category
 9   ART_DESIGNACAO      249013 non-null  category
 10  DATAMOV             168845 non-null  datetime64[ns]
 11  DATAHORA_ADM        271987 non-null  datetime64[ns]
 12  DATAHORA_ALTA       271987 non-null  datetime64[ns]
 13  COD_DIAG_ALTA       271987 non-null  category
 14  INFECAO_CIRANTI     277586 non-null  int32
 15  INFECAO_INTERANTI   277586 non-null  int32
 16  INFECAO_CIRINTER    277586 non-null  int32
 17  INFECAO_INTERINFEC  277586 non-null  int64
 18  INFECAO_REINTER     277586 non-null  int64
 19  INFECAO_CIURG       277586 non-null  int32
 20  N_INTERNAMENTOS     277586 non-null  int64
 21  N_URGENCIAS         277586 non-null  int64
 22  IDADE               277586 non-null  int64
```

Figura 41 – Dataset Infeção_Nosocomial após a criação das 3 colunas do historial hospitalar

Em quinto lugar, foi necessário realizar um *merge* entre os *datasets* Infecao_Nosocomial e Infeções, de modo a conseguir concluir os pacientes que tiveram mesmo infeção ou não. No entanto, essas duas amostras de dados não possuíam nenhuma coluna em comum que possibilitasse a execução do *merge*. Desse modo, foi necessário realizar um mapeamento dos ID's (coluna "NUM_SEQUENCIAL") presentes no *dataset* Infecao_Nosocomial com o número de processo presentes na base de dados do hospital. Isto criou um *dataset* com as colunas "NUM_SEQUENCIAL" e "NUM_PROCESSO". A seguir, realizou-se um *inner merge* entre esse *dataset* e o *dataset* Infeções, sobre a coluna "NUM_PROCESSO", originando uma amostra de dados com todos os pacientes que tivessem correspondência no número de processo. Por fim, efetuou-se o *merge* entre os datasets Infecao_Nosocomial e o *dataset* criado anteriormente, sobre a coluna "NUM_SEQUENCIAL" e com tipo *inner join*. Este último *merge* originou um *dataset* com 27 campos e 44273 registos, como é possível verificar na figura seguinte (Figura 42).

```
Int64Index: 44273 entries, 0 to 44272
Data columns (total 26 columns):
 #   Column              Non-Null Count  Dtype
---  ------              --------------  -----
 0   NUM_SEQUENCIAL      44273 non-null  int64
 1   ADMISSAO            44273 non-null  datetime64[ns]
 2   ALTA                44273 non-null  datetime64[ns]
 3   DTA_NASCIMENTO      44273 non-null  datetime64[ns]
 4   SEXO                44273 non-null  category
 5   CIRURGIA            44273 non-null  int64
 6   DATA_INICIO         42580 non-null  datetime64[ns]
 7   DATA_FIM            42580 non-null  datetime64[ns]
 8   MED_DESIGNACAO      42580 non-null  category
 9   ART_DESIGNACAO      42580 non-null  category
 10  DATAMOV             41971 non-null  datetime64[ns]
 11  DATAHORA_ADM        43735 non-null  datetime64[ns]
 12  DATAHORA_ALTA       43735 non-null  datetime64[ns]
 13  COD_DIAG_ALTA       43735 non-null  category
 14  INFECAO_CIRANTI     44273 non-null  int32
 15  INFECAO_INTERANTI   44273 non-null  int32
 16  INFECAO_CIRINTER    44273 non-null  int32
 17  INFECAO_INTERINFEC  44273 non-null  int64
 18  INFECAO_REINTER     44273 non-null  int64
 19  INFECAO_CIURG       44273 non-null  int32
 20  N_INTERNAMENTOS     44273 non-null  int64
 21  N_URGENCIAS         44273 non-null  int64
 22  IDADE               44273 non-null  int64
 23  NUM_PROCESSO        44273 non-null  int32
 24  DATA_ADMISSAO       44273 non-null  datetime64[ns]
 25  INFECAO             44273 non-null  int32
```

Figura 42 – Dataset Infecao_Nosocomial após o merge

Em sexto lugar, para ter um *dataset* fiável e assertivo, foi necessário verificar se a data de admissão no departamento de infeção foi durante o tempo de internamento no hospital. Assim sendo, concebeu-se uma função, com a utilização das colunas "ADMISSAO" e "DATA_ADMISSAO", que verificasse se houve primeiro registo no internamento e só depois no departamento de infeção, e se foi durante o tempo de internamento no hospital. Se as condições se verificassem como verdadeiras, a linha era mantida no *dataset*, se não, a linha era eliminada. Com isto, o *dataset* Infecao_Nosocomial passou a ter 27 campos e 2512 registos, como é possível visualizar na figura seguinte (Figura 43).

```
RangeIndex: 2512 entries, 0 to 2511
Data columns (total 26 columns):
 #   Column              Non-Null Count  Dtype
---  ------              --------------  -----
 0   NUM_SEQUENCIAL      2512 non-null   int64
 1   ADMISSAO            2512 non-null   datetime64[ns]
 2   ALTA                2512 non-null   datetime64[ns]
 3   DTA_NASCIMENTO      2512 non-null   datetime64[ns]
 4   SEXO                2512 non-null   category
 5   CIRURGIA            2512 non-null   int64
 6   DATA_INICIO         2257 non-null   datetime64[ns]
 7   DATA_FIM            2257 non-null   datetime64[ns]
 8   MED_DESIGNACAO      2257 non-null   category
 9   ART_DESIGNACAO      2257 non-null   category
 10  DATAMOV             2512 non-null   datetime64[ns]
 11  DATAHORA_ADM        2460 non-null   datetime64[ns]
 12  DATAHORA_ALTA       2460 non-null   datetime64[ns]
 13  COD_DIAG_ALTA       2460 non-null   category
 14  INFECAO_CIRANTI     2512 non-null   int32
 15  INFECAO_INTERANTI   2512 non-null   int32
 16  INFECAO_CIRINTER    2512 non-null   int32
 17  INFECAO_INTERINFEC  2512 non-null   int64
 18  INFECAO_REINTER     2512 non-null   int64
 19  INFECAO_CIURG       2512 non-null   int32
 20  N_INTERNAMENTOS     2512 non-null   int64
 21  N_URGENCIAS         2512 non-null   int64
 22  IDADE               2512 non-null   int64
 23  NUM_PROCESSO        2512 non-null   int32
 24  DATA_ADMISSAO       2512 non-null   datetime64[ns]
 25  INFECAO             2512 non-null   int32
```

Figura 43 – Dataset Infecao_Nosocomial após a limpeza de dados

Em sétimo e último lugar, foi necessário realizar uma última verificação ao *dataset* criado (*dataset* Infecao_Nosocomial) para poder prosseguir para a fase de modelação. Desse modo, efetuou-se de novo atividades de processamento de dados, sendo neste caso, a limpeza, transformação e construção de novos dados.

A coluna "ART_DESIGNACAO" foi eliminada, porque não acrescentava informação adicional à informação que a coluna "MED_DESIGNAÇÃO" possui. As colunas "NUM_SEQUENCIAL" e "NUM_PROCESSO" foram eliminadas, uma vez que são colunas constituídas por ID's, logo não iriam acrescentar informação essencial na conceção dos modelos. Todas as colunas referentes a datas, ou seja, as colunas "ADMISSAO", "ALTA", "DTA_NASCIMENTO", "DATA_INICIO", "DATA_FIM", "DATAMOV", "DATAHORA_ADM", "DATAHORA_ALTA" e "DATA_ADMISSAO" foram eliminadas, dado que não possuem informação fundamental, e desso modo, não iriam trazer nenhum benefício para o desenvolvimento de modelos preditivos. As colunas "INFECAO_INTERANTI" e "INFECAO_CIRINTER" foram também eliminadas, uma vez que todas as linhas passaram a ser apenas constituídas pelo valor

"0", logo não iria ter peso nenhum sobre a coluna *target*. As colunas "MED_DESIGNACAO" e "COD_DIAG_ALTA" apresentavam um conjunto de linhas vazias, e optou-se por preenchê-las pelo valor "Not Available".

A coluna "COD_DIAG_ALTA" era constituída por códigos ICD-9 e ICD-10, que possuem um elevado número de valores diferentes. O código ICD, *International Classification of Diseases*, foi desenvolvido pela OMS (Organização Mundial de Saúde) e em Portugal é utilizado no SNS (Sistema Nacional de Saúde), isto é, em todos os hospitais públicos (J. V. Santos et al., 2021). Neste momento o código utilizado é o ICD-10, no entanto, os dados nesta coluna são constituídos pelos dois tipos de código, uma vez que a amostra de dados se encontra entre os anos de 2018 a 2021. Desse modo, foi necessário normalizar a coluna e optou-se por apenas ser constituída pelo código ICD-9 e com os valores mais altos da hierarquia. Assim sendo, foi efetuada uma correspondência entre os códigos ICD-9 e ICD-10, prevalecendo no final os de ICD-9, e no fim, substituímos todos os valores pelos valores mais altos na hierarquia. Isto originou a criação da coluna "COD_IDC", toda constituída por valores de ICD-9, e a eliminação da coluna "COD_DIAG_ALTA". Esta nova coluna possuía algumas linhas vazias, que foram depois preenchidas pelo valor "Not Available".

As colunas categóricas, isto é, as colunas "SEXO", "MED_DESGINACAO" e "COD_IDC" são colunas que representam categorias, como por exemplo, género e tipo de medicamento. Nestas 3 colunas efetuou-se o método de *One-Hot Encoding*, em que cada atributo é mapeado em *L* entradas binárias, relacionando-as com o número de atributos existentes (Matos et al., 2019). Este método originou a criação de várias colunas, e consequentemente, a eliminação das colunas "SEXO", "MED_DESIGNACAO" e "COD_IDC".

Adicionalmente, também foi necessário realizar um *oversampling* ao *dataset*, uma vez que a variável *target* não se encontrava balanceada. Esta técnica consiste em ajustar a distribuição de classes de um certo conjunto de dados, isto é, tem como objetivo igualar o número de *tuples* positivos e negativos (Han et al., 2012).

Posto isto, o *dataset* Infecao_Nosocomial ficou constituído por 59 campos e 2528 registos, como é possível verificar na figura a seguir (Figura 44).

```
RangeIndex: 2528 entries, 0 to 2527
Data columns (total 59 columns):
 #   Column                                Non-Null Count  Dtype
---  ------                                --------------  -----
 0   CIRURGIA                              2528 non-null   int64
 1   INFECAO_CIRANTI                       2528 non-null   int32
 2   INFECAO_INTERINFEC                    2528 non-null   int32
 3   INFECAO_REINTER                       2528 non-null   int32
 4   INFECAO_CIURG                         2528 non-null   int32
 5   N_INTERNAMENTOS                       2528 non-null   int64
 6   N_URGENCIAS                           2528 non-null   int64
 7   IDADE                                 2528 non-null   int64
 8   MED_Amicacina                         2528 non-null   int32
 9   MED_Amoxicilina                       2528 non-null   int32
 10  MED_Amoxicilina + Ácido clavulânico   2528 non-null   int32
 11  MED_Ampicilina                        2528 non-null   int32
 12  MED_Azitromicina                      2528 non-null   int32
 13  MED_CIPROfloxacina                    2528 non-null   int32
 14  MED_Cefradina                         2528 non-null   int32
 15  MED_Cefuroxima                        2528 non-null   int32
 16  MED_Clindamicina                      2528 non-null   int32
 17  MED_Cloranfenicol                     2528 non-null   int32
 18  MED_Cotrimoxazol                      2528 non-null   int32
 19  MED_Eritromicina                      2528 non-null   int32
 20  MED_Ertapenem                         2528 non-null   int32
 21  MED_Fosfomicina                       2528 non-null   int32
 22  MED_LEVOfloxacina                     2528 non-null   int32
 23  MED_Linezolida                        2528 non-null   int32
 24  MED_Meropenem                         2528 non-null   int32
 25  MED_Mupirocina                        2528 non-null   int32
 26  MED_Neomicina                         2528 non-null   int32
 27  MED_Not Available                     2528 non-null   int32
 28  MED_Piperacilina + Tazobactam         2528 non-null   int32
 29  MED_Teicoplanina                      2528 non-null   int32
 30  MED_Vancomicina                       2528 non-null   int32
 31  MED_cefEPIMA                          2528 non-null   int32
 32  MED_cefOXITINA                        2528 non-null   int32
 33  MED_cefTAZIDIMA                       2528 non-null   int32
 34  MED_cefTRIAXONA                       2528 non-null   int32
 35  MED_cefaZOLINA                        2528 non-null   int32
 36  MED_geNTAMICina                       2528 non-null   int32
 37  MED_metRONIDazol                      2528 non-null   int32
 38  COD_1                                 2528 non-null   int32
 39  COD_10                                2528 non-null   int32
 40  COD_11                                2528 non-null   int32
 41  COD_12                                2528 non-null   int32
 42  COD_13                                2528 non-null   int32
 43  COD_14                                2528 non-null   int32
 44  COD_15                                2528 non-null   int32
 45  COD_16                                2528 non-null   int32
 46  COD_17                                2528 non-null   int32
 47  COD_18                                2528 non-null   int32
 48  COD_2                                 2528 non-null   int32
 49  COD_3                                 2528 non-null   int32
 50  COD_5                                 2528 non-null   int32
 51  COD_6                                 2528 non-null   int32
 52  COD_7                                 2528 non-null   int32
 53  COD_8                                 2528 non-null   int32
 54  COD_9                                 2528 non-null   int32
 55  COD_Not Available                     2528 non-null   int32
 56  SEXO_1                                2528 non-null   int32
 57  SEXO_2                                2528 non-null   int32
 58  INFECAO                               2528 non-null   int32
```

Figura 44 - Dataset final Infecao_Nosocomial

5.5 Modelação

Após terminado a fase de preparação de dados, avançou-se para a fase de modelação. Esta fase foi efetuada utilizando a abordagem de Classificação, com o objetivo de resolver problemas de Previsão, sendo neste caso, a previsão de infeção nosocomial (coluna "INFECAO"). Nesta abordagem, as colunas utilizadas foram todas as presentes no *dataset* final de Infecao_Nosocomial e foi necessário dividi-las em dois tipos de variáveis, as *feature columns* e a *target*. As *feature columns* são todas as colunas usadas para o treino do modelo e a *target* corresponde à coluna que queremos prever. Na figura seguinte (Figura 45) é possível visualizar a divisão efetuada.

```
X = infec_hospitalar.drop('INFECAO',axis=1)
y = infec_hospitalar['INFECAO']
```

Figura 45 – Criação das features columns e target

As técnicas de Classificação empregues foram, Árvore de Decisão, *Random Forest*, Redes Neuronais, *Naive Bayes*, *Support Vector Machine* e Regressão Logística.

```
from sklearn.tree import DecisionTreeClassifier
from sklearn.ensemble import RandomForestClassifier
from sklearn.neural_network import MLPClassifier
from sklearn.naive_bayes  import GaussianNB
from sklearn.svm import SVC
from sklearn.linear_model import LogisticRegression

cv_tree = DecisionTreeClassifier(max_depth=400, max_leaf_nodes=5000)
cv_rfc = RandomForestClassifier(n_estimators=200, max_depth=400, max_leaf_nodes=5000)
cv_mlp = MLPClassifier(hidden_layer_sizes=(100,100), max_iter=300)
cv_nb = GaussianNB()
cv_SVM = SVC(C=10, kernel='rbf')
cv_logreg = LogisticRegression(C=10, solver='newton-cg', max_iter=500)
```

Figura 46 - Modelos e os Parâmetros utilizados

Além disso, foi utilizado a técnica de amostragem *10-folds Cross-Validation* (10-*folds* CV) em todas as práticas de Classificação. A técnica normalmente utilizada é a de *train/test split*, que consiste na divisão dos dados em dois subconjuntos, um de treino, que possui um maior volume de dados, e um de teste, que é utilizado para estimar a *target* escolhida, e avaliar o desempenho e qualidade dos modelos. No entanto, optou-se por utilizar o método 10-*folds* CV, uma vez que é uma prática que apresenta bons resultados em relação às estimativas de erro de um modelo, tornando assim, o modelo mais fiável.

```
from sklearn.model_selection import KFold
from sklearn.model_selection import cross_val_predict

cv = KFold(n_splits=10, random_state=1, shuffle=True)

cv_tree_pred = cross_val_predict(cv_tree,X,y,cv=cv)
cv_rfc_pred = cross_val_predict(cv_rfc,X,y,cv=cv)
cv_mlp_pred = cross_val_predict(cv_mlp,X,y,cv=cv)
cv_nb_pred = cross_val_predict(cv_nb,X,y,cv=cv)
cv_SVM_pred = cross_val_predict(cv_SVM,X,y,cv=cv)
cv_logreg_pred = cross_val_predict(cv_logreg,X,y,cv=cv)
```

Figura 47 – 10-folds CV e avaliação dos modelos criados

5.6 Avaliação

Finalizado a fase de modelação, avançou-se para a fase de avaliação, onde é realizado a avaliação de desempenho dos modelos de DM criados. Para avaliar os modelos preditivos, utilizou-se a Matriz de Confusão, que permitiu a definição de sete métricas, sendo elas, Acuidade, Sensibilidade, Especificidade, Precisão, *F1-Score*, Índice Kappa e Curva AUC.

A seguir é possível visualizar os resultados das Matrizes de Confusão de cada modelo e uma tabela (Tabela 17) com todos os valores obtidos para as diferentes métricas e modelos.

Tabela 14 – Matriz Confusão para a técnica Árvore de Decisão

	1	**0**
1	1256	8
0	8	1256

Tabela 15 – Matriz Confusão para a técnica Random Forest

	1	**0**
1	1249	15
0	1	1263

Tabela 16 – Matriz Confusão para a técnica Redes Neuronais

	1	**0**
1	1249	15
0	13	1251

Tabela 17 – Matriz Confusão para a técnica Naive Bayes

	1	**0**
1	1236	28
0	691	573

Tabela 18 – Matriz Confusão para a técnica SVM

	1	**0**
1	1202	62
0	126	1138

Tabela 19 – Matriz Confusão para a técnica Regressão Logística

	1	**0**
1	1157	107
0	112	1152

Tabela 20 – Resultados das métricas nos diferentes modelos

Modelos	Acuidade	Sensibilidade	Especificidade	Precisão	F1-Score	Índice Kappa	AUC
Árvore de Decisão	99.37%	99.37%	99.37%	99.37%	99.37%	98.73%	99.37%
Random Forest	99.37%	98.81%	99.92%	99.92%	99.36%	98.73%	99.37%
Redes Neuronais	98.89%	98.81%	98.97%	98.97%	98.89%	97.78%	98.89%
Naive Bayes	71.56%	97.78%	45.33%	64.14%	77.47%	43.11%	71.56%
SVM	92.56%	95.09%	90.03%	90.51%	92.75%	85.13%	92.56%
Regressão Logística	91.34%	91.53%	91.14%	91.17%	91.35%	82.67%	91.34%

Para a avaliação dos modelos, a Acuidade e a Sensibilidade serão as métricas com maior importância na decisão do melhor modelo, dado que é fundamental na área da saúde, que um modelo tenha a maior quantidade de registos corretamente classificados e que possua uma maior quantidade de registos positivos.

Após a análise da Tabela 17, é possível verificar que todos os modelos, exceto o *Naive Bayes*, apresentam uma elevada Acuidade, acima de 90%. As Acuidades mais altas, quase perfeitas, são as dos modelos Árvore de Decisão, *Random Forest* e Redes Neuronais, com valores de 99,37%, 99.37% e 98.89%, respetivamente. Posto isto, o modelo *Naive Bayes* é já descartado, uma vez que apresenta a Acuidade mais baixa, com valor de 71.56%, em relação aos outros cinco modelos.

Ao comparar os modelos na métrica Índice Kappa, é possível verificar que o SVM e Regressão Logística são os que apresentam valores mais baixos, a rondar entre os 82% a 85%, enquanto que os modelos Árvore de Decisão, *Random Forest* e Redes Neuronais apresentam valores acima dos 90%. Desse modo, pode-se excluir, desde já, os modelos SVM e Regressão Logística.

Relativamente á métrica AUC, o modelo Redes Neuronais é o que apresenta um valor mais baixo, de 98.89%, em relação aos modelos Árvore de Decisão e *Random Forest*, que apresentam uma AUC com o valor igual a 99.37%. A diferença do Redes Neuronais para os outros dois modelos pode ser menor que 1%, no entanto, como a Árvore de Decisão e *Random Forest* possuem um valor mais elevado, têm uma maior capacidade a distinguir as duas classes (0 e o 1). Desse modo, o modelo Redes Neuronais foi descartado.

De modo a perceber qual dos dois modelos restantes, Árvore de Decisão e *Random Forest*, é o melhor modelo, analisou-se o resto das métricas, a Sensibilidade, Especificidade, Precisão e F1-Score. Os dois modelos apresentam valores muito semelhantes nestas quatro métricas, porém, na métrica Sensibilidade, a Árvore de Decisão possui um valor maior, de 99.37%, enquanto que o *Random Forest* possui um valor igual a 98.81%.

Assim, é possível concluir que o melhor modelo e o mais indicado para a previsão de infeção nosocomial é o método Árvore de Decisão. Este modelo é o que demonstra ser mais preciso na identificação de infeção e o que possui uma maior taxa de falsos positivos, o que na área da saúde é importante, devido ao facto de ser fundamental prevenir a propagação de infeção.

5.7 Implementação

O trabalho realizado neste projeto de dissertação demonstra ser importante e fundamental para a gestão e controlo de infeção nos hospitais, e custos associados.

Os resultados obtidos nos modelos de previsão demonstram ser positivos, precisos e capazes de satisfazer os objetivos pretendidos deste projeto.

Desse modo, com base nesta investigação, podemos concluir que estes modelos possuem uma elevada capacidade de previsão, e para além disso, podem, no futuro, contribuir com sucesso, na construção de um protótipo Sistema de Suporte à Decisão Inteligente, com base em modelos de inteligência artificial e *Data Mining*, que permite, em tempo real, otimizar a gestão e controlo de infeção de doentes internados.

Porém, de forma a garantir a eficiência e eficácia dos modelos no protótipo Sistema de Suporte à Decisão Inteligente, é essencial uma monitorização em tempo real dos modelos de DM, uma vez que poderá ser necessário introduzir novos dados e mudar os parâmetros dos modelos de previsão.

6. DISCUSSÃO DE RESULTADOS

Após a realização de todas as fases de trabalho, isto é, de todas as fases do CRISP-DM, avançou-se para a discussão de resultados, onde os resultados finais serão discutidos e confrontados com os objetivos propostos

De todas as fases do CRISP-DM realizadas, a mais demorada foi a Preparação dos dados, uma vez que foi necessário realizar diversas transformações aos dados recolhidos, de forma a que tivessem a melhor qualidade possível, e assim, pudessem ser utilizados na fase de Modelação.

Os resultados apresentados neste projeto de dissertação foram avaliados a partir de sete métricas, a Acuidade, Sensibilidade, Especificidade, Precisão, F1-Score, Índice Kappa e Curva AUC, sendo que a Acuidade e Sensibilidade foram estabelecidas como as mais importantes na decisão do melhor modelo de previsão.

Os resultados obtidos foram bons e mostram uma grande capacidade para prever infeção nosocomial. Os resultados até podem ser considerados demasiados bons, uma vez que as previsões resultaram em Acuidades com valores entre 71.56% a 99.37%. Apenas um dos seis modelos apresenta uma Acuidade abaixo de 90%. Além disso, todas previsões resultaram em Sensibilidades com valores acima de 90%. Estes valores podem justificar-se pela qualidade e tipo de dados que foi fornecido para a realização desta dissertação.

Com a matriz de correlação é possível visualizar o peso que cada variável tem sobre a *target* em questão, e desse modo, conseguiu-se perceber que nenhuma variável tinha um peso significativo na *target*. A maior parte das variáveis possuíam um peso com valor muito próximo de 0, o que pode explicar os resultados dos modelos.

Durante a realização do projeto foram utilizadas diversas técnicas para evitar o *overfitting* dos dados e melhorar a qualidade dos resultados. Foram usadas técnicas como o *oversampling* para balancear a variável *target*, o *One-Hot Encoding* para o processamento das colunas categóricas e o *K-folds Cross Validation* para obter melhores resultados e mais fiáveis. No entanto, nenhuma dessas técnicas mostrou fazer algum efeito, o que pode revelar que o grande problema está mesmo na da qualidade dos dados.

Para além disso, é importante referir que a ferramenta *Python* no ambiente *Jupyter Notebook* mostrou ser bastante adequada e útil para este trabalho. Foram utilizadas diversas bibliotecas para resolver problemas de DM encontrados ao longo do projeto de dissertação.

É também importante referir que estes modelos de previsão podem ser utilizados num Sistema de Suporte à Decisão Inteligente, dado que foram construídos com base em dados reais, logo iriam auxiliar na tomada de decisão e na gestão e controlo de infeção nosocomial.

Em síntese, o modelo de previsão que utiliza a técnica *Naive Bayes* é o que demonstra ser o menos fiável e eficaz, dado que é o que apresenta valores mais baixos em praticamente todas as métricas. No entanto, o resto dos modelos demonstram ser precisos e capazes de realizar previsões de infeção nosocomial. Estes modelos apresentam uma grande capacidade para prevenir a propagação de infeção e auxiliar os profissionais de saúde na tomada de decisão.

7. CONCLUSÃO

7.1 Síntese

A gestão e controlo de infeção é fundamental e importante para um bom funcionamento do hospital, para a segurança dos doentes hospitalizados e dos profissionais de saúde, e para a redução da mortalidade e custos associados. A aplicação de *Predictive Analytics* com a utilização de técnicas de DM e ML, permite a melhoria da gestão e controlo de infeção, uma vez que irá descobrir padrões nos dados, o que por consequente, possibilitará a identificação atempada e automática de infeção nosocomial.

Este projeto de dissertação aborda a problemática antes referida, a gestão e controlo de infeção, mais especificamente, a previsão de infeção nosocomial. O objetivo desta dissertação foi a conceção de modelos de DM com uma boa capacidade de previsão de infeção nosocomial, e considera-se que esse objetivo foi concluído com sucesso. Desse modo, foi possível responder á questão de investigação colado no início deste projeto de dissertação:

"É possível, através do uso de técnicas de Data Mining e Machine Learning, obter modelos que permitam perceber a probabilidade de um doente contrair uma infeção nosocomial?"

Este trabalho envolve dados reais do Hospital da Senhora Oliveira de Guimarães e são dados constituídos por pacientes que foram internados, que foram submetidos a uma ou mais cirurgias, que deram entrada nas urgências e foram administrados com antibióticos. Dado que este projeto tem como objetivo a realização de previsões eficazes de infeção nosocomial, foi utilizado o processo de *Data Mining* (DM) e métodos de *Machine Learning* (ML). Deste modo, a extração de conhecimento e descoberta de padrões nas fontes de dados recolhidas foi concebida a partir de técnicas de DM e ML.

A abordagem de DM aplicada foi a Classificação e para que os modelos de DM fossem criados, foram selecionadas seis técnicas, sendo elas, Árvore de Decisão (AD), *Random Forest* (RF), Redes Neuronais (RN), *Naive Bayes* (NB), *Support Vector Machine* (SVM) e Regressão Logística (RL).

Para além disso, de forma a perceber a qualidade e desempenho dos modelos, foram selecionadas sete métricas, sendo elas, Acuidade, Sensibilidade, Especificidade, Precisão, *F1-Score*, Índice Kappa e Curva AUC. Destas sete, a Acuidade e Sensibilidade, foram selecionadas como as mais importantes na decisão do melhor modelo. Os resultados foram positivos e bons, e em modo geral, todos os modelos apresentam boa capacidade de previsão de infeção nosocomial, exceto o modelo concebido a partir da técnica *Naive Bayes*, uma vez que é o modelo com Acuidade mais baixa. Estes resultados até podem ser

considerados demasiados bons, porque a maior parte dos modelos deram Acuidades e Sensibilidades acima de 90%, mas isso pode justificar-se pela qualidade dos dados fornecida.

Para que a realização deste projeto fosse possível, foram utilizadas diversas ferramentas, sendo elas, *Python* no ambiente *Jupyter Notebook* e *Microsoft Excel 2019* e MS Project. Em relação aos métodos, foram utlizados dois, *Cross-Industrial Standard Process for Data Mining* (CRISP-DM) e o *Design Science Research Methodology* (DSRM). O CRISP-DM foi usado para a aplicação de técnicas de DM, que tornou o processo mais simples e provou ser uma grande ajuda na orientação durante o desenvolvimento deste trabalho. O DSRM foi aplicado na investigação deste projeto de dissertação.

Na realização deste projeto foram encontradas algumas dificuldades, sendo que a primeira o número excessivo de resultados em cada pesquisa sobre informação científica relacionada com o tema. Isto provocou um maior gasto de tempo, uma vez que foi necessário a análise de cada artigo/livro, de modo a selecionar os melhores artigos/livros. Além disso, foram encontradas algumas dificuldades durante a realização das fases Preparação dos Dados e Modelação do CRISP-DM. A Preparação dos Dados foi a que demorou mais tempo a finalizar, dado que foi necessário realizar diversas transformações aos dados, para que no final tivessem a melhor qualidade possível e pudessem ser utilizados na fase seguinte. A fase de Modelação, os resultados iniciais não eram os melhores, pelo que foram aplicadas diversas técnicas para que houvesse uma melhoria nos resultados.

Por fim, é possível concluir que os objetivos propostos foram alcançados com sucesso e que podem oferecer um contributo útil para posteriores investigações e trabalhos. Estes modelos poderão auxiliar os profissionais de saúde na tomada de decisão, ajudando assim na melhoria da gestão e controlo de infeção. Além de mais, espera-se que no futuro, os modelos de previsão criados neste projeto possam contribuir para a construção de um protótipo de Sistema de Suporte à Decisão Inteligente.

7.2 Trabalho Futuro

Para trabalhos futuros que seguirem a mesma linha de investigação deste projeto de dissertação, previsão de infeção nosocomial nos hospitais, é importante indicar um conjunto de orientações e direções a ser tomadas:

- Integrar novas variáveis e novos dados nos modelos de previsão para obter modelos com melhores resultados;
- Incorporar cenários e novas técnicas de DM;
- Implementar um protótipo de Sistema de Suporte à Decisão inteligente para a gestão e controlo de infeção nosocomial de doentes internados.

REFERÊNCIAS

Alharthi, H. (2018). Healthcare predictive analytics: An overview with a focus on Saudi Arabia. In *Journal of Infection and Public Health* (Vol. 11, Issue 6, pp. 749–756). Elsevier Ltd. https://doi.org/10.1016/j.jiph.2018.02.005

Ashfaque, J. M. (2020). *Johar M . Ashfaque*. *August*.

Azevedo, A., & Santos, M. F. (2008). KDD , SEMMA AND CRISP-DM : A PARALLEL OVERVIEW Ana Azevedo and M . F . Santos. *IADIS European Conference Data Mining*, 182–185. http://recipp.ipp.pt/handle/10400.22/136%0Ahttp://recipp.ipp.pt/bitstream/10400.22/136/3/KDD-CRISP-SEMMA.pdf

Boonsiritomachai, W., McGrath, G. M., & Burgess, S. (2016). Exploring business intelligence and its depth of maturity in Thai SMEs. *Cogent Business and Management*, *3*(1). https://doi.org/10.1080/23311975.2016.1220663

Carletta, J. (2008). Assessing agreement on classification taske: the kappa statistic. *Computational Linguistics*.

Chang, Y., Yeh, M., Li, Y., Hsu, C., & Lin, C. (2011). *Predicting Hospital-Acquired Infections by Scoring System with Simple Parameters*. *6*(8). https://doi.org/10.1371/journal.pone.0023137

Dua, S., & Du, X. (2016). Data Mining and Machine Learning in Cybersecurity. In *Data Mining and Machine Learning in Cybersecurity*. https://doi.org/10.1201/b10867

Engelgau, M. M., Khoury, M. J., Roper, R. A., Curry, J. S., & Mensah, G. A. (2019). Predictive Analytics: Helping Guide the Implementation Research Agenda at the National Heart, Lung, and Blood Institute. *Global Heart*, *14*(1), 75–79. https://doi.org/10.1016/j.gheart.2019.02.003

Fayyad, U., Piatetsky-Shapiro, G., & Smyth, P. (1996). From data mining to knowledge discovery in databases. *AI Magazine*, *17*(3), 37–53. https://doi.org/10.1609/AIMAG.V17I3.1230

Fitz-enz, J., & Mattox II, J. (2014). *Predictive Analytics for Human Resources* (1st ed.). Wiley.

Frawley, W. J., Piatetsky-Shapiro, G., & Matheus, C. J. (1992). Knowledge Discovery in Databases: An Overview. In *AI Magazine* (Vol. 13, Issue 3). https://doi.org/10.1609/AIMAG.V13I3.1011

Han, J., Kamber, M., & Pei, J. (2012). Data Mining: Concepts and Techniques. In *Data Mining: Concepts and Techniques*. https://doi.org/10.1016/C2009-0-61819-5

Inweregbu, K., Dave, J., & Pittard, A. (2005). Nosocomial infections. *Continuing Education in Anaesthesia, Critical Care and Pain*, *5*(1), 14–17. https://doi.org/10.1093/bjaceaccp/mki006

Jupyter, P. (2016). *Project Jupyter | Home*. https://jupyter.org/

Khan, H. A., Ahmad, A., & Mehboob, R. (2015). Nosocomial infections and their control strategies. *Asian Pacific Journal of Tropical Biomedicine*, *5*(7), 509–514. https://doi.org/10.1016/j.apjtb.2015.05.001

Larose, D. T., & Larose, C. D. (2015). *DATA MINING AND PREDICTIVE ANALYTICS* (2nd ed.). Wiley.

Maimon, O., & Rokach, L. (2011). Data mining and knowledge discovery handbook. *Choice Reviews Online*, *48*(10), 48-5729-48–5729. https://doi.org/10.5860/choice.48-5729

Matos, L. M., Cortez, P., Mendes, R., & Moreau, A. (2019). Using Deep Learning for Mobile Marketing User Conversion Prediction. *Proceedings of the International Joint Conference on Neural Networks*, *2019-July*(July), 1–8. https://doi.org/10.1109/IJCNN.2019.8851888

Mohammed, M., Khan, M. B., & Bashie, E. B. M. (2016). Machine learning: Algorithms and applications. In *Machine Learning: Algorithms and Applications* (Vol. 7, Issue 13). https://doi.org/10.1201/9781315371658

Negash, S. (2004). Business intelligence. In *Communications of the Association for Information Systems* (Issue Volume13). https://www.researchgate.net/publication/228765967

Nelli, F. (2018). Python Data Analytics. In *Python Data Analytics*. https://doi.org/10.1007/978-1-4842-3913-1

Novakovic, J., Veljovi, A., Iiic, S., Papic, Z., & Tomovic, M. (2017). Evaluation of Classification Models in Machine Learning. *Theory and Applications of Mathematics & Computer Science*, *7*(1), 39–46. https://uav.ro/applications/se/journal/index.php/TAMCS/article/view/158

O'Connell, A. (2011). Logistic Regression Models for Ordinal Response Variables. In *Logistic Regression Models for Ordinal Response Variables*. https://doi.org/10.4135/9781412984812

Peffers, K., Tuunanen, T., Rothenberger, M. A., & Chatterjee, S. (2007). A design science research methodology for information systems research. *Journal of Management Information Systems*, *24*(3), 45–77. https://doi.org/10.2753/MIS0742-1222240302

Pete, C., Julian, C., Randy, K., Thomas, K., Thomas, R., Colin, S., & Wirth, R. (2000). Crisp-Dm 1.0. *CRISP-DM Consortium*, 76.

Pina, E., Ferreira, E., Marques, A., & Matos, B. (2010). Infecções associadas aos cuidados de saúde e segurança do doente. *Revista Portuguesa de Saúde Pública*, *Tematico*(10), 27–39.

Santos, J. V., Novo, R., Souza, J., Lopes, F., & Freitas, A. (2021). Transition from ICD-9-CM to ICD-10-CM/PCS in Portugal: An heterogeneous implementation with potential data implications. In *Health Information Management Journal*. SAGE PublicationsSage UK: London, England. https://doi.org/10.1177/18333583211027241

Santos, M., & Ramos, I. (2009). Business Intelligence: Tecnologias da Informação na Gestão de Conhecimento. *FCA - Editora de Informática*, 25. http://repositorium.sdum.uminho.pt/bitstream/1822/6198/1/Resumo_Livro_BI_MYS_IR.pdf

Saybani, M. R., Wah, T. Y., Aghabozorgi, S. R., Shamshirband, S., Mat Kiah, M. L., & Balas, V. E. (2016). Diagnosing breast cancer with an improved artificial immune recognition system. *Soft Computing*, *20*(10), 4069–4084. https://doi.org/10.1007/s00500-015-1742-1

Silva, E., Cardoso, L., Portela, F., Abelha, A., Santos, M. F., & Machado, J. (2015). *Predicting Nosocomial Infection by Using Data Mining Technologies*. 189–198. https://doi.org/10.1007/978-3-319-16528-8

Sofaer, H. R., Hoeting, J. A., & Jarnevich, C. S. (2019). The area under the precision-recall curve as a performance metric for rare binary events. *Methods in Ecology and Evolution*, *10*(4), 565–577. https://doi.org/10.1111/2041-210X.13140

Witten, I., Frank, E., & Hall, M. (2011). Data mining 2nd. In *Annals of Physics* (Vol. 54, Issue 2). http://www.cs.waikato.ac.nz/~ml/weka/book.html%5Cnhttp://www.amazon.com/Data-Mining-Practical-Techniques-Management/dp/0123748569

World Health Organization. (2002). Prevenção de infecções adquiridas no hospital - Um guia prático. *Insituto Nacional de Saúde Dr. Ricardo Jorge*, 93. https://www.dgs.pt/programa-nacional-de-controlo-da-infeccao/documentos/manuais-de-boas-praticas/prevencao-de-infeccoes-adquiridas-no-hospital-um-guia-pratico-pdf.aspx

Zaitseva, E., Kvassay, M., Levashenko, V., & Kostolny, J. (2015). Introduction to knowledge discovery in medical databases and use of reliability analysis in data mining. *Proceedings of the 2015 Federated Conference on Computer Science and Information Systems, FedCSIS 2015*, *5*, 311–320. https://doi.org/10.15439/2015F327

ANEXO I – CÓDIGO PYTHON COLUNAS DE RISCO DE INFEÇÃO

```
infecao_ns["INFECAO_CIRANTI"] = pd.np.where(
    (infecao_ns["DATA_INICIO"] - infecao_ns["DATAMOV"] <= datetime.timedelta(days=30))
    &
    (infecao_ns["DATA_INICIO"] - infecao_ns["DATAMOV"] >= datetime.timedelta(days=1)), "1", "0")
```

Figura 48 – Criação da coluna "INFECAO_CIRANTI"

```
infecao_ns["INFECAO_INTERANTI"] = pd.np.where(
    (infecao_ns["DATA_INICIO"] - infecao_ns["ADMISSAO"]) == datetime.timedelta(days=3), "1", "0")
```

Figura 49 – Criação da coluna "INFECAO_INTERANTI"

```
infecao_ns["INFECAO_CIRINTER"] = pd.np.where(
    (infecao_ns["ADMISSAO"] - infecao_ns["DATAMOV"] > datetime.timedelta(days=32))
    ,0
mapi = {}
```

Figura 50 – Criação da coluna "INFECAO_CIRINTER"

```
    if infecao_ns["NUM_SEQUENCIAL"][ind] in mapi:
        mdate = datetime.datetime.strptime(str(infecao_ns["ADMISSAO"][ind]), "%Y-%m-%d %H:%M:%S").date()
        mdate1 = datetime.datetime.strptime(str(mapi[infecao_ns["NUM_SEQUENCIAL"][ind]]["data"]), "%Y-%m-%d %H:%M:%S").date()
        if mdate - mdate1 <= datetime.timedelta(days=180):
            mapi[infecao_ns["NUM_SEQUENCIAL"][ind]]["count"] = mapi[infecao_ns["NUM_SEQUENCIAL"][ind]]["count"] + 1
    else:
        elm = {}
        elm["count"] = 1
        elm["flag"] = 0
        elm["data"] = infecao_ns["ADMISSAO"][ind]

        mapi[infecao_ns["NUM_SEQUENCIAL"][ind]] = elm


infecao_ns["INFECAO_INTERINFEC"] = 0

for ind1 in infecao_ns.index:
    if mapi[infecao_ns["NUM_SEQUENCIAL"][ind1]]["count"] > 1 and mapi[infecao_ns["NUM_SEQUENCIAL"][ind1]]["flag"] != 0:
        infecao_ns["INFECAO_INTERINFEC"][ind1] = 1
    else:
        mapi[infecao_ns["NUM_SEQUENCIAL"][ind1]]["flag"] = 1
```

Figura 51 – Criação da coluna "INFECAO_INTERINFEC"

```
mapi1 = {}


for ind in infecao_ns.index:

    if infecao_ns["NUM_SEQUENCIAL"][ind] in mapi1:
        mdate = datetime.datetime.strptime(str(infecao_ns["ADMISSAO"][ind]), "%Y-%m-%d %H:%M:%S").date()
        mdate1 = datetime.datetime.strptime(str(mapi1[infecao_ns["NUM_SEQUENCIAL"][ind]]["data"]), "%Y-%m-%d %H:%M:%S").date()
        if mdate - mdate1 <= datetime.timedelta(days=3):
            mapi1[infecao_ns["NUM_SEQUENCIAL"][ind]]["count"] = mapi1[infecao_ns["NUM_SEQUENCIAL"][ind]]["count"] + 1
    else:
        elm = {}
        elm["count"] = 1
        elm["flag"] = 0
        elm["data"] = infecao_ns["ALTA"][ind]

        mapi1[infecao_ns["NUM_SEQUENCIAL"][ind]] = elm


infecao_ns["INFECAO_REINTER"] = 0

for ind1 in infecao_ns.index:
    if mapi1[infecao_ns["NUM_SEQUENCIAL"][ind1]]["count"] > 1 and mapi1[infecao_ns["NUM_SEQUENCIAL"][ind1]]["flag"] != 0:
        infecao_ns["INFECAO_REINTER"][ind1] = 1
    else:
        mapi1[infecao_ns["NUM_SEQUENCIAL"][ind1]]["flag"] = 1
```

Figura 52 – Criação da coluna "INFECAO_REINTER"

```
infecao_ns["INFECAO_CIURG"] = pd.np.where(
    (infecao_ns["DATAHORA_ADM"] - infecao_ns["DATAMOV"] <= datetime.timedelta(days=30))
    &
    (infecao_ns["DATAHORA_ADM"] - infecao_ns["DATAMOV"] >= datetime.timedelta(days=0)) , "1", "0")
```

Figura 53 – Criação da coluna "INFECAO_CIURG"

ANEXO II – CÓDIGO PYTHON COLUNAS DE HISTORIAL HOSPITALAR

```
mapj = {}

i=0

for index,intern in infecao_ns.iterrows():
    idint = int(intern['NUM_SEQUENCIAL'])
    aux = mapj.get(idint,[])
    aux.append(intern['ADMISSAO'])
    mapj[idint] = aux

res= {}

for key,value in mapj.items():
    res[key] = len(set(value))

infecao_ns['N_INTERNAMENTOS'] =  infecao_ns['NUM_SEQUENCIAL'].map(res)
```

Figura 54 – Criação da coluna "N_INTERNAMENTOS"

```
mapk = {}

i=0

for index,urgente in infecao_ns.iterrows():
    idint = int(urgente['NUM_SEQUENCIAL'])
    aux = mapk.get(idint,[])
    aux.append(urgente['DATAHORA_ADM'])
    mapk[idint] = aux

res= {}

for key,value in mapk.items():
    res[key] = len(set(value))

infecao_ns['N_URGENCIAS'] =  infecao_ns['NUM_SEQUENCIAL'].map(res)
```

Figura 55 – Criação da coluna "N_URGENCIAS"

```
def age(born):
    born = datetime.strptime(str(born), "%Y-%m-%d %H:%M:%S").date()
    today = date.today()
    return today.year - born.year - ((today.month,
                                      today.day) < (born.month,
                                                    born.day))

infecao_ns['IDADE'] = infecao_ns['DTA_NASCIMENTO'].apply(age)
```

Figura 56 – Criação da coluna "IDADE"

ANEXO III – CÓDIGO PYTHON ONE-HOT ENCODING

```
from sklearn.preprocessing import OneHotEncoder

encoder = OneHotEncoder()
encoder_results = encoder.fit_transform(infec_nosocomial[['MED_DESIGNACAO']])
infec_med = pd.DataFrame(encoder_results.toarray(), columns=encoder.categories_)
infec_med.columns = encoder.get_feature_names(['MED'])

infecoes = pd.concat([infec_nosocomial,infec_med],axis=1)
```

Figura 57 – Aplicação do One-Hot Encoding na coluna "MED_DESIGNACAO"

```
encoder_results2 = encoder.fit_transform(infecoes[['COD_ICD']])
infec_cod = pd.DataFrame(encoder_results2.toarray(), columns=encoder.categories_)
infec_cod.columns = encoder.get_feature_names(['COD'])

infec_hospitalar = pd.concat([infecoes,infec_cod],axis=1)
```

Figura 58 – Aplicação do One-Hot Encoding na coluna "COD_ICD"

```
encoder_results3 = encoder.fit_transform(infec_nosocomial[['SEXO']])
infec_sex = pd.DataFrame(encoder_results3.toarray(), columns=encoder.categories_)
infec_sex.columns = encoder.get_feature_names(['SEXO'])

infec_hospitalar = pd.concat([infec_hospitalar,infec_sex],axis=1)
```

Figura 59 – Aplicação do One-Hot Encoding na coluna "SEXO"