NOVA IMS

Information Management School

MAA

Mestrado em Métodos Analíticos Avançados

Master Program in Advanced Analytics

Automated Machine Learning implementation framework in the banking sector

Pedro Carmona

Dissertation presented as partial requirement for obtaining the Master's degree in Data Science and Advanced Analytics, with specialization in Business Analytics

NOVA Information Management School
Instituto Superior de Estatística e Gestão de Informação
Universidade Nova de Lisboa

NOVA Information Management School
Instituto Superior de Estatística e Gestão de Informação
Universidade Nova de Lisboa

AUTOMATED MACHINE LEARNING IMPLEMENTATION FRAMEWORK IN THE BANKING SECTOR

by

Pedro Carmona

Dissertation presented as partial requirement for obtaining the Master's degree in Data Science and Advanced Analytics, with specialization in Business Analytics

Advisor Vitor Manuel Pereira Duarte dos Santos

November 2021

ACKNOWLEDGEMENTS

I would like to use the space dedicated to acknowledgments to thank everyone that supported me through my entire journey of obtaining this degree and completing this dissertation.

Especially, I would like to express my gratitude to my advisor, Professor Doctor Vitor Manuel Pereira Duarte dos Santos, Assistant Professor at NOVA Information Management School (NOVA IMS), that, for his support, the guidance and knowledge, essential contributor without whose guidance and knowledge I would not have completed this journey. Your feedback and inspiration made it possible for me to complete this work.

And, at last, I would like to thank my family for being supportive and loving in every step.

ABSTRACT

Automated Machine Learning is a subject in the Machine Learning field, designed to give the possibility of Machine Learning use to non-expert users, it aroused from the lack of subject matter experts, trying to remove humans from these topic implementations. The advantages behind automated machine learning are leaning towards the removal of human implementation, fastening the machine learning deployment speed. The organizations will benefit from effective solutions benchmarking and validations. The use of an automated machine learning implementation framework can deeply transform an organization adding value to the business by freeing the subject matter experts of the low-level machine learning projects, letting them focus on high level projects. This will also help the organization reach new competence, customization, and decision-making levels in a higher analytical maturity level.

This work pretends, firstly to investigate the impact and benefits automated machine learning implementation in the banking sector, and afterwards develop an implementation framework that could be used by banking institutions as a guideline for the automated machine learning implementation through their departments. The autoML advantages and benefits are evaluated regarding business value and competitive advantage and it is presented the implementation in a fictitious institution, considering all the need steps and the possible setbacks that could arise.

Banking institutions, in their business have different business processes, and since most of them are old institutions, the main concerns are related with the automating their business process, improving their analytical maturity and sensibilizing their workforce to the benefits of the implementation of new forms of work. To proceed to a successful implementation plan should be known the institution particularities, adapt to them and ensured the sensibilization of the workforce and management to the investments that need to be made and the changes in all levels of their organizational work that will come from that, that will lead to a lot of facilities in everyone’s daily work.

KEYWORDS

Machine Learning, Artificial Intelligence, Data Science, Advanced Analytics, automated Machine Learning, banking

INDEX

LIST OF FIGURES

LIST OF TABLES

1. INTRODUCTION

1.1. Context

The word's constant change has aroused the companies needs to streamline and automate their business processes with the least resources expended and the maximum profits obtained, since there is a rising need to achieve competitive advantage over their competitors, and the banking is not an exception. Every bank pretends to become a more data and analytics centric organization, leveraging data to know the customer better than the competition, to produced accurate marketing campaigns and better fulfilling the customer needs. Deep Learning rise led to the development of a new branch, nominated automated Machine Learning, came from the lack of experience personnel in these fields and the constant need for expertise in these area from the organizations. Automated Machine Learning allows non experts in machine learning to use machine learning models and work with their outputs, replacing the human in almost the entire process, from the dataset creation to the analytical model deployment. In the banking sector, there is a lack of skilled analytical personnel, which makes automated machine learning, a great opportunity to change the sector business processes status-quo, automating the business in most of the organization areas.

The concepts of Artificial Intelligence, Machine Learning and Deep Learning are not new, and it has been the goal of most organization to implement them, replacing humans with machines instructed to execute their functions and use of machine learning capabilities for analytical based decision making (Sodhi et al., 2019).

Artificial Intelligence is a computer science branch, pretending to use computer systems to develop programs design to solve complex problems applying human-like reasoning processes and develop intelligent machines and software. One of the most known topics in Artificial Intelligence is Neural Networks, with real world success in language processing and speech recognition (Ren et al., 2017a). Artificial Intelligence is essential in every industry due to its reasoning, planning, learning, communicating and perception capabilities and the manipulating and moving objects ability (Singh, 2017a).

Artificial Intelligence pretends to create programs replicating cognitive skills, as learning and understanding rules to use data for organizational purposes, the created algorithms are known as rules, with specific instructions for each step to complete the tasks and obtain the results as accurately as possible (Haton, 2006).

1.2. Motivation

In the last couple of years, the Artificial Intelligence investment have been increasing, with the emerging technologies, and the Big Data technologies and Internet of Things (IoT) have the creation and growing of new applications and services. AI is considered an engine for future economic growth, since has several application fields. In banking sector, it is currently used in email filtering and fraud detection, mainly (Jones et al., 2018).

An Artificial Intelligence subfield, Machine Learning, evolved from the need to teach computers to learn a problem's solution, making the computers programming themselves (Essinger & Rosen, 2011). Machine Learning differs from the traditional programming paradigm, since the program is created by inputting the existent data and the pretended output, in the computer (Sodhi et al., 2019). There existing types of learning are, supervised learning, the learning algorithm uses the desired output and the labeled data, unsupervised learning, the algorithm uses unlabeled data, semi-supervised learning, with part of outputs included in the training data and reinforcement learning, an algorithm providing dynamic positive and negative feedback (Sodhi et al., 2019). Machine Learning is used to solve different problems, Classification problems, identifying an object category, Regression problems, predicting a numeric value associated with the object, Similarity problems, to find similarities or anomalies in the objects' behavior, Ranking problems, sorting data according to the input and Sequence Prediction problems, predict the next element in a data series (Simeone, n.d.).

The data availability, with the increase in the generation and the reduced storage costs, leveraged the machine learning growth, and the increasing availability for use, with faster and powerful computers lead to the development of new machine learning techniques and innovative algorithms (Choi et al., 2020b). Deep Learning is a machine learning branch, creating algorithms inspired in the human brain function to solve complex prediction functions, the artificial neural networks, with several uses in the Big Data field and its applications (Emmert-Streib et al., 2020).

To allow machine learning use in a daily base for unexperienced works, reducing the demand for knowledge experts, automated Machine Learning has been created. Having several advantages, like the removal of most of human expert intervention, enablement of faster delivery of machine learning projects, and an efficient evaluation of deployed solutions performance, leaving most experienced machine learning workers for high level business projects. Making machine learning accessible in daily tasks, with companies achieving higher competence levels, with significant impacts in the business results (Wever et al., 2018).

AutoML idea and its techniques has been in development, attempting to take the place of humans on identifying the ideal machine learning model configurations with the human expertise removal (Truong et al., n.d.).

With the focus on automatizing business processes and transforming companies into analytic centric organizations, some interesting use cases that will benefit from the use of automated Machine Learning are related with precise product and service targeting. This can help the anticipation of client needs and deepen relationships, creation of precise credit models for risk analysis, management of consultancy of clients portfolio, identification and cooperation of clients in financial difficulties and creation mitigation risk forecast models, leading to competitive advantage, critical for any business´ success (Feurer et al., n.d.).

1.3. OBJECTIVES

This paper pretends to discuss some interesting use cases of application of autoML in the banking sector and develop and propose a strategy for the implementation of a AutoML tool. To achieve this goal, the following intermediate objectives were defined:

- Find out interesting and applicable use cases regarding autoML in the banking sector
- Find out how these use cases can positively impact the banking sector
- Find out the competitive advantage obtained from the implementation of these use cases would be translated in money and time savings
- Develop a framework to the implementation of an automated machine learning solution
- Propose the roadmap for the implementation of this framework

This work has the following structure:

The introduction contains the work context, the motivation, and the goals this work.

The second part contains the artificial intelligence concept and its sub concept's introduction, the automated machine learning concept, function, and advantages are elaborated. The advantages that every institution could benefit with, the normal functioning, and the challenges of its implementation are presented. Also, it is explained the fundamentals of banking.

The third part contains the methodology adopted explanation.

The work proposal, with the strategy design, the case study simulation, its interview validation, and results discussion are presented in the fourth part.

The final part contains the thesis conclusion.

2. LITERATURE REVIEW

2.1. BANKING

Banking is an industry that handles financial transactions, mainly cash and credit, and provides a place to store extra cash and credit, according to security standards. The banks offer a great diversity of products, saving accounts, certificates of deposits, checking accounts and several types of loans, home mortgages, business, and car loans (*Banking: How It Works,Types, How It's Changed*, n.d.).

This industry is one of the developed countries economy drivers, assures liquidity, monetary assets for investment and spending, it can consist in cash, treasury bills, notes, and bonds, either for families or for companies to invest (*Banking: How It Works,Types, How It's Changed*, n.d.).

A bank needs to be licensed to deal with customers, and needs to follow several standard procedures and regulations to accept their deposits and loan applications, also doing lending services, wealth management, currency exchange and safe deposit boxes (*Bank Definition*, n.d.). Banks are the safest place to deposit excess cash, they are insured by a central bank, and return to the clients a percentage of their deposit, according to the current interest rate. The banks obtain profits charging higher interest rates than the ones they pay for the deposits (*Banking: How It Works,Types, How It's Changed*, n.d.).

2.1.1. Opportunities

In the 21st Century, the information became a powerful tool to doing business, the rise in information produced by the entities has given them tons of information to work with, along with the rise and development of Machine Learning have given the opportunity to change the paradigm and improve their strategies of use and posterior analysis of the collected data, change user experience, detect fraud, decrease existing risks, automation of routine tasks, and increase profits. Banks receive data from every entity they are in contact with from customers to investors and partners, and contractors and need to find improved strategies with them to be closer to their customer and achieve competitive advantage.

There are several opportunities in the banking sector to apply the new technologies, to improve the productivity and increase the automation, using the ability of ML and AI to handle mundane tasks, freeing the employees to deal with more complex challenges, contributing to a profit increase (*Machine Learning in Banking - Opportunities, Risks, Use Cases*, n.d.-a).

Having more ways and tools to gather more and analyze the available one will be possible to have a more intimate knowledge of the customer, which will allow the possibility of giving a personalized customer experience by knowing as much as possible about the customers, the creation of chatbots is an useful way to interact with the customers, leading him along the right path and reduce the staff workload, it also performs lock of cards and notifications if the client exceeded the card's limit. The overload of information makes easier to find the needs and wants, and what each client is looking for and since knowing their possessions, gave the ability of offering a personalized offer. It is also possible to predict the client intentions, this knowledge will trigger the measures to increase customer retention. The abundance of information about each customer can also give the banks a higher ability to predict the risk associated with each client, being more accurate than individuals, eliminating any possible human bias.

Since, there is always going to exist criminals trying to commit financial fraud, it exists a wide range of ML based fraud detection methods and techniques, there are a lot of solutions being developed based on anomaly detection or predictive or descriptive analytics, to secure the bank systems from several levels of the transaction threat. With the implementation of fraud detection and prevention systems the customers can experience a greater and improved experience and also be more protected in relation to losses (Huang et al., 2020).

The new technologies are also very important for maintaining a bank institution safe, by the implementation of facial recognition cameras, to check if a card is being used by its owner in the transactions or tracking suspicious IP addresses where is made the order of a transaction to prevent fraudulent intentions (Lima, 2018).

Machine Learning in conjunction with Big Data available also made possible, more than just collecting information, also made possible to research the market and predict the incoming trends by monitoring currency value, studying investment options and credit risks, studying competitors, and identifying security weaknesses. It also made possible the identification of business processes' improvement opportunities and the increment of the overall efficiency, to profits increasing (*Machine Learning in Banking - Opportunities, Risks, Use Cases*, n.d.-b).

The use of AutoML automation will help banks deploying AI to optimize their customer experience, allowing product targeting, profiling customers for specific products and services, deepen the relationships with the bank, anticipating their needs to identify new opportunities. These technologies can also be used to predict customer lending, that can be a high-risk proposition, creating more value, with risk analysis credit models and identify business opportunities according to the forecasted returns, that leads portfolio maximum returns, identifying clients in financial difficulties to proactively give them solutions and creating mitigation risk models. With the correct use of AutoML, AI and Machine Learning, the banks can also give better investment management consultancy to their clients' portfolio, gathering deep insights into market changes and opportunities, spotting key trends and optimize returns, the value is gained by optimizing trades' executions, propose investment opportunities, study market conditions to discover key trends and reducing transaction costs (Balaji & Allen, n.d.).

The bank safety can also be improved, by monitoring fraud and financial crime, by identifying the patterns of money laundering, use previous gathered data for model creation to flag suspicious activity. It is also possible improve the banking client experience with AI, matching the clients' expectation by knowing everything about their needs and present them with solutions, using analytical models from client behavior predictions and client needs previsions, satisfaction, and complaints data to prevent client churn, and predict branch traffic volumes. AI and ML are used to help banks managing customer acquisition, modelling customer patterns, and optimizing marketing campaigns to ensure the highest returns, identify ideal target customers, predict behaviors, identify ideal pricing and define marketing channels, either digital or traditional media to optimal client retention (Hu et al., 2020).

The financial product demand can be forecasted with machine learning and artificial intelligence, to understand the products needed, direction them to the specific markets in a specific time to be more profitable, AI can also be used to forecast demand for all banking products, either, destined for companies or particular customers. (Hu et al., 2020).

The financial investment is optimized with ML and AI, replacing the precarious and time-consuming of understanding and managing modern economy investments, trade routing and execution optimization models, and corresponding validation and develop decision support systems for optimized market impact ensuring compliance (*AI for Banking | Powered by dotData's AutoML 2.0 Platform*, n.d.).

2.2. Artificial Intelligence

Artificial Intelligence (AI) is a computer science's branch defined as the showing of intelligence by a non-human entity, normally, a computer. The Artificial Intelligence's research deals with the implementation of intelligent behavior into machine tasks to be able to act work for humans in tasks of controlling, planning, scheduling, handling diagnostic and consumer questions, handwriting, facial and speech recognition, in general, provide solutions to daily problems. The AI's importance comes to mind when we see their dependence on economics, medicine, engineering, and military (Andreu-perez, 2017).

Artificial Intelligence's main goal is, as mentioned, the automation of human activities requiring intelligence, addressing areas as Perception, building physical world models with visual or audio input, Manipulation, the articulation of appendages (mechanical arms and locomotion devices) to affect a physical word desired state, Reasoning, simulating cognitive functions, as inferential conclusions and planning, for examples, diagnosing and designing, Communication, understanding a problem using transmission of information by language, and Learning, treating automation problems to improve the systems performance with the gained experience over time (Jones et al., 2018).

AI has been in development for many years, and launched a lot of concepts into the market, the functions have a knowledge facts base characteristic of base proficiency system, consisting of independently valid information blocks, automatically organized, and used by the system to solve the problems presented. A process known as goal directed specific search, complex since it is heuristics' controlled, is needed to rank information importance and find ambiguous data. (Chen et al., 2019) It needs an adaptative organization, with computational architectures to represented the obtained knowledge, has a subfield of Knowledge Engineering, related with the architecture acquisition and encoding of knowledge (Internet Society, 2017).

The traditional methods of Artificial Intelligence, nowadays, are classified as Machine Learning, focusing on statistical analysis, like expert systems, reasoning capabilities to get conclusion, with the process of huge information amounts to obtain conclusions, reasoning based on experience and cases, Bayesian networks and behavior-based AI, building AI systems (Cockburn et al., 2018).

Computational Intelligence includes the development and learning iteratively, from empiric data and non-symbolic AI associated, scruffy AI and soft computing, with neural networks, building strong pattern recognition systems, fuzzy systems, to reason under uncertainty, industrial and consumer product control systems, and evolutionary computation, replicating biological concepts to obtain better problem solutions (Haton, 2006).

2.2.1. Applications of AI

Artificial Intelligence, nowadays, is applied in several disciplines, including Game Playing, Speech Recognition, Computer Vision, Expert systems, Heuristics Classification, Neural Network, Robotics and Natural Language Processing. The Computer Vision AI applications help machines recognizing objects, visual information capturing and analyzing, transmission of conversations from analog to digital and processing of digital signals, the learning is done with labelling and classification of various objects, handling the implications and quickly deciphering (Jones et al., 2018). Since the word has three-dimensional objects, and computer's cameras and the human-eye catch just two dimensions, not being able to see through plain objects, machine learning and computer vision try to overcome these limitations. An expert system is a high performance, reliable, highly responsible, and understandable AI applications relying on joining the human experts and programming knowledge into a system, replicating the human's expert decision making ability. Heuristic Classification, placing information, from several sources on some defined categories. Neural Network, getting cognitive science and machines performing tasks, with some nervous system science, using the same structure as the human brain with neurons coded into a system or machine, performing minimal tasks with minimal effort, and automating most of them. Robotics, innovating on the design and development of robots, consolidating science, and engineering fields, managing the systems controlling the robots, deciding the design, production, operation, their results, and data changes. A robot can be defined as a programmed machine performing automatic and semi-automatic actions, and the AI applications makes robots performing even more complex tasks, deploying them to replace people in tasks that these are not able to solve repeatedly. A computer science and AI subfield, called Natural Language Processing, gives the computer the ability to understand and process human language, allowing to have human instructions, by the direct communication with the user, by speech or text. It has the need of a robust and flexible infrastructure to support the new applications scalability and development, also providing the connection of different devices across the work for high network quality for cloud technologies (Singh, 2017b).

2.3. MACHINE LEARNING

Machine Learning is a field related with algorithm development to represent data, it uses subsets of data to create an algorithm using different features and weights combinations, coming from defined principles, a dataset and an algorithm showing how to operate in the dataset are given to a computer, pretending to obtain outputs, differently to classical programming, with an algorithm coded to obtain the known features. There are several learning methods described in the following sections, each with several algorithms and used in different tasks (Sodhi et al., 2019).

The application of Machine Learning techniques solves several types of problems, classification, the predictions of non-continuous or discrete categories to identify the one corresponding to an object, regression, predicting a continuous numeric value associated with an object, similarity, obtain similar objects to find behavior anomalies, ranking, sorting relevant data according to input and sequence prediction, predict next elements in a series of data.

Every Machine Learning algorithm has a label/target, that represent the value that the algorithm is being solved for, features/inputs, the properties of the given instance and the prediction model, where the functionalities occur. The main operations are training, a set of features is paired with labels to build the model, inference, features are the inputs for requesting a prediction, that happens when a model is given features towards resulting a result (Sodhi et al., 2019).

A machine learning algorithm has a set of components, representation, to show the gathered knowledge according to the used algorithm, the hypothesis evaluation or search process, it can be made through measuring accuracy, the hypothesis evaluation, using a set of evaluation metrics according to the chosen algorithm (Sodhi et al., 2019).

2.3.1. Supervised Learning

In a supervised learning algorithm, the existent data is labeled, either the inputted data and output expected in the training process, a mapping function is generated from the algorithm identifying the correspondence between the received input and the corresponding output, and the training is continuously happening when the desired accuracy level is obtained (Sodhi et al., 2019). It trains making comparisons the obtained output with previous ones finding errors and modifying the model, to allow the computer to learn a classification system. The supervised learning algorithms are most used to solve regression and classification tasks, regression dealing with the prediction of numeric data, and classification, executing the transformation of numeric target variables into categorical variables, dividing the variables into ordinal classes, with an associated categories' order (Simeone, n.d.).

The supervised learning algorithms make use of the dataset patterns destined to training to map the properties assigned to predictions (features), making predictions on future datasets, inferring an algorithm with feature-target pairs to determine how to predict correctly. It divides into three datasets, training, validation and testing and the dataset assigned for training will be the one in which the model will perform optimally (Choi et al., 2020a). It maps the features to the target, that learn the mapping function, the performance is evaluated according to the test dataset (unknown to the algorithm), proceeds doing these steps: splitting the original dataset into training, validation, and test datasets, use the datasets assigned for training and validation to learn about the feature-target relationship and evaluating the model on the test dataset according to the defined metrics, that are compared and evaluated in each iteration in the training and validation dataset (Sodhi et al., 2019).

Most of the supervised learning problems are labeled classification and regression tasks, with a dependent variable, continuous response, or class membership. The algorithms learn from training data, and, in classification problems, classify the new observations in binary or categorical responses or, in regression problems, predict the new observations in numerical observations. The binary data regression algorithms can be use in classification problems, thresholding the numerical predicted scores at an appropriate value (Nasteski, 2017).

2.3.1.1. Linear Regression

One of the simplest linear regression algorithms is a linear regression, a regression pretending to specify the relationship between numeric features and a simple numeric target, it is intended to solve the regression problem with a straight line describing the dataset relationship, if it is an univariate regression, it is only used a single feature to predict the target value, takes the form: *y=ax+b* (Choi et al., 2020a).

A linear regression algorithm describes the dataset in a slope-intercept form, where machine must identify the values *a* and *b* to obtain the best fitted line to the relation the values of x and y. The multivariate linear regression algorithm, has multiple weights, describing the degree to each feature influence the target (Choi et al., 2020a).

It is impossible to have a function fitting the dataset perfectly, having the need to calculate residuals to measure the error associated with the fit, the vertical distances between predicted and actual values. A cost function tries to eliminate model errors, and it can be decreased with the application of a gradient descent process, iteratively optimizing the algorithm. For this case, the cost function is a mean squared errors, obtaining the parameters estimated to best model a dataset, by minimizing a function (Rong & Bao-Wen, 2018).

2.3.1.2. Logistic Regression

Logistic regression has the main goal of discovering a relationship between the data features and a particular outcome probability, with the use of numerical and categorical predictor variables, using a sigmoidal curve to estimate the class probability, being obtained a curve converting the numeric features into a single numerical value in the range between 0 and 1, and the probability can be binomial or multinomial, depending on the possible outcomes (Choi et al., 2020a).

Works by extracting input weighted features, converting it into logarithms and combine them linearly, each feature being multiplied by the corresponding and added, it is considered a discriminative classifier, since the probability of an event occurrence is predicted by fitting data to a logistic function.

It extracts a set of weighted input features, transforming it into logarithms and linearly combine them, each feature is multiplied by a weight and added up, it is a discriminative classifier, since an event probability using existing data fitted to a logistic function (Nasteski, 2017).

2.3.1.3. Naive Bayes

Naive Bayes is a classification method for the text classification industry assuming an underlying probabilistic model, with some uncertainty about the principled probabilities outcomes from the obtained model to solve predictive problems. The model provides practical algorithms, inputted data combinations and help in understanding and in the evaluation of learning algorithms. It calculates explicit probabilities for any hypothesis and the data noise robustness, obtaining the equation $P(x1, x2) = P(x1|x2)P(x2)$ without generality loss, and it can be generalized for a set of variables, with two variables having a conditional independence assumption on another variable c, as $P(x \mid c) = \prod P(xi \mid c)$ (Nasteski, 2017).

The Naïve Bayes classification algorithm is a subset of Bayesian decision theory, names from the naïve assumptions made in the formulation, most of its popularity coming from the conditional independence, overfitting, and Bayesian methods. It is a very simple algorithm, useful for documents classification, and can also be used time-storage critical, minimal storage and fast training conditions. It assumes that the features are independent according to the class and the variables can be with optimization problems equations, simplifying learning, from autonomic given class features (*(15) (PDF) Short Survey on Naive Bayes Algorithm*, n.d.).

2.3.1.4. Decision Tree

Decision trees is a reliable classification algorithm, using a rooted node tree, directed without incoming edges, with other containing one incoming edge, internal nodes, the ones with outgoing edges, and leaves, nodes in the tree's bottom. The tree starts having a root node (representing the first dataset split), with a single feature splitting optimally the instance space data in the defined classes according to the feature value or a range, if the attributes are numerical. The splits are connected through an edge to another node having another feature splitting the data into groups or to a terminal node representing the class (this process is called recursive partitioning). The leaves are representations of the classes with the best target value, holding a probability value and every node is labeled with the tested attributes and the branches labeled with the corresponding values (Patel & Prajapati, 2018).

The algorithm recursively partitions the instance space to build tree with the purest nodes as possible, containing single class points. A new point is classified, and moved starting in the root node to the terminal node through the tree branches (Rokach & Maimon, 2006).

Decisions trees are an effective classification algorithm, mainly assigned to small datasets using rule inductions, and to allow better comprehensibility. It controls the complexity with the stopping criteria and the pruning method, improving the accuracy and measuring the total of nodes, leaves, the tree depth, and the used attributes. The computational complexity is not favorable to the number of data dimensions, and large datasets will originate very complex trees, needing a lot of memory space (Custode & Iacca, 2020).

2.3.1.5. Random Forest

Random forest is a decision tree's ensemble method, that generates several decision trees, using part of the features to each decision, with each one predicting the outcome of the class and in the class being one select for the class prediction (Ren et al., 2017b).

It is grouped several not pruned classification or regression trees, with training data random selections, made randomly in an induction process, for classification problems is voted by majority and for regression problems are calculated the average. Every tree is grow by a random n sampling, with n cases belonging to the training set for replacement, from the original dataset, as the training set for adding each leaf, for the random selected input variables, the best split is used in the nodes and the number of input variables remains constant, growing each tree the largest extent possible (*(PDF) Random Forests and Decision Trees*, n.d.).

Random forests' main purpose is to build an ensemble predictor using a set of decision trees to randomly select data subsamples, to build independent and non-identical decision trees randomly distributed (Biau, 2010).

2.3.1.6. Support Vector Machine

Support Vector Machine (SVM) is a supervised model mapping inputs to highly dimensional feature spaces, building a hyper-plane in a dimensional space for classification or regression tasks (Tang, 2013).

SVM are mainly used to optimization problems, using the maximal margin principle, dual theory and kernel trick, for convex and non-convex problems, it maps the training data into a feature space, using a kernel function to separate the data using a hyper plane, creating a margin separating the training and previous examples (Schlag et al., n.d.). This is enlarged to ensure a correct classification, if the decision rule generalizes and the kernels defined positively, it efficiently solves the optimization and the algorithm is interpreted as separated hyperplane in a high dimension feature space (Tian et al., 2012).

The algorithms show its best performance to solve classification problems, after the right parameters are found for the specific problems, to train a new model for all parameters' set, making it difficult with larger datasets, and different models are evaluated in parallel, it is a highly parallelizable process, with larger time complexity to solve optimization problems in large datasets. Optimized SVM algorithms cannot use large datasets, due to possible imbalances, it is possible to deal with it using random sampling, reflecting the training data distribution, and passing on significant testing data. To train a SVM on large dataset problems, needs to be built a framework for problem hierarchy, with each level as a problem instance decreasing in size and reflecting the original problem structure, after, a regular model is trained on the general problem, then projected in the multilevel paradigm hierarchy to reduce computation processing time than other approaches, mainly non-hierarchical for the quality of prediction in larger data sets (Tang, 2013).

Support vector machines are classifiers using a hyperplane to decide the new data point class, mainly used for classification. With a set of labeled data points, having a minority class of all data points with a positive label. The remaining data points belong to the majority, where every training data point is interpreted as a dimensional vector, the algorithm finds the hyperplane separating both the two classes. It is defined the best hyperplane according to the distance from the classes. When the data is not linearly separable in the Euclidean space, it is used the kernel trick, to map points to a dimensional space to separate a class by a hyper plane (Tian et al., 2012).

2.3.2. Unsupervised Learning

An unsupervised learning algorithm, receives unlabeled and unclassified data to generate a function identifying the hidden structures in the dataset patterns, differences and similarities among data untrained and does not make any assessment of the structure level of accuracy (Sodhi et al., 2019).

Pretends to detect dataset patterns and classify the dataset instances to categories, the algorithm must determine all the existent pattens. It is focused on clustering, grouping dataset instance into distant groups (clusters) according to their features' combinations (Choi et al., 2020a).

2.3.2.1. Principal Component Analysis

Principal Component Analysis (PCA) is an algorithm pretending to reduce data dimensionality reduction and preserve variability. Starts by decomposing the design matrix or calculate the covariance matrix data and decompose the eigenvalues. The algorithms searches for a low-dimensional representation of data and uncorrelated factor to determine data variation, retaining the most possible information (Jollife & Cadima, 2016).

The algorithm uses the features' correlation, extremely high between features' subset, to combine highly correlated features and represent a small number of linearly uncorrelated features. The correlation is done by the algorithm to find the directions of maximum variance and project it into a substantially smaller dimensional space, with components called principal components. (Salem & Hussein, 2019) The next step, is to proceed to the original features reconstruction, with the algorithm minimizing the errors occurred during the optimal components search. Most machine learning algorithms are optimized with PCA, due to the reduced data dimensionality and data size obtained (Howley et al., 2006).

It is a statistical discipline, pretending to preserve the most variability possible, so it lead to many algorithm reinventions, trying to optimize the way of successfully find uncorrelated variables, representing the original dataset linear functions, which maximizes the variance, and it leads to an eigenvalue/eigenvector problem (Howley et al., 2006).

The Principal Components Algorithm definition is defined as the search for an eigen problem solution, or decomposition of a data matrix into singular values, and it is balanced in the covariance or correlation matrix. For both cases, the variables depend on the dataset, instead of pre-defined adaptative functions. It is used for descriptive problems, in need of assumptions not distributional, being an exploratory adaptative method to use on several data types and for inferential purposed, in datasets with a multivariate normal distribution (Jollife & Cadima, 2016).

The principal components analysis' goals are the extraction of the most important information from data, compressing the dataset (to reduce the dimensions' number) and simplifying its description, while maintaining as most information as possible (Salem & Hussein, 2019).

2.3.2.2. K-Means

K-Means is a powerful data mining algorithm, however it is associated with some limitations, regarding the random centroids initialization generating unexpected convergence, the clustering definition needed before initiation, without it can be obtained several cluster shapes and influenced by outliers, and the lack of handling for several data types (Khanum et al., 2015).

The algorithm divides the training dataset into different clusters located near each other, assigning to different centroids, each a different value, every training instance is assigned to the cluster with the nearest centroid index, and then updates the mean, alternating these two steps until the convergence is found (Ahmed et al., 2020).

It produces clusters effectively for practical applications, requiring the prior definition of the dataset number of centroids, creating different cluster types according to the initial centroids choice (Capó et al., 2018). Although it is an algorithm affected by the mean points centroid initialization, if it is initialized as a distant point, can be a cluster without associated points, contrarily, different cluster may be connected to a single centroid, and if this centroid initializes more than one cluster, it is a case of poor clustering (Li & Wu, 2012).

2.3.3. Semi-Supervised Learning

Between supervised and unsupervised learning, there is another learning technique called semi-supervised learning, used with datasets with both, labeled and unlabeled data, avoiding time intensive and cost prohibitive process as giving labels to images. Usually works by using a subset of unlabeled data to train the model and then classify the remaining data, returning a labeled dataset to train other models, in theory it performs better than unsupervised models (Ouali et al., n.d.).

2.3.3.1. Generative Models

In machine learning daily applications, the engineers develop their intuition about the datasets, models, and their interactions. The study of raw data (samples, outliers, and classifications) is the first step towards the identification and correction of all the problems in the data, obtaining new models' hypothesis and assigning labels, but it comes with privacy problems, avoided with allowing access only to aggregated outputs. (metrics or parameters) (Ouali et al., n.d.).

A generative model pretends to find existent data issues without inspecting it directly, the engineer would use review the data manually to find errors, propose hypothesis to improve the labelling, making available just the final model parameters and statistics to the engineer (Pang et al., 2020).

The model works, after the manual inspection according to defined criteria, selects the procedure to build a training data set, for federated learning, only selects instances' subset to the model training and filtering the data (Ruthotto & Haber, 2021).

2.3.3.2. Self-Training

Self-training algorithms have the purpose of fitting labels got from another model, are successful learned from neural networks, linear models. Regarding a low probability data subset, it should be expanded to the neighbor having the largest subset relative probability, making assumptions about classes minimal overlap, achieving high accuracy in ground-truth labels (Wei et al., 2020).

It is an algorithm pretending to take advantage of unlabeled data with deep networks, extending to assure predictions' stability under perturbations and data augmentation. Proposed to use unlabeled data with supervised learning success, to obtain a well labeled dataset, generating unlabeled data all the time. The algorithm uses unlabeled data, labeling it according to structure and characteristics of the data, training unsupervised data with a supervised method (Livieris et al., 2018).

2.3.3.3. Transductive Support Vector Machine

Transductive Support Vector Machine (TSVM) is algorithm containing the strength regularization for the unlabeled data, to get the decision boundary, after obtaining a spectrum of models by changing the unlabeled data regularization strength, which translated is changing from supervised SVM to Transductive SVM. The optimal model is found by using the regularization assumption to enable a smooth prediction function over the data space, since the optimal function is a supervised models' and semi-supervised models' linear combination, effectively combining the cluster regularization assumption (Wang, 2005).

2.3.4. Reinforcement Learning

A technique called Reinforcement Learning, assigned to tasks without a predefined correct answer and overall outcome, works learning on a trial and error besides the data, simulating the human learning experience. It is applied in several areas, from medicine to video games training, since it is ideal to situations without a corrected inputs' sequence, just combinations leading to winning and failing, and the model reinforces the behavior when it is tried an input and it is successful until it gets any sequence leading to the entire tasks' goal (Choi et al., 2020a).

It is another case, that is neither supervised nor unsupervised learning, since it has some supervision, although not from the final output from the input data, the feedback is gathered from the environment after each received output, showing the percentage of goals' fulfillment. The learner sequentially interacts with an environment, acting according to the observations feedback of its previous actions (Choi et al., 2020a).

The machines are exposed to a sequential decision making environment, trial and error based from the past actions taken, it stores the information and receives a reinforcement from the correct actions (Sodhi et al., 2019).

2.3.5. Multi-task Learning

Multitask learning uses inductive transfers improving generalization with information contained in related tasks' signals training as inductive bias. The tasks are parallelly learned with a shared representation, each task results supports other tasks' results, with the goals of improving the learning tasks performance using the important information existent (Ruder, 2017). It is a technique used for different machine learning applications due to the capacity of learning to learn, and also learning helped by auxiliary tasks, when a loss function is optimized (Varghese & Mahmoud, 2020).

The algorithm's goals are to have the most accurate learned for every task, according to the useful information in the multiple learning tasks, it is based on the principle that the tasks, or a subset of them are related some way, using several empirically tasks to obtain a better performance in relation with independent learning. According to the tasks' nature, it is classified in multi-task supervised learning, for both, classification and regression problems, predicting new data labels, with training data instances labeled, multitask unsupervised learning, for clustering tasks, finding useful patterns in the training dataset, multi-task semi-supervised learning, including labeled and unlabeled data, multi-task active learning, actively query their labels using unlabeled data, multi-task reinforcement learning, aiming to maximize each task reward, multi-task online learning, for data sequentially ordered with each action and multi-task multi-view learning, dealing multi-view data with multiple features' set describing each data instance (Dobrescu et al., 2020).

This technique is a form to replicate the human learning, by the knowledge transfers between related tasks, making it useful the learning of multiple related learning tasks, it relates to other machine learning areas, like transfer and multi-label learning (Ruder, 2017). Multi-task learning is a generalization of multi-label learning and multi-output regression, some cases, with a larger number of tasks or different tasks assigned to different machines, leveraging the need for parallel and distributed multitask learning models (Varghese & Mahmoud, 2020).

2.3.6. Ensemble Learning

To improve the predictive performance of a statistical method or a learning technique it is possible to use ensemble methods. These methods have the main principle of building a linear combination of model fitting methods, instead of a single fit. In other words, it groups models together to work on solving a problem, instead of looking the best solution, avoiding the weakness of each methods and conveying the advantages of them, making it a less error prone collection (Ganaie et al., n.d.).

Several weak models combined lead to a strong learner, combining, either in a homogeneous ensembled model with a single base learning constant across all models, or a heterogeneous ensemble model, a multiple base learning algorithm differing for each model. Most time, it is used to help achieve regularization with decision trees, to help the reliability of prediction and the stability and robustness of the models (Faußer & Schwenker, 2013).

2.3.6.1. Boosting

Bosting models combine, in an iterative process, base hypotheses (thumb rules) to obtain the prediction. The base hypothesis is generated from a base learner and combined linearly. If we are dealing with a two-class classification problem, it obtains the final prediction by majority, combining the rules to improve the performance. A model related with support vector machines and large margin classification problems, used on high dimensional data sets and adapted to many different examples (Rätsch, n.d.).

The algorithms are developed as ensemble methods, happening sequentially, since the weights are dependent on the previous obtained functions, leading to very accurate classifications (Peter, 2015).

2.3.6.2. Bagging

Bagging models are ensemble methods used to improve estimations and classification problems, it is known as a technique to decrease the variance of certain base procedures (decision trees, variable selection, and linear model fitting methods). The main advantages of the method are the simplicity and the bootstrap methodology popularity (Kotsiantis et al., 2005). A technique working smoothly, giving advantages in terms of regression or classification trees' performance, it also works to reduce the variance and mean squared error (MSE) of decision trees. There is a version cheaper, in terms of computational costs, called by sub bagging (subsample aggregating) (Peter, 2015).

2.3.7. Neural Network

Neural Networks are the most common approach to classifications problem, are computational models based on the nervous systems and the neurons' connectivity happening in the different system types. The Universal Approximation Theorem increased their popularity, works by a neuron, computing its' input weighted sums, and perform a linear function of them. To use nonlinear functions, the network computes the functions, approximating it with mappings of the training patterns to the training targets (Rätsch, n.d.).

2.3.8. Instance Based Learning

Instance based learning algorithms get to know how to train the examples by generalization, according to measures of similarity to the new instances, using the training instances to build hypotheses, can be called memory-based or lazy learning. The algorithm used the sequences of instances given as inputs, each one having a set of pairs attribute-value, with the instances having an equal set of attributes, in a dimensional instance space, with one attribute being categorical (all instances in the instance space with the same category attribute value) and the remaining are the predictors (Graph et al., 2021).

The algorithm has the setback in the moment to decide the instances that should be stored in the generalization phase, too many instances stored lead to high memory requirements, decreases in the execution speed and oversensitivity of noise, working using the original instances as exemplars. In the generalization phase, a distance function is used to determine the closeness between a new input vector and each stored instance, and the newly vector output class is predicted with the nearest instance(s) (Mccallum, n.d.).

A set of classifiers is used to approximate to the target function, instead of doing an entire estimation, making the adaptation to the new data much easier, with higher classification costs and larger memory amounts required to adapt to the new data. It has a time complexity according to the size of the training data, with the worst case of O(n), having n as the training instances' number (Fontana, 2008).

2.3.8.1. K-Nearest Neighbor

K-Nearest Neighbors (KNN) is a clustering algorithm, effective for classification and regression problems, works by collecting data and group it into clusters or subsets, organized according to the data characteristics and the new inputs are classified according to the previously trained data, assigning it to the class with the nearest neighbors (Anava & Levy, 2016).

KNN is an instance-based learning algorithm, since the goal is not the construction of an internal model, the training data instances are stored and the prediction is, then obtained from the most voted nearest neighbor at each point. The algorithm does not make assumptions on the provided dataset, being nonparametric, it provides a training dataset labeled, categorizing the points according to the different classes and predicting the unlabeled data (Anava & Levy, 2016).

The algorithm's main use is the data classification according to the closed or neighboring training data examples in the region, for each new input, calculates the k nearest neighbors and the neighboring data majority classifies the new input, deciding the k value, running the classifier several times with different values to obtain the most effective results and use it in datasets with data in different clusters (Y. Wu et al., 2002).

The classifier functions in a two-step phase, the learning step, that use the training to classify the new data, and classifier assessment, classifying the unlabeled data and determine to which cluster they are included in. The accuracy increases with the separation of the inputted data into different classes, finding the one with the most numbers of points with the least distance of the data point in classification. The defined k defines the number of neighbors, and the classifier defines the new data distance from its neighbors, obtaining its class, with the maximum nearest neighbors. The value of k, increased the precision of the classifier and reduces the noise. If k is predefined as 1, the data is defined as belonging to the nearest neighbor class, with a zero-training data error rare, because the nearest point to any training data is itself, obtaining the best results, considering, however over-fitted results. A good k values is obtained removing the training and validation dataset from the initial data, k increases with the definition of the nearest neighbor region and the smoothness of the boundaries (Taunk, 2019).

The main advantages of KNN are its simplicity, comprehensibility, and scalability, having huge predictive power, being extremely effective and efficient for large training data sets, without complex steps and simpler calculations. In large datasets, although there is a drawback of the expensive k determination, needing higher storage space comparing to an effective classifier and requiring more time to any prediction. It is classified as a lazy learning algorithm since the data use needs come from the new input classification instead of the training data (Taunk, 2019).

2.3.9. Machine Learning workflow

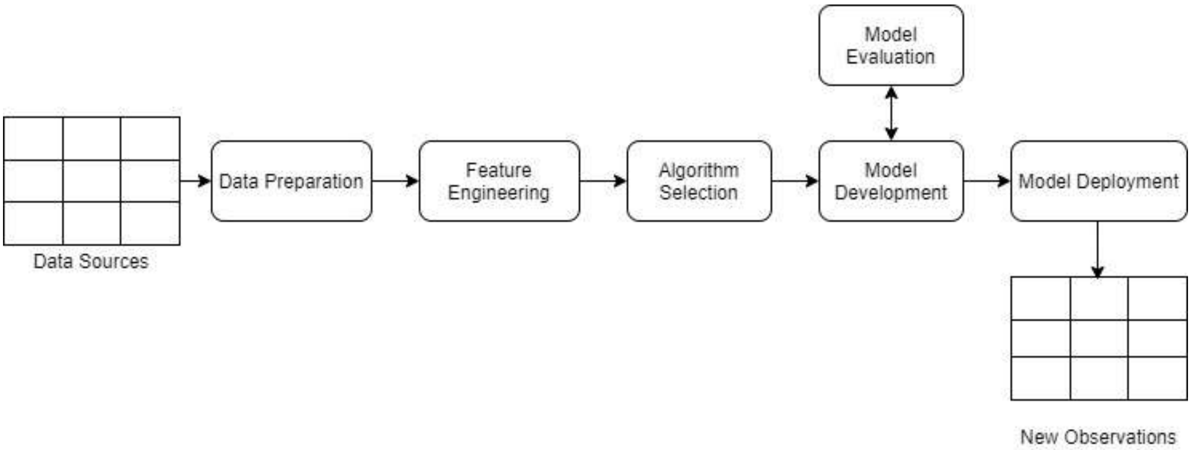


Figure 1: Machine Learning workflow

The workflow starts with data preparation and feature engineering, the ingestion of the data sources in the environment, cleaning of the data and transformations (normalization, scaling) needed to clean and validate the data to the model requirements. The model is chosen according to the problem specifications, it will generate a trained model, iteratively tested for accuracy with the test data, with the ideal accuracy value reached, the model is deployed and is ready to use by the application (Deelman et al., 2019).

2.3.10. Performance Evaluation

The performance evaluation of any algorithm tries to evaluate its works under the maximum effectiveness in unseen data, splitting the training dataset into a smaller one and a validation dataset. The process works by assigning evaluation metrics to each model, depending on the model itself and the phase that it is in (training or testing). The validation dataset has the same purpose of the test dataset, to tune an algorithm and verify when it works better in a new dataset, in generalization terms. The validation dataset is a population subset, since it contains part of a population, the generalizability and performance are not accessed by validation set performance, since it is impossible to create a validation dataset without bias. It has great performance in the validation dataset, but not in the test dataset since it is not generalizable (Choi et al., 2020a).

Each model performance is evaluated by the accuracy of the training and validation dataset, checking if the accuracy in the data is increasing or converging each training iteration. If training and validation data sets' accuracy converges and decreases, the model is not learning enough as expected about the features and target relationship and could be underfitting. If the training performance increases comparatively with the validation performance, the model is overfitting, learning due to population generalizability, since it is learning specific training dataset features. The main goal of the validation is tuning it iteratively, not being a good indication of model performance on unknown data (Choi et al., 2020a).

The ideal output of the training phase is highly generalizable trained model needing confirmation of the testing dataset. For supervised learning models, in classification problems the performance is evaluated according to the accuracy and in regression problems with errors and residuals. Since the test dataset has original instances that the model did not face in the training phase, if it has high predictive power on the training dataset, but poor on the test dataset, the model created is overfitting for training data patterns, which means it memorized patterns instead of learning a generalizable model. Contrarily, if a model underperforms in both, training, and test dataset, it is underfitting and means it neither has learned or memorized the training dataset and cannot be considered generalizable. The goal is to create a model performing well in training and testing dataset and generalizable for other datasets (Choi et al., 2020b).

For regression models, the average mean squared error (MSE) indicates the performance by measuring the closeness between the targeted and predicted value, the final value is obtained from the sum between all the predicted and targeted values squared and divided by the number of instances (Choi et al., 2020a). In case of binary classification models, the desired output is a class, with the probability of belonging to either class A or B, predetermined with a threshold at 0.5. The ROC curve evaluates the model trye positive rate (TPR)- sensitivity and recall,- meaning the instances correctly classified as positive divided by the total of positive samples, the false positive rate (FPR)- 1-specifity- meaning the incorrectly identified positive samples divided by the negative samples (Simeone, n.d.).

Another indicator is the precision recall curve to evaluate the quality of positive prediction value of a model (precision, recall) and the number of the correctly positive identified samples divided by the number of positive samples (Choi et al., 2020a). The evaluation of the curves is done a probability model threshold, with a range from 1 to 0, from left to right, the ROC (receiver operating characteristics) curve starts at TPR=0 and FPR=0, with a decision threshold of 1, classifying the samples as negative without false or true positives. The final curve's point is at FPR=1 and TPR= 1, having a decision threshold of 0, with all samples as positives, being the points either true or false positives. Between the curve extremes, the points are obtained calculating TPR and FPR for different threshold between 1 and 0, pretending to minimize false negatives (sensitivity) with false positives (specificity). The existent area under the ROC curve (AUROC) evaluates a classifier overall performance, taking as an assumption a balanced dataset, contrarily, the area under the precision-recall curve (AUPR) is preferable for overall classifier performance evaluation, since it is an adjusted threshold (Choi et al., 2020b). If a dataset contains ¾ of class A and ¼ of class B, the defined threshold is 0.75, with an AUROC of 0.5 indicating a model performing like a coin toss, and an AUC of 1.00 indicating models performing perfectly, meaning that higher the AUC, the model performance will be stronger, and an AUPR value at the threshold shows that a model is no better than a coin toss and at 1.00 indicates a perfect one (Choi et al., 2020a).

2.4. Deep Learning

Deep learning is a machine learning branch, building models with huge data amounts, with neural networks as base. Deep Learning models do not need any guidance to improve, it uses the feedback to adapt the weights to the input parameters. The artificial neurons in conjunction create a deep neural network, with a perceptron as a human brain replica, connected between each other forming a deep neural network, with input nodes, human brain dendrites replica, a decision function and output nodes (human brain axons) (Cockburn et al., 2018).

A deep neural network is formed by input layers, the first network layer receives and process input in raw form, transferring it, next to the other neurons' layers, intermediates according to the complexity of the problems, called hidden layers, can exist from one to hundreds of layers. The information is processed from the input, through the hidden layers until the output layer, which transfers the output next to the user (Jones et al., 2018).

An artificial neural network (ANN) is a biological neural network inspired algorithm, with nodes (replicas of biological cell bodies) communicating between each other through connections (axons and dendrites replicas), that are weighted upon the ability to reach pretended output, replicating biological neural networks synapses strengthen the neurons relations and their related outputs (Choi et al., 2020b). The output is reached, starting with set of inputs and multiply each other by the corresponding weight to sum their weighted combination and creating a non-linear activation function (Lecun et al., 2015).

Deep Learning models have been a revolution in the daily machine learning activities, for the most difficult actions, medical imaging Neural Networks working in tasks like injury detection, segmentation, text-to-speech and text-to-image, natural language processing, optics, image processing and computer vision. Being a powerful force improving fields like autonomous driving, face recognition, anomaly detection, text understanding and art technologies (Lecun et al., 2015).

To select the adequate feature with input instances having the pretended properties to solve a particular problem. In cases that the input space does not directly fill a property, it is possible to the map the instances in an intermediate feature space, with linearly separable classes. The feature space occupying the intermediate, could have features hand-coded, passing the design responsibility for the user, increasing the computational time and expert cost, mostly in highly dimensional input spaces or be implicitly defined by a kernel function, or learned the features automatically from deep architectures (with multiple nonlinear processing layers). (Lecun et al., 2015) The nonlinear functions representation is done by the deep architectures parameters numbers, with a proved parity function to n-bit inputs to be in a hidden layer of O (log n) complexity and O(n) neurons in a feed-forward neural network, instead of just a single hidden layer feed-forward neural network needing an exponential neurons number for the same task. For the highly varying functions, the learning algorithms local generalization based suffer with the dimensionality, using distributed representations as an alternative (Arnold et al., 2019).

Deep architectures are hard to be trained since the classic methods suffer a lack of effectiveness when they need to be adapted to deep architectures. The addition of a layer does not improve the results since higher the number of neural network layers, decreases the first layer's backpropagation impact. A descent gradient stuck a local minima, making the preferred choice the neural networks limitation in a maximum of two hidden layers and the addition an unsupervised layer deep architectures pre training. Meaning that, every layer is successively greedy trained, after the previous ones, and a new one is trained according to the previous layer data input. Afterwards, performing a supervised tuning of the entire network (Arnold et al., 2019).

In the deep learning area, a perceptron receives, as inputs, several features, and their targets, aiming to discover a line, plane, or hyperplane, dividing the existing class into a highly dimensional space. It is used a sigmoid function, to allow class association, instead of the probability of an instance belonging to a class. While connected, the model is called multilayer perceptron algorithm or artificial neural networks. It contains a layer with input nodes, another with output nodes, and a hidden layer between the previous layers (Emmert-Streib et al., 2020).

2.4.1. Feedforward Neural Networks

Most tasks requiring artificial neural networks, use them to feed the information forward, moving the information from one node in the current layer to each node in the next layer, and to the next, successively. Defined as feedforward neural networks are the less artificial neural networks, working with the input data travelling in one direction, using the input nodes to enter the layer and the output nodes to exiting it, sometimes there is no hidden layers (*A Brief Review of Feed-Forward Neural Networks*, 2014).

A neural network has a set number of nodes, with the output layer number of nodes being the number of predicted classes, for multiclass classification problem, in regression problems, is used a linear activation function, and, in binary classification problems, it is used a single node with a sigmoidal function. The mentioned activation functions convert a node's input into the pretended output, with every node having a nonlinear and noncomplex activation function, applying a several transformations, a linear to inputs ≥ 0, and setting inputs < 0 equal to 0. The inputs are modified each layer and at the final one, no longer look similar to the original state, and the final state is the best prediction of the desired result (Dai et al., n.d.).

The feedforward ANN has some disadvantages, in some tasks, like image recognition, each ANN is an image pixel, longer from the ideal scenario without layer node connections, losing the image features spatial context (in images, close pixels are most correlated than distant pixels), a fact not considered by feedforward ANN (Yamashita et al., 2018).

2.4.2. Convolutional Neural Networks

Convolutional Neural Networks are known for the image spatial relationship preserving, with every single pixel as input, feeding image patches to every next layer nodes (instead of all existing), keeping the spatial context where the features are extracted, the nodes' patches extract specified features and are known as convolutional filters (Yamashita et al., 2018).

This type of neural networks is mostly used in image processing, with the goal of, blur or sharpen images and edge detection. Since a digital image is a grayscale image matrix or three stacked color (red, green, and blue) matrices, which each contain a value between 0 and 255, in representation of the image pixels and each channel color intensity. A convolutional filter is a small square matrix, that goes through the image, and performs every element matrix multiplication, at each position. Maps the output in a new feature map matrix, with a value indication if the convolutional filter is defined as a feature of interest (Yamashita et al., 2018).

For each Convolutional Neural Network, the filters' training goal is the specific features extraction from images and the feature map identification of their location. A deep CNN maps the features as next layers input and using new filters to create the next map of features. A process repeated for every layer, becoming abstract extracted features, and making high useful predictions. The final feature map is obtained using their compressed square representations as input to a feedforward ANN, classifying an image according to the obtained features and textures. It can also be used in image segmentation tasks, object identification in an image and individual pixels classification according to the existent identification (Indolia et al., 2018).

These neural networks are defined as deep architectures with known advantages as obtaining visual inputs good generalization, great in the recognition of digits, are also inspired in biology, pretend to classify the cortical cells, and extract specific information from the orientations, their information and composition. It is based on the principle of combining local computations and pooling, transferring invariance, having weights dependent on spatial separation, without positioning concerns. The goal of Pooling is to use a nonlinear combination of the previous level features to build a features' set, depending on the unput data topology. With the alternance between convolution and pooling layers, local features are extracted by the network to achieve a representation of the input. The CNN connectivity, used every unit in a convolution or pooling layer to connect it to preceding layer subset, training networks with the maximum of seven hidden layers (Albawi & Mohammed, 2017).

2.4.3. Radial basis function Neural Network

Radial basis function neural networks take into account the points' distance to the layer's center, with two layers, in the inner layer containing the features combined with the radial basis function, and memorizing the features' output for the next layer's computed output (Article et al., 2016).

This type of networks main use is in artificial neural networks assigned to approximation problems, with advantages in learning speed and the universal approximation. It is composed by three layers, specifically the input, the hidden and the output layer, having each specific tasks assigned. The training goal is to reach the calculated error desired value and the iterations numbers, to, then chose a specific nodes' number in the hidden layer (Sanjeev et al., 2016).

2.4.4. Kohonen Self Organizing Neural Network

Kohonen Self Organizing Neural Networks uses input vectors to create discrete neurons map, create its own data organization after training, with at most two dimensions. In the map training process, each neuron location is constant, however with different weights according to value. The self-organization process has different steps, a neuron is initialized with a weight and the input vector, the closest neuron to the desired point is chosen and the ones connected will be moved towards that point. The method chosen to the calculate the neurons distance to the points is the Euclidean and the neuron with the smallest distance is selected. In every iteration, it occurs a point clustering, each neuron representing a cluster. It is very useful in data pattern recognition, mainly by categorizing the data (Lobo, n.d.).

The Self-Organizing Maps have significant differences compared to Artificial Neural Networks, expressed in the architecture and the algorithm, since the structure is a two dimensional single-layer linear grid of neurons. The grid nodes are directly related to the input vector, instead of to one another, the neighbors in unknown to each node, updating the connections weights according to the given inputs. The grid is a self-organized map, working as input function of the input data every iteration, after clustering, each node has its own coordinates, to use the Euclidean distance formula to calculate the nodes distances with the Pythagoras theorem (Miljković, 2017).

This Neural Networks uses competitive learning, adjusting the weight, activating every single node at each iteration, the instance features are used as input vector of the present neural network, all nodes trying to answer to the input. The desired node choice is done by the current input values and the grid nodes similarity (Miljković, 2017).

After the calculations, the smallest Euclidean distance obtained between the nodes and the input vector is used to choose a node, with their neighboring nodes with a defined radius, to adjust the position according to the input vector. This process happens for all the grid nodes, to obtain a grid entirely matching the input dataset, clustering similar nodes in different areas (D. Wu et al., 2018).

2.4.5. Recurrent Neural Network

Recurrent Neural Networks pretend to save the layer output and return it as input to predict the next layer outcome. The first layer is built like a feed forward neural network, multiplying the weights sum and the features. After it is computed, each neuron remembers the previous step information, acting as a computing memory cell, and the neural network performs the front propagation and saves the latter needed information. For wrong predictions, both learning rate and error correction perform small changes towards the right prediction in the back propagation phase (Sherstinsky, 2020).

RNN is an artificial neural network using sequential and time series data, for ordinal and temporal problems, in language, speech and image, learning from the training data. Have the advantage of having previous input information influencing the current input and output, instead of having assumptions of inputs and outputs being independent of one another, the output is dependent of sequence prior elements. Next events are useful for an output given sequence output determination, except for unidirectional recurrent neural networks (Lipton, 2015).

2.4.6. Modular Neural Network

Modular Neural Networks contain different networks independently working to reach the same output goal, every network has a unique input set, different from the remaining networks with sub tasks assigned, not interacting between them to perform the tasks desired goal (Kamel & Raafat, 1996). Compared to other artificial neural networks, breaks the computational processes into small components, reducing the complexity, posteriorly, the connections numbers and remove the networks interactions, improves the computational speed, making the processing time dependent on the neurons number and their effect on the results' computation (Mining & Wasilewska, n.d.). This artificial neural networks type is mainly known for their problem solving approach, using several neural networks working as a module to solve a problem part, benefiting from modules response integrations to get each output and integrate them all into a final output (Xue et al., n.d.).

2.4.7. Belief Networks (DBNs)

Deep Belief Networks are known as a problem solver in neural networks training in deep layered networks, like slow learning, poor parameter selection and several dataset training requirements. The first introduction was as probabilistic generative models, as alternatives to traditional neural networks discriminative nature. Generative models obtain input data and labels probability distribution, easing their estimation (Hua et al., 2015).

A generative model contains stacked Restricted Boltzmann Machines modules, an energy based two layers visible model, hidden layers, and layers' connections. The modules are trained unsupervised once at each time, in a divergent procedure, with the learned features, which means each stage output working as the subsequent input stage. The entire network is supervised learning trained in a tuning method to improve the performance of the classification, training in a greedy-layer manner using weight tuning abstraction of the hierarchical input data features, the design pretends the design a distribution in the input and hidden layers' space, to obtain direct connection in layer nodes and indirect connections at the upper layer nodes (Emmert-Streib et al., 2020). The training is done layer wise with the weight parameters being adjusted in convergence establishing balanced learning probability estimates. The input samples conditional probability distribution is obtained by the abstract features learning, not being affected by transformation and noise due to their robustness (Hinton & Osindero, 2006).

2.5. Data Science

Data Science is a discipline focused on getting to know the origin, representation, and transformation of information into a business and IT strategy resource, it can be done with the application of different machine learning models, tools, and algorithms to execute the several tasks needed for extraction of information and knowledge. Involves several disciplines, from statistics to data modeling, data analytics and algorithms, pretending to research and optimize companies business processes (Zhu & Xiong, 2015).

A Data Scientist works in the data exploratory analysis, pretending to gather business insights, and applies several machine learning algorithms to discover future occurrences of different events. Its goal is to give the best tools to the decision makers by predicting future events' occurrence, building decision maker models that can be modified by dynamic parameters, suggesting actions and respective outcomes, pattern discovery to make predictions, training machines using transactional data (Mesquita, n.d.).

The typical data science project workflow includes objective identification, importation of data, data pre-processing (exploration and cleaning), modelling and application of the model (Zhu & Xiong, 2015).

2.5.1. Objective Identification

Any data science project starts with the identification of the objective, it needs to be clearly stated the business problems that needs to be solved, without any tasks related with modelling and evaluation metrics, just defining the company problem and the fields in which is needed to be gathered more information. It can only happen after a clear definition from the client, who can be a group or individuals in the organization or even outside the organization, needing their output for the work. The problem is a need from the client, however it should be discussed with the data scientist or data science team, since due to some inexperience and hype around data science, the organizations could lean towards thinking about using the data in some way without desired goals and the requesting data science projects would not be feasible and not provide any business value. Prior to the entire problem definition, it should be made any consideration on the existing data, otherwise, should define the needed data (Zhu & Xiong, 2015).

2.5.2. Importing Data

Next to the objective definition, there is time study the existent data, and collect the data to analyze, should be collected the most data with different types and variables as possible to make the project easier, is beneficial to have data not used than needing data not stored (Zhu & Xiong, 2015).

This step faces setbacks related to the tracking of the data origin, in the tracking of the different sources, since usually there is a need for reacquiring the data for different experiments, with updated data sources or to test new hypothesis. Another topic to be considered in this step is data management, all the files should be clearly named and organized, besides that when it is collected or created another version of the data, it needs to be properly named to immediately know their main differences. Data storage is also a subject to consider, since it must be used remote servers, due to the limitation in memory of hard drives (Zhu & Xiong, 2015).

The different types of data available to be import are structured data, the usual way of thinking about data, structured tables, with clearly defined variables and their values, variables as the columns and rows as the cases, more accessible and used for statistical and machine learning models training or unstructured data, everything that is not structured, like video, audio, images, text, websites (Mesquita, n.d.).

2.5.3. Data Pre-processing

Data pre-processing treats the data's needs to be prepared to be used in machine learning or statistical models, it the first step of the proper data scientist work, and the one requiring the most time. It involves cleaning data, removing inaccurate and wrong values, and making the data homogeneous, treating different subsets in a specific form, and labeling nominal valued variables, which requires domain knowledge. This happens since real life data sets are incomplete, and the algorithms cannot work with incomplete data, giving the data scientist the options of completing data, estimating the missing values or remove it (Zhu & Xiong, 2015).

2.5.4. Data analysis

Previous to the modelling of the project solution, there is an important step that should be taken, a complete data analysis should be performed, leading to several benefits, build familiarity with the data and start the insights extraction, when this step is not performed could lead to the generation of inaccurate models and the use of insignificant variables in the modulation (Mesquita, n.d.).

Data analysis is a process happening iteratively to become closer to a solution, having a cost associated with each iteration. Pretends to help in the choice of the features to be used in the model, the most important in the data characterization. For unstructured data, the data analysis process decides the extracted features from the data, with machine learning techniques to avoid a time-consuming and difficult feature extraction process (Mesquita, n.d.).

The features, either in structured or unstructured data, could not be the most useful to use in the model, but be useful for another, could have quality or redundancy problems, so it comes the importance of the selection of the maximum informative features. Using feature selection techniques will help to find the appropriate set of features to use in the model, taking an approach of transforming the original features into a small number of highly informative features (Zhu & Xiong, 2015).

2.5.5. Modeling

The modelling data science phase starts with the gathering of the possible techniques for the solution achievement, enumerating the needed steps and calculations to achieve it and choosing the technique according to the requirement. After the data creation, a set of models is built, to discover the best performing approach according to the data. The main restrictions affecting the choice of the best selecting model are the problem type, the data available, and the pretended model characteristics (Weihs & Ickstadt, 2018).

2.5.6. Model application

Since the selection and evaluation of the model is performed, it needs to be evaluated outside the scope, in cases of a model with high performance, it can be used for the tasks that lead to the its design, which needs access to the new data, achieve through building an application, making the results accessible to humans and programs (Weihs & Ickstadt, 2018).

2.5.7. Results presentation

The remaining step is the presentation of results, everyone involved and with responsibility in the organization wants to know the results, so the different stakeholders can hear the conclusions and the critical findings. Since the explanation is done to a non-technical audience, the data scientist should work on their communication skills, specifically simplifying the technical language, making a simple and understandable presentation to help the creation of a business plan from the data science project (Zhu & Xiong, 2015). The data scientist must gather some information about the audience, adapt the language, mention the values, outcomes and advantages, the data origin, mentioning some credible sources to build confidence in the results and state the assumptions and limitations encountered (Weihs & Ickstadt, 2018).

2.6. Automated Machine Learning

Automated Machine Learning was born due to the need to ease the application of machine learning techniques, while do not increasing the number of subject matter experts. It pretends to remove the human from the implementation, speed up the deployment of machine learning application in the organizations, making the organization benefit of an effective validation and deployed solutions benchmarking and keep the subject experts focused on the machine learning applications and the business value added by them. Implementing automated machine learning in any organization, would exponentially increase the availability of machine learning, having a significant business impact, by helping the organization reach new competence and customization levels (Gijsbers et al., n.d.).

It is considered a subfield of machine learning, with proven applications both in machine learning and computer vision, using different libraries to ensure the availability of the correct classifier to each learning problem, having a collection of classifiers available to deal with a new problem and obtain a final prediction, searches for correct model and their ideal hyper-parameters to, instantly find correct models from the machine learning classifications tools, without the intervention of the human (He et al., 2021). Neutral architecture search, is the design of main performance improvement source in the deep learning, it is essential to automated the design of neutral architectures to obtain an acceptable learning performance and automated feature engineering, influencing the features' quality, really important for the model performance, obtaining a new set of features, leading to the improvement of subsequent machine learning models' performance, instead of manually performing the feature selection tasks by humans by trial and error with just human knowledge (Yao et al., 2018). It can perform automated feature engineering, another booster of the learning performance, having a more extensive set of features, with the use of the experience and the models performance is obtained from the constructed features, automated model selection, with the input training data being heavily influenced by the previous experience, in classification tasks measured by the assigned tasks performance, with the use of libraries that select the ideal classifiers to find the correct hyper-parameters without human-support (Zöller & Huber, 2021).

Automated Machine Learning is created from automation, the use of different methods of control operating underneath components, and machine learning, the computer programs learning from experience according to tasks evaluated by their performance, resulting in Automated Machine Learning being the machine learning configurations tools adapted to the input data task, with a proper generalization performance on the given tasks and input data (Yang & Shami, 2020). Pretends to have high level controlling approaches over the underneath learning tools to obtain the ideal configurations, get a good learning performance (the machine learning's objective) with the least possible assistance from humans (the automation's perspective). A simplified AutoML description can be simplified like an attempt to obtain a partial identification of machine learning configurations, with least possible human assistance and within the defined budget (computational and monetary). The main differences comparing to the classical machine learning are related with the human involvement in the feature engineering, model selection and algorithm selection tasks (Gijsbers et al., 2019).

The main goals are the obtention of a good generalization performance across different input data to be used in different learning tasks, the removal of human intervention, the automatic obtention of the ideal configurations for machine learning models with high computational efficiency and the return of the best possible output, all this with the lowest budget possible. This leads to the decrease the time to deploy machine learning solutions across the organizations, the validation and benchmarking of deployed solutions, releasing humans from these tasks, shifting their focus to tasks regarding problem definition, data collection, deployment, and the accessibility of the machine learning application (He et al., 2021).

The precursor process lets human tune the configurations, and when a learning problem is defined, finds some learning tools to handle it, which target different parts of the pipeline and to get a good learning performance, a configuration will be set using the personal experience with data and tools. According to the learning tools performance feedback, changes the configurations to improve the performance, ending the process with the achievement of a desired performance or without computational budget (Bezrukavnikov & Linder, 2021).

2.6.1. AutoML Framework

A framework for AutoML has following conditions: a controller replaces the human intervention, trying to find ideal learning tools configuration. Including an evaluator to measure the performance of the learning tools, receiving, from the optimizer the proper configurations and returning the feedback, trains the model based, in a time-consuming process, on the input data or directly from external knowledge, simulating human experience (Real et al., 2020). The evaluator measures efficiently and accurately the configurations' performance, updating it and generating it for learning tools, with a search space determined by the target learning process, improving the configurations' performance after each iteration, offering feedback, according to the optimizer's type chosen according to the search space and the defined learning process, to be able to choose and apply the most adequate optimization methods, preferably generic and efficient (Gijsbers et al., 2019). The approach to choose the problems to optimize passes for general learning problems, uses the full scope of machine learning applications, feature selection and engineering, model training and selection and hyper parameters choice to select the best model. The listing of Neural Architectural Search is different, since it targets deep models, configuring features, models, and algorithms simultaneously (Balaji & Allen, 2018).

To solve an AutoML problem there is a set of techniques, divided in basic and experienced techniques. The basic techniques, work with the optimizer and evaluator as ingredients, categorized according depending on the targeted ingredient, the optimizer is focused on the search and optimization of the configurations, with different methods (grid search, random search, and reinforcement learning) and the performance evaluation measures with the current configurations' parameters. The experienced techniques accumulate the knowledge from previous experiences and existing data to enhance the optimizer and the evaluator, in methods like meta-learning and transfer learning (Zöller & Huber, 2021).

2.6.2. Data Preparation

The Machine Learning pipeline starts with the preparation of data, that is divided in three steps, data collection, essential to create a new data or extend an existing one, data cleaning, the process of filtering junk data to ensure the correct model training afterwards, and data augmentation, to improve model robustness and performance (He et al., 2021).

2.6.2.1. Data Collection

To every Machine Learning process, must be quality data available and it can be obtained by Data Searching and posterior, Data Synthesis. Data Searching, taking advantage of the Internet as a giant data source, it can be used to search for data to create a dataset, however there are some drawbacks, the search results must match the given keywords, must be filtered unrelated data, the labels need to be checked and reviewed, could be done by a learning-based self-labeling method, the dataset imbalance, solved using a synthetic minority over-sampling technique (SMOTE), synthesizing a minority sample from the current samples, instead of up-sampling minority or down-sampling majority samples or using boosting with data generation to increase the model adaptability to different data sets (Elshawi et al., 2019). Data Synthesis to generate data, with a data simulator as one of the most used methods, using a simulator matching the real word as closely as possible, applying a reinforcement learning-based methods to optimize the data simulator parameters, controlling the synthesized data distribution. It can be used a technique that derives synthetic data called GANs (Generative Adversarial Networks), that generates images, tabular and text data (He et al., 2021).

2.6.3. Data Cleaning

The obtained data certainly contains noise, which can negatively affect the model training, and raises the need to carry the data cleaning process. The process usually requires specialists knowledge, however due to the limited access to the specialists, shifting the effort from crowdsourcing to automation (Elshawi et al., 2019). To obtain a more efficient there was some proposals of cleaning just a small data subset and keep comparable results to the full dataset cleaning, although it requires a data scientist to be responsible for the design of the data cleaning operations that need to apply to the dataset. There are several libraries to do this process, BoostClean automates the process, solving the problem as a boosting one, the data cleaning operation is added to the input downstream ML model, using boosting and feature selection, with several cleaning operations for a generated ML model with better performance, AlphaClean turns a problem related with data cleaning into a hyperparameter optimization problem, increasing the automation, the final combinatorial operation has different data preparation operations to be executed in the search space. There are also some proposals regarding the continuous process of cleaning data, mainly with workflows that can learn from cleaning tasks previously done (Yao et al., 2018).

2.6.4. Data Augmentation

Data Augmentation works as the regulator avoiding the model training overfittings, the different techniques can be grouped according to the data type (image, audio, and text). Regarding image data, the affine transformations are divided into rotation, scaling, random cropping, and reflection; the advanced transformations include random erasing, image blending, cutout, and mix-up; the elastic transformations involve contrast shift, brightness shift, blurring, and channel shuffle, and there is noise, GAN technique and neural transfer for the neural-based transformations. The textual can have two data augmentation approaches, data warping, by generating samples by applying transformations to the data space, and synthetic over-sampling, generating samples in the features' set, by synonyms' insertion or foreign language translation (He et al., 2021).

2.6.5. Feature engineering

Since the quality of the data and the features are determinant to the models' performance, feature engineering tries to maximize the feature extraction from the data in raw format to the use in algorithms and models, maximizing the gain of information, it consists in feature transformation and feature construction. Feature transformation includes feature extraction and feature construction, create new features set, with feature extraction to decrease the features dimensionality with mapping functions and feature construction expanding the original feature spaces. Feature construction proposed to decrease the redundancy by the selection of important features (He et al., 2021).

2.6.6. Feature Selection

Feature selection builds a features' subset from the original set, deleting the irrelevant and redundant ones (usually, divergent, and highly correlated values), simplifying the model, decreasing the risk of overfitting, and improving the performance (Yao et al., 2018). It is an iterative process, starting generation of a new features' dataset according to the defined search strategy and the evaluation if the criteria are met and there is no need to execute another search.

There are different types of algorithms involved in the search strategy, complete search, using exhaustive and non-exhaustive searching, with techniques breadth-first, branch and bound, beam, best-first, heuristic, with sequential forward selection or sequential backward selection adding the features from an empty set or removing from an entire set, bidirectional, using the previous algorithms to get the same subset and random, with simulated annealing and genetic algorithms (Elshawi et al., 2019).

The subset evaluation can be done with the filter method, scores features according to their divergence or correlation, selecting features depending on a threshold, calculating the variance, correlation, chi-square and mutual information, with the wrapper method, classifying the sample with the selected subset, and the accuracy measures the feature subset quality, the embedded method, performing variable selection in the learning procedure (Zöller & Huber, 2021).

2.6.7. Feature Construction

Constructs new features coming from the initial data, for model robustness and increase the generalizability for different dataset types. Pretends to features' dataset representative ability, in a process needing intervention from human to use techniques as standardization, normalization, or feature discretization (Elshawi et al., 2019).

There were proposed some automatic feature construction methods, helping with the improvement of efficiency, the searching process automation and operation evaluation combination and to show the achievement of results as good or superior of the ones achieved with human expertise. There is a set of feature construction methods, like decision tree-based and genetic algorithms, needing a predefined operation space, contrarily to the annotation approaches, since it uses domain knowledge and the training back relate to the protocol of feature space construction, the learner can identify inadequate feature space regions and using existing sematic resources to add descriptiveness, after the constructing of a new feature, the new feature is evaluated with feature-selection techniques (He et al., 2021).

2.6.8. Feature Extraction

Feature extraction is a process attempting to the reduce the dimensionality, with mapping functions, extracting features, avoiding the non-informative and redundant), according to defined metrics, with approaches like principal and independent component analysis, dimensionality reduction, and linear discriminant analysis (Bezrukavnikov & Linder, 2021).

2.6.9. Model Generation

AutoML divides this step into search space, the structures designed and optimized and optimization methods, split into ML models. The optimization methods contains training hyperparameters and model design parameters (He et al., 2021).

In the search space it is defined the design principles of neural architectures, it changes according to the scenarios, it can be defined different search space's types, one containing the tools' hyper-parameters, covering dimensionality reduction and feature enhancing methods, and another containing the generated and selected features, with operations made on the original features (Elshawi et al., 2019).

The architecture optimization method defines the guide to efficiently search to be able to discover the highest performance model architecture after the search space definition and the model evaluation method, after the model generation, needs to be done an evaluation of the performance, training the model to approximate to the training set, and estimate the validation set performance (He et al., 2021).

After the search space definition (entire-structured, cell-based, hierarchical), needs to be found the ideal architecture according to the performance, and there were proposed several methods to free this process from the human interventions and make it more time friendly, like evolutionary algorithms, a robust algorithm with broad applicability, inspired from the biological evolution, unlimited by the problem's nature, reinforcement learning, typically using a recurrent neural network, updating the sampling strategy according to each action performed, gradient descent, neural architectures search, grid and random search (Elshawi et al., 2019).

Instead of using the same hyperparameters set in the search stage, it is needed to redesign the set of hyperparameters, with grid search, random search, Bayesian optimization and gradient based optimization (Yang & Shami, 2020).

2.6.10. Model Evaluation

After the generation of the new neural network, the performance needs to be evaluated, according to intuitive methods to train the network convergence, with computationally expensive methods as low fidelity, weight sharing, surrogate and early stopping (Bezrukavnikov & Linder, 2021).

2.6.11. Automated Machine Learning tools

The goal of this paper is to study the better tools of Automated Machine Learning to apply in a banking environment, and there a lot of different tools that can be used to enable general workers to work on machine learning use cases and find optimal and efficient solutions for the problems presented, in this chapter it is described several automated machine learning tools description and a table comparing their characteristics.

2.6.11.1.TransmogrifAI

TransmogrifAI is an AutoML library written in Scala, running on Apache Spark, with the goal of speeding up the machine learning developer/data scientist productivity, leveraging machine learning automation, allowing compile time type safety, modularity, and speeding to build of machine applications, not needing degrees, just domain knowledge and modular and robust machine learning workflows. TransmogrifAI support an unstable data type detection, supervised models training, and hyper parameters choice with Bayesian search (Truong et al., 2019).

2.6.11.2.H2O-AutoML

H2O-AutoML is an automated machine learning tool designed for the advanced users, works by performing modelling tasks from H2O platform with a function. It is very useful in the machine learning workflow automation, in the automatic training, models tuning and several explain ability techniques for the application of different methods (Balaji & Allen, 2018). The advantages of H2O-AutoML are the automatic detection of numerical, categorical and time-series datatypes or schemas, high variety of supervised models, optimized hyperparameters choice, with parameter spaces random search and stacked ensembles random grid search, to achieve high accuracy levels (Zöller & Huber, 2021).

2.6.11.3.H2O-Driverless AI

H2O-Driverless AI pretends to help in the automation of analytical workflows, pretends to obtain the achieve quickly a reasonable predictive accuracy, while offering visualizations and models interpretability (Truong et al., 2019).

This tool provides the functionality of detection and processing time-series data with an interactive user interface helping the customers to quickly experiment different machine learning tasks, exports an object for model deployment in any platform supporting Java. Comparing with other tools, H2O-Driverless AI offers the possibility to detect basic (numerical, categorical and time-series) data types and schemas and from the extracted features supports different algorithms to model building, either for supervised and unsupervised methods, also using random and Bayesian search for hyper-parameter optimization. For the presentation of results, this tool provides model dashboarding, description of feature importance, visualization methods and outlier highlighting. The advantage of this tool is from the extended functionality offered in data processing, by detecting the schema of data, running feature engineering and in the interpretation of the model results (*Overview — Using Driverless AI 1.10.0 Documentation*, n.d.).

2.6.11.4.Darwin

Darwin is an open source automated machine learning tool, making use of Spark Cognition machine learning platform to build this tool cloud based, having access to model generation, data science automating process stages, data cleansing, data standardization, data filling, data homogenization, solve size imbalances, automatically generating the needed features according to the problem, to have the ones maximizing the important information to an optimized solution, in model optimization, using the faster supervised and unsupervised learning models. The machine learning process is all done in a faster speed, comparing to the usual data science pipeline, the prototyping and development of use cases, the implementation and tuning of machine learning applications, also helped by evolutionary algorithms and deep learning methods. The models are iteratively improved with several customizations, use of genetic algorithms and creation of deep learning models, easing the generation and maintenance of models (Truong et al., 2019).

Can be inserted in data science pipeline, in datatype detection stages, for numerical, categorical and time-series data types and schemas, however need intervention on feature engineering, to select the desired extracted features, to train different supervised or unsupervised learning models, or genetic algorithms, with previously trained datasets features, having a learner, with a dataset marked as target, to find a similar on the inputted meta features and use the best result as the initial model (Truong et al., 2019).

2.6.11.5.DataRobot

DataRobot is another open-source automated machine learning tool, created with the goal of allowing non expert workers finding the most adequate machine learning models to deploy in every situation, to give business decision makers the needed information for their work, uses SageMaker auto modelling capabilities to apply the best model on each problem, selecting through several models and parameters. DataRobot benefits from the faster detection and processing of time-series data, with an interface to try different machine learning models, besides that, allows the numerical, categorical and time series data types or schemas detection and uses the features in supervised and unsupervised methods (Truong et al., 2019).

2.6.11.6. Google Cloud AutoML

Google offered an Automated Machine Learning tool as a Service, on their cloud platform, with capability of developing model according to the business, neural network search execution in image and text, and neural architecture search to find the best models. All of this can be done by workers without any machine learning experience (*Cloud AutoML: Modelos de Machine Learning Personalizados*, n.d.).

2.6.11.7.Auto-sklearn

Auto-sklearn is tool based on scikit-learn, giving capability of applying supervised machined learning to non-experts, works by searching the correct algorithm and the most fitted hyperparameters (Balaji & Allen, 2018). Built upon the idea of having a machine learning framework that could be globally optimized, builds and ensembles tested models, with meta-learning recognizing almost equal datasets according to past knowledge, speeding the process (Gijsbers et al., 2019).

The creation was based on some AutoML principles, algorithm selection and optimization of hyperparameters selection conjunction to deal with classification and regression problems, composed by meta-learning, Bayesian optimization and ensemble construction (Balaji & Allen, 2018).

For data processing and feature engineering tasks executes features on the data and its inputted characteristics, converting categorical input into numeric data, and training supervised learning models with the extracted features, using Bayesian search for hyperparameters choice optimization and parameter space pruning with previous datasets learned preprocessed meta features (Gijsbers et al., 2019).

2.6.11.8.MLjar

MLjar is an automated machine learning python package destined to tabular, saving data scientists time and helping expert workers. It allows abstract machine learning models construction, in the data preprocessing and hyper parameter tuning to reach the desired model, also gives a report of the built pipeline generating a file with each model description. Has functionalities of automatic exploratory data analysis, to data explaining and understanding, algorithm selection and hyper parameters tuning, to obtain the fittest machine learning models, generating models analysis, saving, loading and analyzing the generated models. The tool contains different work modes, explain, understanding the data, with decision trees visualization, linear model's coefficient displaying, perform, producing ML pipelines, compete, ML models training and tuning, with ensembling and stacking, reaching the goals without computations limitations in terms of time and performance. This automated machine learning tool uses scikit-learn, along TensorFlow, being able to detect numerical, categorical and time-series data types or schemas, using the extracted data to train supervised models, applying, on the parameter spaces, random search for hyperparameter optimization (Truong et al., 2019).

2.6.11.9. Auto_ml

Auto_ml is an open-source python package, designed to be used in production and analytics, based on scikit learn, offers column categories segmentation with different models. It is an automated machine learning tool that executes feature engineering according to the user specifications, column data type, using the features extracted for supervised machine learning models, and random and Bayesian search for tool hyper parameter optimization (Balaji & Allen, 2018).

2.6.11.10. TPOT

Tree-Based Optimization Tool (TPOT) pretends to automate the ML pipelines construction, with genetic programming algorithms, build from scikit-learn package. TPOT is a data scientists' assistant to find the best pipeline, from thousands of explored ones, generating the most fitted Python code, finding solutions for different machine learning problems. Works as an evolutionary algorithm trying to obtain the best parameters and model ensembles, and offers code exportation (Balaji & Allen, 2018). When using TPOT it is required from the user to perform the data pre-processing tasks, accepting numerical feature matrices, and it also does not help in the feature engineering, the extracted features are used manually for different supervised machine learning and genetic algorithm models (Truong et al., 2019).

2.6.11.11. Auto-keras

Auto-keras is an open source automated machine learning tool, based on Keras, Tensorflow and Scikit-learn, using neutral architecture search, changing its own architecture without affecting the network functionality, Bayesian optimization for efficient neural network search, and provides end of pipeline documentation (Jin et al., 2019). The package starts by conducting image and text data neural network search, after requiring user manual data preprocessing and feature engineering user, pretends to find the best neural network for the existent features, and hyper parametrizes the model with random or Bayesian search (Truong et al., 2019).

2.6.11.12. Ludwig

Ludwig is an automated machine learning tool built from TensorFlow, to allow training and testing of advanced analytical models without subject knowledge. To run successfully it needs an inputted data file, with the input columns list and the desired output columns, training the models locally in distributed programming, has an available API having visualization tools with model training and performance analysis (Truong et al., 2019).

The main goal behind Ludwig's development was the use from non-expert personnel without coding skills, to the tasks of model training, obtaining predictions, and having generalized approaches to deep learning models design, allowing the use for several use cases, datasets and variables datatypes, allowing the user to control the model training, independently of the domain experience, extensibility, with an easy and understandable addition of model architectures and new feature datatypes, and the visualizations use to understand and compare deep learning models performance (Truong et al., 2019).

Ludwig contain different model architectures, used together to build end-to-end models for the use cases, the first step is the data preprocessing, next the feature engineering according to user input, using the results to the training of supervised learning models, using random and Bayesian search for the model parameters hyper-parametrization (Truong et al., 2019).

2.6.11.13. Auto-Weka

Auto-Weka is an open-source automated machine learning tool, with an interactive user interface, searches automatically the learning algorithm's space and its set of hyperparameters to achieve the best possible performance, with a Bayesian optimization method (Truong et al., 2019).

2.6.11.14. Azure ML

Microsoft AzureML is an automated machine learning solution available from basic machine learning to deep learning, either supervised or unsupervised, it is integrated with deep learning solutions as PyTorch, TensorFlow or scikit-learn, and available for both programmers or workers without programming knowledge (*MachineLearningNotebooks/How-to-Use-Azureml/Automated-Machine-Learning at Master · Azure/MachineLearningNotebooks · GitHub*, n.d.). Makes use of Azure's algorithms, providing tools for different experience levels, low code programmers, for people nonexpert in machine learning, Jupiter Notebooks and R scripts for programmers, model training modules, machine learning models operation and managing. With a drawback of data preprocessing and feature engineering dependent of the user specified inputs (Truong et al., 2019).

Tool	*Input data sources*	*Data Pre-processing*	*Data types detected*	*Feature engineering*	*ML Tasks*	*Model selection and Hyperparameter optimization*	*Quick start/ early stop*	*Model evaluation/Result analysis*
TransmogrifAI	(a)	Yes (however fails for some datasets)	(d)	(j)	(p)	(r)	(z)	(f1)
H2O-AutoML	(a)	Yes	(e)	(k)	(p)	(s)	(z)	(g1)
Darwin	(a)	Yes	(e)	(j)	(q)	(t)	(a1)	(g1)
DataRobot	(b)	Yes	(e)	(j)	(q)	(u)	(a1)	(g1)
Google AutoML	(c)	Yes		(l)	(q)	(v)	(b1)	(g1)
Auto-sklearn	(a)	No	(f)	(j)	(p)	(r)	(b1)	(g1)
MLjar	(a)	Yes	(g)	(m)	(p)	(s)		(f1)
Auto_ml	(a)	No	(f)	(j)	(p)	(r)		(g1)
TPOT	(a)	No	(f)	(n)	(p)	(w)	(c1)	(f1)
Auto-keras	(b)	No	(f)	(o)	(p)	(x)	(a1)	(h1)
Ludwig	(b)	Yes (however fails for some datasets)	(h)	(l)	(p)	(y)	(d1)	(f1)
Auto-weka	(a)	No	(i)	(o)	(p)	(r)	(z)	(j1)
Azure ML	(b)	Yes (needs column types)	(e)	(j)	(p)	(r)	(z)	(f1)
H2O-Driverless AI	(b)	No	(d)	(j)	(q)	(u)	(e1)	(g1)

Table 1: AutoML tools characteristics

(a)-Spreadsheet datasets
(b)-Spreadsheet datasets, image, text, in case of MLjar and H2O-Driverless AI the datasets must include headers
(c)-Image, text
(d)-Numerical, categorical, datetime, time series and other data types
(e)- Numerical, categorical, datetime and time series data types

(f)-Does not detect data types
(g)- Numerical and categorical data types
(h)- Numerical, categorical and time series data types
(i)- Numerical and categorical data types
(j)-Datetime, categorical processing, imbalance, missing values, feature selection, reduction, and advanced feature extraction
(k)- Datetime, categorical processing, imbalance, missing values, feature selection, and reduction
(l)- Imbalance, missing values, feature selection, reduction, and advanced feature extraction
(m)- Datetime, categorical processing, imbalance, missing values
(n)- imbalance, missing values, and advanced feature extraction
(o)- imbalance, missing values, feature selection, reduction
(p)- Supervised learning
(q)- Supervised and unsupervised learning
(r)-Ensemble, random and Bayesian search
(s)- Ensemble, random search
(t)- Ensemble, generic algorithm, and neutral architecture search
(u)-Ensemble, generic algorithm, random and Bayesian search
(v)-Generic algorithm, random, Bayesian and neutral architecture search
(w)-Ensemble and generic algorithm
(x)- random, Bayesian and neutral architecture search
(y)-Ensemble, random, Bayesian and neutral architecture search
(z)-Allow maximum limit search time, restrict time consuming combination of components
(a1)- Quick finding of starting model, Allow maximum limit search time
(b1)- Quick finding of starting model, allow maximum limit search time, restrict time consuming combination of components
(c1)- Allow maximum limit search time
(d1)- Quick finding of starting model
(e1)- restrict time consuming combination of components
(f1)- Model dashboard, feature importance
(g1)- Model dashboard, feature importance, model explicability and interpretation and reason code
(h1)- Model dashboard, feature importance
(j1)- Model dashboard

3. METHODOLOGY

After the search and discussion about the best methodological process in an autoML tool implementation framework, the best choice is a design science methodology, it was the fittest solution according to the requirements and the pretended works. The choice was made according to theoretical approach advantages, namely the design fundamental and scientific background analytical point of view combination (Baskerville, Kaul, & Storey, 2015). This master thesis pretended output is a roadmap theory to automated machine learning implementation in the banking sector, a known way to object knowledge classify from Design Science Research, without any physical output, either product or service, just a conceptual output, a design theory, construct, method, model, design principle or technological rule (Gregor & Hevner, 2013). The theory is built according to a problem-solving approach, with the first step defined as the business needs identification, to define the most suitable roadmap to a product implementation in an organization. The work conclusions are used as an implementation plan to construct a solution for the organization work (Hevner, March, Park, & Ram, 2004).

3.1. DESIGN SCIENCE RESEARCH

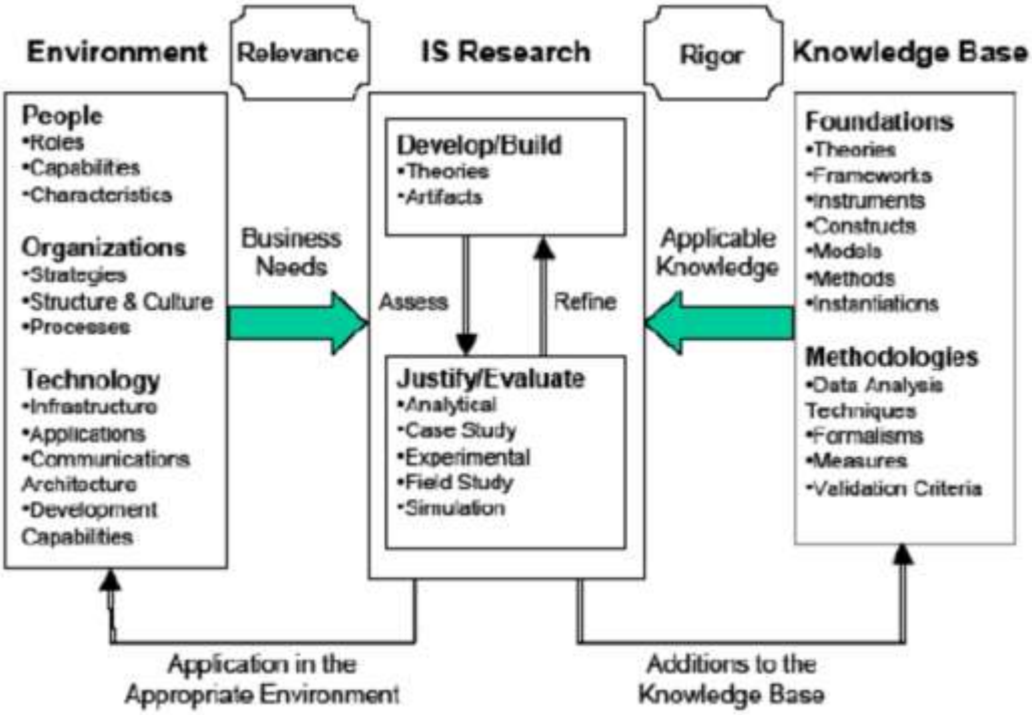


Figure 2: Design Science Research diagram

Design Science Research is a research methodology for creating an innovative solution, according to identified organization's problems (Hevner, March, Park, & Ram, 2004), particularly, this master thesis pretends to create the implementation roadmap of an automated machine learning tool in the banking sector.

The solution creation occurs through scientific research ensures the proposed theory coherence and credibility, and it is essential an effective communication of the final output, to ensure a correct use in the future (Hevner, March, Park, & Ram, 2004).

Some assumptions are made behind the methodology application, it is not creating anything existent currently, coming from an existent created idea or previous idea (Gregor & Hevner, 2013). The result is adapted and improved from an existing solution, changing the characteristics and the context, in this field, would consider the knowledge base inputs, specifically the workforce characteristics, organization strategy, structure, culture and process, and the knowledge on the machine learning deep learning, artificial intelligence and automated machine learning development to achieve a solution adapted to the current banking reality.

Problem identification and motivation

A Design Research study is research about a specific problem, to reason about a proposed solution, defining the research questions, identify the main research problem, to justify an effective solution to the stakeholders and shareholders and the implementation potential benefits (Peffers, Tuunanen, Rothenberger, & Chatterjee, 2007).

Objective's definition and solution

It is specified, qualitatively or quantitatively, objectives and requirements defining the solution base to the defined problem and the following actions (Peffers, Tuunanen, Rothenberger, & Chatterjee, 2007). For the work and the achievement of a usable final solution it is essential the definition of a clear and objective problem, and their detail in sub problems to plan a phased roadmap (Gregor & Hevner, 2013).

Design and Development

In this phase it is pretended to design a useful product for future applications (Gregor & Hevner, 2013), starting with the problem division into smaller ones (Hevner, March, Park, & Ram, 2004). The advantages need to be justified and explained, along with the theory reasoning to achieve a consistent and usable design (Peffers, Tuunanen, Rothenberger, & Chatterjee, 2007). The developed work must answer all the identified needs (Hevner, March, Park, & Ram, 2004), and use the theoretical reasoning to acquire expertise on the problem state and possible solutions, with direct and indirect analysis on their efficiency (Peffers, Tuunanen, Rothenberger, & Chatterjee, 2007), to create and achieve the artifact business proposed objectives (Hevner, March, Park, & Ram, 2004).

Evaluation

The result is evaluated by the interested parties according to the practical application efficiency (Peffers, Tuunanen, Rothenberger, & Chatterjee, 2007) considering the evaluation methods appropriate definition of the problem and the business requirements. Essential to align the business requirements to obtain a solution performing greatly in the initial problem proposed (Hevner, March, Park, & Ram, 2004), evaluating by comparing between the thesis initial proposal and final application (Peffers, Tuunanen, Rothenberger, & Chatterjee, 2007), clearing that the main objective is the artifact usability in different banking sectors (Hevner, March, Park, & Ram, 2004). The criteria for finishing this step ends are the final artifact effectiveness and efficiency in solving the initially defined problems solution (Peffers, Tuunanen, Rothenberger, & Chatterjee, 2007).

Communication

The final methodology subpart is the communication to the stakeholders, showing the effectiveness of the artifact in the identified problems (Peffers, Tuunanen, Rothenberger, & Chatterjee, 2007), clarify the entire of building and validating the artifact process (Hevner, March, Park, & Ram, 2004). The communication should be directed to the entire workforce personnel to gather feedback information to improve the solution, and collect information for future implementations (Hevner, March, Park, & Ram, 2004).

3.2. Strategy implementation

This chapter describes how the methodology parts will be used in the study and the which tasks are involved. The study begins with the objectives design and the solution stages definition according to the solution objective and the presented problem, in this case, the implementation of an automated machine learning tool in several banking departments. There was needed to define the objectives and a solution centered on the study main purpose, to know the study focus areas and a solution focused on the identified subject. The stages destined to design, and development are focused on research around different subjects: the existent automated machine learning tools, the banking problems, how they could benefit from the implementation of machine learning and the existing processes that could be replaced. Posteriorly involves the design of the roadmap implementation in different departments according to the previous gathered knowledge and the requirements previously defined in conjunction with the banking personnel, and the following presentation to the stakeholders for evaluation. The final stage before implementation should several presentations to the entire organizations with separated ones for every department to get everyone onboard with the upcoming changes.

3.3. Expert Interviews

The expert interview is the way of collecting the stakeholder's perception without using general surveys (Devault, 2018), the advantages of the interviews are the fact that promotes the discussion of alternatives, solutions and small changes between the implementation team and the subject field experts, gathering relevant insights about how some changes can increase the solution effectiveness and acceptance, since it requires a deeper though compared to surveys (Prasad, 2017). Using this technique, could be targeted the workers providing the best feedback of the artifact (Prasad, 2017). Since it a qualitative study, the evaluation method is also a qualitative one (Gill, Stewart, Treasure, & Chadwick, 2008), mainly, since it generates an honest and instant reaction from the stakeholders to the proposed artifact. The interviews should be done privately and should took around 30 minutes, in case of being presential, if it is the case of written interviews the estimated time to answer the planned questions should be the same, keeping the subject on the essential matters of the artifact, trying to avoid any loss of attention (Prasad, 2017). Considering that every interview should start with the presentation of work that led to the artifact creation and the artifact itself, allowing every expert to gather their thoughts and opinions about the artifact, being able to deliver accurate and usable feedback, ensuring the freedom of every expert to express their thoughts (Gill, Stewart, Treasure, & Chadwick, 2008). The number of interviewed experts should be defined according to the time allocated for the interviews and the experts' availability to participate, in case of presential interviews, or the availability to be part of the study, in purely written interviews (Gill, Stewart, Treasure, & Chadwick, 2008). The people are chosen to take part into the interviews according to their previous work experience and the roles occupied during the career (Morgan, 2019). This method of gathering feedback and evaluate the artifact is considered successful if it is gathered a lot of usable information that lead to improvements in the final proposed solution. (Devault, 2018). The questions should have different objectives, gathering feedback on the overall solution, check the improvements proposed by the experts in specific artifact areas, going from general to specific, without yes/no answers (Prasad, 2017). In the context of this master's thesis, interview participants were chosen based on their previous roles and experience in the banking sector.

4. PROPOSAL

This chapter presents proposes a strategy of the implementation of an automated machine learning tool in the banking sector. The above sections describe the use case implementation to maximize the company productivity. There are referenced the assumptions taken as basis of the artefact, evaluating, and discussing afterwards.

4.1. Assumptions

Assuming the existence of a database with all the necessary client data and the institution willingness to take the next step to the get machine learning available to most users, so it is possible to say:

- Most part of the institution workers do not have specific knowledge in the subjects of machine learning and artificial intelligence, making an automated machine learning tool the ideal form to have them being able to gather some knowledge from the application models

- Most of the institution workers, also does not have an advanced knowledge of technology, so they would need to have formation to be able to gather useful knowledge from the tool

- The data that would be injected to the tool is in raw format, so it is needed a tool with a huge data preparation capacity

- The workers do not have machine learning models knowledge, so there is a need to provide simplified documentation to choose the adequate model for each problem

- The institution already has some tools involved in the data pipeline, there will be needed a team to integrate the autoML tool with these tools

- The tool will need maintenance and will generate bugs, so there will be needed some experts to solve the technological problems

- Every contract analyzed by the institution, is done manually by the assigned workers, taking an extensive amount of time to analyze them by hand and detail every important subject

- The institution has financial consultants and strategists, each one having a detailed list of clients assigned, being responsible to repeatedly analyze in detail the movements in the respective clients' accounts, making them accountable of their obligations

- The interaction between the institution and their clients is limited to the access to the website and mobile application, or the contact to their workers, a process not user friendly, that lacks easiness in problem solvability

- The institution dealing with money transactions have fraud prevention systems, the accounts users flagging the movements with any suspicious activity, mostly from the shopping made on the Internet

4.2. Strategy Design

The assumptions previously mentioned helped the elaboration of the proposed implementation roadmap for automated machine learning in the institution.

The framework pretends to extend the used of machine learning technologies to several institution areas as possible, and need, to guide the business through data-based decision making, identifying the areas with the more expressive need for the implementation of an automated machine learning tool, and the corresponding tool that will better fulfill the identified needs, having the necessary information to propose an implementation roadmap for the organization.

The graph above, has the macro vision of the designed artifact.

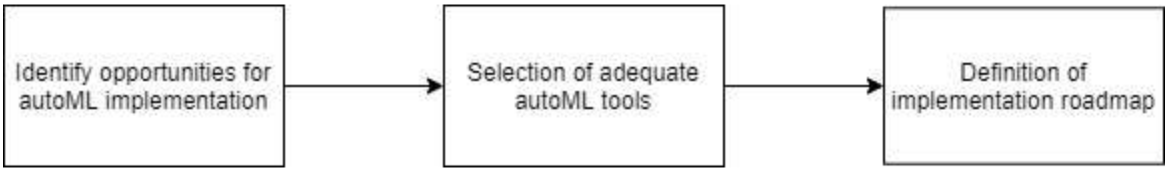

Identify opportunities for autoML implementation
Selection of adequate autoML tools
Definition of implementation roadmap

Figure 3: Macro implementation strategy

4.2.1. Identification of opportunities of autoML implementation

The first step in this framework is the gathering of all the opportunities in the organization's daily activity where machine learning could be used to lead to better and faster performance and decision making. The needs will vary according to the use cases assigned for each department, the data that is being treated, however the models applied after will be similar in most of the entity units.

The selection of the ideal department to the implementation of an automated machine tool comes from the work executed in a daily basis in the area, and the problems faced, collecting data from the expert personnel, identifying the existing business processes and their expected value, the type of tools available in the market, comparing each one of them.

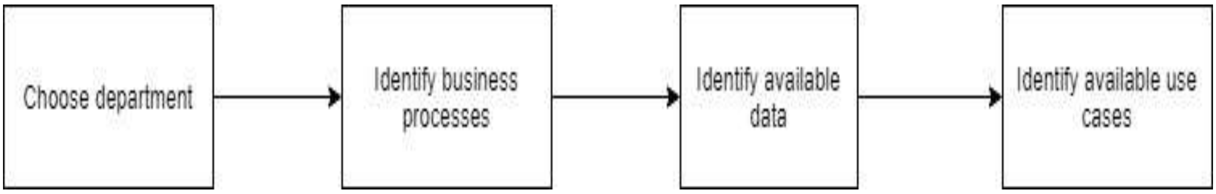


Figure 4: First step implementation strategy

The process of gathering information from the different department will happen in the form of a questionnaire to a specialist from the area, the questionnaire will go around every process of a machine learning pipeline, from the type of data that is treated and collected in the corresponding department, to have the most adequate tool to proceed to detect every type and proceed to their pre-processing, this is going to later influence the needs for feature engineering in the preprocessed data. According to the business purpose and the department objectives, there is different needs for machine learning tasks, being a parameter of exclusion of automated machine learning tools, since each one supports different types of algorithms, having different features to select the correct model to apply and to optimize the hyper parameters choice and also the evaluation of the models, through the presentation of a dashboard, showing the importance of the different model features and with an explanation of the model process.

The graph above shows the process of gathering information, to obtain a list of machine learning/analytics projects available in an organization, the process is going to be transformed in a questionnaire to be presented to the specialist from different departments, the workflow will assure that is defined correctly the department in which is applicable the implementation of an automated machine learning tool and the departments that does not have the maturity to the application of this technology yet.

First of all, it is necessary to know the essential information about the department, from the department name, the main goals of the department, the business processes executed to achieve the goals and the type of use cases, if the department is going to deal with analytical or data related use cases, since there is no point in applying an automated machine learning tool to a department that is not dealing with analytical and data related use cases. If there is a positive answer to the previous step, the use cases also need to deal with business data to be possible to be selected to the implementation of an autoML tool. The following step involves the clarification if another institution department or an external contractor was hired or is planned to be hired to develop an analytical project, in a positive case it arises the need to check if there are any department worker with knowledge of the entire process, as, in a negative case, arises the need to check if any department worker has knowledge of the data involved in the use cases, a negative answer in any of the cases will identified a department as not fit to proceed to the implementation. For both cases, it is also needed to verify the existence of a person semi-qualified in informatics that could manage the result of the implementation. The last five steps of the process relate with the detection of the analytical needs of the department to, posteriorly, combine every department with the most adequate automated machine learning tool.

The questionnaire consists in the following questions:

1. Department Identification
    a. Name
    b. Main goals
    c. Main business processes

    Questions: What is the name of your department? What are the main goals of your department? What is the main business process to the department achieve those goals?
2. Need of analytical output

    Questions: What type of use cases are you going to be treated in the next months? Analytical or data related use cases? And regarding these use cases what is the analytical output?

    (Yes/No)
3. Existence of business data?

    (Yes/No)

4. Are the Machine Learning projects done internally by the department?
5. Does anyone in your department have knowledge of the process? (If the previous answer was No)
6. Does anyone in your department have informatics knowledge? (If the previous answer was Yes)
7. What are the use cases objectives? (In every use case mentioned previously)
8. What are the analytical and machine learning needs to fulfill the mentioned objectives? (In the tasks of data treatment, data processing, feature selection and engineering, the type of learning used- supervised and unsupervised, the types of model selection and training techniques and project dash boarding and feature importance description or other specific functionalities)

The presentation of this questionnaire to different members of an organization will uncover needs for the implementation of automated machine learning in their departments, due to the non-existence of personnel qualified to the implementation of machine learning in every department, the implementation of an automated machine learning tool will solve these two problems. As said before, autoML allows the use of machine learning by people untrained in machine learning.

The responses gathered from different organization workers, will be comprised in a list, identifying the different use cases that could benefit from the use of automated machine learning, and their respective detailed objectives. The questionnaire will also identify which are the departments suitable for the autoML implementation as is explained in the flowchart.

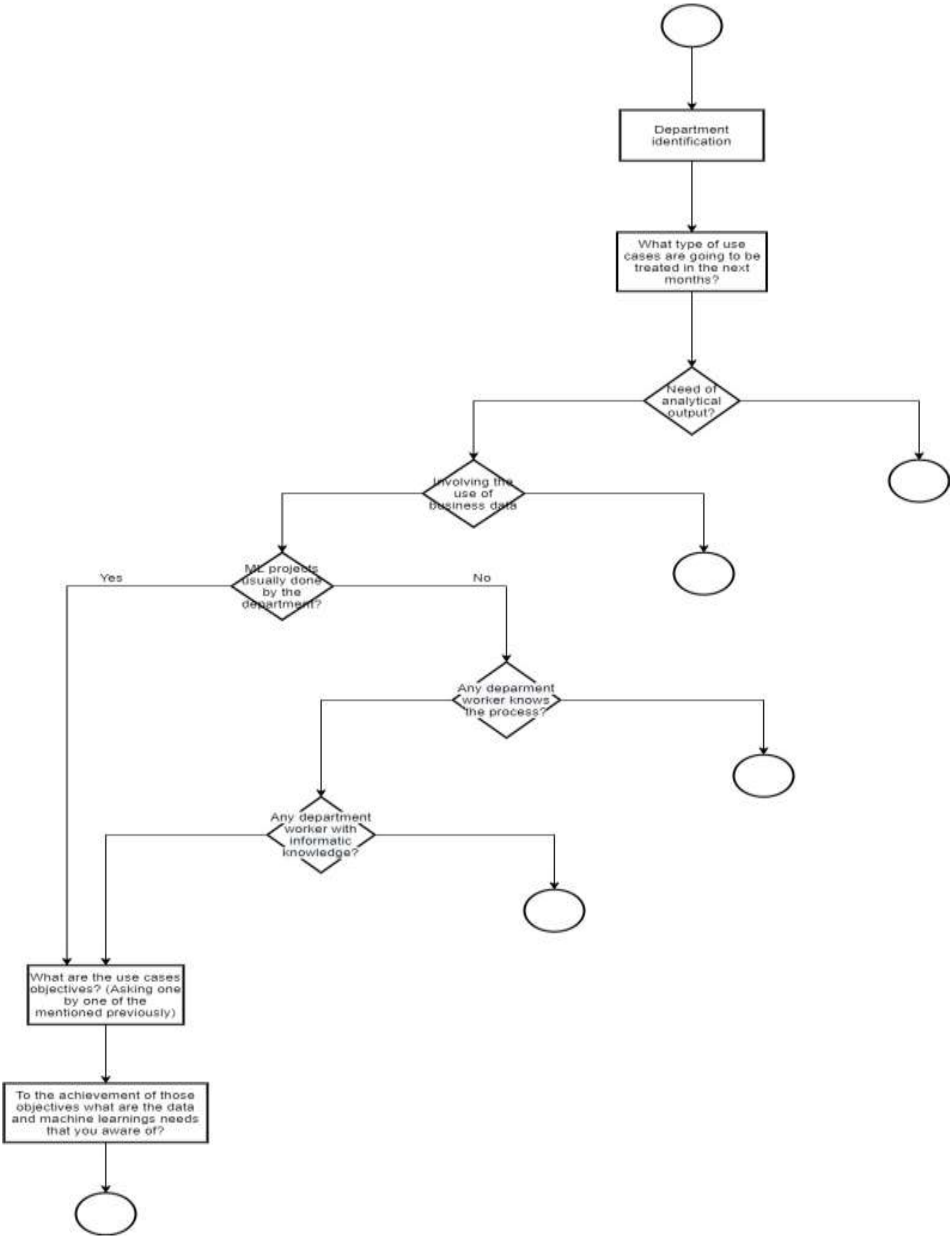

Department identification
What type of use cases are going to be treated in the next months?
Need of analytical output?
Involving the use of business data
ML projects usually done by the department?
Yes
No
Any deparment worker knows the process?
Any department worker with informatic knowledge?
What are the use cases objectives? (Asking one by one of the mentioned previously)
To the achievement of those objectives what are the data and machine learnings needs that you aware of?

Figure 5: Interviews guideline flow

The previous flowchart shows the process of obtaining, through an interview a list of projects, with the corresponding objectives, in the departments suitable for the implementation of an autoML tool. The interviews process starts with the gathering of general information about the department, its designation, the objectives proposed by the upper management to the department and the existing business processes, with that data collected, the interview will start gathering information about every use case and project that will be done within the department in the following couple of months, it will be checked if there are projects that will produce an analytical output, in a negative case it will not be possible to proceed to the implementation of autoML in that specific department and the interview must end right away. Case the answer is positive, there are eliminatory questions, the next step in the interview, regarding the use of business data in the use cases, since without data there is no need for machine learning and analytical use case, it is also necessary to check if the analytical and machine learning use cases are being currently done inside the department. If all the analytical use cases are done inside the department the interviewer should start about the use cases and the list of proposed objectives, trying to as much useful information as possible about the use cases' objectives and the data and machine learning needs that the interviewed is currently aware of to the fulfillment of those goals. On the opposite case before getting the detailed information about the use cases' objectives and the analytical and machine learning needs, the interviewer should check the familiarization of the department workers with the process and their informatics knowledge of the department workers, finishing the interview by collecting a detailed list of the use cases' objectives.

4.2.2. Selection of adequate autoML tools

For each project identified in the list created previously, the next figure shows the steps leading to the choice of an adequate automated machine learning tool, it will be made with the help of the information and conclusions learned in the previous phase.

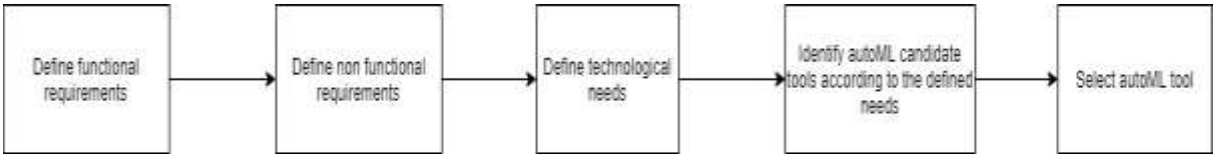

Define functional requirements
Define non functional requirements
Define technological needs
Identify autoML candidate tools according to the defined needs
Select autoML tool

Figure 6: Second step implementation strategy

Every department has a determined set of goals to achieve in defined time periods, having personnel working collectively to the fulfillment of that department goals. To achieve these

The functional requirements describe the pretended solution in a functional form, relatively to the actions and behavior that should happen, the non-functional requirements describe the properties or restrictions considered in the final solution conception.

The current banking environment could influence the selection of the technology, according to the current department personnel, their skillset, the agreements with current autoML suppliers and the current deadlines, having the availability to soften the deadlines to implement a change that will decrease a use case development and improve the work quality further down the line. Another important subject that should be considered is the possible existence of the European Central Bank or Portuguese Central Bank regulations regarding the implementation of automated machine learning tools.

The identification of requirements comes from the answers gotten from the made questionnaire in the previous step, the environment in the department, the number of persons involved in the process, the technical and machine learning knowledge level of the involved workers, according to these variables there will lead to different types of requirements, identified in a table like the above.

Type of requirement	Examples of requirements
Functional requirement	• Identify economic trends • Identify political and regulatory issues • Analyze organization characteristics • Systems and technology • Financial health • Define strategic vision • Develop business strategy • Adequate with long term business strategy • Create functional user experience • Understand target population • Identify target segments • Define and manage strategy and plans • Integration with different types of data • Integration with different data sources • Integration with different file formats • Compatibility with existing systems • Generalizable across different target datasets • Work with big data • Handle both cross-sectional and longitudinal data • Work with supervised model • Work with unsupervised model

	• Feature extraction and transformation • Split data set into training, validation, and test data • Determine best model evaluation criteria • Train set of different models (hyperparameter optimization) • Choose best model • Accurate models • Model deployment • Dashboarding results
Non-functional requirement	• Scalability • Formation • Computational power • Easy use of resulting model • Explain ability of machine learning process

Figure 7: Different types of requirements

The table presents the general requirements identified in an automated machine learning tool implementation, and next will be presented some recommendations according to the common requirements:

- Integration with different types of data: The tool chosen to implement independently of the tasks that is meant to replace in the pipeline, it will be necessary to adapt to different types of data, having the capacity to process all types of data to avowing being a bottleneck to the machine learning pipeline

- Integration with different data sources: To integrate an automated machine learning tool, in a department that is already working in a fully developed pipeline, it is essential to create automatizations mechanisms with the current organization data sources, which is very important to ensure that a smooth integration and avoid to the maximum the interaction with humans in the initial pipeline stages

- Integration with different file formats: A machine learning pipeline usually is feed with different types of files coming from different sources or departments, so another essential requirement for a machine learning pipeline is capacity of integration with a set of different file formats

- Compatibility with existing systems: The departments in their daily work, usually use a set of different of tools, and department that need or work with machine learning are not an exception, having the need to have developers in the roadmap to integrate the automated machine learning tool with the existing tools in the department pipeline

- Generalizable across different target datasets: Since the data received daily by an organization could come from many different contexts, so it also could be of different types (categorical, numerical, …) and it is needed to generate different types of data to the current problems faced

- Work with big data: The data used in the daily activity is from many different types and it could happen that department is dealing with huge amounts of data, what is also needed from the automated machine learning tool is the ability to deal with datasets of different sizes

- Work with supervised model: According to the business needs and the presented use cases, the proposed use cases involving machine learning will have many different outcomes, so it is necessary to be able to apply a variety of supervised learning models

- Work with unsupervised model: For the same reasons mentioned above, it is also needed to apply a variety of unsupervised learning models

- Feature extraction and transformation: Every machine learning use case has a step, after the preparation when it is needed to transform the existing features and create new ones, so it useful the capacity of automatically create new features and select the ones that will have a better relevancy for the model

- Split data set into training, validation, and test data: Before the application of machine learning model, the data set is split into three different data sets, one assigned to train the model, another assign to validate the trained model and the other to test the model, making it an important feature of an automated machine learning tool

- Determine best model evaluation criteria: An essential step of the deployment of any machine learning criteria is the choice of the best model to implement afterwards, according to the defined criteria that should be defined by the implemented tool

- Train set of different models (hyper parameter optimization): To achieve the best model to implement it is necessary to train a set of various models, and to do that, the data scientist usually give to the model different hyper parameters, however the automated machine learning tool should combine instantly different sets of hyper parameters to achieve a model with that best fits the defined criteria

- Choose best model: After the execution of the hyper parameter optimization of the machine learning models, there is a need to choose the set of hyperparameters that will lead to the best and most optimized model according to the problem specifications defined previously and the criteria defined for each model

- Accurate models: The created machine learning models need to be evaluated according to their accuracy, having a threshold identifying the minimum accuracy available to accept a machine learning model

- Model deployment: The last phase of the machine learning lifecycle is related with the deployment of the chosen model in production to use the updated data and be able to keep delivering updated results

- Dashboarding results: Since the automated machine learning tool it is pretended to not be used by machine learning experts, so it is useful to produce a report and a dashboard with the description of the entire process and the obtained results, helping the users getting familiarized and understanding what has been done

After the identification of the requirements that need to be fulfilled by the automated machine learning tool, and the identification of the technological support needed, will be identified a set of machine learning tools adequate to the pretended implementation, however, to finalize the choice there is a need to define a set of acceptance criteria to be assure that the correct choice has been made.

Due to the impact of the implementation of autoML in the business processes, there are several criteria considered essential, as accuracy, assuring the selection of features to use and the creation of new ones, the comparison and selection of a variety of new models and their automatic tuning, setting up validation procedures and model rank by performance, it is the most important criteria, without accuracy there is no reason to start the implementation of autoML, productivity, taking from the experienced workers the constant necessity of finding the best model, doing this automatically, having the ability to find the best algorithm in the smaller amount of time, handling gracefully the needs of each algorithm, ease of use, should be of easy use, easily integrated in the process flow and understandable for experienced data scientists, and must have intuitive explanations and visualizations for non-data scientists, reducing as most as possible the necessary knowledge of machine learning to produce an effective output, understanding and learning, by improving the analyst capacity to understand the context of the problem, visualizing the interaction between features and the target and explain support the analyst to present the findings to the management and answer the presented questions, allowing the combination, of the uncovering of useful findings with the subject matter experts to use correctly the finding, resource availability, should be assured the compatibility between the business systems and autoML system and the other business tools, meaning the existence of connection between existing databases and file formats for data ingestion, besides that, should also be assured an easy use of the resulting model, through an application programming interface (API) or code placed into the organizational workflow, addressing memory issues, storage space and processing capabilities, process transparency, allowing users who do not have extended machine learning knowledge to improve their knowledge and understand the decisions made by the systems, context generalizability, working for different target data types, data sizes and different types perspectives, predicting either numerical or categorical targets, handling small, medium and big data and either cross-sectional and longitudinal data and recommend actions, the system should be able to transfer a probability into action and generalization across different contexts.

	Functionalities	*Costs*	*Complexity*	*Open Source*
TransmogrifAI	(a) (j) (p) (r) (z) (f1)	Low	Simple	Yes
H2O-AutoML	(a) (k) (p) (s) (z) (h1)	Low	Simple	Yes
Darwin	(a) (j) (q) (t) (a1) (g1)	High	Hard	No
DataRobot	(b) (j) (q) (u) (a1) (g1)	High	Simple	No
Google AutoML	(c) (j) (q) (v) (b1) (g1)	High	Simple	No
Auto-sklearn	(a) (j) (p) (r) (b1) (g1)	Low	Medium	Yes
MLjar	(a) (m) (p) (s) (f1)	High	Hard	No
Auto_ml	(a) (j) (p) (r) (g1)	Low	Simple	Yes
TPOT	(a) (n) (p) (w) (c1) (f1)	Low	Simple	Yes
Ludwig	(b) (l) (p) (y) (d1) (f1)	Low	Simple	Yes
Auto-keras	(b) (o) (p) (x) (a1) (h1)	Low	Medium	Yes
Auto-weka	(a) (o) (p) (r) (z) (j1)	Low	Medium	Yes
Azure ML	(b) (j) (p) (r) (z) (f1)	High	Simple	No
H2O-Driverless AI	(b) (j) (q) (u) (e1) (g1)	Low	Hard	Yes

Table 2: Study of the AutoML tools characteristics based on Table 1

This phase will be done when there is completed the previously done list of projects along with their respective requirements (either functional or non-functional) and the automated machine learning more adequate to each project, to be possible to make a choice.

4.2.3. Definition of AutoML implementation roadmap strategy

The implementation roadmap of an automated machine learning tool will consist in the several phases:

1. Project Preparation

The first phase is the preparation of the project, it is formally initiated, and the project starts being planned, and are defined the project milestones and work packages, and sets the first definition of the project scope, defining the standards and procedures of the project.

1.1. Prioritize automated machine learning implementation needs

The project starts with the scope definition, and according to the list of the projects with their respective requirements and adequate automated machine learning tool, this list will be update with the implementation priority, ranking by tool to reach more divisions of the organization.

1.2. Assign resources

According to the previously defined implementation priorities, the implementation team will need several resources to achieve the proposed goal, so the next step consists in the assignment of the necessary resources, either technical resources or personnel. This step will consider the organization current financial state, probably affecting the next step if the organization is not available to allocate all the identified necessary resources.

1.3. Set up a calendar

After the assignment of resources, the project team can define the methodology of work, and define the calendar to deliver the several project iterations or the final iteration, according to the defined project methodology. As mentioned previously, this phase is affected by the previous one, since the final delivery of the project is highly dependent from the project team and the conditions given to the team to perform their work.

1.4. Set a control and monitoring structure

To monitor the correct execution of the work and ensure that the work is going to be done when it is agreed to the client and stakeholders, there is a need to define a set of milestones to keep track of the progress and keep showing progress to the client. It is also necessary to set up a team of workers in the affected company divisions to stay updated of the project progress, iterate any need changes during the execution and to check if the acceptance criteria are being met.

2. Business Blueprint

In the second macro phase of the project the requirements are reviewed and for each part of the project.

2.1. Review of identified requirements

The second phase of the project initiates with the gathering of the requirements previously identified for the select tool, according to this tool and the corresponding different areas of implementation and reviewing them according to the defined project characteristics.

2.2. Update of requirements definition

After the list of requirements revision, there is the possibility of arising changes in the requirements previously defined, so it is necessary to iterate with the business teams regarding the update of the requirements.

3. Realization

The third macro phase of the project involves the realization of the project proposed work, involving the execution of all the proposed work to reach all the points identified in the project scope. This step will occur for the different departments that have the same automated machine learning tool identified as the one that fulfills the identified requirements

3.1. Integration in the pipeline defined steps

These phase starts with the technical integration of the automated machine learning tool functionalities in the machine learning pipeline, for each department, to replace the defined steps according to the requirements to create a new machine learning pipeline for each identified department, also defining the configurations of the new pipeline and how it will be executed and tested.

3.2. Test integration

With one of the new pipelines created integrated with the automated machine learning chosen tool, the implementation team must do a series of tests to ensure the pipeline is working correctly and can start being tested in a real-life environment.

3.3. Train end users

After the creation and testing of a new integrated pipeline, there is a need to train end users and produce documentation regarding the new way of work to change to the new pipeline, just getting them through the changes and teach them how they can perform their work in a new and easiest way.

3.4. Develop use case in parallel in both ways

With the users correctly informed and trained about the newly created machine learning pipeline, the next step should be the development of a chosen use case with the two pipelines in parallel, in conjunction by the implementation team and a business team, at two different speeds since one pipeline has created with automated steps. This step also tests the environment hardware capabilities to the new pipeline

3.5. Evaluate results

When the two pipelines reach to an end, the comparison of the obtained results will let the implementation team know if the new pipeline has been correctly developed and if it obtains similar results. According to these conditions it is decided if the department is moving forward with the new pipeline and the implementation team is going to start working with another department or if there is a need to perform the work another time in a different form for the same department.

3.6. Iterate progress according to defined criteria

After the results evaluation, the resultant pipeline is going to be evaluated according to the defined acceptance criteria to decide, one more time, if there is time to start working with a new department or iterate over the work done for the current department

4. Go live and support

When the created pipeline passes all the acceptance conditions defined by the implementation team and the end users, the pipeline can be used lively by the departments and there is one element from the implementation that will be assigned to each department to support the end users to provide support, when needed.

4.3. Use case simulation

The development of these use case would with the gathering of information from the different institution departments, in this case the choice was to interview coordinators from the Department of Companies Marketing, the Department of Particular Marketing, the Department of Distance Management and the Department of Information Systems.

Firstly, there is the answers from the Department of Business Marketing coordinator,

1. What is the name of your department? Department of Companies Marketing
2. What are the main goals of your department? The department works with the companies' clients of the institution to strengthen the relationship and improve the profitability that each company gives to the institution. The department achieves those goals combining the analytical knowledge from the analytical suppliers inside the institution, filtering them with our business knowledge and passing the usable information to the commercials to help them doing their job.
3. What is the main business process to the department achieve those goals? The main business process in the department involves the generation of leads to the commercials with the clients more that will most likely acquire some product, this way certain company will increase their profitability to the company

   What type of use cases are you going to be treated in the next months? Analytical or data related use cases? And regarding these use cases what is the analytical output? In the next months, there will be treated use cases regarding the churn of automatic payment terminals, the equipment leasing churn, the churn of companies related products, capturing and binding per product, profitability by product and a lead regarding meal cards. In a general way, the outputs will be list of clients that are more likely to acquire or quit of certain product.
4. Does your department work with business data? Yes
5. Are the Machine Learning projects done internally by the department? No
6. Does anyone in your department have knowledge of the process? Yes
7. Does anyone in your department have informatics knowledge? Yes
8. What are the use cases objectives? (In every use case mentioned previously)
9. What are the analytical and machine learning needs to fulfill the mentioned objectives? (In the tasks of data treatment, data processing, feature selection and engineering, the type of learning used- supervised and unsupervised, the types of model selection and training techniques and project dash boarding and feature importance description or other specific functionalities) The data goes through several stages, so it is needed to treat and process several types of inputted data, process them, select the best features for the model, allow

supervised learning models, ensemble and neutral architecture search and preferably feature importance description and end of project dashboarding.

Secondly, there is the answers from the Department of Particular Marketing coordinator,

1. What is the name of your department? Department of Particular Marketing
2. What are the main goals of your department? The department works with the clients of the institution to strengthen the relationship and improve the profitability that each client segment gives to the institution. The department achieves those goals combining the analytical knowledge from the analytical suppliers inside the institution, filtering them with our business knowledge and passing the usable information to the Department of Distance Management to help them doing their job.
3. What is the main business process to the department achieve those goals? The main business process in the department involves the generation of leads to the Department of Distance Management with the clients more that will most likely acquire some product, this way certain client segment will increase their profitability to the company.
4. What type of use cases are you going to be treated in the next months? Analytical or data related use cases? And regarding these use cases what is the analytical output? In the next months, there will be treated use cases regarding the credit cards, saving accounts, capturing, and binding per client and profitability by client segment. In a general way, the outputs will be list of clients, per segment that are more likely to acquire or quit of certain product.
5. Does your department work with business data? Yes
6. Are the Machine Learning projects done internally by the department? No
7. Does anyone in your department have knowledge of the process? Yes
8. Does anyone in your department have informatics knowledge? Yes
9. What are the use cases objectives? (In every use case mentioned previously)
10. What are the analytical and machine learning needs to fulfill the mentioned objectives? (In the tasks of data treatment, data processing, feature selection and engineering, the type of learning used- supervised and unsupervised, the types of model selection and training techniques and project dash boarding and feature importance description or other specific functionalities) The data goes through several stages, so it is needed to treat and process several types of inputted data, process them, select the best features for the model, allow supervised learning models, random search and preferably feature importance description and end of project dashboarding.

Thirdly, there is the answers from the Department of Management Distance coordinator,

1. What is the name of your department? Department of Management Distance
2. What are the main goals of your department? The department works with the clients who are flagged by the Department of Particular Marketing and their relationship with the institution is done at the distance without dislocation to the agencies

3. What is the main business process to the department achieve those goals? The main business process in the department involves the generation of leads and reception of usable information from the Department of Distance to Management with the clients more that will most likely acquire some product, this way certain client segment will increase their profitability to the company

4. What type of use cases are you going to be treated in the next months? Analytical or data related use cases? And regarding these use cases what is the analytical output? In the next months, there will be treated use cases regarding the credit cards, saving accounts, capturing, and binding per client and profitability by client segment. In a general way, the outputs will be list of clients, per segment that are more likely to acquire or quit of certain product
5. Does your department work with business data? Yes

6. Are the Machine Learning projects done internally by the department? No
7. Does anyone in your department have knowledge of the process? No

Finally, there is the answers from the Department of Information Systems coordinator,

1. What is the department that you work for? Department of Information Systems
2. What are the main goals of your department? The department works with different organization departments, answering their analytical necessities.
3. What are the main business processes involved in your department? The main business process in the department involves the transformation of other department problems in analytical problems to provide them usable information to perform their work
4. What type of use cases are you going to be treated in the next months? Analytical or data related use cases? And regarding these use cases what is the analytical output? In the next months, there will be treated use cases from Department of Business Marketing, Department of Particular Marketing and Department of Management Distance. In a general way, the outputs will be list of clients that are more likely to acquire or quit of certain product.
5. Does your department work with business data? Yes
6. Are the Machine Learning projects done internally by the department? Yes
7. What are the use cases objectives? (In every use case mentioned previously)
8. What are the analytical and machine learning needs to fulfill the mentioned objectives? (In the tasks of data treatment, data processing, feature selection and engineering, the type of learning used- supervised and unsupervised, the types of model selection and training techniques and project dash boarding and feature importance description or other specific functionalities) The data goes through several stages, so it is needed to treat and process several types of inputted data, process them, select the best features for the model, allow supervised learning models, ensemble and random search and preferably feature importance description and end of project dashboarding.

As it is seen for the answers to the questionnaires, there are three departments dependent of the Advanced Analytics department, so the implementation of an automated machine learning tool in the Department of Companies Marketing and in the Department of Particular Marketing, would free the department to deal with high level use cases, these would help organizations reach another level of artificial intelligence maturity. For the Department of Distance Management, it is not feasible to implement an automated machine learning tool since there is not an employee with knowledge of the process.

The realized interviews allowed the identification of available use cases to start the use of an automated machine learning tool, the churn of automatic payment terminals, the equipment leasing churn, capturing, and binding per product and profitability by client segment.

According to the defined objectives of the use cases, it is possible to identify the related functional requirements, like integration with different types of data, data sources, file formats, compatibility with existing systems, generalization across different target datasets, ability to work with supervised models mainly, feature extraction and transformation, training, validation, and test data splitting of the original dataset, hyperparameter optimization and dashboarding results. With the non-functional requirements identified being scalability, computational power, easy use of resulting model and explicability of machine learning process.

The gathering of requirements allows to narrow the choices to two automated machine learning tools, Darwin and DataRobot. For the Department of Companies Marketing with would be useful a tool able to treat and process different data types, able to apply supervised learning models and ensemble and random search, since it is a department with the possibility of having more machine learning needs over time, it would be better to implement Darwin, as far as for the Department of Particular Marketing, the needs are similar, however there is a different maturity in terms of data and informatics knowledge, what would make DataRobot the most adequate tool to start the implementation.

The implementation will start by the prioritization of the use cases, with the churn of automatic payment terminals in first place, followed by the equipment leasing churn, profitability by client segment, capturing per product and binding per product, the team assigned to this project should include data engineers, data scientists, data analysts, business analysts and business experts from the different departments, setting a limit of 2 months from implementation in each different area, testing the process with one use case relative to each area, the work is going to be done in sprints iterating the project each 2 weeks to the Department Directors, who will the project owners.

After the closing of the entire project team and calendar, along with the calendar, the monitoring structure and the milestones, the team will be reviewing the previously identified requirements and iterate if there is any need to update them. Starting the project execution afterwards, the pipeline constructed for the automatic payments terminal churn is reviewed and checked according to the acceptance criteria defined, and it is created another pipeline integrating the autoML tool, the use case is developed in the two pipelines and checked regarding the acceptance criteria and the pretended results, assuring that the final users have the necessary knowledge to deal with the implemented technology, according to the results, the process is going to be repeated from the same department or for the remaining ones, until the department are fully working live with the new technologies.

4.4. Validation

The proposed framework for the implementation of automated machine learning was evaluated by industry experts, attempting to recognize its utility, improvement suggestions and validation. According to the defined methodology, interviews to some experts to seize technical and non-technical improvements to the framework. The literature review served as a base to the questions made during the interviews, which were sent by email to the interviewees and the questions answered in return.

To validate the proposed framework, specialists with a life dedicated to the banking sector were selected: Drº. Paulo Zeferino (PZ), former coordinator of the department of Big Companies of Banco Espirito Santo, with previous experience in another banking institutions, Drª. Cristina Isabel (CI), an economist working currently for Santander, also with previous experiences in another banking institutions and Dr. º Miguel Torres (MT), who has working in different departments of Santander for his entire life. After a small presentation of the framework (Attachment 1), the participants were prompted with some questions, to collect their insights, knowledge, and constructive feedback (Table 3).

Q1.	Do you consider the proposed framework useful? Why or why not?
Q2.	Do you have any comments, either criticism or recommendations towards the proposed framework? Please explain.
Q3.	Would you consider using the proposed framework? Why or why not?

Table 3: Validation questions to be made to the interviewed experts

These questions allowed the acknowledge of the frameworks' utility among the banking institutions, the overall acceptance across the different institution departments and relevant improvements for future work.

Next, it is possible to find the answers to the questionnaire questions, since it was answered by email it was not necessary to worry about biased information, the complete answers can be found above.

Regarding Q1, the answers were the following:

PZ: From what you presented me in the academical point of view, this framework makes a lot of sense in terms of application and usefulness. As I can see, from an academic and organizational high level point view, it makes sense. The implementation of automated Machine Learning in an organization, theoretically, has lot of benefits. Although it is necessary to understand the economical and organizational impact of a transversal implementation in an organization. All the involved personnel should be involved in the discussion.

The framework would allow an implementation that could meet several requirements currently in the banking activity, so it would really benefit from having the end user involved from the beginning of the process. Probably the end users have a more accurate knowledge of the real needs, how to integrate the framework and to smoothen the transition.

Finally, I think this framework has the potential to be useful in the banking sector, if it passes some possible drawbacks, the implementation needs to be agreed by all the users, mostly the older workers that are not really familiarized with the technological changes and the used to work just with the outputs of the machine learning process, being imperative a good training and sensibilization of these workers. If that possible drawback is correctly addressed the frameworks could solve and leverage the institution and their profits since it proposes to drastically reduce the output of analytical and machine learning use cases, besides the fact of allowing the analytical teams to work on higher level use cases that could lead to major improvements and changes into the organization.

CI: My opinion, based on the experience of working a lot with the results and outputs of analytical and machine learning use cases. I am usually in contact with analytical teams and my work includes some outputs of the use cases developed by them. The concept of Machine Learning and data and analytical based decision making has been here for a lot of time, however it was the first time that I heard about automated Machine Learning, and it seems a great opportunity to increase the analytical and data knowledge across the organization, since it an objective of all companies but the reality is that most of times is just talk, due to delays in the deliveries or lack of knowledge on how to use the output.

This framework would really facilitate the process, by giving a template for banking institutions implementation of autoML, and the further implementation would leads to an analytical and data based organization, allowing the most decisions to be made according to the existing data what would make a safer and more trustworthy decision making, and this is would be possible since as it is mentioned in your proposal, it will significantly reduce the time of delivery of each machine learning use case.

MT: It is a very useful idea; however, it could face some challenges, aligning the complexity and diversity of changes, organizational, in personnel, technologically, that this implementation could lead to. I believe that would be useful and successful if all the involved personnel would be aware of the potential benefits of the autoML implementation and the new responsibilities and opportunities that this will lead to. In my organization, I would support this implementation and the use of this framework, mainly because of the benefits that would give to the organization in the long run. It will give an organization a competitive advantage across their competitors, and it would make a data centric organization which would make a lot of decisions more trustworthy,

Regarding Q2, the answers were the following:

PZ: Most academic works have a great economic potential, but how this potential is translated into organizational profits and competitive advantage. For example, with the implementation of analytical teams, working for different organizations' departments will solve the fact of the organization not being analytical based, the hiring of several analytical experts to different department to being the responsible worker for the connection with the analytical team increases the analytical knowledge across the organization. However, this is conditioned by the fact of being a team working for several others, the team needs to robust and be available to different departments at the same time, working in several completely different use cases in parallel. Is that possible? I do not believe that most organizations have reached that analytical maturity level. What could be done to avoid that? I think you have the solution; however, the framework and the implementation process need to consider all the transversal organizational changes, all the different routes of work that are going to be created. Another problem that I can see is the implementation costs, how is the implementation going be done, by an external contractor? The internal workers are busy keeping the workflow of the organization use cases, what are going to be the implementation costs? How long it would take to cover the implementation costs? How long it would to the organization to make profits of this implementation? And the fact that I mentioned previously of having everyone affected on board to the changes.

CI: From my experience, I never saw any idea similar implementation in the banking institutions that I have been part of. The framework seems to be well built and you defined all the steps to a successful implementation. However, I would like to see more sensibilization, more open discussions between all the involved personnel, the parameters to choose the implemented tool should be clearer and how the expert personnel will have a word in that choice. I would also be interested to see what implementation tests are going to be done, how the possible implemented tools can be integrated with the existing systems, if there are any systems that can compromise the implementation.

The flowchart that you showed seems interesting, but I was wondering if only the director and coordinators are going to be involved, what about the workers of the departments that are daily in contact, building all the organizations use cases, and working with the outputs of the projects. So, I think it should be added in the last step of the roadmap a step to design the new workflow of the affected departments, having all the involved personnel on board.

MT: There are some interesting ideas in this framework. The implementation of a possibility to have machine learning across the organization without the need to hire expert personnel in that area, I have to say that I question myself how it is possible, so I would suggest, first, in the future it would be preferable to have someone with field knowledge to help in the development and make sure that the results are being correctly interpreted and the process is being correctly developed. I also add that I do not really think it would lead toa great result if the implementation framework only involved the direction and coordination of departments, the workers are what makes a company move forward, so it would be a smarter choice to include them in the questionnaire, or if it is not possible make an open discussion with all the workers involved in the changes so that everyone could speak their mind and give their opinion regarding the identified requirements and the changes that are about to happen in the organization and will affect or change their work and functions. I would also add that the project roadmap seems a great template, since it specifies what needs to be done in each part and how it will be reached a good implementation, in my opinion, could be added some formations to help the workers getting familiarized with machine learning and being completely independent in their future work.

Regarding Q3, the answers were the following:

PZ: Over the years, I have worked and most recently coordinated some teams that had to make decisions regarding banking products (credits, …), and possible new products. Most of the times my teams would not have all the data needed to decide with the most accuracy possible. Recently, my team have been working an analytical team, however they help us in just a few use cases, since their time is limited, and they work with another teams. The implementation of this framework would really let my team have the base to make a lot of decisions according to the data available in the organization. I would also add that it would be preferable to have some people with knowledge of what has being done to allow a correct interpretation and use of the results.

CI: The implementation of the framework would solve part of the problems that I am currently dealing with, for example, I worked with several outputs of the data analytics teams, however it is just part of my work that is affected by their work. So, since most of the cases are not of a big analytical complexity, and the autoML implementation would make my life much easier. However, I believe that most of the necessities are not really in the mind of the coordination, it would make a more accurate requirements definition if the analytical team workers and the people that uses their work outputs were involved. My answer is yes, although I would recommend the changes that I mentioned before.

MT: I have been working with analytical teams a lot during recent years, mostly using their use case outputs to help in my current work. Firstly, I would think that most people would be more comfortable with this implementation if they could be involved in the process and have some training to better understand the process. Secondly, I would also think that for the use cases that I am not currently used to work with analytical outputs, it would be preferable to have an expert support in the first developments, which could extend the time projections for the implementation. Besides that, I would suggest the framework implementation as it is, adding the suggestions that I already pointed.

4.5. Discussion

After the evaluation of the artifact by different banking sector workers, the feedback was carefully analyzed, with the focus on the utility of the artifacts for the organizations, improvements that could be made and observations and adaptations of the participants during the interviews.

In terms of the artifact's utility, the participants stated that the usefulness of the framework for a banking institution, understood and highlighted the benefits that the implementation of automated machine learning tool could present to every organization. The need to make data and analytical based decisions across the institutions around the world and lack of formation in these topics from the general workers increases the demand for solutions that would not require machine learning knowledge to get all the benefits to the organization use cases.

The interview participants highlighted the challenges that could presented itself in the framework implementation, the organizational challenge (the adaptation of different organizational departments, the acceptance from the department workers in adapting to the function changes), the economical challenge (there are different autoML tools, some are open-source, however this implementation will have the some costs, like the tool cost, the surrounding infrastructure, the implementation team and their assignments), the knowledge challenge (most part of the workers assigned to work on the resultant of the implementation will see their changes in their daily functions, having to deal with different problems, probably outside of their knowledge base, it will arise an adaptation and lack of knowledge problem, what could be lead to problems interpreting the results and choosing the correct process to apply), the complexity challenge (with the framework implementation, the machine learning use cases will be done by people without machine learning experience, and they will need an adaptation and formation to the new tasks) and the security and trustworthiness challenge (the workers used the work with machine learning use cases outputs will have access to all the related data, since they will perform the entire process, having access to confidential information, should be ensured some sensibilizations regarding these challenges).

By analyzing the improvements and recommendations suggestions made by the users, it is shown an overall concern regarding the involvement of the affected department workers in the process. All the participants stated that the implementation of autoML would be great for every banking organization and the framework is the ideal to be adopted in the implementation process. However, every participant referred that the implementation process should include meetings including the personnel that will see their work affected with the implementation changes, it was a concern of the experts that these people could reject these kinds of structural changes in their work, although the proven benefits of these changes. Therefore, should be added department meeting, over the select ones to discuss the implementation and have everyone’s opinion, instead of just relying on the department’s director and coordinator opinion.

Within the defined process there were pointed some skepticism, regarding the possible of the implementation process and solution costs, the fact that some automated machine learning tools come with associated costs, the need of outsourcing a consulting team to implement the solution and manage the process, these will elevate the costs of using the framework, so the suggestion was regarding the inclusion of a detailed costs plan, previewing how the new way of work will increase the company profits, and how it will take for the organization to regain the implementation costs. This point will be important to persuade the managers who does not support the process, since it will make their department increase the profits which help increase their recognition across the organization.

The suggestions made could be implemented by increasing the complexity of the roadmap, previously to the initial interviews it could be made a study with the Analytical teams and the ones using their outputs, to quantify the profits the analytical use cases are providing to the organization. The addition of these step could increase the acceptance of these implementation across the organization management, it could have the same effect in the other personnel if the organization uses the increase in profits and all the benefits of the changes to make everyone happier and more satisfied with their work, increasing or reactivating the prizes share for every worker.

Also, regarding the employee’s satisfaction and productivity, the goal should be to make the transition as smoothly as possible, the suggestions made by the interviewed of give formation classes and sensibilization across the organization and the added step of redesigning the organizational departments workflow fits this goal perfectly. The new workflow definition will work as a guidance for the employees making them used to the new way of work, helping them with the transition

Finally, a suggestion that was also made was that some classes should be of introduction to the topics of Machine Learning and Data Science, since it would help the workers have some knowledge of the process that is being done by them. Despite, the use of automated machine learning main benefit being the fact that the ones using do not need to have machine learning knowledge, it would help the workers getting the familiarized with the terms and concepts to help them make better decision regarding the choice of models and giving them knowledge to make informed decisions regarding these subjects.

Regarding the choice that the participants would make if they had the possibility to use the proposed framework to the implementation of automated machine learning, it was a positive answer from everyone involved. All of them also stated that would implement with the suggestions that they have made previously, to make the implementation more adequate to their organizational environment, a main general concern was the communication across the organizational different levels, hoping that it worked to increase the acceptance. In terms of usability and complexity, seemed like all participants understood the frameworks and the involved steps.

Additionally, the participants shared that the use of this framework in their past experiences would make their work much easier, since they would have data to make to work on and make knowledge-based decisions and suggest that the supervision of a subject expert in the first analytical outputs from the automated machine learning tool to help obtaining a correct interpretation.

5. CONCLUSION

In this chapter can be found the conclusions of the developed dissertation work, including the most import conclusions, limitations in the developed work and future work. It allows the understanding of the completion of the proposed objectives. From the received feedback, the proposed can be marked as completed, and the created artifact can be used in every banking institution for the implementation of an automated machine learning tool across the organization.

5.1. DEVELOPED WORK SYNTHESIS

This investigation obtained an overview of different machine learning subjects and the encapsulated subject, to Artificial Intelligence, Deep Learning, Data Science and Automated Machine Learning. The gathered knowledge made possible the definition of an implementation framework to support a roadmap to the implementation of an automated machine learning tool, that was posteriorly validated by area specialists to check the usefulness in the banking organizations' environment, meaning the completion of all the objects proposed in the beginning of the work.

5.2. LIMITATIONS

The investigation had limitations, mainly related the framework validation, it was approved by some organization experts, however it could have been interviewed more people, to provide a more universal validation and more suggestions of improvement. The pandemic reduced the availability and disposition of several participants, and the questions were considered uninviting by some participants.

The participants were a little reluctant to include information about their organizations, the existent workflows, the current technological conditions, the type of use cases, so the questionnaire tried to be less intrusive as possible.

There were made efforts to include workflows and use cases from several organizations, and feedback from workers in different organizations, however, due personnel unavailability, there was the impossibility of gathering information from the all the different national banking organizations. The dissertation work assumed that, in general, the type of analytical use cases is going to be similar in the different organizations.

Lastly, the framework practical application would benefit, as said before, if it was possible to study different scenarios leading to collect more insights to widen the framework's scope. The framework refers different technological tools since it is always evolving and what is used and available today may not be adequate in the future. Therefore, should be considered the rising and obsolete technologies in future uses of this work.

5.3. FUTURE WORK

Regarding future work, the validation process could be improved with more insights from a larger number of people, making the framework as much universal and accessible as possible. Different users in diverse banking organizations should be interviewed to collect information from different scenarios.

Due to the constant evolution of the technological world and the exponential growth of knowledge in the machine learning field, it would benefit from the continuous update of technologies and automated machine learning tools. As automated machine learning and machine learning are growing and the new developments are being launched almost daily, inevitably it would be new tools, a follow-up would be essential for future work. Apart from that, a larger list of possible tools to be implemented will widen the range of options in every implementation process and give more options to the implementation team.

This work would be also benefit from a study field application, different testing scenarios, types of users, use cases and technologies and the obtained would be an important reinforcement input of this work.

This investigation would also benefit from a public communication and a field application to study and test different scenarios, types of users and technologies. The retrieved data from that operation would represent an important input to the enhancement of this work.

Finally, turning this investigation available to the academic world through, for instance, a publication, would also made it more accessible to the public and allow the arise of other investigations.

BIBLIOGRAPHY

(15) (PDF) Random Forests and Decision Trees. (n.d.). Retrieved March 2, 2021, from https://www.researchgate.net/publication/259235118_Random_Forests_and_Decision_Trees

(15) (PDF) Short Survey on Naive Bayes Algorithm. (n.d.). Retrieved March 2, 2021, from https://www.researchgate.net/publication/323946641_Short_Survey_on_Naive_Bayes_Algorithm

18 Types of Bank Services. (n.d.). Retrieved February 8, 2021, from https://www.iedunote.com/bank-services

A brief review of feed-forward neural networks. (2014). *January 2006*, 10–17. https://doi.org/10.1501/0003168

Ahmed, M., Seraj, R., & Islam, S. M. S. (2020). The k-means algorithm: A comprehensive survey and performance evaluation. *Electronics (Switzerland)*, *9*(8), 1–12. https://doi.org/10.3390/electronics9081295

AI for Banking | Powered by dotData's AutoML 2.0 Platform. Retrieved February 7, 2021, from https://dotdata.com/banking/#1569241283009-def5e51e-5163

Albawi, S., & Mohammed, T. A. (2017). *Understanding of a Convolutional Neural Network*.

Anava, O., & Levy, K. Y. (2016). *k ⇤ -Nearest Neighbors : From Global to Local*. *Nips*, 1–9.

Andreu-perez, J. (2017). Artificial Intelligence and Robotics Javier Andreu Perez , Fani Deligianni , Daniele Ravi and Guang-Zhong Yang. *ResearchGate*, *June*.

Arnold, L., Rebecchi, S., & Chevallier, S. (2019). *An Introduction to Deep Learning An Introduction to Deep Learning To cite this version : HAL Id : hal-01352061*. *January 2011*.

Article, R., Access, O., Behera, A. K., Dehuri, S., & Cho, S. (2016). *Radial basis function neural networks : a topical state-of-the-art survey RBFNs architecture*. 33–63. https://doi.org/10.1515/comp-2016-0005

Balaji, A., & Allen, A. (n.d.). *Benchmarking Automatic Machine Learning Frameworks*.

Balaji, A., & Allen, A. (2018). *Benchmarking Automatic Machine Learning Frameworks*. http://arxiv.org/abs/1808.06492

Bank Definition. (n.d.). Retrieved February 7, 2021, from https://www.investopedia.com/terms/b/bank.asp

Banking: How It Works,Types, How It's Changed. (n.d.). Retrieved February 7, 2021, from https://www.thebalance.com/what-is-banking-3305812

Banking Services. (n.d.). Retrieved February 8, 2021, from https://www.practicalbusinessskills.com/managing-a-business/financial-management/banking-services

Bezrukavnikov, O., & Linder, R. (2021). A Neophyte With AutoML: Evaluating the Promises of Automatic Machine Learning Tools. In *Proceedings of* (Vol. 1, Issue 1). Association for Computing Machinery. http://arxiv.org/abs/2101.05840

Biau, G. (2010). Analysis of a Random Forests Model. *Journal of Machine Learning Research*, *13*, 1063–1095. http://arxiv.org/abs/1005.0208

Capó, M., Pérez, A., & Lozano, J. A. (2018). *An efficient K -means clustering algorithm for massive data*. *14*(8), 1–14. http://arxiv.org/abs/1801.02949

Chen, M., Challita, U., Saad, W., Yin, C., & Debbah, M. (2019). Artificial Neural Networks-Based Machine Learning for Wireless Networks: A Tutorial. *IEEE Communications Surveys and Tutorials*, *21*(4), 3039–3071. https://doi.org/10.1109/COMST.2019.2926625

Choi, R. Y., Coyner, A. S., Kalpathy-Cramer, J., Chiang, M. F., & Peter Campbell, J. (2020a). Introduction to machine learning, neural networks, and deep learning. *Translational Vision Science and Technology*, *9*(2), 1–12. https://doi.org/10.1167/tvst.9.2.14

Choi, R. Y., Coyner, A. S., Kalpathy-Cramer, J., Chiang, M. F., & Peter Campbell, J. (2020b). Introduction to machine learning, neural networks, and deep learning. *Translational Vision Science and Technology*, *9*(2), 14–14. https://doi.org/10.1167/tvst.9.2.14

Cloud AutoML: modelos de machine learning personalizados. (n.d.). Retrieved October 17, 2021, from https://cloud.google.com/automl/

Cockburn, I. M., Henderson, R., & Stern, S. (2018). NBER WORKING PAPER SERIES - The Impact of Artificial Intelligence on Innovation. *National Bureau of Economic Research WORKING PAPER SERIES*, *Working Pa*. http://www.nber.org/papers/w24449%0Ahttp://www.nber.org/papers/w24449.ack

Custode, L. L., & Iacca, G. (2020). *Evolutionary learning of interpretable decision trees CC-BY-NC-ND 4.0 http://creativecommons.org/licenses/by-nc-nd/4.0*.

Dai, D., Tan, W., & Zhan, H. (n.d.). *Understanding the Feedforward Artificial Neural Network Model From the Perspective of Network Flow Abstract :*

Deelman, E., Mandal, A., Jiang, M., & Sakellariou, R. (2019). *The role of machine learning in scientific workflows*. https://doi.org/10.1177/1094342019852127

Different Types of Banks - SmartAsset. (n.d.). Retrieved February 7, 2021, from https://smartasset.com/checking-account/types-of-banks

Dobrescu, A., Giuffrida, M. V., & Tsaftaris, S. A. (2020). *Doing More With Less : A Multitask Deep Learning Approach in Plant Phenotyping*. *11*(February), 1–11. https://doi.org/10.3389/fpls.2020.00141

Elshawi, R., Maher, M., & Sakr, S. (2019). *Automated Machine Learning: State-of-The-Art and Open Challenges*. http://arxiv.org/abs/1906.02287

Emmert-Streib, F., Yang, Z., Feng, H., Tripathi, S., & Dehmer, M. (2020). An Introductory Review of Deep Learning for Prediction Models With Big Data. *Frontiers in Artificial Intelligence*, *3*(February), 1–23. https://doi.org/10.3389/frai.2020.00004

Essinger, S. D., & Rosen, G. L. (2011). An introduction to machine learning for students in secondary education. *2011 Digital Signal Processing and Signal Processing Education Meeting, DSP/SPE 2011 - Proceedings*, *February 2011*, 243–248. https://doi.org/10.1109/DSP-SPE.2011.5739219

Faußer, S., & Schwenker, F. (2013). *Neural Network Ensembles in Reinforcement Learning*. *November*. https://doi.org/10.1007/s11063-013-9334-5

Feurer, M., Klein, A., Jost, K. E., Springenberg, T., Blum, M., & Hutter, F. (n.d.). *Efficient and Robust Automated Machine Learning*.

Fontana, P. (2008). *A Combination of Decision Trees and Instance-Based Learning Master ' s Scholarly Paper*.

Ganaie, M. A., Hu, M., Tanveer, M., & Suganthan, P. N. (n.d.). *Ensemble deep learning : A review*.

Gijsbers, P., Ledell, E., Thomas, J., Poirier, S., Bischl, B., & Vanschoren, J. (n.d.). *An Open Source AutoML Benchmark*. Retrieved January 30, 2021, from https://aws.amazon.com/ec2/instance-types/m5/

Gijsbers, P., LeDell, E., Thomas, J., Poirier, S., Bischl, B., & Vanschoren, J. (2019). *An Open Source AutoML Benchmark*. 1–8. http://arxiv.org/abs/1907.00909

GitHub - datarobot/datarobot-sagemaker-examples: This repository contains some sample notebooks illustrating the use of DataRobot and SageMaker. (n.d.). Retrieved October 17, 2021, from https://github.com/datarobot/datarobot-sagemaker-examples

GitHub - ludwig-ai/ludwig: Ludwig is a toolbox that allows to train and evaluate deep learning models without the need to write code. (n.d.). Retrieved October 17, 2021, from https://github.com/ludwig-ai/ludwig

GitHub - salesforce/TransmogrifAI: TransmogrifAI (pronounced trăns-mŏgˈrə-fī) is an AutoML library for building modular, reusable, strongly typed machine learning workflows on Apache Spark with minimal hand-tuning. (n.d.). Retrieved October 17, 2021, from https://github.com/salesforce/TransmogrifAI

Graph, P., Feb, L. G., Talamantes, A., & Chavez, E. (2021). *Instance-based learning using the Half-Space*. 1–17.

Haton, J. P. (2006). A brief introduction to artificial intelligence. In *IFAC Proceedings Volumes (IFAC-PapersOnline)* (Vol. 9, Issue PART 1). IFAC. https://doi.org/10.3182/20060522-3-fr-2904.00003

He, X., Zhao, K., & Chu, X. (2021). AutoML: A survey of the state-of-the-art. *Knowledge-Based*

Systems, *212*, 106622. https://doi.org/10.1016/j.knosys.2020.106622

Hinton, G. E., & Osindero, S. (2006). *A fast learning algorithm for deep belief nets ∗500 units 500 units*.

Howley, T., Madden, M. G., O'Connell, M. L., & Ryder, A. G. (2006). The effect of principal component analysis on machine learning accuracy with high-dimensional spectral data. *Knowledge-Based Systems*, *19*(5), 363–370. https://doi.org/10.1016/j.knosys.2005.11.014

Hu, L., Chen, J., Vaughan, J., Yang, H., Wang, K., Sudjianto, A., & Nair, V. N. (2020). *Supervised Machine Learning Techniques: An Overview with Applications to Banking Corporate Model Risk, Wells Fargo*.

Hua, Y., Guo, J., & Zhao, H. (2015). *Deep Belief Networks and deep learning*. 1–4.

Huang, J., Chai, J., & Cho, S. (2020). Deep learning in finance and banking: A literature review and classification. *Frontiers of Business Research in China*, *14*(1). https://doi.org/10.1186/s11782-020-00082-6

Indolia, S., Kumar, A., Mishra, S. P., & Asopa, P. (2018). ScienceDirect Conceptual Understanding of Convolutional Neural Network- A Deep Learning Approach. *Procedia Computer Science*, *132*, 679–688. https://doi.org/10.1016/j.procs.2018.05.069

Internet Society. (2017). Artificial Intelligence and Machine Learning: Policy Paper. *Artificial Intelligence*, *April*. https://www.internetsociety.org/resources/doc/2017/artificial-intelligence-and-machine-learning-policy-paper/?gclid=CjwKCAjw8qjnBRA-EiwAaNvhwHSr9CPjaPfF-_p9bD8HmtUsO0PR2Yy-_SQrFw-Ruia94PHsro4STRoCi7IQAvD_BwE#_ftn7

Jin, H., Song, Q., & Hu, X. (2019). Auto-keras: An efficient neural architecture search system. *Proceedings of the ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, 1946–1956. https://doi.org/10.1145/3292500.3330648

Jollife, I. T., & Cadima, J. (2016). Principal component analysis: A review and recent developments. *Philosophical Transactions of the Royal Society A: Mathematical, Physical and Engineering Sciences*, *374*(2065). https://doi.org/10.1098/rsta.2015.0202

Jones, L. D., Golan, D., Hanna, S. A., & Ramachandran, M. (2018). Artificial intelligence, machine learning and the evolution of healthcare: A bright future or cause for concern? In *Bone and Joint Research* (Vol. 7, Issue 3, pp. 223–225). British Editorial Society of Bone and Joint Surgery. https://doi.org/10.1302/2046-3758.73.BJR-2017-0147.R1

Kamel, M. S., & Raafat, H. (1996). *Modular neural network architectures for classification*. *July*. https://doi.org/10.1109/ICNN.1996.549082

Kasabov, N., Pang, S., Engineering, K., Bag, P., & Zealand, N. (2003). *TRANSDUCTIVE SUPPORT VECTOR MACHINES AND APPLICATIONS IN*. 1–6.

Khanum, M., Mahboob, T., Imtiaz, W., Abdul Ghafoor, H., & Sehar, R. (2015). A Survey on Unsupervised Machine Learning Algorithms for Automation, Classification and Maintenance. *International Journal of Computer Applications*, *119*(13), 34–39. https://doi.org/10.5120/21131-4058

Kotsiantis, S., Tsekouras, G. E., & Pintelas, P. E. (2005). *Bagging Model Trees for Classification Problems . December 2013*.

Lecun, Y., Bengio, Y., & Hinton, G. (2015). Deep learning. In *Nature* (Vol. 521, Issue 7553, pp. 436–444). Nature Publishing Group. https://doi.org/10.1038/nature14539

Li, Y., & Wu, H. (2012). A Clustering Method Based on K-Means Algorithm. *Physics Procedia*, *25*, 1104–1109. https://doi.org/10.1016/j.phpro.2012.03.206

Lima, S. (2018). *Deep learning for fraud detection in the banking industry*. *December*.

Lipton, Z. C. (2015). *A Critical Review of Recurrent Neural Networks for Sequence Learning*. *June*.

Livieris, I. E., Kanavos, A., Tampakas, V., & Pintelas, P. (2018). An auto-adjustable semi-supervised self-training algorithm. *Algorithms*, *11*(9), 1–16. https://doi.org/10.3390/a11090139

Lobo, V. J. A. S. (n.d.). *Application of Self-Organizing Maps to the Maritime Environment*.

Machine Learning in Banking - Opportunities, Risks, Use Cases. (n.d.-a). Retrieved February 8, 2021, from https://spd.group/machine-learning/machine-learning-in-banking/

Machine Learning in Banking - Opportunities, Risks, Use Cases. (n.d.-b). Retrieved February 13, 2021, from https://spd.group/machine-learning/machine-learning-in-banking/

MachineLearningNotebooks/how-to-use-azureml/automated-machine-learning at master · Azure/MachineLearningNotebooks · GitHub. (n.d.). Retrieved October 17, 2021, from https://github.com/Azure/MachineLearningNotebooks/tree/master/how-to-use-azureml/automated-machine-learning

Mccallum, R. A. (n.d.). *Instance-Based State Identification for Reinforcement Learning*.

Mesquita, S. (n.d.). *Introduction to Data Science Kelleher and Tierney (2018) Data Science . Boston : The MIT Press Essential Knowledge Series . URL :*

Miljković, D. (2017). *Brief Review of Self-Organizing Maps*. *October*. https://doi.org/10.23919/MIPRO.2017.7973581

Mining, D., & Wasilewska, A. (n.d.). *Modular Neural Networks*.

Nasteski, V. (2017). An overview of the supervised machine learning methods. *HORIZONS.B*, *4*, 51–62. https://doi.org/10.20544/horizons.b.04.1.17.p05

Ouali, Y., Hudelot, C., & Tami, M. (n.d.). *An Overview of Deep Semi-Supervised Learning*.

Oussidi, A., & Elhassouny, A. (2018). Deep generative models: Survey. *2018 International Conference on Intelligent Systems and Computer Vision, ISCV 2018*, *2018-May*, 1–8. https://doi.org/10.1109/ISACV.2018.8354080

Overview — Using Driverless AI 1.10.0 documentation. (n.d.). Retrieved October 17, 2021, from https://docs.h2o.ai/driverless-ai/latest-stable/docs/userguide/index.html

Pang, T., Xu, K., Li, C., Song, Y., Ermon, S., & Zhu, J. (2020). Efficient learning of generative models via finite-difference score matching. *Advances in Neural Information Processing Systems*, *2020-Decem*(NeurIPS), 1–25.

Patel, H. H., & Prajapati, P. (2018). Study and Analysis of Decision Tree Based Classification Algorithms. *International Journal of Computer Sciences and Engineering*, *6*(10), 74–78. https://doi.org/10.26438/ijcse/v6i10.7478

Peter, B. (2015). *Bagging , Boosting and Ensemble Methods Bagging , Boosting and Ensemble Methods* (Issue January 2012). https://doi.org/10.1007/978-3-642-21551-3

Rätsch, G. (n.d.). *A Brief Introduction into Machine Learning*. Retrieved March 2, 2021, from http://www.tuebingen.mpg.de/˜raetsch

Real, E., Liang, C., So, D. R., & Le, Q. V. (2020). Automl-zero: Evolving machine learning algorithms from scratch. *37th International Conference on Machine Learning, ICML 2020*, *PartF16814*, 7963–7975.

Ren, Q., Cheng, H., & Han, H. (2017a). ARTICLES YOU MAY BE INTERESTED IN Comparison of artificial neural network, random forest and random perceptron forest for forecasting the spatial impurity distribution AIP Conference Research on Machine Learning Framework Based on Random Forest Algorithm. *The Journal of Chemical Physics*, *1820*, 170901. https://doi.org/10.1063/1.4977376

Ren, Q., Cheng, H., & Han, H. (2017b). Comparison of artificial neural network, random forest and random perceptron forest for forecasting the spatial impurity distribution AIP Conference Research on Machine Learning Framework Based on Random Forest Algorithm. *The Journal of Chemical Physics*, *1820*, 170901. https://doi.org/10.1063/1.4977376

Rokach, L., & Maimon, O. (2006). Decision Trees. In *Data Mining and Knowledge Discovery Handbook* (pp. 165–192). Springer-Verlag. https://doi.org/10.1007/0-387-25465-x_9

Rong, S., & Bao-Wen, Z. (2018). The research of regression model in machine learning field. *MATEC Web of Conferences*, *176*. https://doi.org/10.1051/matecconf/201817601033

Ruder, S. (2017). *An Overview of Multi-Task Learning in Deep Neural Networks ∗ arXiv : 1706 . 05098v1 [cs . LG] 15 Jun 2017*. *May*.

Ruthotto, L., & Haber, E. (2021). An introduction to deep generative modeling. *GAMM Mitteilungen*, *44*(2), 1–26. https://doi.org/10.1002/gamm.202100008

Salem, N., & Hussein, S. (2019). Data dimensional reduction and principal components analysis. *Procedia Computer Science*, *163*, 292–299. https://doi.org/10.1016/j.procs.2019.12.111

Sanjeev, C., Dash, K., Behera, A. K., & Dehuri, S. (2016). *Radial basis function neural networks : A* *topical state-of-the-art survey Radial basis function neural networks : a topical state-of-the-art survey*. *January*. https://doi.org/10.1515/comp-2016-0005

Schlag, S., Schmitt, M., & Schulz, C. (n.d.). *Faster Support Vector Machines **.

Sherstinsky, A. (2020). *Fundamentals of Recurrent Neural Network (RNN) and Long Short-Term Memory (LSTM) Network*. *404*(March), 1–43.

Simeone, O. (n.d.). *A Very Brief Introduction to Machine Learning With Applications to Communication Systems*.

Singh, J. (2017a). Research Paper on Artificial Intelligence. *International Journal of Scientific Research and Management*, *6*, 7–14. https://doi.org/10.18535/ijsrm/v5i11.10

Singh, J. (2017b). Research Paper on Artificial Intelligence. *International Journal of Scientific Research and Management*, *05*(11), 7411–7417. https://doi.org/10.18535/ijsrm/v5i11.10

Sodhi, P., Awasthi, N., & Sharma, V. (2019). Introduction to Machine Learning and Its Basic Application in Python. In *SSRN Electronic Journal*. https://doi.org/10.2139/ssrn.3323796

Tang, Y. (2013). *Deep Learning using Linear Support Vector Machines*. http://arxiv.org/abs/1306.0239

Taunk, K. (2019). *A Brief Review of Nearest Neighbor Algorithm for Learning and Classification*. *Iciccs*, 1255–1260.

Tian, Y., Shi, Y., & Liu, X. (2012). Recent advances on support vector machines research. *Technological and Economic Development of Economy*, *18*(1), 5–33. https://doi.org/10.3846/20294913.2012.661205

Truong, A., Walters, A., Goodsitt, J., Hines, K., Bruss, B., & Farivar, R. (n.d.). *Towards Automated Machine Learning: Evaluation and Comparison of AutoML Approaches and Tools*.

Truong, A., Walters, A., Goodsitt, J., Hines, K., Bruss, C. B., & Farivar, R. (2019). Towards automated machine learning: Evaluation and comparison of AutoML approaches and tools. *Proceedings - International Conference on Tools with Artificial Intelligence, ICTAI*, *2019-November*. https://doi.org/10.1109/ICTAI.2019.00209

Types of Banks: Which Is Right for Your Needs? - TheStreet. (n.d.). Retrieved February 7, 2021, from https://www.thestreet.com/personal-finance/education/types-of-banks-14934713

Vaccaro, L., Sansonetti, G., & Micarelli, A. (2021). An Empirical Review of Automated Machine Learning. *Computers*, *10*(1), 11. https://doi.org/10.3390/computers10010011

Varghese, N. V., & Mahmoud, Q. H. (2020). *A Survey of Multi-Task Deep Reinforcement Learning*.

Wang, J. (2005). *On Transductive Support Vector Machines ∗*. *1998*.

Wei, C., Shen, K., Chen, Y., & Ma, T. (2020). *Theoretical Analysis of Self-Training with Deep Networks on Unlabeled Data*. 1–30. http://arxiv.org/abs/2010.03622

Weihs, C., & Ickstadt, K. (2018). Data Science : the impact of statistics. *International Journal of Data Science and Analytics*, *6*(3), 189–194. https://doi.org/10.1007/s41060-018-0102-5

Wever, M., Mohr, F., & Hüllermeier, E. (2018). ML-Plan for Unlimited-Length Machine Learning Pipelines. *International Conference on Machine Learning AutoML Workshop*.

What Are the Different Types of Banks? (n.d.). Retrieved February 7, 2021, from https://www.thebalance.com/types-of-banks-315214

Wu, D., Shang, M., Wang, G., & Li, L. (2018). A self-training semi-supervised classification algorithm based on density peaks of data and differential evolution. *ICNSC 2018 - 15th IEEE International Conference on Networking, Sensing and Control*, 1–6. https://doi.org/10.1109/ICNSC.2018.8361359

Wu, Y., Ianakiev, K., & Govindaraju, V. (2002). *Improved k -nearest neighbor classiÿcation*. *35*, 2311–2318.

Xue, S., Ma, Y., Yi, N., & Dodgson, T. E. (n.d.). *Approach for MIMO Signal Detection*. *1*, 1–12.

Yamashita, R., Nishio, M., Kinh, R., Do, G., & Togashi, K. (2018). *Convolutional neural networks : an overview and application in radiology*. 611–629.

Yang, L., & Shami, A. (2020). On hyperparameter optimization of machine learning algorithms: Theory and practice. *Neurocomputing*, *415*, 295–316. https://doi.org/10.1016/j.neucom.2020.07.061

Yao, Q., Wang, M., Chen, Y., Dai, W., Li, Y.-F., Tu, W.-W., Yang, Q., & Yu, Y. (2018). *Taking Human out* *of Learning Applications: A Survey on Automated Machine Learning*. 1–20. http://arxiv.org/abs/1810.13306

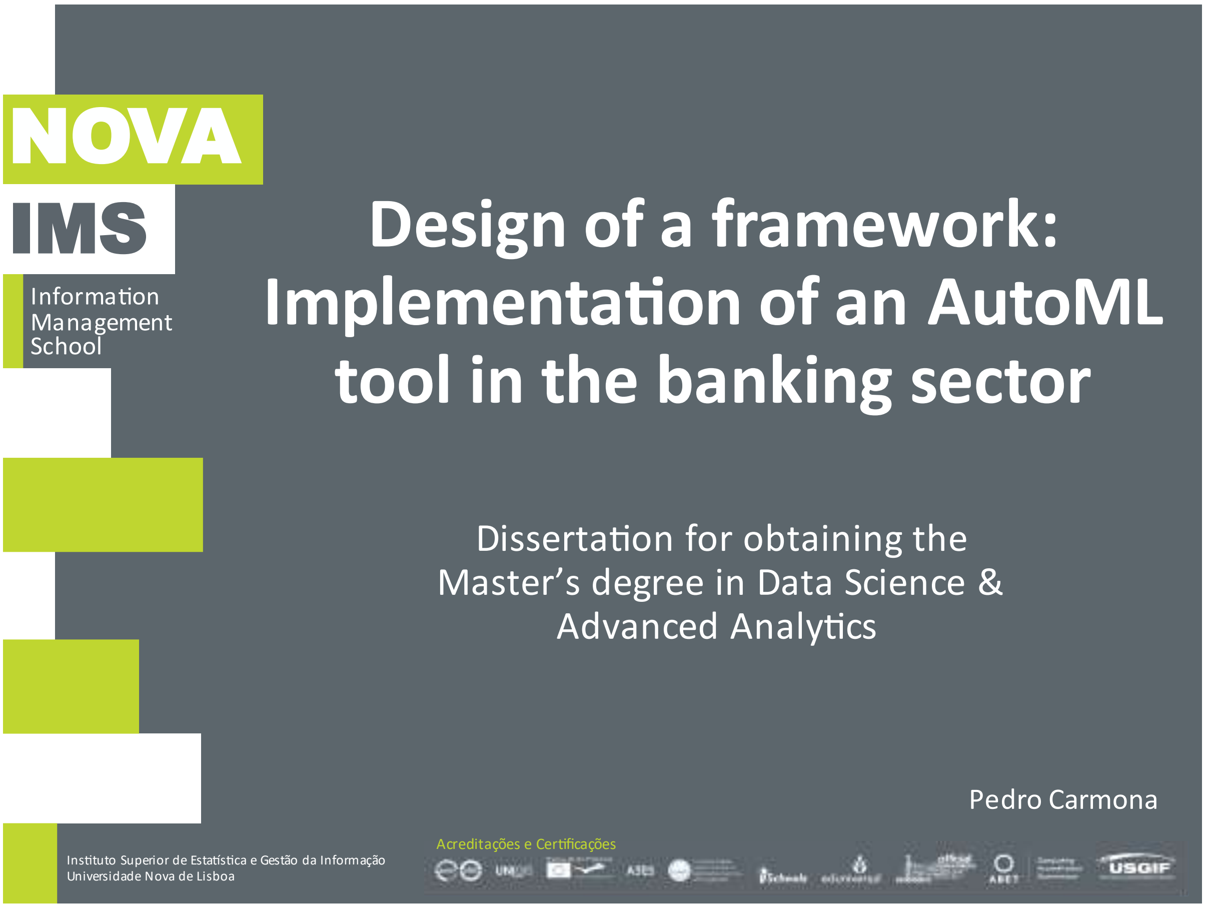
Zhou, Y., Kantarcioglu, M., & Thuraisingham, B. (2012). Self-training with selection-by-rejection. *Proceedings - IEEE International Conference on Data Mining, ICDM*, 795–803. https://doi.org/10.1109/ICDM.2012.56

Zhu, Y., & Xiong, Y. (2015). *Towards Data Science*. 1–7.

Zöller, M. A., & Huber, M. F. (2021). Benchmark and Survey of Automated Machine Learning Frameworks. *Journal of Artificial Intelligence Research*, *70*, 409–472. https://doi.org/10.1613/JAIR.1.11854

ANNEXES

Presentation used to explain the work to the interviewed experts



NOVA IMS

Information Management School

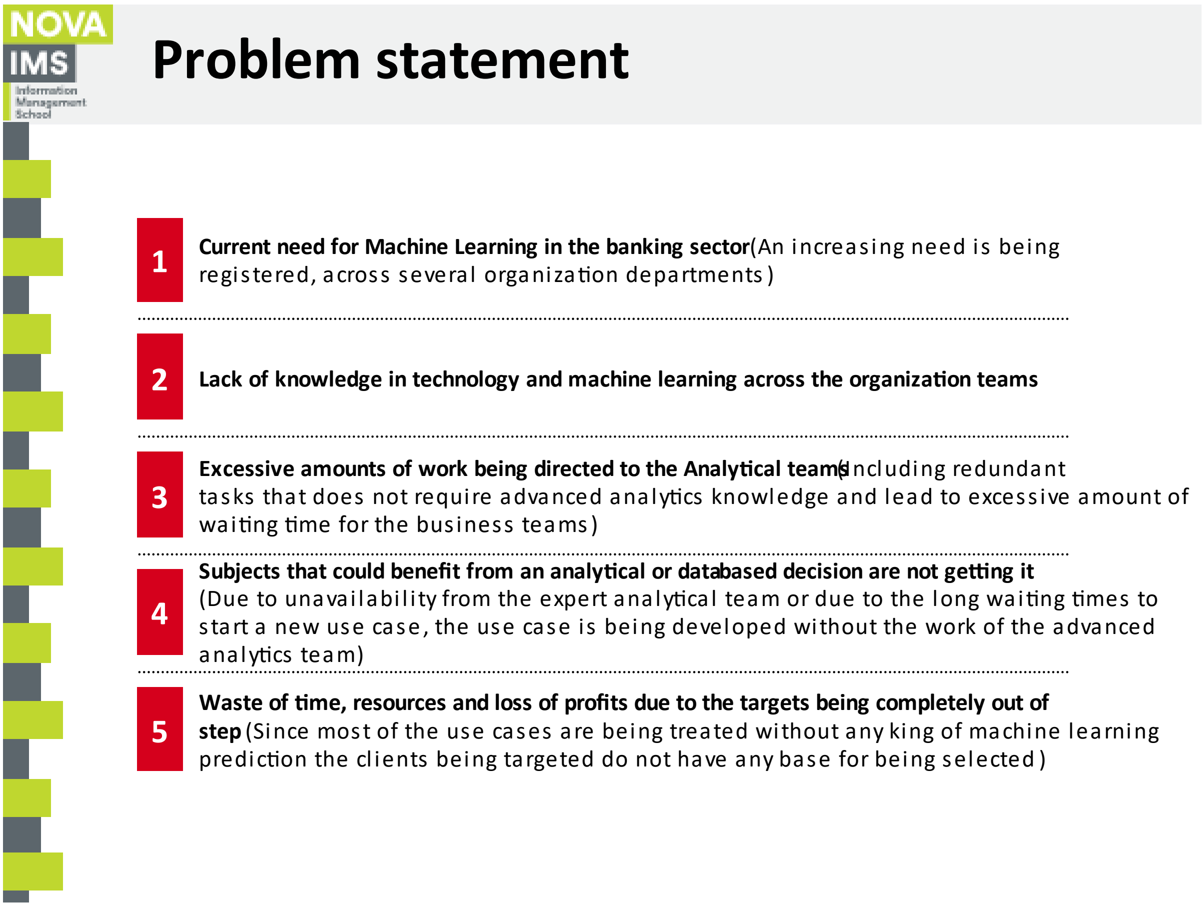
Design of a framework: Implementation of an AutoML tool in the banking sector

Dissertation for obtaining the Master's degree in Data Science & Advanced Analytics

Pedro Carmona

Instituto Superior de Estatística e Gestão da Informação
Universidade Nova de Lisboa

Acreditações e Certificações



Problem statement

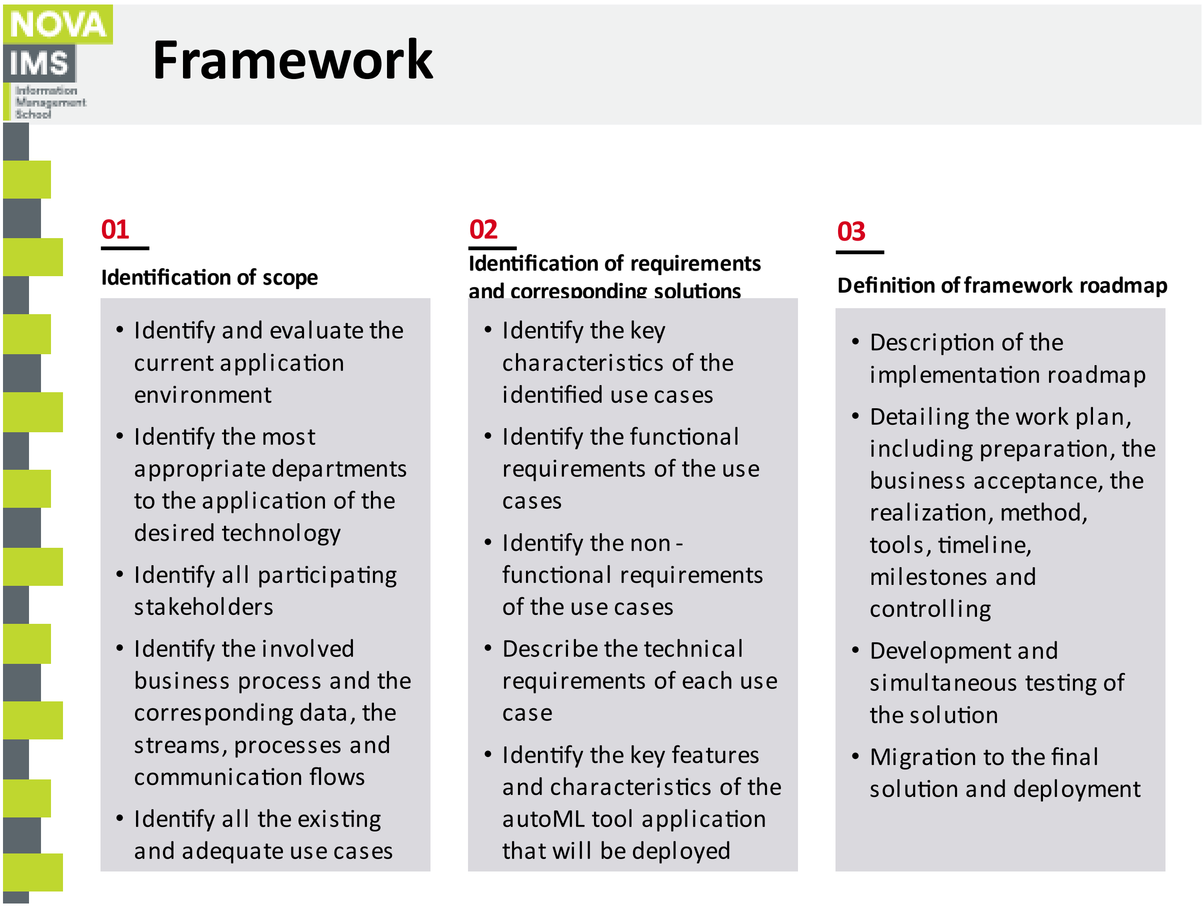
1 **Current need for Machine Learning in the banking sector**(An increasing need is being registered, across several organization departments)

2 **Lack of knowledge in technology and machine learning across the organization teams**

3 **Excessive amounts of work being directed to the Analytical teams**(including redundant tasks that does not require advanced analytics knowledge and lead to excessive amount of waiting time for the business teams)

4 **Subjects that could benefit from an analytical or databased decision are not getting it** (Due to unavailability from the expert analytical team or due to the long waiting times to start a new use case, the use case is being developed without the work of the advanced analytics team)

5 **Waste of time, resources and loss of profits due to the targets being completely out of step** (Since most of the use cases are being treated without any king of machine learning prediction the clients being targeted do not have any base for being selected)



Framework

01

Identification of scope

- Identify and evaluate the current application environment
- Identify the most appropriate departments to the application of the desired technology
- Identify all participating stakeholders
- Identify the involved business process and the corresponding data, the streams, processes and communication flows
- Identify all the existing and adequate use cases

02

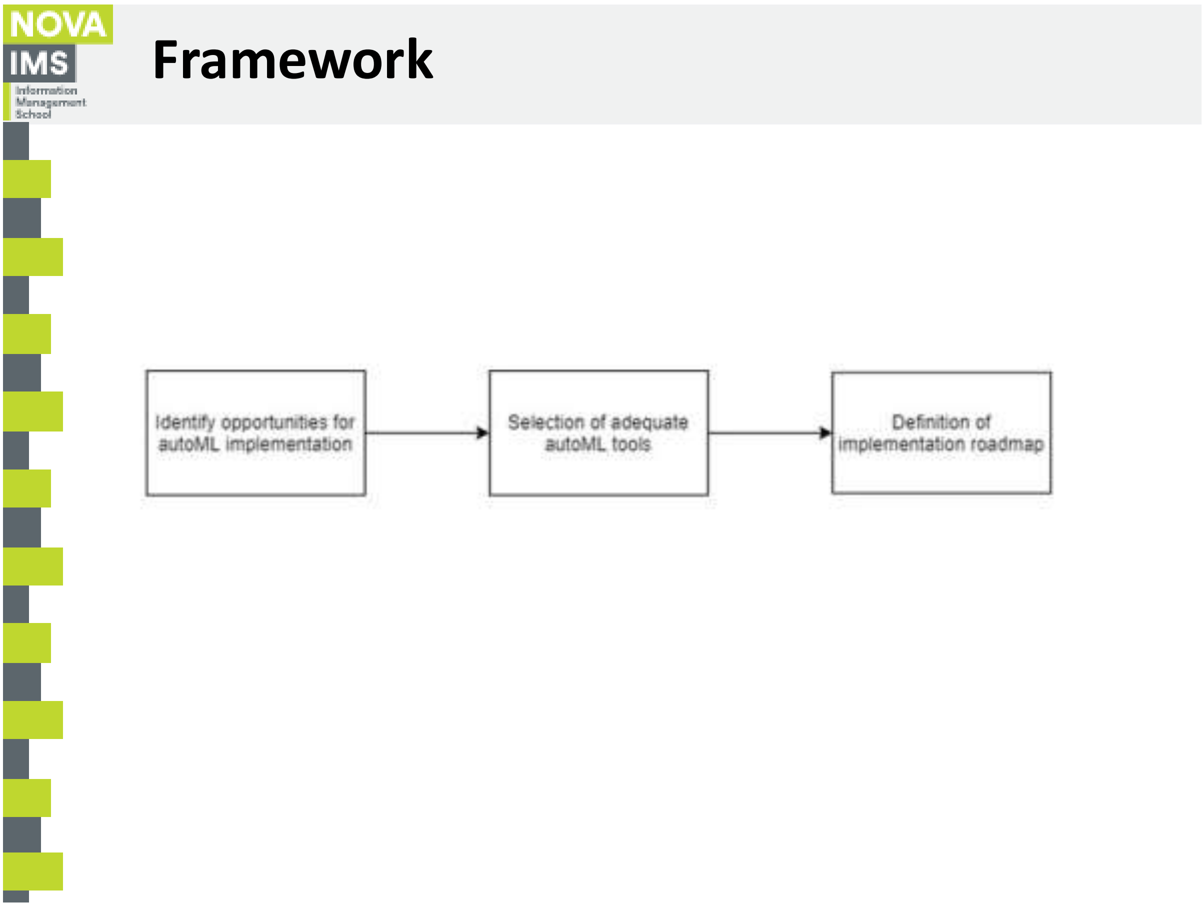
Identification of requirements and corresponding solutions

- Identify the key characteristics of the identified use cases
- Identify the functional requirements of the use cases
- Identify the non-functional requirements of the use cases
- Describe the technical requirements of each use case
- Identify the key features and characteristics of the autoML tool application that will be deployed

03

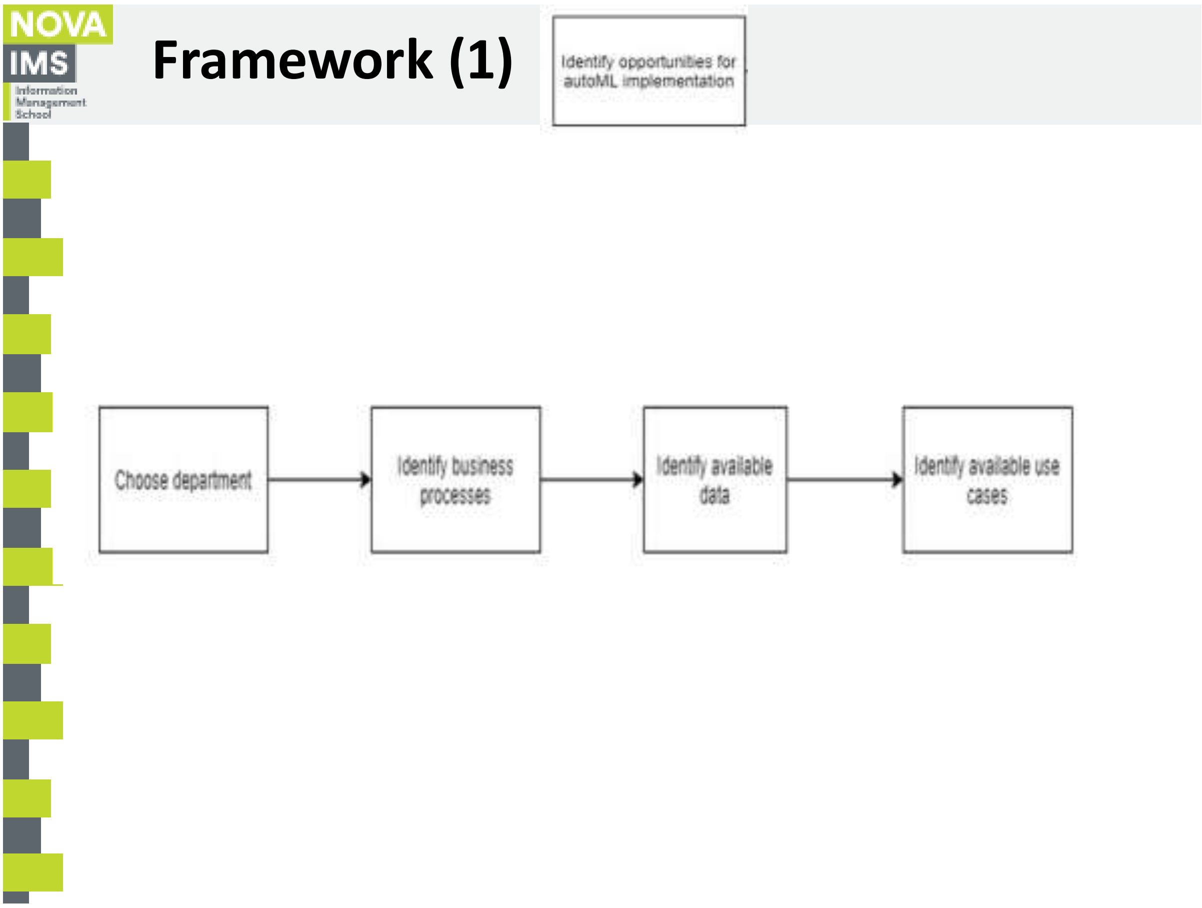
Definition of framework roadmap

- Description of the implementation roadmap
- Detailing the work plan, including preparation, the business acceptance, the realization, method, tools, timeline, milestones and controlling
- Development and simultaneous testing of the solution
- Migration to the final solution and deployment



Framework

Identify opportunities for autoML implementation → Selection of adequate autoML tools → Definition of implementation roadmap

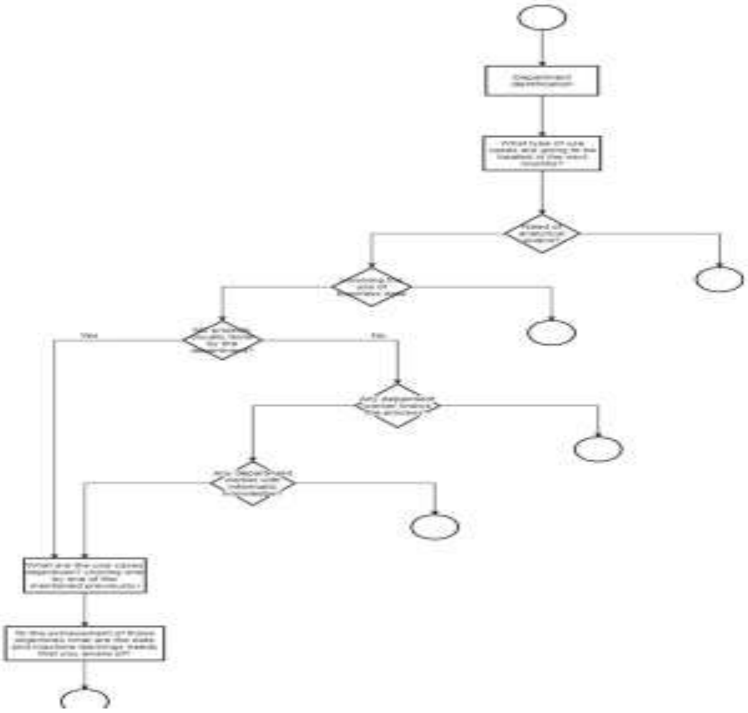


Framework (1)

Identify opportunities for autoML implementation

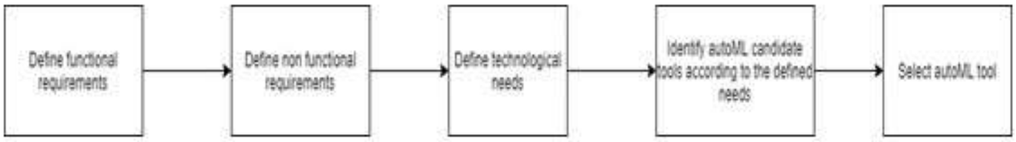
Choose department → Identify business processes → Identify available data → Identify available use cases

Framework (1.1)-Interviews flowchart



Framework (2)

Selection of adequate autoML tools



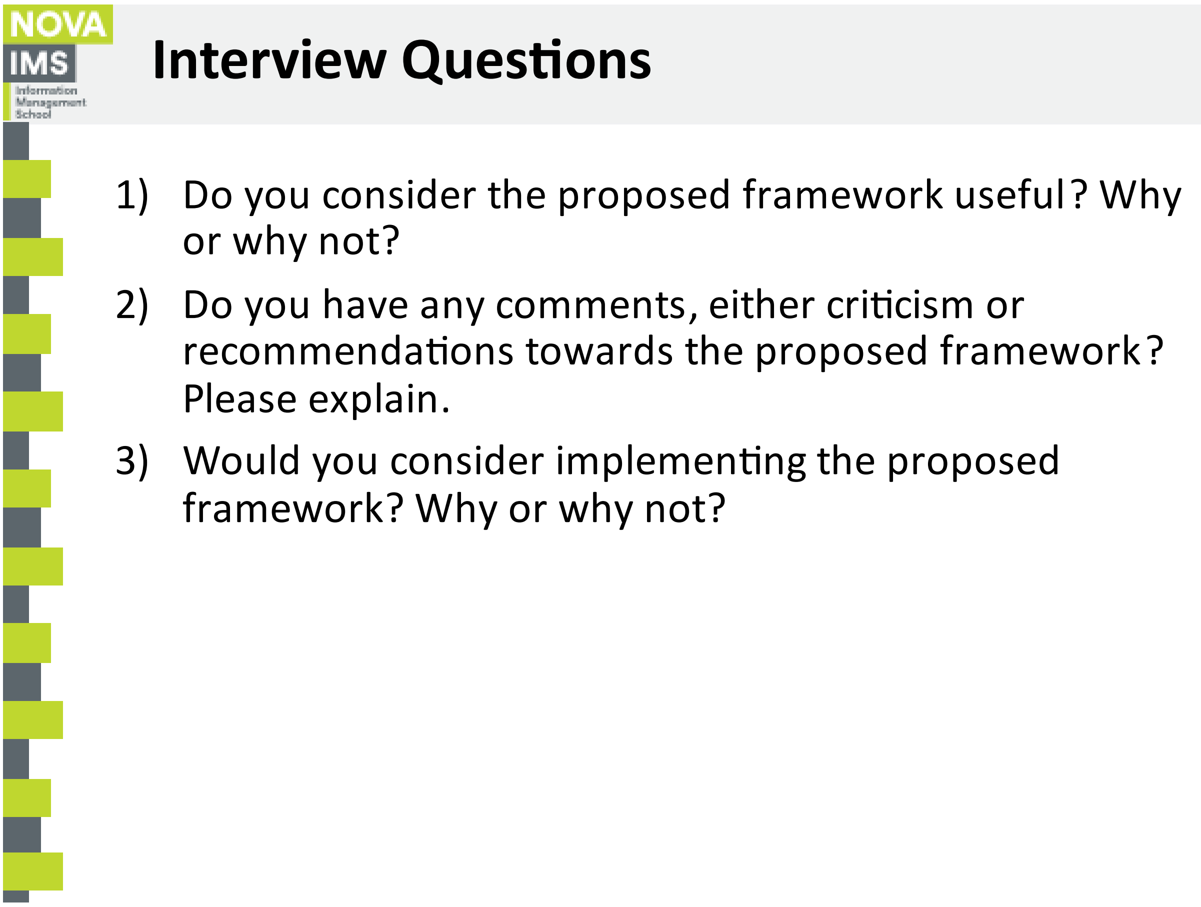
Define functional requirements

Define non functional requirements

Define technological needs

Identify autoML candidate tools according to the defined needs

Select autoML tool



NOVA IMS Information Management School

Interview Questions

1) Do you consider the proposed framework useful? Why or why not?
2) Do you have any comments, either criticism or recommendations towards the proposed framework? Please explain.
3) Would you consider implementing the proposed framework? Why or why not?

Thank you for your time and expertise!

Address Campus de Campolide, 1070-312 Lisboa, Portugal
Phone +351 213 828 610 Fax:+351 213 828 611

Acreditações e Certificações

Instituto Superior de Estatística e Gestão da Informação
Universidade Nova de Lisboa