ISLA Santarém

Fatores que influenciam a retenção de recursos humanos nas empresas

Relatório de Projeto Aplicado em Data Science realizado no âmbito da Pós-Graduação em Data Science

Julho de 2023

Resumo

Num mundo altamente competitivo e fortemente globalizado a procura pela vantagem competitiva tornou-se uma missão premente no atual contexto socioeconómico. Para isso é fundamental garantir a retenção dos melhores recursos humanos, na medida em que a perda dos elementos *core*, provoca uma perda de competências, conhecimento e contactos relevantes. É fundamental conhecer na sua génese os fatores que poderão potenciar a *attrition* dos colaboradores, procurando antecipar medidas estratégicas e mitigar os efeitos negativos. Neste estudo foram utilizados vários algoritmos de aprendizagem supervisionada e não supervisionada, como também a técnica estatística de análise de sobrevivência com o objetivo de conhecer os principais fatores que influenciam a *attrition* utilizando o *dataset* criado pela *IBM Watson Analytics* disponibilizado na plataforma *online Kaggle*. O algoritmo de agrupamento *K-means* foi aplicado aos colaboradores em *attrition* através das variáveis *MonthlyIncome e Age*, obtendo-se 2 clusters. Dos resultados do algoritmo *Apriori*, também aplicado aos colaboradores em *attrition*, destaca-se a regra de associação com suporte = 0.31, confiança = 0.87 e *lift* = 2.26, em que os colaboradores que trabalham há mais anos com a mesma chefia são também aqueles que exercem funções há mais anos na empresa. Relativamente aos algoritmos de classificação, o que apresenta melhores resultados após *cross-validation* é o *Support Vector Machine – Kernel Linear* com *F1- score* = 88%, *Accuracy* = 90%, *Precision* = 90%, *Recall* = 88% *e AUC* = 0.86. Esta análise avançada, com recurso a *machine learning* permite às empresas identificar fatores que influenciam a rotatividade e/ou demissões dos trabalhadores, prever tendências e desenvolver estratégias eficazes para a retenção do talento, implementando políticas de recursos humanos personalizadas, promover um ambiente de trabalho mais produtivo e reduzir os custos.

Palavras-chave: Algoritmos aprendizagem não supervisionada, Algoritmos aprendizagem supervisionada, Análise de sobrevivência, Gestão de recursos humanos, Retenção

Página em branco

Abstract

In a highly competitive and globally interconnected world, the pursuit of competitive advantage has become a pressing mission in the current socio-economic context. It is crucial to ensure the retention of the best human resources, as the loss of core elements leads to a loss of competencies, knowledge, and relevant contacts. Understanding the factors that may contribute to employee attrition at its core is essential in order to anticipate strategic measures and mitigate the negative effects. In this study, various supervised and unsupervised learning algorithms were utilized, along with the statistical technique of survival analysis, to identify the key factors influencing attrition using the dataset provided by IBM Watson Analytics on the Kaggle online platform. The K-means clustering algorithm was applied to employees in attrition using the variables MonthlyIncome and Age, resulting in 2 clusters. Among the results of the Apriori algorithm, which was also applied to employees in attrition, a notable association rule with support = 0.31, confidence = 0.87, and lift = 2.26 emerged, indicating that employees who have been working with the same manager for a longer period of time are also those who have been in their roles for a longer time within the company. Regarding classification algorithms, the Support Vector Machine - Linear Kernel achieved the best results after cross-validation, with an F1-score of 88%, accuracy of 90%, precision of 90%, recall of 88%, and AUC of 0.86. This advanced analysis utilizing machine learning enables companies to identify factors influencing employee turnover and dismissals, predict trends, and develop effective strategies for talent retention. It facilitates the implementation of personalized human resources policies, promotes a more productive work environment, and reduces costs.

Keywords: unsupervised learning algorithms, supervised learning algorithms, survival analysis, human resource management, retention

Página em branco

Agradecimentos

Página em branco

Índice

Lista de figuras

Lista de tabelas

Abreviaturas e Símbolos

Adaboost	*Adaptative Boosting*
ANN	*Artificial Neural Network*
AUC	*Area Under the Roc Curve*
FN	*False Negative*
FP	*False Positive*
IDE	*Integrated Development Environment*
K-NN	*K-Nearest Neighbors*
ROC	*Receiver Operating Characteristic*
SMOTE	*Synthetic Minority Over-sampling Technique*
SVM	*Support Vector Machine*
TN	*True Negative*
TP	*True Positive*
XGBoost	*Extreme Gradient Boosting*

Capítulo 1

Introdução

Num mundo altamente competitivo, complexo e em constante evolução, a retenção de colaboradores tem vindo a tornar-se uma necessidade premente para as empresas. A capacidade de assegurar a permanência dos colaboradores com as competências técnicas e comportamentais *core*, é fundamental para garantir o sucesso e o crescimento sustentável de uma Organização. Nesse contexto, a compreensão dos principais fatores que influenciam a retenção dos colaboradores e o desenvolvimento de estratégias eficazes torna-se uma das principais prioridades no mundo empresarial. Estes fatores encontram-se assim intrinsecamente associados às necessidades dos trabalhadores e como tal, deverão ser percecionadas para mitigar eventuais efeitos prejudiciais às Organizações (Haldorai *et al.*, 2019).

Para Yahia *et al*. (2021) o desafio da retenção dos colaboradores é relevante para as Organizações, uma vez que o *turnover* dos mesmos significa a perda de competências, experiências, pessoas e consequentemente incapacidade de assegurar novas oportunidades de negócio. Se por um lado a perda contínua de colaboradores para além de afetar a estrutura base da formação base da equipa e consequente redução da satisfação dos clientes devido à falta de estabilidade interna, por outro lado a sua saída resultará numa possível "fuga" de informação relevante para a constituição de vantagem competitiva (Ersoz Kaya & Korkmaz 2021).

Neste sentido torna-se imprescindível promover uma análise cuidada dos fatores que efetivamente influenciam a saída dos colaboradores com o objetivo de promover um planeamento analítico das estratégias a serem implementadas (e.g. através da implementação de um conjunto de medidas que visem a diminuição das demissões e a preservação do talento) e auxiliar no processo de tomada de decisão através da recolha de um elevado conjunto de dados.

É aqui que a Inteligência Artificial assume um papel preponderante no apoio à tomada de decisão com a máxima eficácia, possibilitando uma interpretação e a análise cuidadas e criteriosas dos dados de forma eficiente em termos de tempo e custo.

Neste trabalho irão ser implementados diferentes algoritmos de *machine learning*, nomeadamente algoritmos de agrupamento, associação, classificação e, por fim, análise de sobrevivência.

O principal objetivo do presente trabalho consiste em identificar os principais fatores que motivam o abandono na empresa e qual o seu comportamento/probabilidade de ocorrência ao longo do tempo.

Agrupamento

Os algoritmos de agrupamento procuram explorar a distribuição dos dados, definindo regras para os conjuntos com características semelhantes (Ahmed, Seraj, & Islam, 2020). Desta forma, os dados são divididos de acordo com os critérios de agrupamento, agrupando os dados automaticamente, uma vez que na aprendizagem não supervisionada, os dados de input não são rotulados.

O k-means caracteriza-se por ser um algoritmo de agrupamento bastante utilizado devido ao seu bom desempenho para conjuntos grandes de dados (Ghazal *et al.*, 2021).

Neste algoritmo, inicialmente, são identificados *k* pontos como centroides e cada um representa um conjunto de objetos. Cada conjunto contém objetos que se encontram a uma distância mínima do respetivo centroide (Ghazal *et al.*, 2021).

A distância euclidiana é a distância mais utilizada neste algoritmo, sendo calculada iterativamente. No entanto, poderão também ser utilizadas as distâncias de Manhattan e a de Minkowski (Ghazal *et al.*, 2021).

Associação

O algoritmo de associação *Apriori* recorre a regras de associação para encontrar padrões e relações frequentes entre itens de um *dataset* (Kong, Tian, Wu, & Wei, 2020). Neste método são realizadas iterações hierárquicas onde são gerados candidatos teste (Kong, Tian, Wu, & Wei, 2020). Destes, são selecionados os mais frequentes através de varrimento em todo o *dataset* para calcular o nível de suporte dos *itemsets* candidatos. Depois, de acordo com valor de suporte, confiança ou lift mínimo definido, são geradas as regras de associação que satisfaçam os valores definidos (Han, Kamber, & Pei, 2012).

As regras de associação que não sejam relevantes, ou seja, que não se incluam no limite mínimo definidos são descartadas (Han, Kamber, & Pei, 2012). O suporte e a confiança, respetivamente, refletem a utilidade e o grau de confiança das regras que foram geradas (Han, Kamber, & Pei, 2012). Assim, o suporte é definido pela frequência relativa de uma determinada combinação de itens em todo o conjunto de dados, sendo utilizado para filtrar regras pouco frequentes ou não relevantes. A confiança é a probabilidade condicional de que um item apareça em uma determinada transação, dado que todos os itens antecedentes também aparecem nessa mesma transação. O lift é a razão entre a confiança da regra e o suporte do item consequente.

Classificação

A classificação é um dos processos associados à aprendizagem supervisionada mais implementados nos sistemas inteligentes (Osisanwo *et al.*, 2017).

Os algoritmos de classificação são essenciais em *machine learning*, na medida em que viabilizam a automatização da classificação dos dados, prevendo resultados desconhecidos, segmentar grupos e ajudar na análise e interpretação dos dados (Hastie *et al.*, 2009).

Na classificação, o modelo é treinado e testado através da utilização de dados de treino e dados de teste, respetivamente, antes de ser utilizado para a realização de eventuais previsões.

No presente relatório serão analisados sete algoritmos de classificação, sendo que alguns serão analisados com diferentes variantes.

Naive Bayes

É um algoritmo que calcula a probabilidade de um evento ocorrer com base no acontecimento de um evento relacionado, através da aplicabilidade do Teorema de Bayes (Raina & Shafi, 2015). Por exemplo, na classificação de um e-mail como sendo spam ou não spam, o algoritmo usará a probabilidade de certas palavras ou frases para fazer a classificação spam ou não spam. Existem dois modelos, sendo que cada um possui as suas próprias suposições e restrições, designadamente: o *Naive Bayes Multinominal* que é mais adequado para atributos discretos e o *Naive Bayes Gaussian* é o mais adequado para atributos contínuos.

Decision Trees

É um algoritmo poderoso cujo processo de classificação é realizado estruturalmente em forma de árvore para modelar as diferentes relações entre as características para os dados de saída (Pineda-Jaramillo, 2019), procedendo à divisão dos dados em dois ou mais conjuntos homogéneos. Para a implementação deste modelo utilizou-se a classe de modelo *DecisionTreeClassifier* do módulo *sklearn.tree*.

Random Forest

Baseia-se em várias árvores de decisão, agregando os votos de diferentes árvores de decisão com o objetivo de determinar a classe final do objeto de teste. É assim um modelo mais robusto e com um melhor desempenho de generalização, sendo menos suscetível ao *overfitting* (Raschka & Mirjalili, 2019). Para a implementação deste modelo utilizou-se a classe de modelo *RandomForestClassifier* do módulo *sklearn.ensemble*.

Logistic Regression

É um algoritmo de classificação binária ou multiclasse de fácil implementação, e procura modelar a relação entre uma variável dependente (binária ou multiclasse) e uma ou mais variáveis

independentes (Raina & Shafi, 2015). Para a implementação deste modelo utilizou-se a classe de modelo *LogisticRegression* do módulo *sklearn.linear_model*.

Support Vector Machine

Consiste num classificador de aprendizagem da máquina supervisionada desenvolvido para a separação de duas classes (Raina & Shafi, 2015) . O objetivo do algoritmo consiste em encontrar um hiperplano entre diferentes classes do conjunto de dados, para que as mesmas possam ser classificadas. Uma boa separação pode ser alcançada se houver uma maior distância entre o hiperplano e o ponto mais próximo de qualquer classe, portanto, se houver uma margem maior, haverá menos erro na classificação. Maximizar a distância da margem fornece a garantia para que os pontos de dados futuros possam ser classificados com maior confiança (Ghandi, 2018). Para a implementação deste modelo utilizou-se a classe de modelo SVC do módulo *sklearn.svm*.

K-Nearest Neighbors (K-NN)

É de fácil implementação e perceção, onde o processo de classificação é realizado com base dos seus vizinhos mais próximos (Cunningham & Delany, 2020). Exige uma forte capacidade computacional podendo ficar mais lento à medida que os dados utilizados aumentam (Mahesh, 2020). O algoritmo K-NN é sensível à distribuição dos dados e à escolha do valor de K. A escolha incorreta de K pode levar a resultados indesejados, como superajuste ou subajuste do modelo. Para a implementação deste modelo utilizou-se a classe de modelo *KNeighborsClassifier* do módulo *sklearn.neighbors*.

ADABoost

Promove a combinação de múltiplos classificadores fracos, para aumentar a precisão dos mesmos. A Combinação de múltiplos algoritmos de classificação fracos ou de baixo desempenho (Hertzmanny, Fleet, Brubacker, 2015). O *AdaBoost* atribui pesos às instâncias de treino, ajustando-os ao longo das iterações, dando mais importância às instâncias mal classificadas anteriormente. Para a implementação deste modelo utilizou-se a classe de modelo *AdaBoostClassifier* do módulo *sklearn.ensemble*.

XGBoost

Segundo Chen e Guestrin (2016) o *XGBoost*, também conhecido como *Extreme Gradient Boosting*, é um algoritmo de aprendizagem da máquina que se baseia no método de *boosting* para construir modelos preditivos mais robustos, sendo amplamente utilizado para resolver problemas de classificação e regressão. O mesmo procura estabelecer uma combinação de várias árvores de decisão, sendo que através da utilização do método "bloqueio exato", consegue explorar todas as possibilidades e encontrar a melhor combinação de forma otimizada, em termos de consumo de tempo. O mesmo tem igualmente a capacidade de evitar o *overfitting*, utilizando uma função de perda regularizada. Para a implementação deste algoritmo utilizou-se a biblioteca *XGBOOST*.

Artificial Neural Network – ANN

São ferramentas computacionais que têm sido amplamente utilizadas em diversas disciplinas para modelar problemas complexos do mundo real (Liao & When, 2007). Para Dell' Aversana (2019) o Artificial Neural Networks (ANN) é inspirado nas redes neurais biológicas que formam o cérebro. No entanto, enquanto os algoritmos matemáticos são adequados para programação linear, cálculos aritméticos e lógicos, o ANN é mais eficaz para resolver problemas relacionados com o reconhecimento e correspondência de padrões, agrupamento e classificação. É assim mais eficazes para resolver problemas linearmente separáveis. Para a implementação deste algoritmo utilizou-se a classe do modelo *KerasClassifier* do módulo *tensorflow.keras.wrappers.sckikit_learn*.

Análise de Sobrevivência

A análise de sobrevivência consiste numa técnica estatística habitualmente utilizada para avaliar o tempo até à ocorrência de um determinado evento de interesse (Kartsonaki, 2016), procurando assim analisar os dados em forma de "tempo", ou seja, a partir de um período de origem até à ocorrência de um evento específico (Rai *et al.*, 2021). Poderá ser aplicada, por exemplo, na análise de dados como a morte, recidiva de uma doença, desenvolvimento de uma reação adversa ou de uma nova entidade de doença. Neste sentido o principal objetivo para a elaboração de uma análise de sobrevivência consiste em determinar a probabilidade de ocorrência de um determinado resultado num evento de interesse, observando o tempo até que esse evento ocorra e explorando as suas relações com diferentes fatores (Sobreiro, 2023).

A principal dificuldade prende-se com o facto de que, num determinado momento, apenas alguns indivíduos observaram o evento enquanto outros não. No entanto a análise de sobrevivência considera os indivíduos que experienciaram e os que não experienciaram a ocorrência de um determinado evento, utilizando os conceitos observação e censura, respetivamente. Considerando que a observação integra os dados associados ao tipo de eventos em análise, a censura integra os dados que não estão completamente relacionadas com o evento de interesse. Por exemplo, na análise de rotatividade de uma empresa, e no campo censura estarão integrados os trabalhadores que ainda não abandonaram a empresa num dado período, no entanto são igualmente considerados na respetiva análise. Isto significa que existem trabalhadores que ainda exercem funções na empresa dos quais não sabemos se o evento de abandono ocorreu e como tal a designação aplicada para estes casos é censura. Os modelos de sobrevivência têm assim em consideração, para além dos dados de observação, a censura e incorporam essa incerteza (Sobreiro, 2023).

Um dos objetivos da análise de sobrevivência consiste em estimar e representar graficamente a função de sobrevivência, a partir de um conjunto de dados, nomeadamente através da implementação da curva de *Kaplan-Meier* (Kartsonaki, 2016). A respetiva curva procura demonstrar a probabilidade de um evento não ocorrer (ou seja, a probabilidade de sobrevivência) em diferentes pontos no tempo. A curva começa em 1 no tempo inicial e, à medida que o tempo passa, a probabilidade de sobrevivência diminui à medida que os eventos ocorrem. Este é um método não paramétrico de estimativa da função de sobrevivência.

Para Katsonaki (2016) a análise de sobrevivência *Kaplan-Meier* consiste numa técnica relativamente simples de ser implementada, sendo útil para resumir os dados de sobrevivência e realizar comparações simples, não sendo tão adequada para lidar com situações mais complexas.

Sistema a Elaborar

O sistema que será implementado servirá para aplicar vários algoritmos de inteligência artificial sobre um conjunto de dados para analisar o seu funcionamento.

Serão implementados os algoritmos de aprendizagem não supervisionada, nomeadamente o de Agrupamento e Associação, os de aprendizagem supervisionada como os algoritmos de Classificação e, finalmente, a análise de sobrevivência.

Finalmente, e para apresentação dos principais resultados alcançados no presente estudo, foi desenvolvido um *dashboard* onde se incluem os elementos visuais correspondentes aos algoritmos e análise de sobrevivência implementados através de aplicação de código aberto *Jupyter Notebook*[1] e da extensão *Voilà*[2].

[1] Disponível em: https://jupyter.org/

[2] Disponível em: https://voila-gallery.org/

Caracterização do Conjunto de Dados

Os dados em análise são compostos por 1.470 observações, contendo 35 variáveis as quais representam os fatores intrinsecamente associados à saída permanente dos colaboradores, cujo *dataset* foi criado pela empresa *IBM Watson Analytics,* tendo o conjunto de dados sido obtido na plataforma *Kaggle*[3].

Na Tabela 1 - Tipos de dados do *dataset* podemos observar as variáveis em análise, como também o tipo de dados em estudo.

Tabela 1 - Tipos de dados do *dataset*

Variável	Tipo de dados	Tipo
Age	int64	Quantitativo
Attrition	object	Qualitativo
BusinessTravel	object	Qualitativo
DailyRate	int64	Quantitativo
Department	object	Qualitativo
DistanceFromHome	int64	Quantitativo
Education	int64	Quantitativo
EducationField	object	Qualitativo
EmployeeCount	int64	Quantitativo
EmployeeNumber	int64	Quantitativo
EnvironmentSatisfaction	int64	Quantitativo
Gender	object	Qualitativo
HourlyRate	int64	Quantitativo
JobInvolvement	int64	Quantitativo
JobLevel	int64	Quantitativo
JobRole	object	Qualitativo
JobSatisfaction	int64	Quantitativo
MaritalStatus	object	Qualitativo
MonthlyIncome	int64	Quantitativo
MonthlyRate	int64	Quantitativo
NumCompaniesWorked	int64	Quantitativo

[3] Conjunto de dados obtido em: https://www.kaggle.com/datasets/pavansubhasht/ibm-hr-analytics-attrition-dataset

Over18	object	Qualitativo
OverTime	object	Qualitativo
PercentSalaryHike	int64	Quantitativo
PerformanceRating	int64	Quantitativo
RelationshipSatisfaction	int64	Quantitativo
StandardHours	int64	Quantitativo
StockOptionLevel	int64	Quantitativo
TotalWorkingYears	int64	Quantitativo
TrainingTimesLastYear	int64	Quantitativo
WorkLifeBalance	int64	Quantitativo
YearsAtCompany	int64	Quantitativo
YearsInCurrentRole	int64	Quantitativo
YearsSinceLastPromotion	int64	Quantitativo
YearsWithCurrManager	int64	Quantitativo

As variáveis anteriormente identificadas e para melhor interpretação/perceção dos dados em análise encontram-se seguidamente apresentadas na Tabela 2:

Tabela 2 - Descrição das variáveis

Variável	Descrição
Age	Idade do trabalhador, variando no intervalo dos 18 aos 60 anos
Attrition	Indica se o trabalhador saiu (*Yes*) ou não (*Não*) da empresa
BusinessTravel	Número de viagens que o trabalhador realiza em trabalho, podendo ser "*Travel_Rarely*" (raramente), "*Travel_Frequently*" (frequentemente) ou "*Non-Travel*" (sem viagens)
DailyRate	O salário diário do trabalhador, cujos valores variam entre os 102 dólares e 1.499 dólares
Department	Departamento do trabalhador, integrando os departamentos *Research & Development*, *Human Resource* e *Sales*
DistanceFromHome	Distância entre a casa do trabalhador e o local de trabalho, variando entre 1 e os 29
Education	Nível de habilitações literárias do trabalhador, variando entre 1 e 5
EducationField	Área de estudo do trabalhador: *Human Resources, Life Sciences, Marketing, Medical,*

	Technical Degree ou *Other*
EmployeeCount	Corresponde ao trabalhador objeto do estudo, cujo valor é constante, ou seja, é igual a 1
EmployeeNumber	Número do colaborador, variando do 1 ao 2068
EnvironmentSatisfaction	Satisfação do trabalhador com o ambiente no trabalho, variando de 1 a 4
Gender	Género do trabalhador, ou seja, Homem ou Mulher.
HourlyRate	Salário por hora do trabalhador, variando de 30 a 100 dólares
JobInvolvement	Nível de envolvimento do trabalhador com o trabalho, variando de 1 a 4
JobLevel	Nível hierárquico do cargo do trabalhador, variando entre 1 e 5
JobRole	Função do trabalhador, nomeadamente *Research Scientist, Laboratory Technician, Manufacturing Director, Sales Executive, Healthcare Representative, Manager, Research Director, Human Resources, Sales Representative*
JobSatisfaction	Satisfação do trabalhador com as suas funções, variando de 1 a 4
MaritalStatus	Estado civil do trabalhador, nomeadamente *Single, Married, Divorced*
MonthlyIncome	Salário mensal do trabalhador, variando de 1.009 a 19.999 dólares
MonthlyRate	Taxa de compensação mensal bruta do trabalhador e varia entre 2.094 a 26.999€
NumCompaniesWorked	Número de empresas em que o trabalhador exerceu funções, variando entre 0 e 9
Over18	Indica se os trabalhadores possuem pelo menos 18 anos. Todos apresentam pelo menos essa idade sendo identificado como Y
OverTime	Execução ou não de horas extra, cujos dados se encontram identificados como *Yes* ou *No*
PercentSalaryHike	Aumento percentual no salário do trabalhador variando de 11 a 25%
PerformanceRating	Classificação do desempenho do trabalhador, variando de 3 e 4
RelationshipSatisfaction	Nível de satisfação do trabalhador com o ambiente laboral, variando de 1 a 4

StandardHours	Número de horas padrão trabalhadas, sendo constante e apresentando 80 horas
StockOptionLevel	Número de ações da empresa detidas pelo trabalhador, cujo valor varia entre 0 e 3
TotalWorkingYears	Anos de experiência de trabalho do trabalhador, variando de 0 a 40 anos
TrainingTimesLastYear	Número de formações que o trabalhador participou no último ano, variando de 0 a 6 formações
WorkLifeBalance	Equilíbrio entre trabalho e vida pessoal do trabalhador, variando de 1 a 4
YearsAtCompany	Número de anos trabalhados na empresa, variando de 0 a 40 anos
YearsInCurrentRole	Número de anos que o trabalhador está no cargo atual, variando de 0 a 18 anos
YearsSinceLastPromotion	Número de anos desde a última promoção do trabalhador, variando de 0 a 15 anos
YearsWithCurrManager	Número de anos que o trabalhador exerce funções com o atual Manager, variando de 0 a 17 anos

Na Tabela 3, são apresentados os dados estatísticos do *dataset* em análise.

Tabela 3 - Caracterização do *dataset*

Variáveis	count	mean	std	min	25%	50%	75%	max
Age	1470	36,92	9,14	18,00	30,00	36,00	43,00	60,00
DailyRate	1470	802,49	403,51	102,00	465,00	802,00	1157,00	1499,00
DistanceFromHome	1470	9,19	8,11	1,00	2,00	7,00	14,00	29,00
Education	1470	2,91	1,02	1,00	2,00	3,00	4,00	5,00
EmployeeCount	1470	1,00	0,00	1,00	1,00	1,00	1,00	1,00
EmployeeNumber	1470	1024,87	602,02	1,00	491,25	1020,50	1555,75	2068,00
EnvironmentSatisfaction	1470	2,72	1,09	1,00	2,00	3,00	4,00	4,00
HourlyRate	1470	65,89	20,33	30,00	48,00	66,00	83,75	100,00
JobInvolvement	1470	2,73	0,71	1,00	2,00	3,00	3,00	4,00
JobLevel	1470	2,06	1,11	1,00	1,00	2,00	3,00	5,00
JobSatisfaction	1470	2,73	1,10	1,00	2,00	3,00	4,00	4,00
MonthlyIncome	1470	6502,93	4707,96	1009,00	2911,00	4919,00	8379,00	19999,00
MonthlyRate	1470	14313,10	7117,79	2094,00	8047,00	14235,50	20461,50	26999,00
NumCompaniesWorked	1470	2,69	2,50	0,00	1,00	2,00	4,00	9,00
PercentSalaryHike	1470	15,21	3,66	11,00	12,00	14,00	18,00	25,00
PerformanceRating	1470	3,15	0,36	3,00	3,00	3,00	3,00	4,00

RelationshipSatisfaction	1470	2,71	1,08	1,00	2,00	3,00	4,00	4,00
StandardHours	1470	80,00	0,00	80,00	80,00	80,00	80,00	80,00
StockOptionLevel	1470	0,79	0,85	0,00	0,00	1,00	1,00	3,00
TotalWorkingYears	1470	11,28	7,78	0,00	6,00	10,00	15,00	40,00
TrainingTimesLastYear	1470	2,80	1,29	0,00	2,00	3,00	3,00	6,00
WorkLifeBalance	1470	2,76	0,71	1,00	2,00	3,00	3,00	4,00
YearsAtCompany	1470	7,01	6,13	0,00	3,00	5,00	9,00	40,00
YearsInCurrentRole	1470	4,23	3,62	0,00	2,00	3,00	7,00	18,00
YearsSinceLastPromotion	1470	2,19	3,22	0,00	0,00	1,00	3,00	15,00
YearsWithCurrManager	1470	4,12	3,57	0,00	2,00	3,00	7,00	17,00

Seguidamente será apresentada uma análise exploratória dos dados por forma a serem observadas algumas relações entre as variáveis independentes com a variável alvo *attrition*. O *pairplot* completo pode ser consultado no Anexo A do presente estudo.

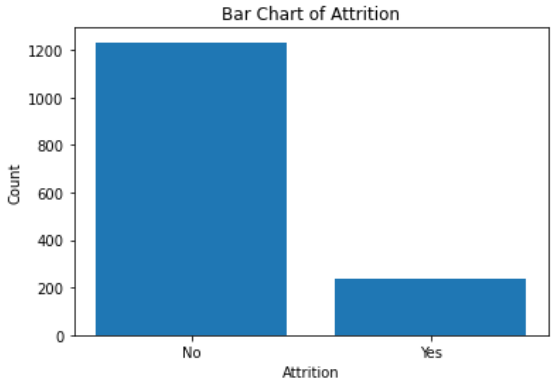


Figura 1 - Nº de trabalhadores com base na variável A*ttrition*

Considerando a variável alvo deste estudo, a Figura 1 demonstra um maior número de permanências na empresa (1.233, cerca de 84%) do que saídas (237, aproximadamente 16%).

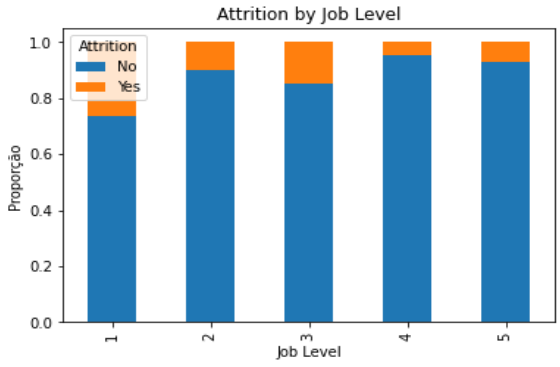


Figura 2 - Relação entre as variáveis *Job level* com *Attrition*

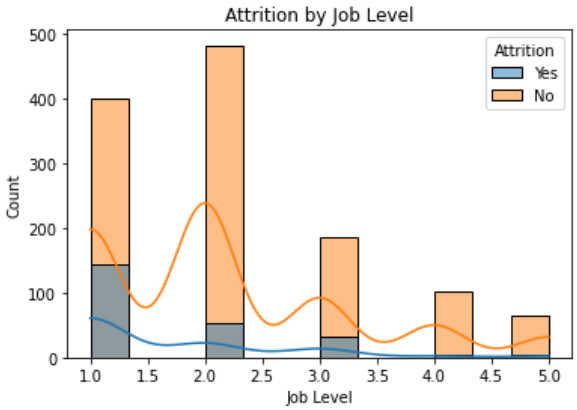


Figura 3 - Distribuição das variáveis *Job level* e *Attrition*

Em termos de proporção, as Figura 2 e Figura 3 evidenciam que no nível de carreira mais baixo há um maior número de saídas, contrastando com os níveis 4 e 5 em que se verifica um número inferior.

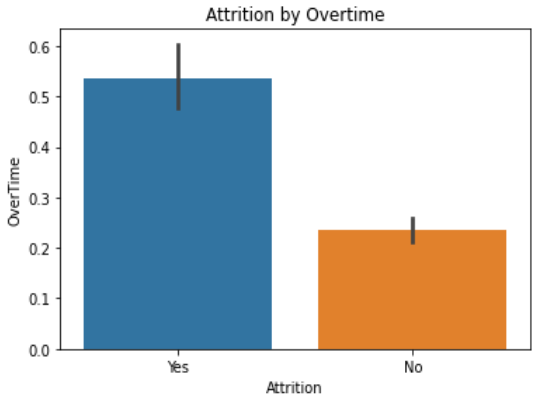


Figura 4 - Relação das variáveis *Overtime* e *Attrition*

Considerando a Figura 4, constata-se que os colaboradores que trabalham mais horas, tendem a não permanecer na empresa, traduzindo-se num maior número de *yes* na variável *Attrition*.

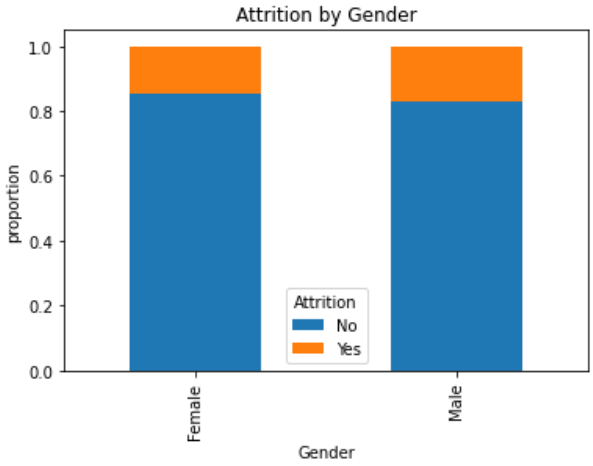


Figura 5 - Relação das variáveis *Gender* com *Attrition*

Em termos de proporção, não existem diferenças significativas entre género feminino e masculino em relação à *Attrition*, apresentando-se apenas um valor ligeiramente mais alto no género masculino (Figura 5).

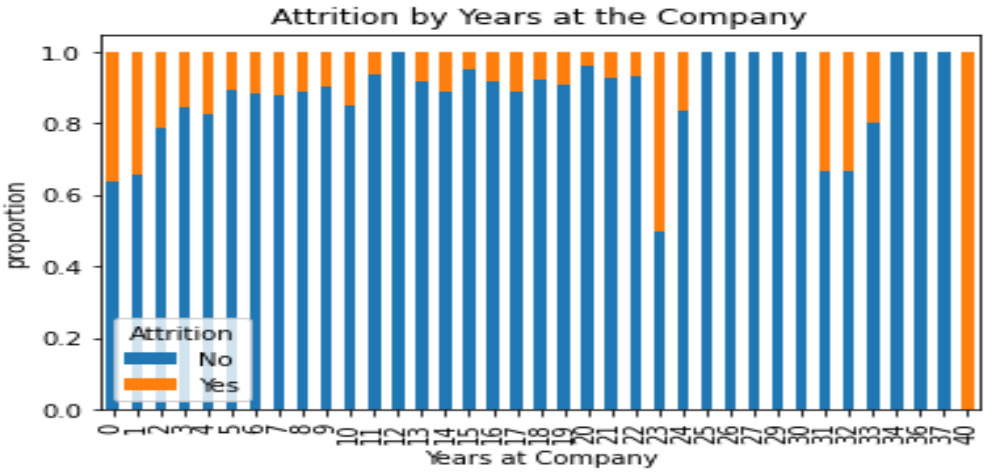


Figura 6 - Relação das variáveis *Years at company* com *Attrition*

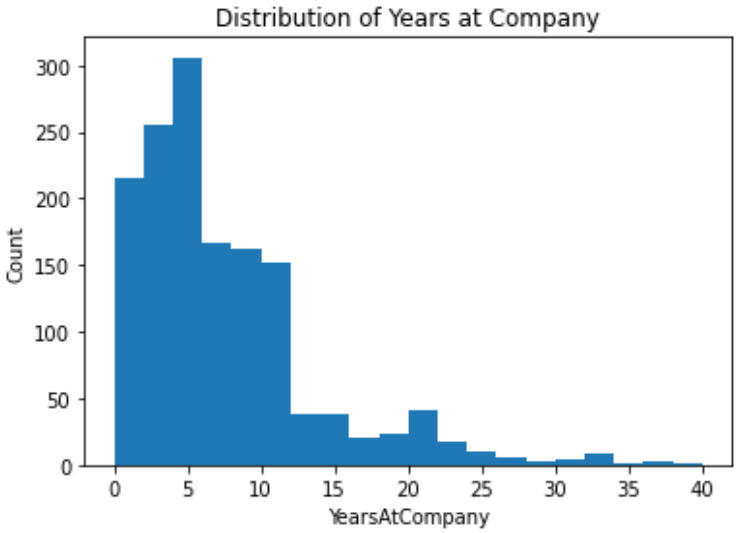


Figura 7 - Histograma da variável *Years at Company*

Na Figura 6 constata-se que 50% dos colaboradores com 23 anos de exercício de funções terminaram vínculo com a empresa e que ao fim de 40 anos cessam funções. Após análise da Figura 7 verifica-se que a variável *YearsAtCompany* não apresenta uma distribuição normal, observando-se uma assimetria à direita.

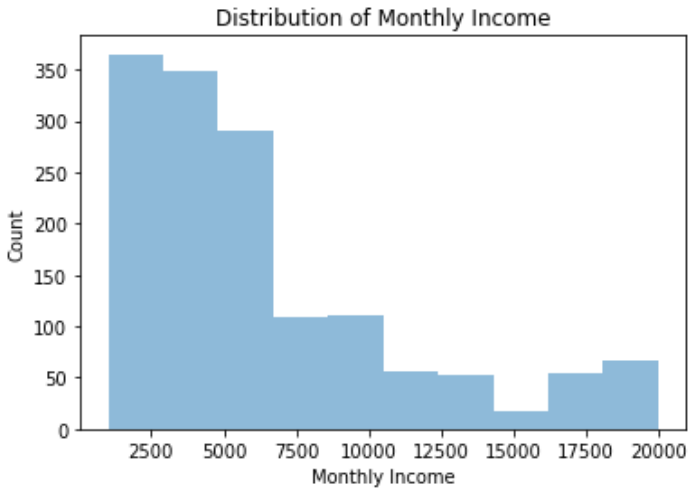


Figura 8 - Histograma da variável *MonthlyIncome*

Por observação da Figura 8, verifica-se que a variável *MonthlyIncome* não apresenta uma distribuição normal, revelando uma assimetria à direita.

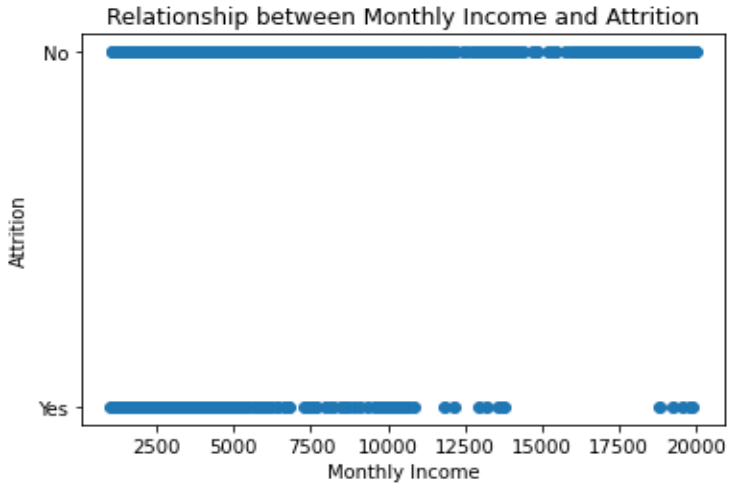


Figura 9 - Relação entre a variável *Monthly Income* com a *Attrition*

Na Figura 9 constata-se que os colaboradores que tendem a rescindir seu vínculo laboral com a empresa apresentam rendimentos mais baixos, verificando-se uma maior concentração de pontos abaixo dos 12.500 dólares.

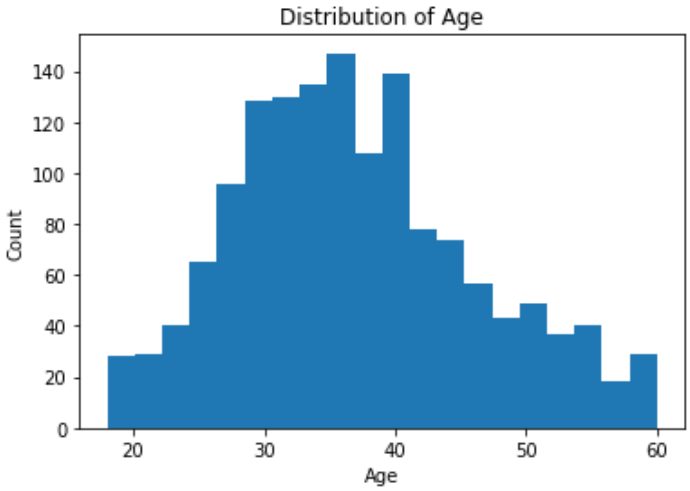


Figura 10 - Histograma da variável *Age*

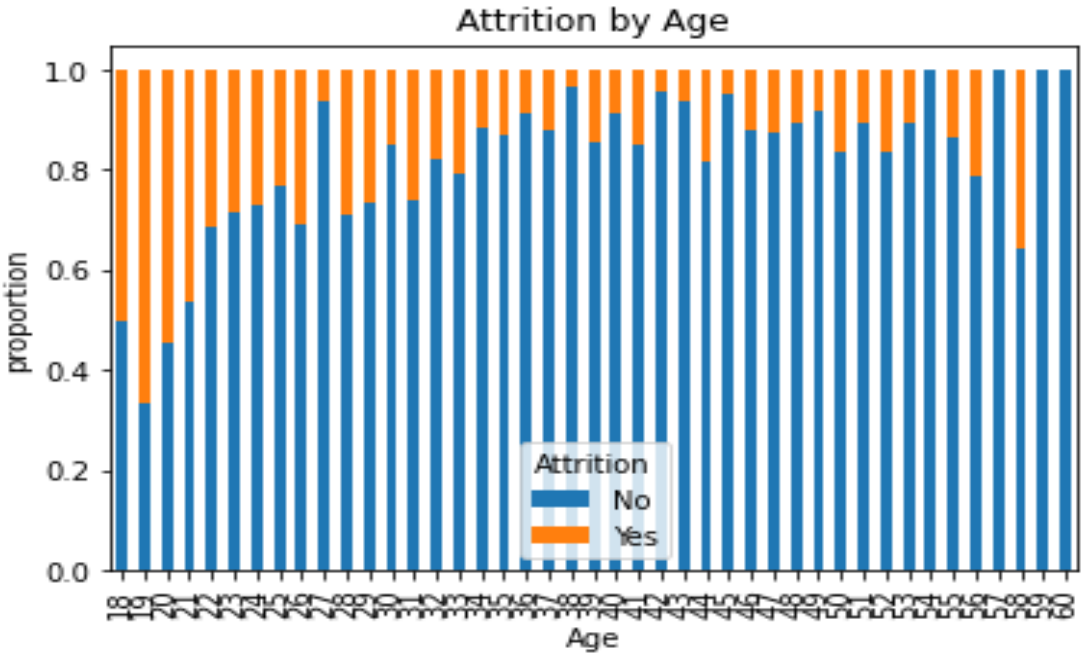


Figura 11 - Relação entre a variável *Age* com *Attrition*

Na Figura 10 verifica-se uma distribuição aproximadamente normal da variável *Age*. Por forma a estabelecer-se a relação da mesma com a variável alvo, observa-se que em termos de proporção existe um maior número de saídas permanentes dos 18 até aos 21 anos, conforme apresentado na Figura 11.

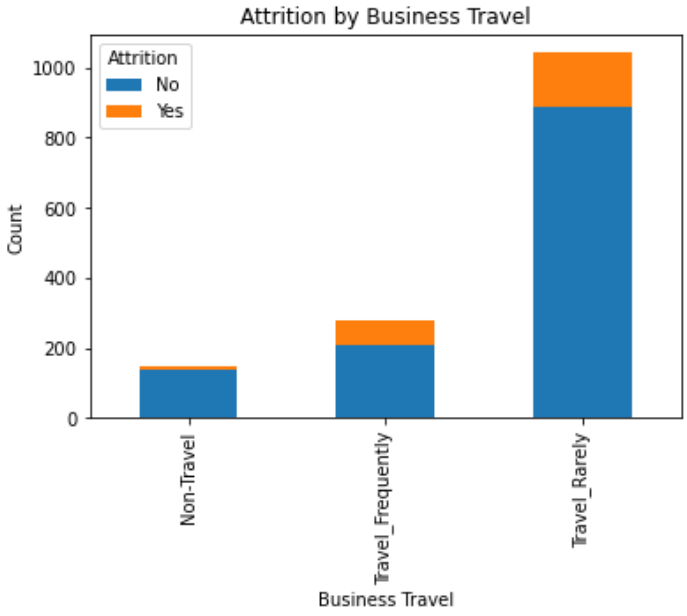


Figura 12 - Relação entre a variável Business *Travel* e *Attrition*

A Figura 12 reflete que a maior parte dos trabalhadores viaja ocasionalmente em negócios, sendo neste campo que se observa uma maior *attrition*.

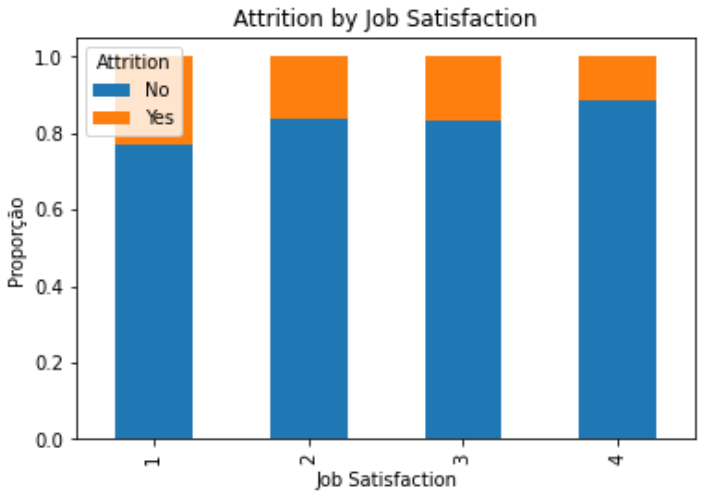


Figura 13 - Relação das variáveis *Job Satisfaction* e *Attrition*

Considerando os dados apresentados na Figura 13 observa-se que existem uma maior *attrition* quando os colaboradores apresentam índices de satisfação com o trabalho menores, contrariamente ao que acontece quando se encontram plenamente satisfeitos com o seu trabalho.

Desenvolvimento do Tema

O *Machine Learning* e o *Deep Learning* são um campo da Inteligência Artificial que procura desenvolver algoritmos capazes de aprender e tomar decisões automaticamente, a partir de experiências passadas e sem serem explicitamente programados. Essa abordagem revolucionária permite que os computadores adquiram uma aprendizagem a partir de dados e melhorem seu desempenho ao longo do tempo, sem intervenção humana direta (Hastie *et al*., 2009).

Para este trabalho foram implementadas algumas abordagens de aprendizagem da máquina e intrinsecamente associadas à aprendizagem supervisionada e não supervisionada, a aprendizagem profunda como também o método estatístico de análise de sobrevivência.

Metodologia

Para a elaboração do presente estudo foi utilizada a linguagem de programação *Python* (versão 3) através de IDE (*Integrated Development Environment*), nomeadamente *Spyder* (*Anaconda*), utilizando diferentes bibliotecas para desenvolver o projeto.

Para que os dados em análise pudessem ser implementados para a criação dos diversos algoritmos, e respetivos métodos estatísticos foi necessária a sua preparação através do pré-processamento. Em primeira instância foram identificadas algumas variáveis cuja sua génese não teria qualquer interferência nos resultados e na análise de estudo.

Neste sentido foram excluídas desta observação as variáveis *EmployeeCount*, *EmployeeNumber*, *Over18*, *StandardHours*, tendo assim sido consideradas 31 variáveis no total.

Aos dados analisados, não foram detetados valores nulos (*Nan Values*) e como tal não houve necessidade da sua remoção.

Posteriormente procedeu-se à transformação dos dados qualitativos em dados quantitativos para a aplicação de técnicas estatísticas e algoritmos de aprendizagem de máquina para análise. Este processo poderá ser observado na Tabela 4:

Tabela 4 - Variáveis qualitativas adaptadas

Variáveis	Dados transformados
Attrition	Yes = 1 Não = 0
BusinessTravel	Non-Travel = 0 Travel_Rarely = 1 Travel_Frequently =2
DailyRate	Research & Development = 0, Human Resources = 1 Sales = 2
EducationField	Human Resources = 0 Life Sciences = 1 Marketing = 2 Medical = 3 Technical Degree = 4 Other = 5
Gender	Female = 0 Male = 1
JobRole	Research Scientist = 0, Laboratory Technician = 1, Manufacturing Director = 2, Sales Executive = 3, Healthcare Representative = 4, Manager = 5, Research Director = 6, Human Resources = 7, Sales Representative = 8
MaritalStatus	Single = 0 Married = 1 Divorced = 2
OverTime	Yes = 1 No = 0

Nesta sequência e considerando a importância de identificar a correlação das variáveis para a tomada de decisão, seguidamente será apresentada a matriz de correlação das 31 variáveis em estudo (Figura 14).

Das variáveis em estudo existe correlação positiva moderada entre as variáveis *YearsAtCompany*, com *YearWithCorrentManager* (0.77) e *YearsinCurrentRule* (0.76), *MonthlyIncome* com *TotalWorkingYears* de 0.77, *JobLevel* com *TotalWorkingYears* de 0.78 e finalmente entre *PercentSalaryHike* com o *PerformanceRating* de 0.77.

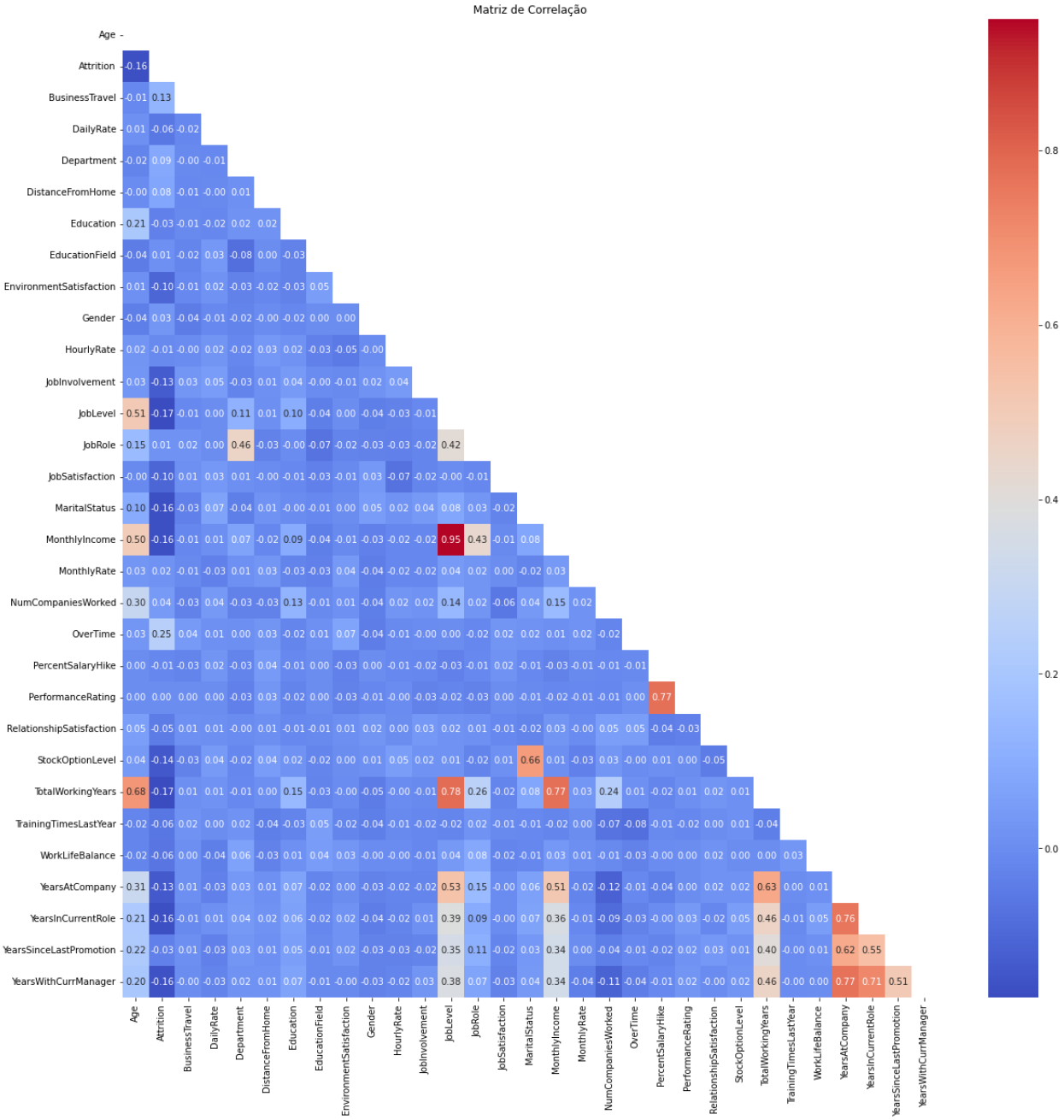


Figura 14 - Matriz de Correlação

Finalmente, é possível constatar que as variáveis que apresentam uma correlação significativa positiva mais forte (0.95) foram *JobRole* e *MonthlyIncome*.

No entanto, considerando a importância das mesmas no presente estudo e as suas características específicas e diferenciadas, optou-se por mantê-las por forma a percecionar o tipo de características e como poderão influenciar a saída (*attrition*) dos colaboradores.

Posteriormente, foi realizada uma análise para detetar a presença de *outliers* do *dataset*. Das variáveis analisadas foram observados *outliers* nas variáveis presentes na Figura 15.

Dos *outliers* identificados, os mesmos foram removidos da análise (Figura 16), tendo sido definido o limite de 3 desvios-padrão, a fim de determinar os valores atípicos. Assim, de um

total de 1.470 observações obteve-se 1.363, ou seja, verificou-se uma redução de 107 observações.

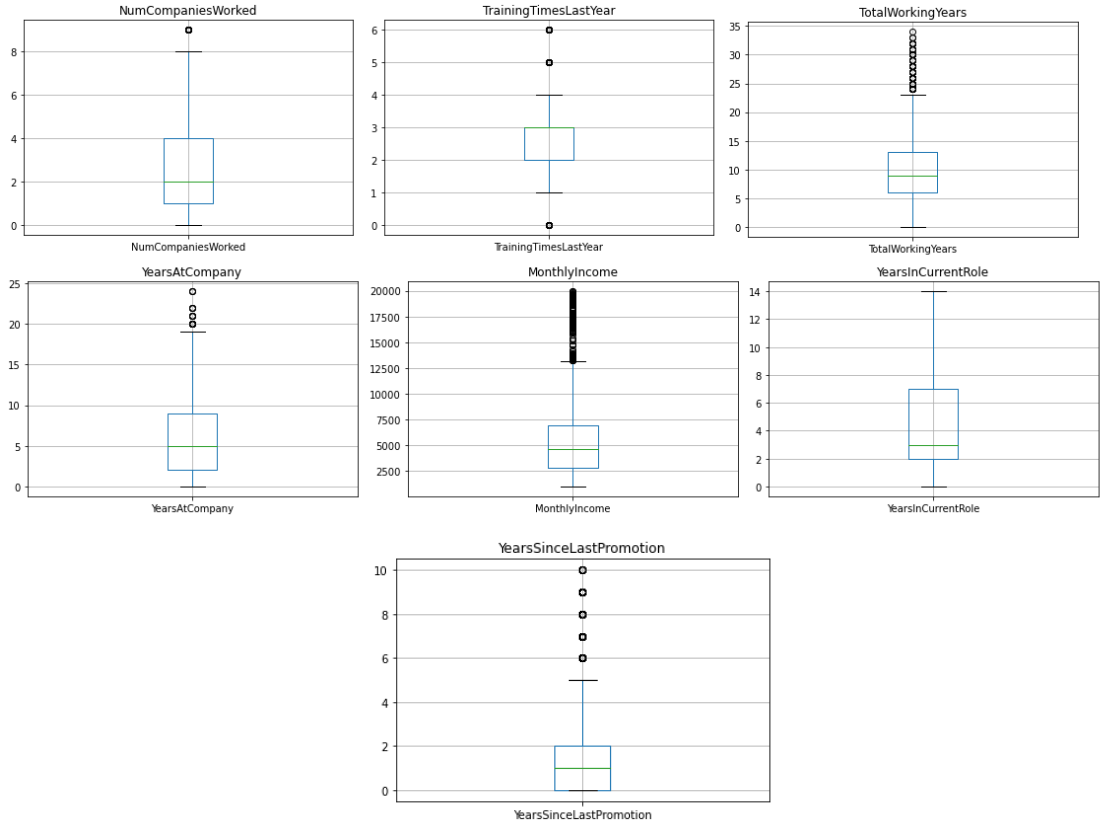


Figura 15 - *Outliers* detetados

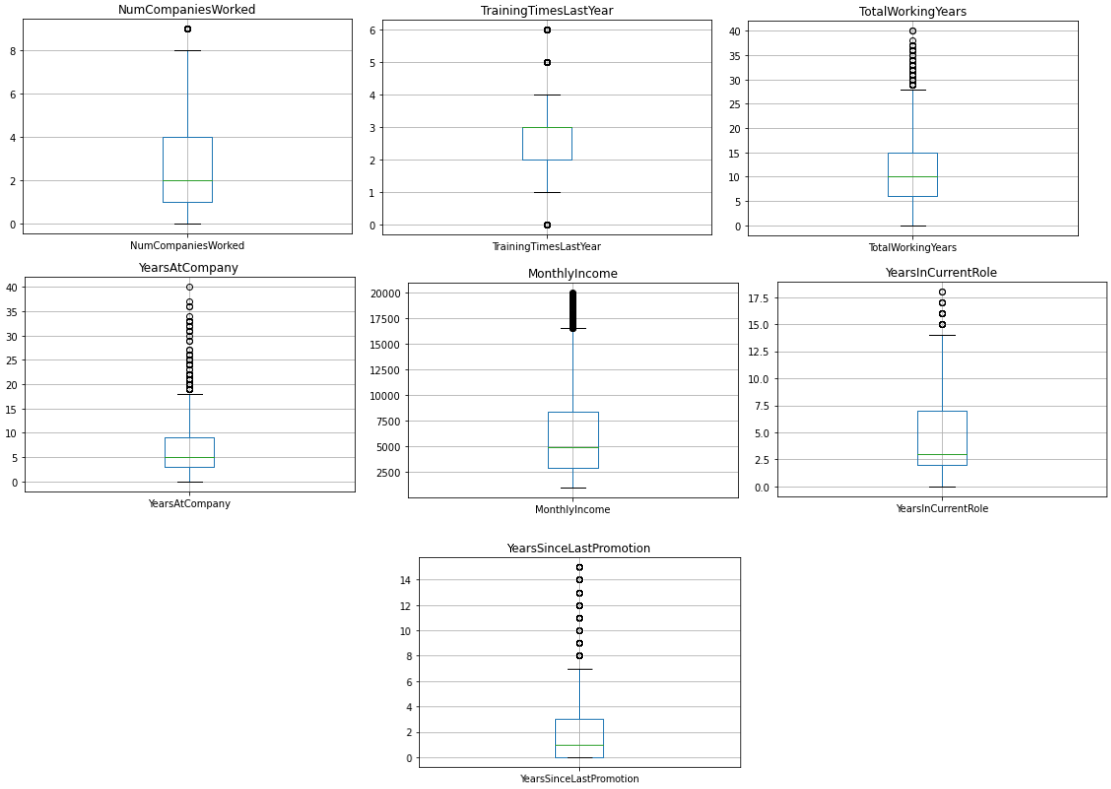


Figura 16 - Remoção dos *outliers*

Implementação dos algoritmos e análise de sobrevivência

Agrupamento

K-means

Para a implementação do *k-means* recorreu-se à biblioteca *Python sklearn.cluster* e foram excluídas as variáveis binárias e as variáveis numéricas discretas com menor variabilidade. Após observação e análise da matriz de correlações (Figura 14) e do *pairplot* (**Anexo A**) do conjunto de dados, selecionaram-se as variáveis *Age* e *MonthlyIncome*. Antes da implementação do algoritmo, foram removidos os *outliers* e foi aplicado um filtro ao conjunto de dados para a variável *Attrition* com o objetivo de conter apenas os dados em que se verificou a existência de *attrition* (*Attrition* =1). De forma a homogeneizar as escalas das variáveis selecionadas, procedeu-se à sua normalização através do método *MinMaxScaler*.

Associação

Apriori

Para a implementação deste algoritmo, foi aplicado um filtro para perceber quais as regras de associação geradas quando existe *attrition (Attrition=yes)*. Neste sentido, foram utilizados os atributos *Age*, *Attrition*, *DistanceFromHome*, *EnvironmentSatisfaction*, *Gender*, *JobInvolvement*, *JobSatisfaction*, *MonthlyIncome*, *NumCompaniesWorked*, *RelationshipSatisfaction*, *TotalWorkingYears*, *WorkLifeBalance*, *YearsAtCompany*, *YearsInCurrentRole*, *YearsSinceLastPromotion* e *YearsWithCurrManager*.

Posteriormente à seleção das variáveis, procedeu-se à transformação das colunas com dados binários, tendo-se definido o valor 0 para valores abaixo da média e 1 para valores acima da média, respetivamente. O algoritmo foi aplicado através do módulo *mlxtend.frequent_patterns* e implementadas as classes *apriori* e *association_rules*, com o objetivo de encontrar os conjuntos de itens mais frequentes no *dataset*.

Assim, após diversos testes foi definido um valor mínimo de suporte de 0.2, um valor mínimo de confiança de 0.5 e um valor mínimo de *lift* de 1.5.

Classificação

Antes de se proceder à avaliação da eficácia dos algoritmos de classificação, foi realizada uma análise da correlação entre as variáveis. Por forma a analisar a correlação entre as diferentes variáveis em estudo, utilizou-se a biblioteca Pandas e procedeu-se à visualização gráfica, através de um *heatmap*, utilizando a biblioteca *Seaborn*. Neste sentido constatou-se que apenas as variáveis que possuem uma correlação significativa entre elas foram as *"MonthlyIncome"* e *"JobRole"*. Não obstante de se ter verificado uma correlação forte entre respetivas variáveis, optou-se por mantê-las no processo de análise de classificação dos

algoritmos, na medida em que apresentam características diferentes e potencialmente informativas para o respetivo estudo.

Foi igualmente desenvolvida uma análise, ao nível de correlações entre a variável alvo e as restantes variáveis independentes.

Posteriormente, foi realizada a normalização dos dados utilizando a classe *MinMaxScaler* do módulo *sklearn.preprocessing* nos algoritmos que requerem este método (*Naive Bayes*, *Adaboost*, *XGBoost*, SVM, K-NN e ANN).

A divisão dos dados em dados de treino e dados de teste foi realizada na proporção de 70% para 30% respetivamente. Os algoritmos de classificação anteriormente referidos foram aplicados e treinados com base nesta divisão.

No processo de classificação e com o objetivo de reforçar a robustez dos respetivos algoritmos de classificação, implementou-se o *cross-validation* para garantir a eliminação do *Underfitting* e *Overfitting*. O *cross-validation* é uma ferramenta utilizada para estimar o verdadeiro erro de predição dos modelos de classificação, procurando ajustar os respetivos parâmetros (Berrar, 2018). Para isso o modelo é treinado *k* vezes, utilizando *k-1 folds* para treino e o *fold* restante para teste.

Para este processo foi utilizada a técnica *K-Fold,* dividindo o conjunto de dados em 5 partes iguais. Nesta análise foi utilizado a classe *KFold* do módulo *sklearn.model_selection*.

Finalmente, através da implementação das classes *accuracy_score*, *confusion_matrix, precision_score, recall_score* e *f1_score* do módulo *sklearn.metrics*, avaliou-se cada um dos resultados obtidos em cada algoritmo.

Naive Bayes

Para a implementação destes modelos utilizaram-se as classes de modelo *GaussianNB* e *MultinomialNB* do módulo *sklearn.naive_bayes*.

Decision Trees

Para a implementação deste modelo utilizou-se a classe de modelo *DecisionTreeClassifier* do módulo *sklearn.tree*.

Random Forest

Para a implementação deste modelo utilizou-se a classe de modelo *RandomForestClassifier* do módulo *sklearn.ensemble*.

Logistic Regression

Para a implementação deste modelo utilizou-se a classe de modelo *LogisticRegression* do módulo *sklearn.linear_model*.

Support Vector Machine

Para a implementação deste modelo utilizou-se a classe de modelo SVC do módulo *sklearn.svm*.

K-Nearest Neighbors (K-NN)

Para a implementação deste modelo utilizou-se a classe de modelo *KNeighborsClassifier* do módulo *sklearn.neighbors*.

ADABoost

Para a implementação deste modelo utilizou-se a classe de modelo *AdaBoostClassifier* do módulo *sklearn.ensemble*, tendo sido igualmente incluído como *classifier base* o modelo da *decision tree*.

XGBoost

Para a implementação deste algoritmo utilizou-se a biblioteca XGBOOST.

Artificial Neural Network (ANN)

Para a implementação deste algoritmo utilizou-se a classe do modelo *KerasClassifier* do módulo *tensorflow.keras.wrappers.sckikit_learn*. Foram identificados 10 *epochs*, *batch size* = 32 e *verbose* = 1.

Análise de Sobrevivência

A análise de sobrevivência foi realizada através da implementação da curva de *Kaplan-Meier*, utilizando a classe *KaplanMeierFitter* da biblioteca l*ifelines*.

Considerando que a análise de sobrevivência de *Kaplan-Meier* fornece estimativas da função de permanência ao longo do tempo, através da representação gráfica das curvas de sobrevivência, inicialmente estabeleceu-se a variável *YearsAtCompany* como a variável tempo de permanência na organização (variável T do modelo) e a variável *Attrition* como a probabilidade de permanência na empresa (variável C do modelo).

É igualmente relevante mencionar que para a análise de sobrevivência os *outliers* identificados no presente estudo, não foram removidos, ou seja, considerou-se mantê-los, na medida em que a sua remoção afetava negativamente o respetivo modelo. Desta forma poderia comprometer a qualidade no processo de análise, na medida em que existia uma forte alteração na probabilidade de sobrevivência e no tempo de permanência na empresa.

Capítulo 2

Resultados

Agrupamento

Após implementação do algoritmo *K-means*, o número ideal de clusters foi selecionado com base no método do cotovelo (*elbow method*), em que o código realiza diversas iterações sobre diferentes valores de k (número de clusters) e executa o algoritmo *K-means* para cada valor, sendo calculada a soma das distâncias quadráticas no cluster, de modo a minimizar simultaneamente o *Within Sum of Squares* (WSS) e o número de clusters (Duda, Hart & Stork, 2001).

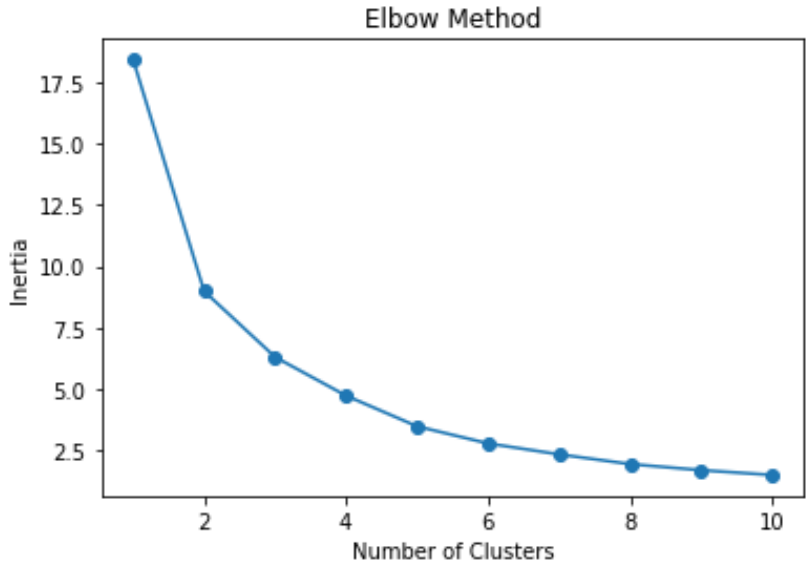


Figura 17 - *Elbow Method*

A análise da Figura 17 sugere que o número de *clusters* ótimo é de dois *clusters* (*k*=2).

A **Erro! A origem da referência não foi encontrada.** representa o resultado após aplicação do algoritmo de agrupamento, onde é possível visualizar a distribuição dos *clusters* com base nas variáveis *Age* e *MonthlyIncome*.

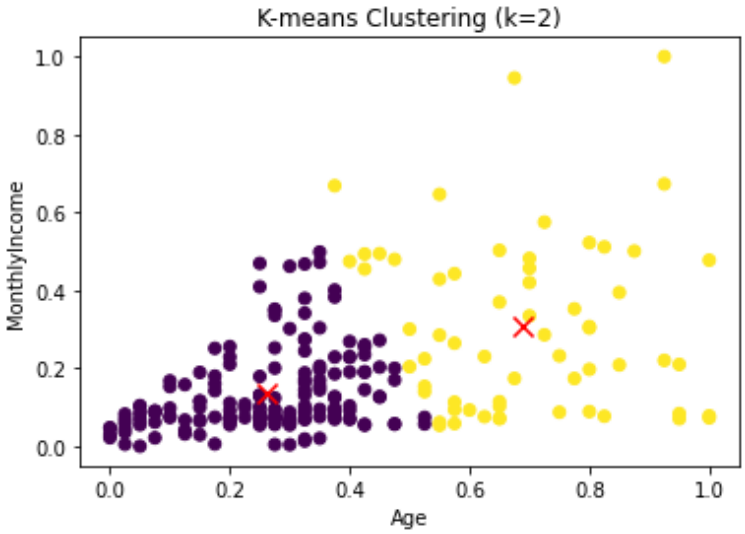


Figura 18 - K-means Clustering

No gráfico acima é possível visualizar a presença de dois *clusters*, tendo por base a existência de *attrition*. As cruzes vermelhas assinalam os centróides, isto é, o ponto que representa o centro médio dos pontos de dados atribuídos a cada um dos *clusters*. Através da Figura 18, observa-se que o cluster 1 se situa na parte superior em que o seu centroide se localiza sensivelmente no ponto (0.26,0.13) e o *cluster* 2 nas coordenadas (0.69, 0.31). O *cluster* 1 corresponde aos dados de trabalhadores que apresentam idade inferior, sendo a sua distribuição predominantemente localizada em valores mais baixos de vencimento. O *cluster* 2 representa os dados de colaboradores com idade superior, em que os dados apresentam uma distribuição heterogénea em termos salariais. No entanto, este último grupo apresenta, em média, salários mais elevados que o anterior.

Associação

Durante a implementação verificou-se que quando os valores mínimos definidos para o suporte, confiança e *lift* são aumentados, é gerado um menor número de regras de associação, sendo o valor do suporte aquele que apresenta um maior impacto na redução de regras geradas. Por exemplo, definindo um valor mínimo de suporte de 0.01 gera 3003426 regras e optando pelo valor mínimo de 0.2, apenas são geradas 71 regras de associação. Neste sentido, o ajuste dos referidos parâmetros permite filtrar apenas as regras de associação mais

fortes, isto é, aquelas que apresentam uma maior proporção no *dataset*, uma maior probabilidade condicional e aquelas que apresentam uma maior ocorrência conjunta dos itens selecionados, definindo valores mínimos para as respetivas métricas.

Na Tabela 5 encontram-se espelhadas as regras de associação geradas ordenadas pelo valor de suporte, por ordem decrescente.

Tabela 5 - Resultados Associação para *Attrition=Yes*

Conjunto	Suporte	Confiança	*Lift*	Antecedentes	Consequentes
1	0.3125	0.87	2.26	*YearsWithCurrMan*	*YearsAtCompany*
2	0.3125	0.81	2.26	*YearsAtCompany*	*YearsWithCurrMana*
3	0.2969	0.90	2.35	*YearsInCurrentRol*	*YearsAtCompany*
4	0.2969	0.77	2.35	*YearsAtCompany*	*YearsInCurrentRole*
5	0.2917	0.79	1.87	*MonthlyIncome*	*TotalWorkingYears*
6	0.2912	0.76	1.79	*YearsAtCompany*	*TotalWorkingYears*

Após análise da tabela acima, verifica-se que a regra de associação mais forte, isto é, com maior nível de suporte apresenta como antecedente a variável *YearsWithCurrManager* e consequente *YearsAtCompany*. Nesta regra, constata-se que a combinação dos itens ocorre em cerca de 31% no *dataset*, ou seja, verifica-se a existência de trabalhadores que se encontram há mais tempo com a mesma chefia e há mais anos na mesma empresa. Para esta regra observa-se uma confiança de 0.87, revelando que quando um colaborador está há mais tempo com a mesma chefia, verifica-se também um maior número de anos a desempenhar funções na mesma empresa, em cerca de 87% das ocorrências. O *lift* de 2.26 indica que existe uma relação positiva forte entre os itens referidos, sendo que quando se verifica o antecedente é 2.26 vezes mais provável a ocorrência do consequente. Esta mesma lógica é aplicada às restantes regras geradas pelo algoritmo *apriori*. Em simultâneo, conclui-se também que os itens identificados nas regras geradas estão bastante associados, trocando a sua posição enquanto antecedentes e consequentes. Importa ainda fazer referência às duas últimas regras apresentadas, mostrando uma forte associação entre as variáveis *MonthlyIncome* e *TotalWorkingYears, bem como entre as YearsAtCompany* e *TotalWorkingYears*. A regra 5 demonstra que, nos casos em que se verifica *attrition*, quando os colaboradores auferem um vencimento acima da média também revelam maior experiência laboral. Por outro lado, nestes colaboradores que evidenciam *attrition,* a regra 6 demonstra associação entre o número de anos que o colaborador está na empresa e um maior número de anos de carreira profissional.

Classificação

Ao nível da correlação entre as variáveis, e tal como é possível observar-se na Tabela 6, não existe correlação forte entre as mesmas. A inclusão de todas as variáveis independentes permite que o algoritmo de classificação tenha acesso a todas as informações disponíveis e possa realizar seu próprio processo de seleção de recursos.

Tabela 6 - Correlações com a variável *Attrition*

Variáveis Independentes	Variável Alvo - *Attrition*
Age	0.16
BusinessTravel	0.13
DailyRate	-0.06
Department	0.09
DistanceFromHome	0.08
Education	-0.03
EducationField	0.01
EnvironmentSatisfaction	-0.10
Gender	0.03
HourlyRate	-0.01
JobInvolvement	-0.13
JobLevel	-0.17
JobRole	0.01
JobSatisfaction	-0.10
MaritalStatus	-0.16
MonthlyIncome	-0.16
MonthlyRate	0.02
NumCompaniesWorked	0.04
OverTime	0.25
PercentSalaryHike	-0.01
PerformanceRating	0.00
RelationshipSatisfaction	-0.05
StockOptionLevel	-0.14
TotalWorkingYears	-0.17
TrainingTimesLastYear	-0.06
WorkLifeBalance	-0.06
YearsAtCompany	-0.13
YearsInCurrentRole	-0.16
YearsSinceLastPromotion	-0.03
YearsWithCurrManager	-0.16

Os resultados da implementação das métricas de avaliação para cada algoritmo poderão ser observados na Tabela 7.

Tabela 7 - Algoritmos de classificação e respetivas métricas de avaliação

Algoritmo	Variante	*Accuracy*	*Precision**	*Recall**	*F1-Score**
Naive Bayes	*Gaussian*	77%	84%	77%	79%
	Multinomial	53%	76%	53%	60%
Decision Tree		78%	80%	78%	79%
Random Forest		86%	85%	86%	82%
Adaboost		79%	82%	80%	81%
SVM	*Kernel Linear*	90%	90%	90%	88%
	Kernel Polinomial	83%	83%	83%	83%
	Kernel RBF	87%	86%	87%	84%
	Kernel Sigmoid	87%	85%	87%	85%
KNN		85%	82%	85%	82%
Logistic Regression		85%	87%	85%	79%
XGBoost		88%	86%	88%	86%
ANN		89%	89%	89%	88%

*Foi considerada a *Weighted avg* (média ponderada) nas métricas de avaliação por se considerar a existência de *imbalance* na variável alvo.

Na Tabela 7 as métricas de avaliação *accuracy*[4], *precision*[5], *recall*[6] *e F1-Score*[7] de todos os algoritmos de classificação analisados no presente relatório, com *cross-validation*. Os algoritmos que apresentaram melhor desempenho das métricas de avaliação foram o *SVM Kernel Linear* com 90% de *accuracy*, *precision* e *recall* e 88% do *F1-Score* e o ANN com 89% de *accuracy*, precision e *recall* e 89% e um *F1-Score de 88%*.

[4] A *accuracy* mede a proporção de exemplos classificados corretamente, ou seja, *accuracy* = (TP+TN) / (TN+FN+FP+FN)

[5] A *precision* mede a proporção de exemplos positivos classificados corretamente (evita falsos positivos), ou seja, *precision* = TP / (FP + TP)

[6] O *recall* mede a proporção de exemplos positivos identificados corretamente, medindo a sensibilidade do modelo (evita falsos negativos), ou seja, *recall* = TP / (FN + TP)

[7] O F1-Score é a métrica que combina *precision* e *recall* de maneira equilibrada, ou seja, 2*((*precision**recall*)/(*precision*+*recall*))

Considerando que o algoritmo de classificação que melhor desempenho demonstrou, tendo em conta os resultados apresentados nas métricas de avaliação, torna-se assim pertinente informar que:

- O algoritmo tem a capacidade de identificar corretamente 90% dos casos considerados como verdadeiros positivos e negativos (*accuracy*);
- O algoritmo é capaz de identificar corretamente 90% dos casos positivos em relação ao número total de casos positivos presentes num conjunto de dados, ou seja, verdadeiros positivos e falsos negativos (*recall*);
- O algoritmo tem a capacidade de classificar corretamente os verdadeiros positivos em relação ao número total de casos que foram classificados como positivos (*precision*);
- Em 88% dos casos o modelo é capaz de classificar corretamente a maioria dos casos positivos e negativos (*F1-score*).

Para além das métricas de avaliação anteriormente obtidas pelos algoritmos de classificação analisados, é igualmente relevante poderem ser identificadas as variáveis com maior importância relativa na separação das classes face à variável *Attrition*, no algoritmo *SVM Kernel Linear*. Na Tabela 8 são assim apresentadas as variáveis e respetivos valores associados ao coeficiente de peso:

Tabela 8 - Algortimo *SVM Kernel Linear*

Variáveis	Coeficiente de Peso[8]
TotalWorkingYears	-1.2418409246055067
OverTime	1.1090945843905544
JobInvolvement	-1.0273020742457426
BusinessTravel	0.8959947382021518

Relativamente ao algoritmo *ANN* as quatro variáveis uma maior sensibilidade face à variável *Attrition* foram (Tabela 9):

Tabela 9 - Algoritmo ANN

Variáveis	Análise de Sensibilidade[9]
BusinessTravel	0.02200488997555012
OverTime	0.01955990220048900
MonthlyRate	0.00977995110024450

[8] O **coeficiente de peso** corresponde à importância atribuída em cada variável na separação das classes no problema de classificação.

[9] A **análise de sensibilidade** é uma abordagem que possibilita entender o impacto individual de cada variável de entrada na ativação neuronal nas camadas seguintes.

Department	0.00733496332518337

É relevante mencionar que os algoritmos de classificação que apresentaram piores resultados em termos das métricas de avaliação acima identificadas foram os algoritmos *Naive Bayes Gaussian* e *Naive Bayes Multinomial*.

Finalmente é igualmente importante referir que o algoritmo de classificação *Decision Tree* também não alcançou percentagens elevadas associadas às métricas de avaliação e como tal utilizou-se como *classifier base* para a implementação do algoritmo *Adaboost*. Ainda assim, não se pode observar melhorias efetivamente significativas.

Receiver Operating Characteristic Curve (ROC)

Os estudos que utilizam a curva ROC procuram demonstrar o equilíbrio entre a taxa de acerto e a taxa de erro (Sobreiro *et al.*, 2022). A *Area Under the ROC Curve* (AUC) é um valor numérico entre 0 e 1 que combina a sensibilidade, ou seja, a taxa de verdadeiros positivos, também conhecida como *recall* e a especificidade, neste caso a taxa de verdadeiros negativos (Sobreiro *et al.*, 2022). Tem como objetivo identificar o desempenho dos algoritmos em análise, sendo que quanto maior for o valor AUC, ou seja, quando mais próximo de 1 e a curva ROC encontrar-se mais próxima do canto superior esquerdo do gráfico, melhor será o seu desempenho relativamente à classificação das duas classes, neste caso, na existência ou não de *attrition*.

Neste sentido serão apresentados as matrizes de confusão e os gráficos da curva ROC, através da implementação das classes *auc* e *roc_curve* do módulo *sklearn.metrics*, para todos os algoritmos de classificação, com *cross-validation*.

Para a implementação da *curva ROC* foi utilizado o módulo *metrics* da biblioteca *sklearn*, e para a visualização gráfica os módulos *express* e *pyplot* das bibliotecas *plotly* e *matplotlib*, respetivamente.

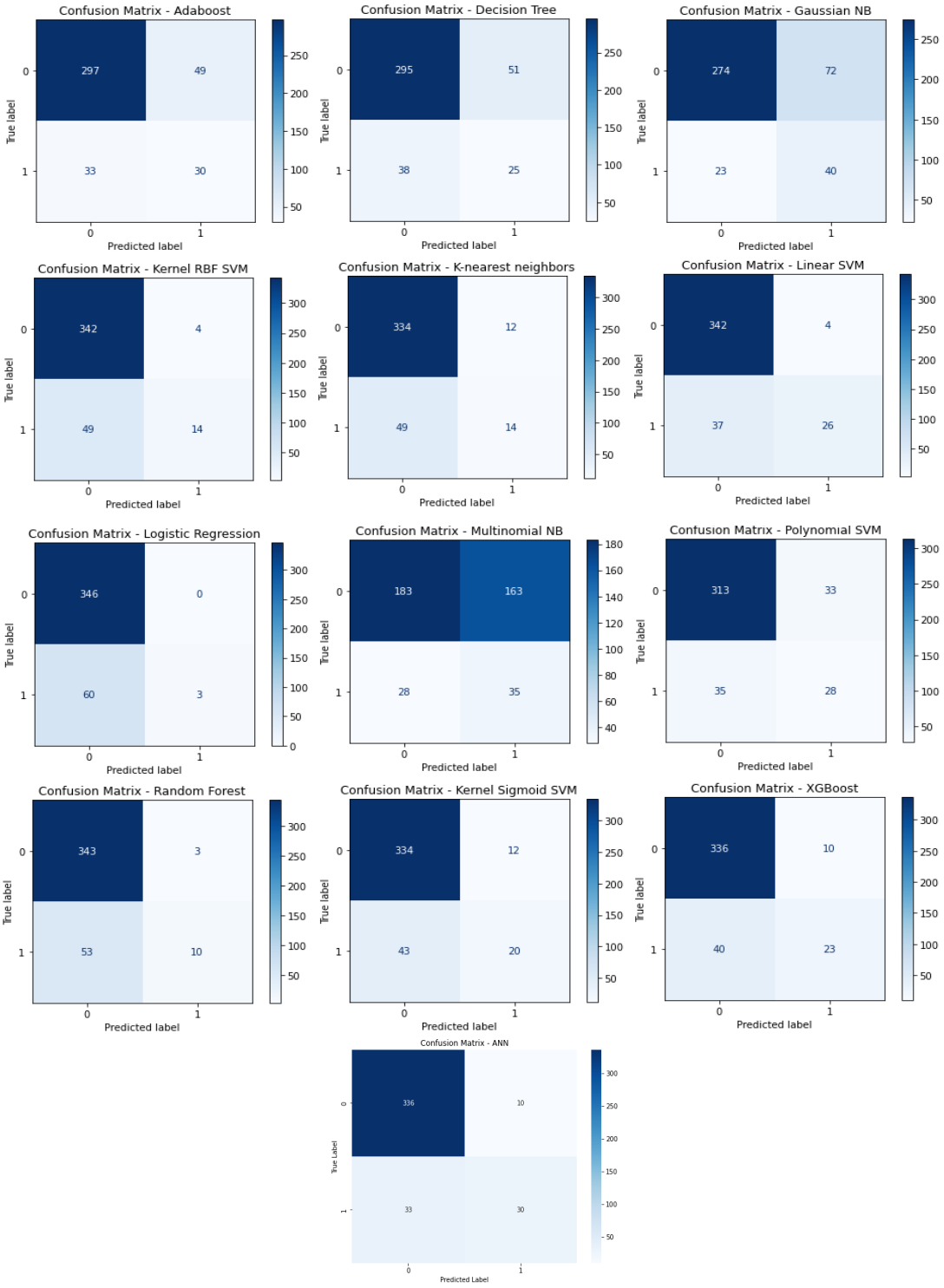


Figura 19 - Matrizes confusão dos algoritmos de classificação

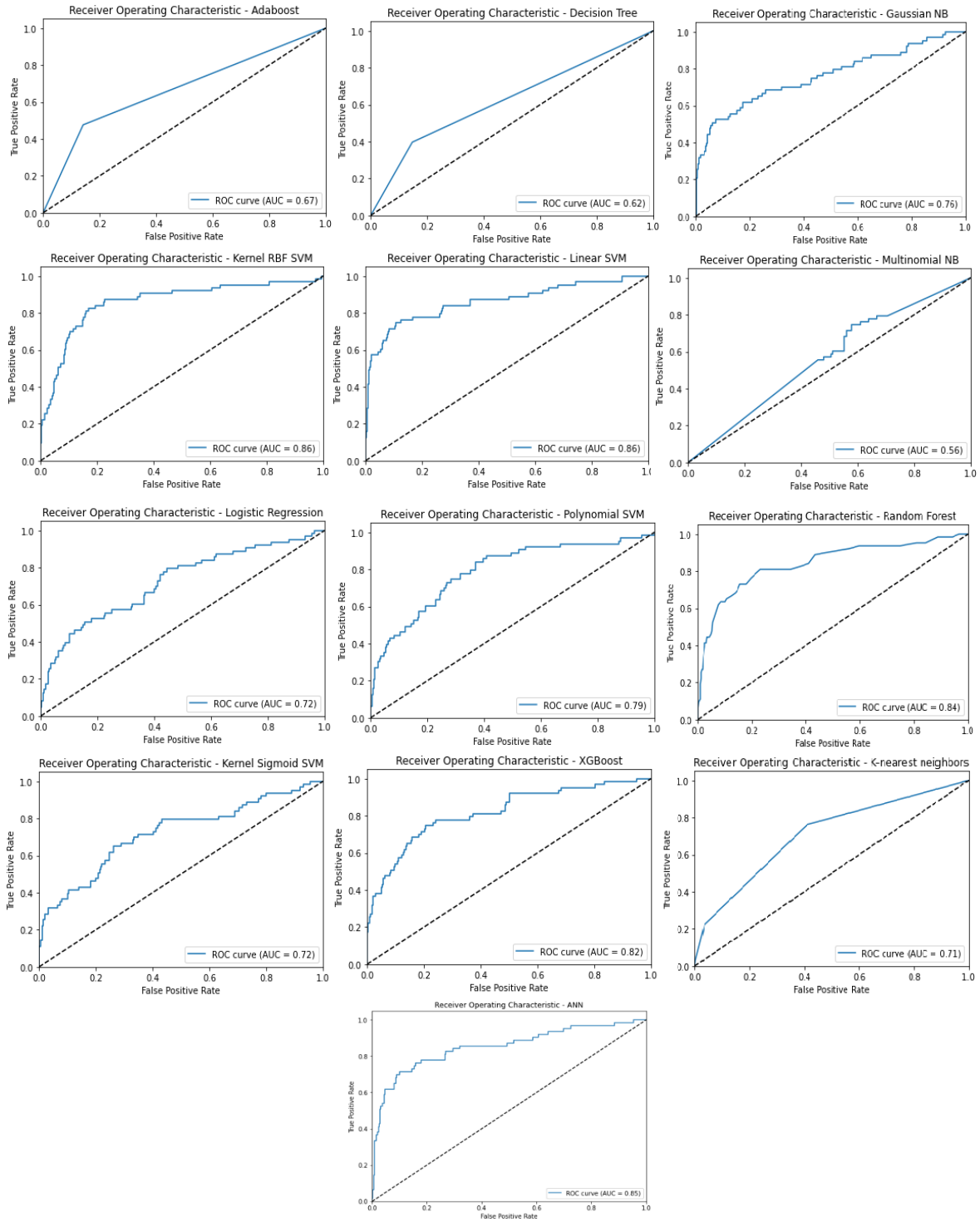


Figura 20 - Curva ROC e AUC dos algoritmos de classificação

No cômputo geral e considerando os resultados das métricas de avaliação, matrizes de confusão (Figura 19) e curvas ROC e AUC (Figura 20), os quarto algoritmos que apresentaram melhor desempenho foram os *SVM Kernel Linear*, *ANN*, *o SVM Kernel RBF* e *XGBoost*, sendo que, dos quatro algoritmos, o que se destacou na sua generalidade, foi *SVM Kernel Linear*.

Análise de Sobrevivência

Depois de analisados os algoritmos de classificação, procedeu-se à análise de sobrevivência. Um dos objetivos da análise de sobrevivência consiste em estimar e representar graficamente a função de sobrevivência, a partir de um conjunto de dados.

Neste sentido, na Figura 21 é possível observar a curva *Kaplan-Meier*, sendo que irremediavelmente a probabilidade de sobrevivência na empresa diminui com o avançar do tempo na empresa, por exemplo, ao fim de 10 anos a probabilidade de sobrevivência dos trabalhadores na empresa é inferior a 80%.

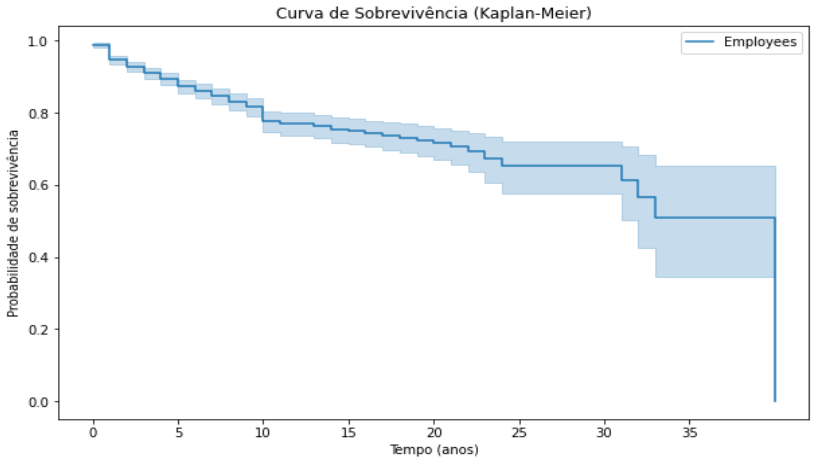


Figura 21 - Análise de Sobrevivência

Por forma a estabelecer uma análise detalhada ao respetivo modelo, e considerando as variáveis que apresentam maior correlação com a variável alvo, como também um maior coeficiente de peso (*SVM Kernel Linear*) e sensibilidade (ANN) na separação das classes, serão igualmente apresentadas algumas variáveis, recorrendo igualmente às curvas de *Kaplan-Meier*, para as variáveis *TotalWorkingYears*, *JobInvolvment*, *BusinessTravel*, *OverTime*, *JobLevel*, *Age* e *MonthlyRate*.

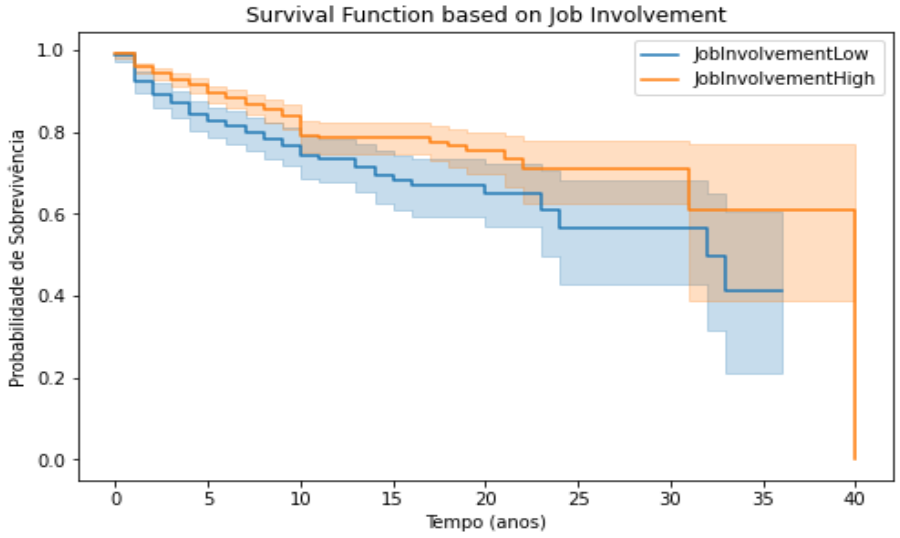


Figura 22 - Análise de sobrevivência para a variável *Job Involvement*

Por observação da Figura 22 é possível verificar que quanto menor o envolvimento nas funções laborais exercidas, menor a probabilidade de permanecerem na empresa.

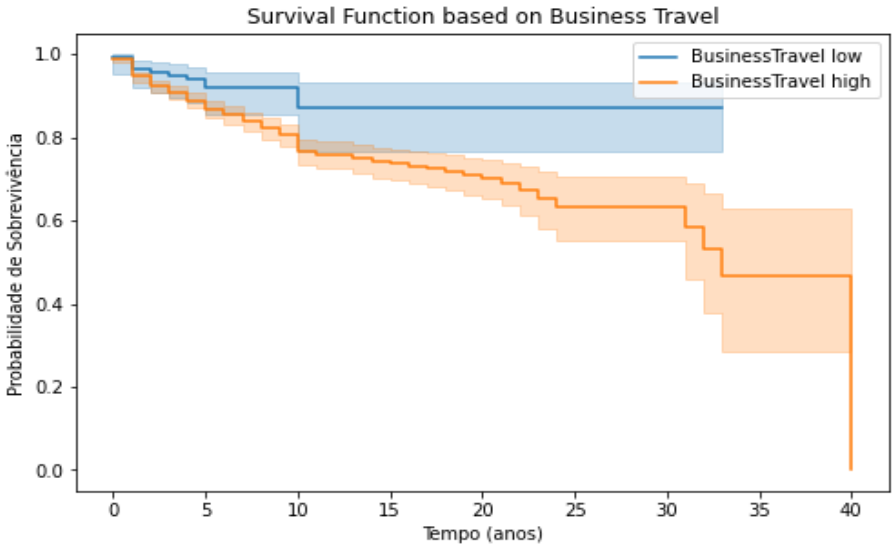


Figura 23 - Análise de Sobrevivência para a variável *Business Travel*

Tendo em conta a Figura 23 constata-se que quanto maior for o número de viagens, aumenta a probabilidade de os trabalhadores saírem da empresa.

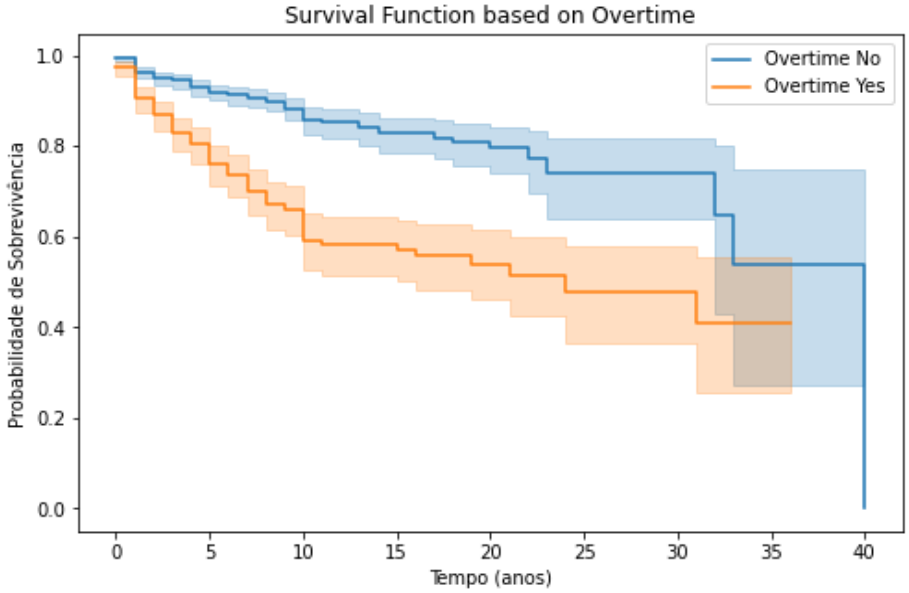


Figura 24 - Análise de sobrevivência para a variável *Overtime*

Tendo em conta a Figura 24 constata-se que quanto maior for o número de horas desempenhadas para além do horário normal de trabalho, aumenta a probabilidade de os trabalhadores saírem da empresa. Por exemplo, ao fim de 10 anos, a probabilidade dos trabalhadores que desempenham mais horas, para além do período normal de trabalho, é de aproximadamente 60%, face aos 85% dos trabalhadores que não desempenham funções, para além do horário normal de trabalho.

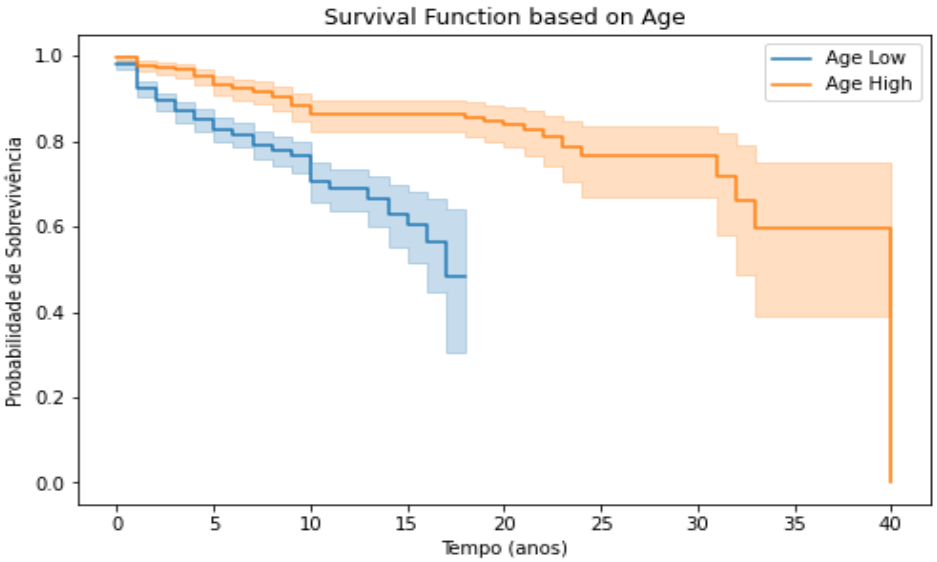


Figura 25 - Análise de sobrevivência para a variável *Age*

Relativamente à Figura 25 observa-se que quanto menor a idade (<37 anos) dos trabalhadores, maior a probabilidade de os trabalhadores cessarem funções na empresa, permanecendo menos tempo na mesma. Por exemplo, ao fim de 10 anos, a probabilidade de a população mais jovem permanecer na empresa é de, aproximadamente, 65% face aos cerca de 85% da população mais sénior.

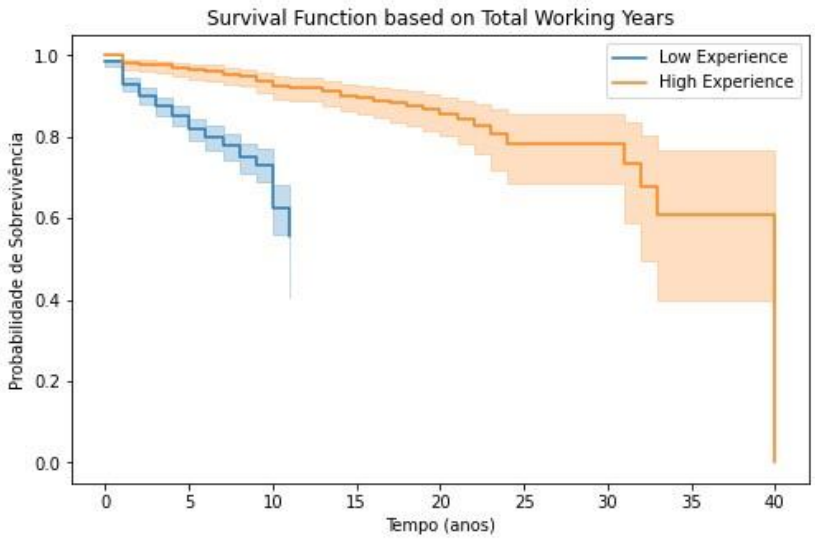


Figura 26 - Análise de sobrevivência para a variável *Total Working Years*

A Figura 26 demonstra que os trabalhadores com menos experiência profissional (menos de 11 anos de experiência), a probabilidade de permanecerem na empresa é menor. Por exemplo, Ao fim de 10 anos a probabilidade de um trabalhador permanecer na empresa é de aproximadamente 70%, sendo que no mesmo período de referência, a probabilidade de permanecerem é quase 100%.

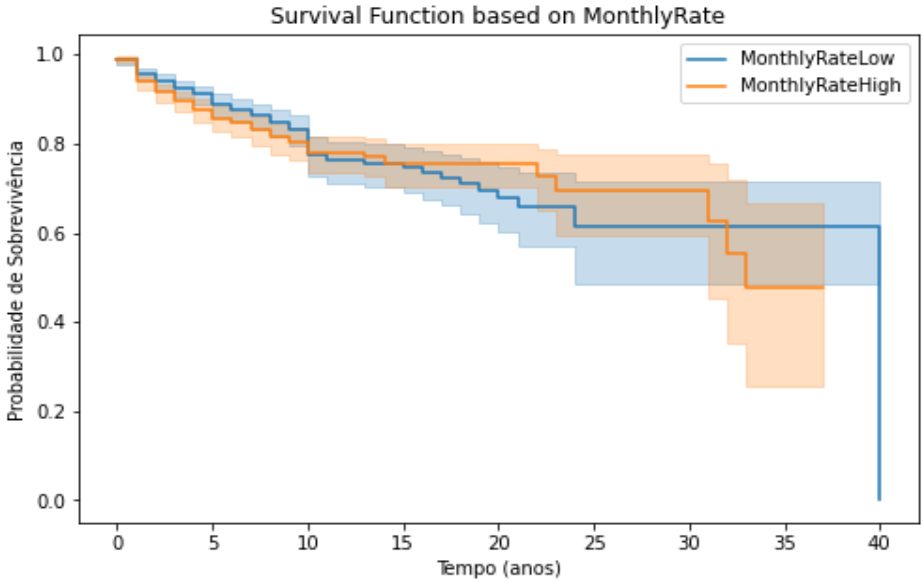


Figura 27 - Análise de sobrevivência à variável *Monthly Rate*

A Figura 27 indica maior variação relativamente à variável *MonthlyRate*. Neste sentido verifica-se que até aos 10 anos, os trabalhadores que auferem um vencimento superior (>14.313 dólares) apresentam uma maior probabilidade de saírem da empresa. Após este período as curvas invertem-se até cerca dos 32 anos. A partir dos 32 anos colaboradores que revelam um vencimento mais elevado, são também aqueles que apresentam uma maior probabilidade de cessar as suas funções na empresa.

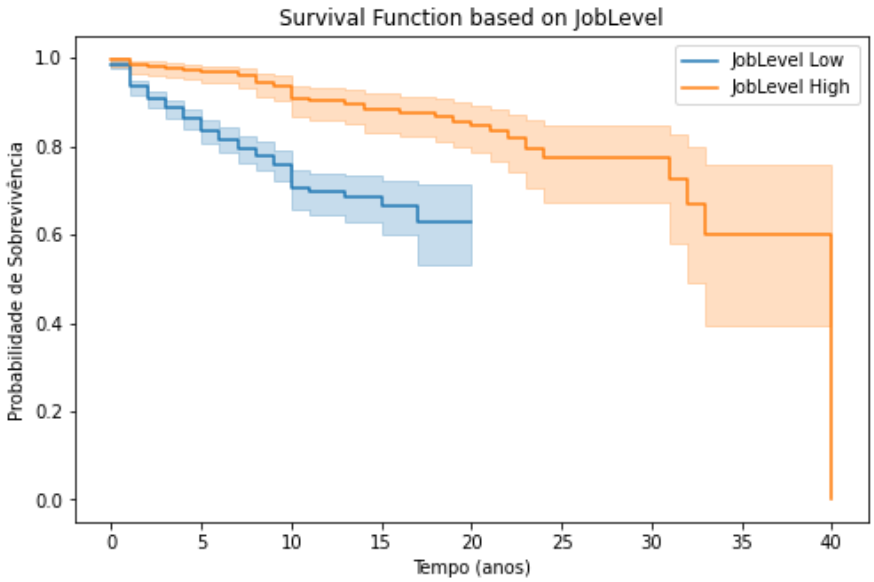


Figura 28 - Análise de sobrevivência à variável *Job Level*

Na Figura 28 é possível observar-se que quanto menor for o nível hierárquico relativo à função que desempenham, maior a probabilidade de os trabalhadores saírem da empresa. Por exemplo, ao fim de 10 anos de atividade e caso apresentem um nível de carreira baixo, a probabilidade de permanecerem é de 65%, face aos 85%, caso os trabalhadores possuam uma categoria profissional superior.

Dashboard

O *Dashboard* permite visualizar informação relevante acerca dos dados para a concretização de um determinado objetivo. Desta forma, torna-se um meio de comunicação importante e eficaz na interação com o público-alvo (Matheus, Janssen, & Maheshwari, 2018).

Com base na informação disponibilizada no *dashboard*, é possível a monitorização e análise de dados que se traduz em tomadas de decisão mais rápidas e objetivas, o que conduz a uma maior eficiência e eficácia das ações (Matheus, Janssen, & Maheshwari, 2018; Sarikaya *et al.*, 2018). Para tal, recorre-se ao uso de gráficos, tabelas, pictogramas, entre outras ferramentas visuais, como uma forma de visualização dos dados simples e fácil de compreender (Matheus, Janssen, & Maheshwari, 2018).

A implementação do *dashboard* no presente trabalho foi realizada através do *Jupyter Notebook* e da extensão *Voilà*. O *Jupyter Notebook* é uma ferramenta comumente utilizada na criação e execução de código interativo e o *Voilà* é uma extensão que permite transformar o *notebook* numa aplicação web independente, em que o *dashboard* é executado num servidor sem que para isso seja necessária a instalação do *Jupyter Notebook*. Assim, o *dataset* foi importado e submetido a um pré-processamento com o objetivo de os dados serem apresentados através de diferentes elementos visuais no *dashboard*.

Para a visualização dos dados através da geração de gráficos foram utilizadas bibliotecas *python* como *matplotlib* e *seaborn*, e tabelas através da biblioteca Pandas e de HTML.

O *dashboard* inclui:

1. Apresentação do *dataset* utilizado e respetivos dados estatísticos;
2. Resultados da implementação da análise de sobrevivência;
3. Resultados da implementação do algoritmo de agrupamento *K-means*;
4. Resultados da implementação do algoritmo de associação *Apriori*;
5. Apresentação dos resultados do melhor algoritmo de classificação – *SVM Kernel Linear*.

Após a implementação de um *dashboard* através das ferramentas utilizadas importa sublinhar a sua capacidade de partilha de dados e informação de forma acessível e eficaz, utilizado a linguagem *Python* através das suas bibliotecas.

Em primeiro lugar é possível observar as primeiras 5 linhas do *dataset* (Figura 29) e o gráfico *pairplot* (Figura 30) com a relação das variáveis e a sua distribuição.

DASHBOARD EMPLOYEES ATTRITION

Dataset

	Age	Attrition	BusinessTravel	DailyRate	Department	DistanceFromHome	Education	EducationField	EnvironmentSatisfaction	Gender	...	PerformanceRating	RelationshipSatisfaction	StockOptionLevel	TotalWork
0	41	1	1	1102	2	1	2	1	2	0	...	3	1	0	
1	49	0	2	279	0	8	1	1	3	1	...	4	4	1	
2	37	1	1	1373	0	2	2	5	4	1	...	3	2	0	
3	33	0	2	1392	0	3	4	1	4	0	...	3	3	0	
4	27	0	1	591	0	2	1	3	1	1	...	3	4	1	

5 rows × 31 columns

Figura 29 - *Dashboard: Dataset*

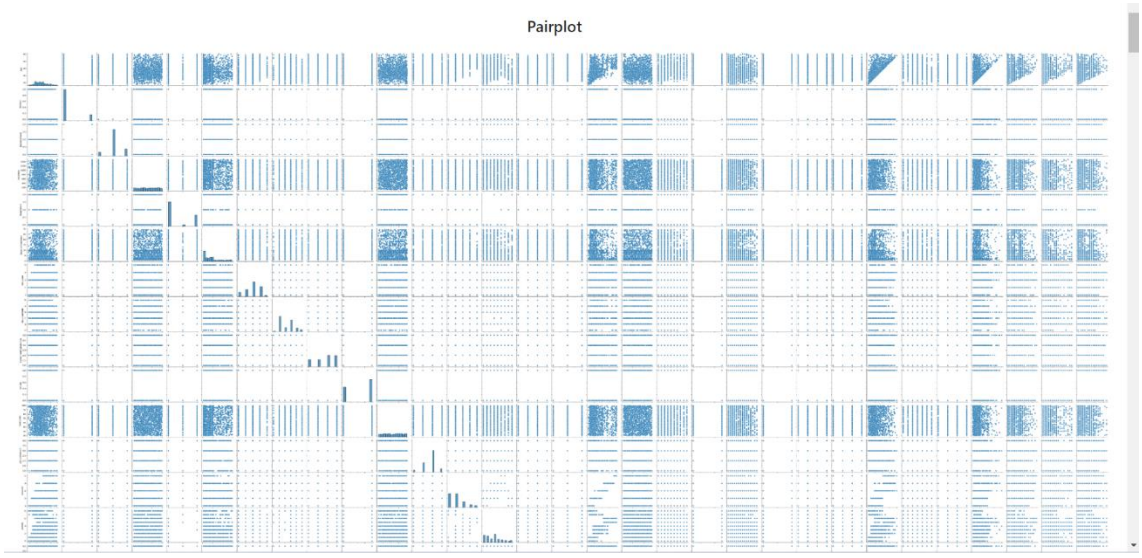


Figura 30 - *Dashboard: Pairplot*

De seguida é apresentado um resumo estatístico com recurso a uma tabela onde se inclui a contagem de observações, média, desvio-padrão, mínimo, máximo e os quartis, 25%, 50% e 75% (Figura 31). O gráfico de correlações entre as várias variáveis é também exibido como demonstra a Figura 32.

Statistical Summary

	Age	Attrition	BusinessTravel	DailyRate	Department	DistanceFromHome	Education	EducationField	EnvironmentSatisfaction	Gender	...	PerformanceRating	RelationshipSatisfaction	Stock
count	1470.000000	1470.000000	1470.000000	1470.000000	1470.000000	1470.000000	1470.000000	1470.000000	1470.000000	1470.000000	...	1470.000000	1470.000000	
mean	36.923810	0.161224	1.086395	802.485714	0.649660	9.192517	2.912925	2.213605	2.721769	0.600000	...	3.153741	2.712245	
std	9.135373	0.367863	0.532170	403.509100	0.913768	8.106864	1.024165	1.272029	1.093082	0.490065	...	0.360824	1.081209	
min	18.000000	0.000000	0.000000	102.000000	0.000000	1.000000	1.000000	0.000000	1.000000	0.000000	...	3.000000	1.000000	
25%	30.000000	0.000000	1.000000	465.000000	0.000000	2.000000	2.000000	1.000000	2.000000	0.000000	...	3.000000	2.000000	
50%	36.000000	0.000000	1.000000	802.000000	0.000000	7.000000	3.000000	2.000000	3.000000	1.000000	...	3.000000	3.000000	
75%	43.000000	0.000000	1.000000	1157.000000	2.000000	14.000000	4.000000	3.000000	4.000000	1.000000	...	3.000000	4.000000	
max	60.000000	1.000000	2.000000	1499.000000	2.000000	29.000000	5.000000	5.000000	4.000000	1.000000	...	4.000000	4.000000	

8 rows × 31 columns

Figura 31 - *Dashboard: Statistical Summary*

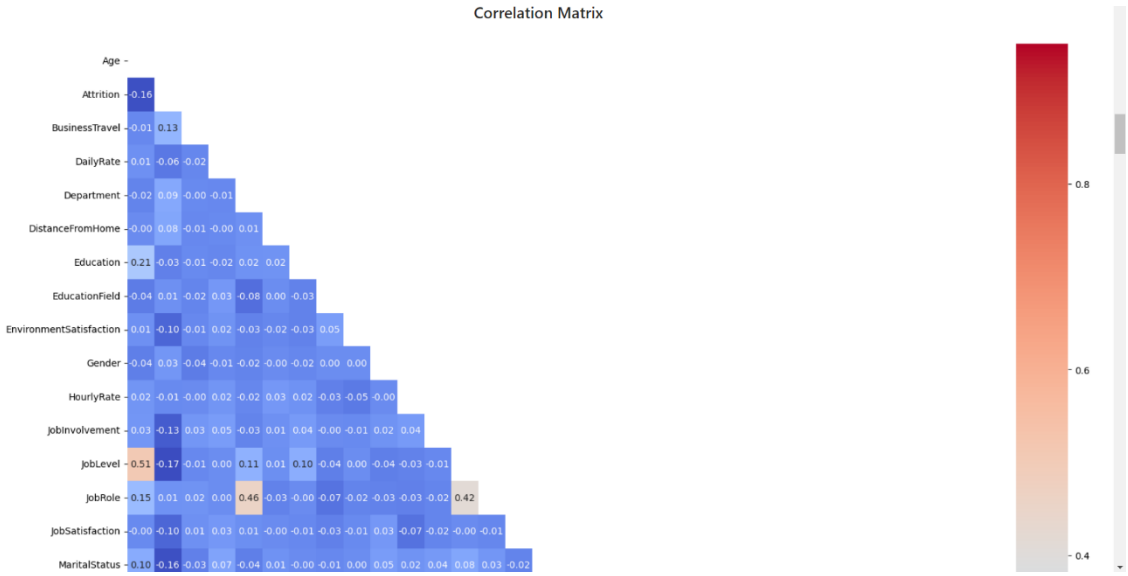


Figura 32 - *Dashboard: Correlation Matrix*

A Figura 33 mostra a análise de sobrevivência, sendo seguida no *dashboard* pelos gráficos que incluem as variáveis *JobInvolvement*, *Overtime*, *BusinessTravel*, *Age*, *MonthlyRate*, *JobLevel* e *TotalWorkingYears*.

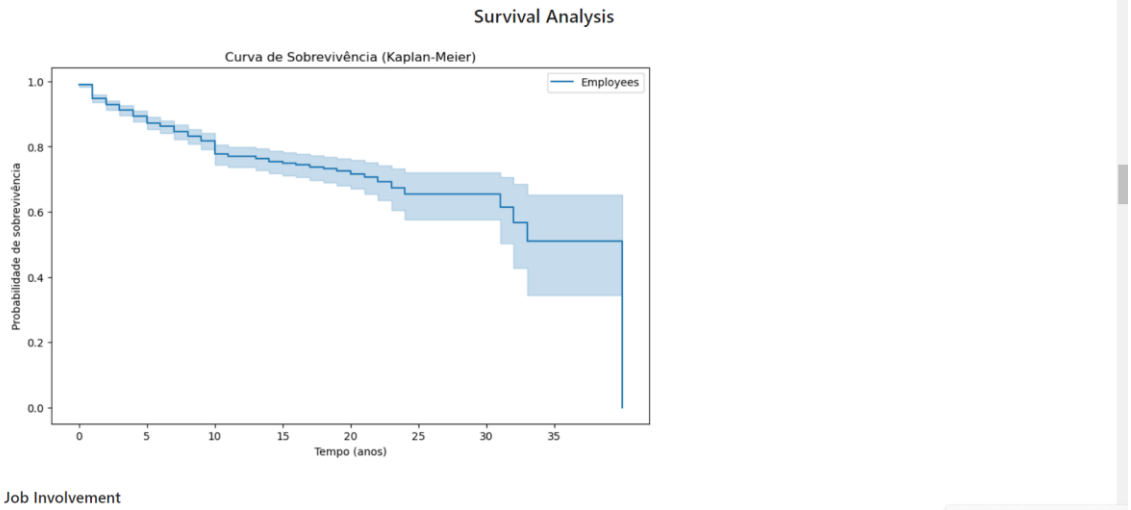


Figura 33 - *Dashboard: Survival Analysis*

Na Figura 34 é apresentado o resultado da implementação do algoritmo de agrupamento *K-means*, onde são exibidos os gráficos do método cotovelo e da representação dos clusters.

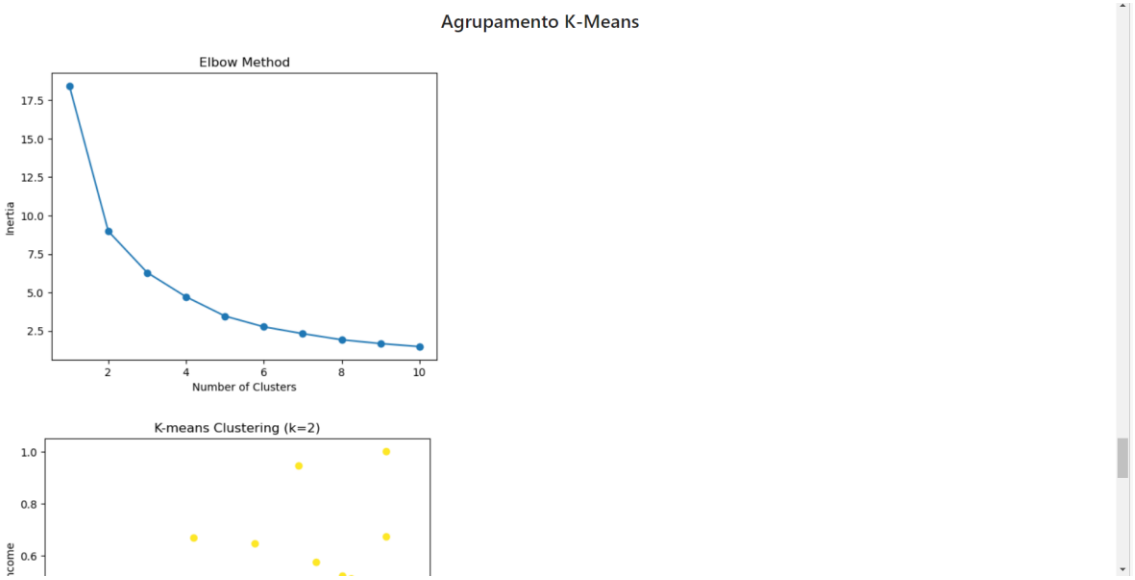


Figura 34 - *Dashboard: K-Means*

A Figura 35 revela a tabela com as principais regras de associação geradas pelo algoritmo de associação *Apriori*.

	Antecedentes	Consequentes	Suporte	Confiança	Lift
1	YearsWithCurrManager_binary	YearsAtCompany_binary	0.312500	0.869565	2.256169
2	YearsAtCompany_binary	YearsWithCurrManager_binary	0.312500	0.810811	2.256169
3	YearsInCurrentRole_binary	YearsAtCompany_binary	0.296875	0.904762	2.347490
4	YearsAtCompany_binary	YearsInCurrentRole_binary	0.296875	0.770270	2.347490
5	MonthlyIncome_binary	TotalWorkingYears_binary	0.291667	0.788732	1.869588

Figura 35 - *Dashboard: Apriori*

Finalmente, na Figura 36 é apresentado o resultado da implementação do algoritmo de classificação *SVM Kernel Linear* pelo facto de ser o algoritmo com melhor desempenho no presente estudo.

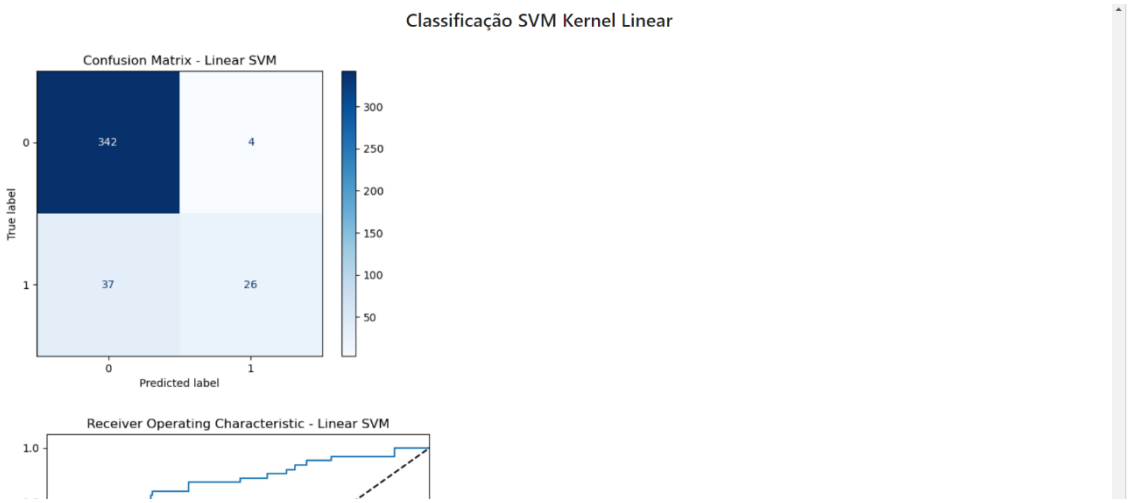


Figura 36 - *Dashboard*: Classificação – *SVM Kernel Linear*

Discussão e Conclusão

A informação desempenha um papel fundamental no mundo atual, tornando-se o principal recurso para indivíduos, sociedade e Organizações. Caracterizado como a era digital, o contexto atual possibilita o acesso célere e facilitado à informação, no entanto a forma como conseguimos analisá-la e aplicá-la é preponderante para garantir o sucesso nas mais diversas áreas, nomeadamente no mercado empresarial.

Sendo a informação um ativo intangível estratégico de elevado valor para as empresas, o recurso ao *Machine Learning* é cada vez mais pertinente, na medida em que possui a capacidade de analisar grandes quantidades de dados de forma eficiente e identificar padrões complexos para a tomada de decisão consciente.

No contexto específico dos recursos humanos, o *Machine Learning* também ele deverá ser considerado, através da interpretação de dados relacionados com os colaboradores. Essa análise avançada permite às empresas identificar fatores que influenciam a rotatividade e/ou demissões dos trabalhadores, prever tendências e desenvolver estratégias eficazes para a retenção do talento, implementando políticas de recursos humanos personalizadas, promover um ambiente de trabalho mais produtivo e reduzir os custos. De acordo com Singh (2019), a retenção promove o crescimento e estabilidade da empresa, sendo que a saída de colaboradores e contratação de novos acarreta custos elevados.

A implementação de diversos algoritmos no presente trabalho permitiu analisar os dados dos colaboradores sob diferentes perspetivas, observando-se tendências e padrões associados às variáveis em estudo.

No algoritmo de agrupamento, *K-means*, verificou-se a existência de dois clusters distintos associados ao fator *attrition*, e analisando os seus centróides, conclui-se que um *cluster* se refere aos colaboradores com idade mais jovem e salários mais baixos, contrastando com o grupo de colaboradores com mais idade e vencimentos heterogéneos, mas, ainda assim, com um vencimento médio superior aos anteriores.

Relativamente às regras de associação geradas pelo algoritmo *Apriori*, verifica-se que as variáveis *YearsWithCurrManager*, *YearsAtCompany* e *YearsInCurrentRole* apresentam uma forte associação neste *dataset*. De facto, é interessante analisar a tendência de que colaboradores que apresentam *attrition*, se encontram há mais tempo com a mesma chefia e estão também há mais tempo na empresa. É fulcral analisar o papel das chefias no contexto laboral, sendo que uma má relação com a liderança assume forte impacto na saída de recursos humanos nas empresas (Singh,2019).

No entanto, importa salientar que, quando se verifica *attrition,* existe também uma forte associação entre colaboradores que apresentam vencimentos acima da média em simultâneo com o facto de estarem no mercado laboral há mais anos.

Posto isto, os resultados obtidos vão de encontro ao que é referido por Singh (2019), em que uma das principais dificuldades das empresas é precisamente aplicar ações para prevenir a fuga de colaboradores qualificados. Neste mesmo estudo, alguns autores referem que as alterações demográficas associadas à idade vão ter impacto na maior procura de talento. Assim, é fulcral que as empresas adotem um planeamento estratégico transversal aos diferentes grupos etários, capaz de agir eficazmente sobre a retenção de colaboradores qualificados e com maior experiência profissional.

Considerando os algoritmos de classificação analisados no presente estudo, foram implementados os seguintes: *Naive Bayes (Gaussian e Multinomial), Decision Tree, Random Forest, Logistic Regression, Adaboost, XGBoost, Support Vector Machine (Kernel Linear, Polinomial, RBF e Sigmoid), K-Nearest Neighbors e o Artificial Neural Network*. Para estes algoritmos e por forma a ser promovida uma análise comparativa do desempenho alcançado por cada um, estabeleceu-se a implementação das métricas de avaliação (*accuracy*, *recall*, *precision* e *F1-Score*), e das respetivas curvas ROC e AUC. Dos resultados alcançados os dois algoritmos que apresentaram um melhor desempenho foram o *Support Vector Machine Kernel Linear* e o *Artificial Neural Network,* sendo que o primeiro evidenciou uma *accuracy = 90%, precision= 90%, recall= 90% e F1-Score= 88%* e o segundo uma *accuracy = 89%, precision= 89%, recall= 89% e F1-Score= 88%*. Quanto à curva *ROC AUC* ambos os algoritmos demonstraram uma boa capacidade de distinguir corretamente as classes positivas e negativas (ou seja, de haver ou não *attrition*), sendo que para *Support Vector Machine Kern*el Linear foi de 0.86 e para o *Artificial Neural Network foi de 0.85*. Face ao exposto, foi assim possível concluir que o algoritmo de classificação que destacou pelo desempenho apresentado foi o *Support Vector Machine Kern*el Linear.

Foi igualmente implementada a técnica estatística de análise de sobrevivência, através da curva de *Kaplan-Meier* que consiste em analisar o tempo até à ocorrência de um determinado evento (neste caso, a probabilidade dos trabalhadores poderem cessar funções ao longo do tempo). Nesta decorrência foram aplicadas um conjunto de variáveis consideradas mais "relevantes" nos algoritmos de classificação (coeficiente de peso e análise de sensibilidade), como também as que apresentavam uma maior correlação face à variável alvo *Attrition*, por exemplo, *TotalWorkingYears*, *JobInvolvment, BusinessTravel, OverTime, JobLevel*, *Age* e *MonthlyRate*. Desta análise foi possível destacar alguns pontos que poderão ser relevantes. Em primeira instância a necessidade de adoção de um conjunto de estratégias que viabilizem a retenção de talento mais jovem, na medida em que, e através da análise das variáveis *Age* e *TotalWorkingYears* a probabilidade de permanência desta população, comparativamente com os trabalhadores mais "séniores" é significativamente menor. Consideramos, que talvez possa ser pertinente a adoção de estratégias que visem a sua

retenção, considerando a aplicabilidade de práticas associadas ao *mentoring* e *coaching*, possibilidade de reajustar a política de retribuições e/ou a criação de mecanismos que procurem fomentar o envolvimento dos jovens em novos projetos e desafiar a aprendizagem contínua. Finalmente, a empresa deverá igualmente considerar o *Overtime* como algo que poderá afetar o processo de retenção, pois a probabilidade de permanecerem na empresa ao fim de 10 anos é significativamente inferior. Será assim fundamental perceber as causas da existência do *Overtime*, sendo que para a sua mitigação poderão ser analisadas novas formas de organização do trabalho e/ou reajustar o processo de recrutamento e seleção.

No entanto é importante referir que o *imbalance* na variável alvo (*attrition*) pode ter influenciado os resultados, nomeadamente na análise das métricas de avaliação associada aos algoritmos de classificação. Neste sentido, e para estudos futuros, o balanceamento dos dados deverá ser considerado para obter resultados mais equilibrados e uma melhor deteção da classe minoritária (e.g. número de *yes* da variável *attrition*). Existem várias abordagens que podem ser utilizadas nos casos de *imbalance* como por exemplo através da técnica de SMOTE (*Synthetic Minority Over-sampling Technique*) como é referido por Azeem *et al*.(2017) como sendo uma técnica eficaz, aumentando a quantidade de amostras da classe minoritária, gerando novos exemplos através de amostras semelhantes. Outra técnica que também poderia ser aplicada seria o *Random Over-sampling*, mas segundo os mesmos autores os resultados demonstraram não ser uma técnica eficaz no processo de equilíbrio das classes.

Referências

Ahmed, M., Seraj, R. & Islam, S. (2020). The k-means Algorithm: A Comprehensive Survey and Performance Evaluation. *Electronics 2020*, 9, 1295. https://doi.org/10.3390/electronics9081295

Azeem, M., Usman, M., & Fong, A. C. M. (2017). A churn prediction model for prepaid customers in telecom using fuzzy classifiers. *Telecommunication Systems*, 66(4), 603–614. https://doi.org/10.1007/s11235-017-0310-7

Bauer, T., Erdogan, B., Caughlin, D., & Truxillo, D. (2019). *Human resource management: People, Data, and Analytics*. SAGE Publications, Incorporated.

Berrar, D. (2019). Cross-Validation. *Elsevier eBooks*, 542–545. https://doi.org/10.1016/b978-0-12-809633-8.20349-x

Chen, T., & Guestrin, C. (2016). XGBoost: A Scalable Tree Boosting System. *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining - KDD '16*, 785–794. https://doi.org/10.1145/2939672.2939785

Cunningham, P., & Delany, S. J. (2020). k-Nearest Neighbour Classifiers - A Tutorial. *ACM Computing Surveys*, *54*(6), 1–25. https://doi.org/10.1145/3459665

Dell'Aversana, P. (2019). *Artificial neural networks and deep learning*. A simple overview. Disponível em: https://www.researchgate.net/publication/333263211

Duda, R. O., Hart, P. E., & Stork, D. G. (2001). *Pattern Classification (2nd ed.)*. Wiley-Interscience.

Ersoz Kaya, İ., & Korkmaz, O. (2021). Machine Learning Approach for Predicting Employee Attrition and Factors Leading to Attrition. *Çukurova Üniversitesi Mühendislik Fakültesi Dergisi*, 913–928. https://doi.org/10.21605/cukurovaumfd.1040487

Gandhi, R. (2018, June 07). Support Vector Machine — Introduction to Machine Learning Algorithms. *Medium*. Retrieved April 15, 2023, from https://towardsdatascience.com/support-vector-machine-introduction-to-machine-learning-algorithms-934a444fca47

Ghazal, T., Hussain, M., Said, R., Nadeem A., Hasan M., Ahmad M., Khan M., & Naseem M. (2021). Performances of K-Means Clustering Algorithm with Different Distance Metrics. *Intelligent Automation & Soft Computing*, *30*(2), 735-742. https://doi.org/10.32604/iasc.2021.019067Haldorai, K., Kim, W., Pillai, S., Park, T. & Balasubramanian, K. (2019). Factors affecting hotel employees' attrition and turnover: Application of pull-push-mooring framework. *International Journal of Hospitality Management*, 83, 46–55. https://doi.org/10.1016/j.ijhm.2019.04.003

Han, J., Kamber, M., & Pei, J. (2012). *Data mining: concepts and techniques*. Choice Reviews Online. https://doi.org/10.5860/choice.49-3305

Hastie, T., Tibshirani, R., & Friedman, J. (2009). *The elements of statistical learning, second edition : data mining, inference, and prediction* (2nd ed.). Springer.

Hertzmann, A., Fleet, D., & Brubaker, M. (2015). *Machine Learning and Data Mining Lecture Notes*. Retrieved April 15, 2023, from https://tensorflowkorea.files.wordpress.com/2016/09/cscc11_lecture_note.pdf

Kartsonaki, C. (2016). Survival analysis. *Diagnostic Histopathology*, 22(7), 263–270. https://doi.org/10.1016/j.mpdhp.2016.06.005

Kong, G., Tian, H., Wu, Y., & Wei, Q. (2020*). Improvement of Association Rule Algorithm Based on Hadoop for Medical Data*. Communications in Computer and Information Science. https://doi.org/10.1007/978-981-15-7984-4_38

Liao, S.-H., & Wen, C.-H. (2007). Artificial neural networks classification and clustering of methodologies and applications – literature analysis from 1995 to 2005. *Expert Systems with Applications*, *32*(1), 1–11. https://doi.org/10.1016/j.eswa.2005.11.014

Mahesh, B. (2020). Machine learning algorithms — *A review*. *Int. J. Sci. Res*. 9, 381–386.

Matheus, R., Janssen, M. & Maheshwari, D. (2018). Data science empowering the public: Data-driven dashboards for transparent and accountable decision-making in smart cities. *Government Information Quarterly*, (), S0740624X18300303–. https://doi.org/10.1016/j.giq.2018.01.006

Mishra, S., Sarkar, U., Taraphder, S., Datta, S., Swain, D., Saikhom, R., Panda, S., & Laishram, M. (2017). Principal Component Analysis. *International Journal of Livestock Research*, *7*(5), 1. https://doi.org/10.5455/ijlr.20170415115235

Osisanwo, F., Akinsola, J., Awodele, O., Hinmikaiye, J., Olakanmi, O. & Akinjobi, J. (2017). Supervised Machine Learning Algorithms: Classification and Comparison. *International Journal of Computer Trends and Technology*, *48*(3), 128–138. https://doi.org/10.14445/22312803/ijctt-v48p126

Pineda-Jaramillo, J. (2019). A review of Machine Learning (ML) algorithms used for modeling travel mode choice. *Dyna-colombia*, *86*(211), 32–41. https://doi.org/10.15446/dyna.v86n211.79743

Rai, S., Mishra, P. & Ghoshal, U. (2021). Survival analysis: A primer for the clinician scientists. *Indian J Gastroenterol* 40, 541–549. https://doi.org/10.1007/s12664-021-01232-1

Raina, H., & Shafi, O. (2015). *Analysis Of Supervised Classification Algorithms*. Retrieved July 16, 2023, from https://www.ijstr.org/final-print/sep2015/Analysis-Of-Supervised-Classification-Algorithms.pdf

Raschka, S., & Mirjalili, V. (2019). Python machine learning : machine learning and deep learning with Python, scikit-learn, and TensorFlow. *Packt Publishing eBooks*. http://202.62.95.70:8080/jspui/handle/123456789/12650

Sarikaya, A., Correll, M., Bartram, L., Tory, M., Fisher, D. (2018). What Do We Talk About When We Talk About Dashboards?. *IEEE Transactions on Visualization and Computer Graphics*, (), 1–1. doi:10.1109/TVCG.2018.2864903

Singh, D. (2019). A Literature Review on Employee Retention with Focus on Recent Trends. *International Journal of Scientific Research in Science, Engineering and Technology*, 425–431. https://doi.org/10.32628/ijsrst195463

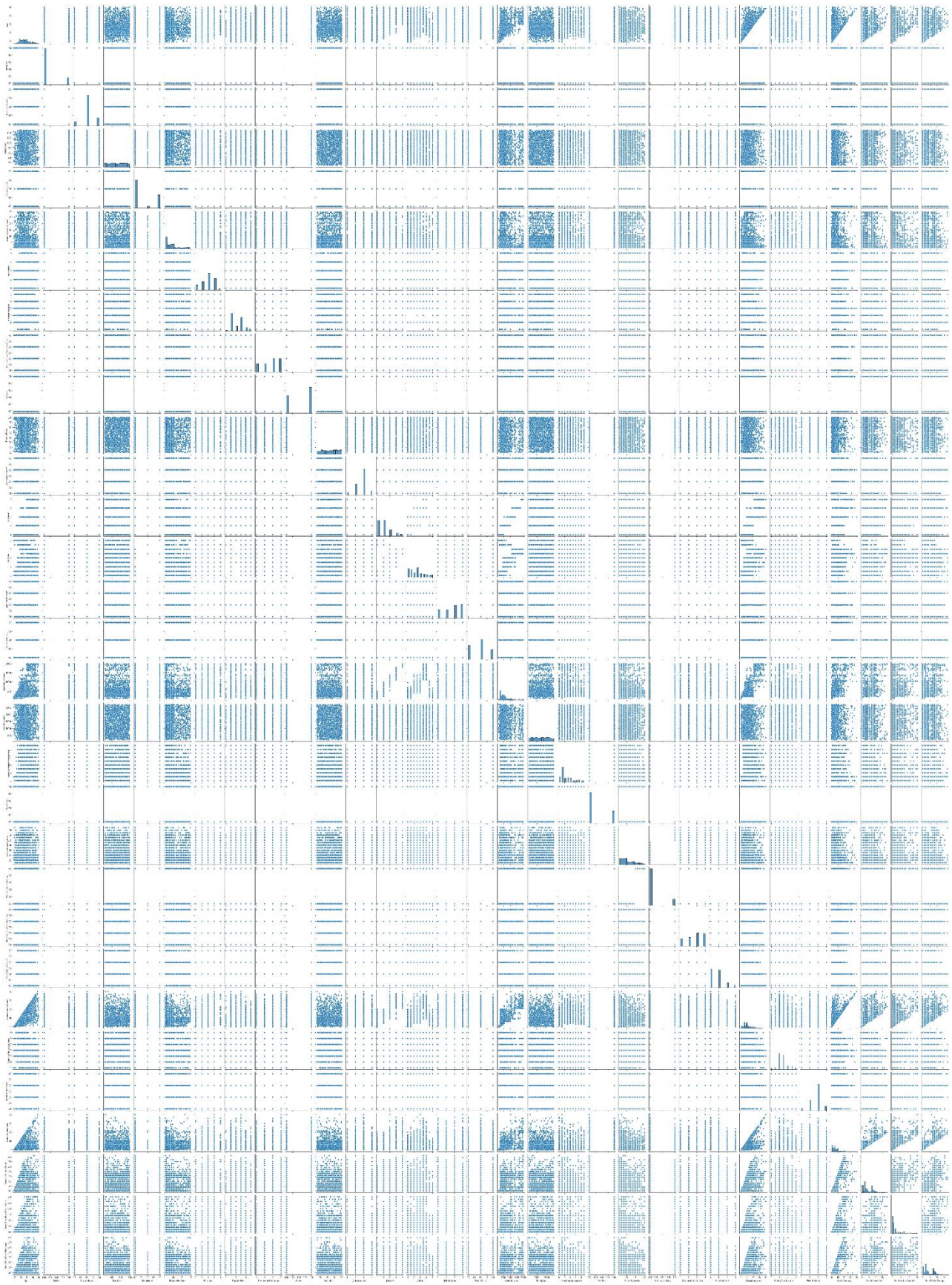
Sobreiro, P. (2023). *Customer dropout prediction using machine learning hybrid survival models*. Dialnet.unirioja.es. https://dialnet.unirioja.es/servlet/dctes?codigo=314884

Sobreiro, P., Martinho, D., Alonso, J. M., & Berrocal, J. (2022). A SLR on Customer Dropout Prediction. *IEEE Access*, 1. https://doi.org/10.1109/access.2022.3146397

Yahia, N., Hlel, J., & Colomo-Palacios, R. (2021). From Big Data to Deep Data to Support People Analytics for Employee Attrition Prediction. *IEEE Access, Access, IEEE*, *9*, 60447–60458. https://doi.org/10.1109/ACCESS.2021.3074559

Anexos

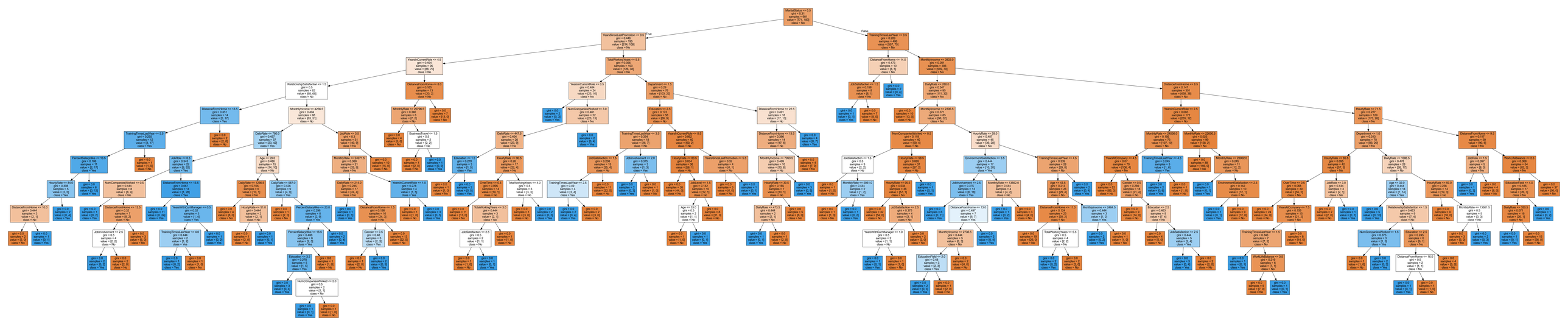
Anexo A - Pairplot



Anexo B – Decision Tree


OverTime <= 0.5
gini = 0.285
samples = 954
value = [790, 164]
class = No
True
False

Anexo C – Random Forest


MaritalStatus <= 0.5
gini = 0.481
samples = 931
value = [771, 160]
class = No
True
False
YearsSinceLastPromotion <= 0.5
gini = 0.499
samples = 195
value = [114, 109]
class = No
TrainingTimesLastYear <= 0.5
gini = 0.327
samples = 736
value = [657, 75]
class = No
YearsInCurrentRole <= 4.5
gini = 0.494
samples = 95
value = [58, 79]
class = Yes
TotalWorkingYears <= 5.5
gini = 0.356
samples = 100
value = [126, 38]
class = No
DistanceFromHome <= 4.0
gini = 0.165
samples = 72
value = [97, 12]
class = No
DistanceFromHome <= 14.0
gini = 0.473
samples = 10
value = [5, 9]
class = Yes
MonthlyIncome <= 2802.0
gini = 0.201
samples = 394
value = [540, 70]
class = No
RelationshipSatisfaction <= 1.5
gini = 0.5
samples = 64
value = [54, 54]
class = No
MonthlyIncome <= 4266.5
gini = 0.494
samples = 31
value = [20, 31]
class = Yes
YearsInCurrentRole <= 0.5
gini = 0.484
samples = 23
value = [23, 16]
class = No
Department <= 1.5
gini = 0.29
samples = 77
value = [103, 22]
class = No
JobRole <= 3.5
gini = 0.494
samples = 9
value = [7, 9]
class = Yes
MonthlyRate <= 20796.5
gini = 0.346
samples = 72
value = [97, 27]
class = No
BusinessTravel <= 1.5
gini = 0.5
samples = 2
value = [2, 2]
class = No
gini = 0.0
samples = 7
value = [13, 0]
class = No
TrainingTimesLastYear <= 5.5
gini = 0.231
samples = 15
value = [11, 11]
class = No
DistanceFromHome <= 13.5
gini = 0.061
samples = 63
value = [93, 3]
class = No
DailyRate <= 780.0
gini = 0.477
samples = 27
value = [25, 42]
class = Yes
JobRole <= 3.5
gini = 0.49
samples = 50
value = [72, 55]
class = No
gini = 0.0
samples = 2
value = [2, 0]
class = No
gini = 0.0
samples = 2
value = [0, 2]
class = Yes
gini = 0.0
samples = 2
value = [2, 0]
class = No
gini = 0.0
samples = 2
value = [0, 2]
class = Yes
NumCompaniesWorked <= 3.0
gini = 0.461
samples = 21
value = [33, 13]
class = No
gini = 0.0
samples = 2
value = [0, 2]
class = Yes
EducationField <= 2.5
gini = 0.172
samples = 61
value = [96, 9]
class = No
DistanceFromHome <= 22.5
gini = 0.491
samples = 16
value = [11, 13]
class = Yes
PercentSalaryHike <= 23.5
gini = 0.48
samples = 13
value = [9, 14]
class = Yes
JobSatisfaction <= 3.5
gini = 0.444
samples = 9
value = [9, 4]
class = No
Age <= 29.0
gini = 0.459
samples = 15
value = [14, 10]
class = No
DailyRate <= 447.0
gini = 0.494
samples = 20
value = [23, 19]
class = No
TrainingTimesLastYear <= 2.5
gini = 0.49
samples = 18
value = [16, 22]
class = Yes
YearsSinceLastPromotion <= 0.5
gini = 0.32
samples = 43
value = [80, 20]
class = No
DistanceFromHome <= 13.5
gini = 0.305
samples = 14
value = [17, 4]
class = No
gini = 0.0
samples = 4
value = [0, 7]
class = Yes
gini = 0.0
samples = 2
value = [2, 0]
class = No
gini = 0.0
samples = 2
value = [0, 2]
class = Yes
DailyRate <= 1170.5
gini = 0.346
samples = 11
value = [15, 4]
class = No
NumCompaniesWorked <= 0.5
gini = 0.444
samples = 9
value = [6, 12]
class = Yes
DailyRate <= 1170.5
gini = 0.346
samples = 11
value = [15, 4]
class = No
Education <= 1.5
gini = 0.219
samples = 3
value = [1, 12]
class = Yes
HourlyRate <= 90.5
gini = 0.26
samples = 17
value = [22, 4]
class = No
Education <= 1.5
gini = 0.219
samples = 3
value = [1, 12]
class = Yes
HourlyRate <= 95.5
gini = 0.26
samples = 17
value = [22, 4]
class = No
JobSatisfaction <= 1.5
gini = 0.219
samples = 3
value = [1, 12]
class = Yes
JobInvolvement <= 2.5
gini = 0.219
samples = 3
value = [1, 12]
class = Yes
gini = 0.0
samples = 2
value = [0, 2]
class = Yes
gini = 0.0
samples = 2
value = [0, 2]
class = Yes
gini = 0.0
samples = 2
value = [0, 2