# Análise de Desempenho

Gabriel Pelosi
*DEI – Departamento de Engenharia Informática*
*(do Instituto Politécnico do Porto)*
*ISEP - Instituto Superior de Engenharia do Porto*
*(do Instituto Politécnico do Porto)*
Porto, Portugal
1180017@isep.ipp.pt

Sofia Costa
*DEI – Departamento de Engenharia Informática*
*(do Instituto Politécnico do Porto)*
*ISEP - Instituto Superior de Engenharia do Porto*
*(do Instituto Politécnico do Porto)*
Porto, Portugal
1191063@isep.ipp.pt

***Abstrato*—Neste artigo, encontra-se presente uma análise de dados relacionados com os clientes de uma operadora móvel que se encontra numa situação de decréscimo de número de clientes. Esta análise tem por base o uso de algoritmos de Machine Learning, de regressão e classificação. Para avaliar o desempenho dos modelos utilizados, foram utilizadas métricas de desempenho como a MAE e RMSE pelo que no final deste estudo, concluímos que o modelo com melhor desempenho em termos de precisão tem o nome de rede neuronal.**

***Palavras-Chave—operadora móvel, modelo de regressão, modelo de classificação, regressão linear, árvore de decisão, rede neuronal, knn***

## I. Introdução

Este artigo foi realizado com o objetivo de documentar a investigação e aplicação de algoritmos de aprendizagem automática na exploração de dados e realizar uma comparação entre os resultados utilizando os testes estatísticos mais adequados. Ao longo deste artigo, serão encontradas secções que irão contextualizar o problema abordado, um breve enquadramento teórico sobre os algoritmos utilizados, de modo a facilitar a compreensão do leitor; a análise e discussão dos resultados obtidos e por fim, conclusões sobre os mesmos.
A metodologia adotada durante a realização deste estudo, teve inicialmente um caráter de pesquisa sobre os algoritmos a utilizar, de forma a entender como estes são aplicados em relação aos dados em estudo. De seguida, foram aplicados diferentes modelos de Machine Learning a dados treino e por fim feita uma análise e conclusão sobre os resultados obtidos.

## II. Contextualização do Problema

Uma Operadora Móvel, encontra-se numa situação anormal de perda de clientes, o que esta a ter um impacto direto no desempenho da organização. Assim, tornou-se necessário identificar as razões subjacentes ao problema e identificar eventualmente medidas preventivas direcionadas. Pretende-se então realizar a análise de dados estruturados referentes aos clientes da Operadora Móvel através de modelos de classificação/regressão usando algoritmos de aprendizagem automática estudados: regressão linear, árvores de decisão, k-vizinhos-mais-próximos e redes neuronais.

## III. Estado de Arte

### A. Machine Learning

*Machine Learning* surgiu como um subcampo de Inteligência Artificial com especial foco no desenvolvimento de algoritmos para que os computadores possam aprender automaticamente modelos (preditivos) a partir de dados . Estes algoritmos podem ser divididos em três áreas: Supervisionados, Semisupervisonados e de Reforço.

O estudo em foco, enfatiza a aprendizagem e utilização de algoritmos supervisionados nos quais se inserem modelos de classificação e modelos de regressão, ou seja, modelos que operam com dados de treino identificados (*labelled*) que são exemplos daquilo que se quer prever.

### B. Regressão Linear

A Regressão Linear, insere-se no modelo de Regressão referido anteriormente, e trata-se de um algoritmo utilizado para prever o valor de algo com base noutros dados (históricos). Existem dois tipos de regressão linear: Simples e Múltipla. Na Regressão Linear Simples, para prever algo é utilizada apenas uma variável preditora (independente), contrariamente, a Regressão Linear Múltipla utiliza várias variáveis independentes para realizar a predição.

### C. Árvores de Decisão

Uma Árvore de Decisão, é um algoritmo que pode ser utilizado tanto em modelos de Regressão, sendo a árvore chamada de Árvore de Regressão e que utiliza variáveis dependentes contínuas, como ou em modelos de Classificação, com o nome de Árvore de Classificação que utiliza variáveis dependentes categóricas .

A sua representação assimila bastante a uma árvore, contendo o nó da raiz, nós intermédios e nós-folha, que possuem o valor do resultado final. Nos ramos, que separam um nó decisão de outro ao qual está conectado, existem regras de "*if-then*" que sendo respondidas de forma diferente, irão produzir também caminhos diferentes até finalmente se chegar a um nó-folha.

### D. Redes Neuronais

O modelo Rede Neuronal Artificial é inspirado nas estruturas de neurónios biológicos, do cérebro, capazes de realizar *Machine Learning* e reconhecer padrões. Um neurónio artificial tem como papel principal, receber sinais que provem de conexões de entrada e calcula um valor de saída que é enviado para outro neurónio, sendo que todos os neurónios estão conectados uns aos outros e cada ligação possui um certo peso. Existem duas fases associadas a este modelo: cálculo da soma ponderada dos valores de entrada e cálculo do neurónio de ativação usando uma função que determina se a soma anteriormente calculada é suficiente para ativar o neurónio .

### E. *Cross Validation*

A metodologia Cross Validation é uma técnica proposta para avaliar o desempenho de modelos de *Machine Learning* com pequenos aglomerados de dados (*k-sets* ou *k-folds*).

A técnica consiste em dividir o *dataset* em *k-folds* aleatoriamente, aplicar uma parte para treino e outra para teste, testar o modelo após o treino e reservar os valores obtidos no *predict*. É importante ressalvar que esse passo pode ser repetido diversas vezes. Finalmente, recorre-se aos valores obtidos nos testes e assim pode-se avaliar o modelo e compará-lo com outros modelos existentes .

### F. *ML Models Evaluation Metrics*

#### 1) *Métricas de Desempenho em Modelos de Regressão*

Para avaliar os modelos de regressão, é necessário recorrer a métodos estatísticos para quantificar o seu desempenho. Para que isso seja possível, são utilizadas métricas como: R2, que é definido como a correlação dos resultados com os valores previstos, quanto maior o R2 melhor será o modelo [4]; RMSE, o qual quantifica o erro médio da previsão, ou seja, é a diferença média entre os resultados obtidos através das previsões e os conhecidos pelo modelo. Quanto menor o RMSE, melhor o desempenho do modelo; MAE, que é considerada uma alternativa ao RMSE pois é menos sensível a *outliers*. O MAE define a diferença média absoluta entre os valores previstos e observados, assim como o RMSE, quanto menor o seu valor, melhor o desempenho do modelo treinado [4].

#### 2) *Métricas de Desempenho em Modelos de Classificacao*

Com o intuito de avaliar modelos de classificação, é necessário que para concretizar a diferença de desempenho deve ser desenvolvida uma matriz de confusão. Matriz essa que obtêm os valores de Falso Positivo, Falso Negativo, Verdadeiro Positivo e Verdadeiro Negativo, que são valores numéricos correspondentes ao número de vezes que o modelo errou um valor positivo e um valor negativo e acertou um valor positivo e um valor negativo [4].

Com a matriz de confusão construída, através de seus valores é possível calcular fatores como *Accuracy*, *Precision*, *Sensitivity*/*Recall*, *Specificity* e F1. Através desses cálculos, é possível avaliar os modelos de classificação [4].

### G. *K-Vizinhos Mais Próximos*

O modelo de aprendizagem automática K-Vizinhos Mais Próximos classifica os dados de acordo com seus vizinhos mais próximos, ou seja, a o valor de classificação de um dado qualquer será tido como base o valor de classificação que seus vizinhos obtiveram.

De acordo então com sua funcionalidade, é possível identificar que sua forma de funcionamento é no reconhecimento de padrões é tanto pode ser usado com "Nearest Neighbor", como "K Nearest Neighbor", uma vez que, o valor de vizinhos é configurável.

Para que o modelo consiga classificar os dados, é necessário calcular a distância entre eles (distância euclidiana). Tendo isso em conta, é necessário realizar sempre um pré processamento dos dados para convertê-los em numéricos, para que possibilite o algoritmo a calcular as distâncias e classificar os dados.

## IV. Análise e Discussão dos Resultados

Esta secção é dedicada à análise e tratamento dos dados utilizados para a resolução do problema assim como a análise dos resultados obtidos com a adição de uma breve discussão sobre os mesmos.

### A. *Regressão*

Antes de aplicar algoritmos e técnicas de regressão, foi necessário realizar um tratamento dos dados do ficheiro utilizado (Clientes_DataSet.csv). Foi efetuado um pré-processamento dos dados com o propósito de identificar todos os campos "NA" e consequente exclusão dos mesmos assim como a identificação de quaisquer dados inconsistentes ou *outliers*. Os dados não numéricos, foram tratados de modo a se tornarem binários, o que resultou na adição de novas colunas adicionalmente, foi excluída a coluna relativa ao atributo "ClienteID" uma vez que não possuiu dados relevantes para o estudo em causa.

Assim, com os dados num estado adequado de utilização, o primeiro passo realizado foi criar um diagrama de correlação entre todos os atributos (Fig. 1).

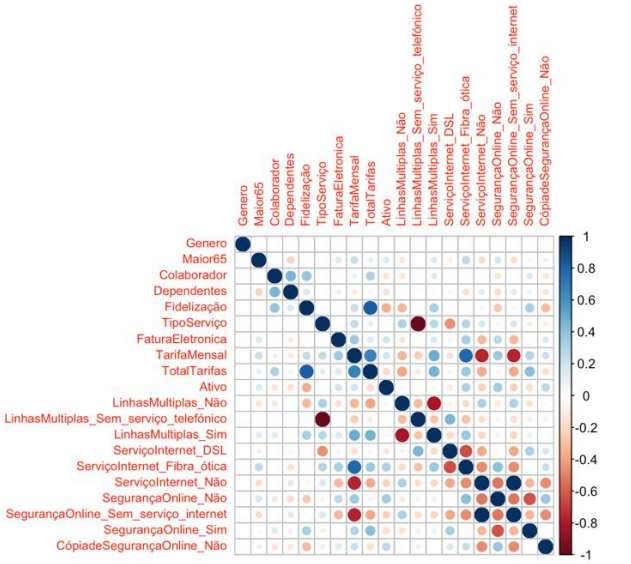


*Figura 1 - Diagrama de Correlação*

Uma vez que existem vários atributos, o diagrama representado na Figura 1 apenas inclui a correlação entre metade dos atributos, porém estes são mais que suficientes para retirar conclusões significativas.

Assim, através da observação do diagrama anterior, é possível verificar uma forte correlação, principalmente, entre os atributos "Fidelização" e "TotalTarifas" assim como entre "Tarifa Mensal" e "TotalTarifas". Tal é comprovado pelo tamanho dos círculos e a sua cor azul (superior a 60% de correlação) que indica uma correlação positiva entre os atributos, enquanto a cor vermelha representa uma correlação negativa.

Antes de aplicar técnicas de regressão, foram criados novos *data frames*, um para os dados treino e outro para os dados teste que irão ser utilizados ao longo do estudo. Os dados treino servirão para criar os modelos e os dados treino para realizar alguns testes de avaliação de desempenho.

Uma vez que os atributos "Fidelização" e "TarifaMensal" estão altamente correlacionados, recorrendo ao Modelo de Regressão Simples, é possível determinar o período de fidelização com o auxílio dos valores da tarifa mensal. Assim, obteve-se uma função linear (1), onde $Fid$ corresponde à variável objetivo (Fidelização) e $T_m$ à variável preditora (TarifaMensal).

$$Fid = 0.2079T_m + 19.0079 \quad (1)$$

Tal função, pode ser representada num diagrama de dispersão e verificar o seu comportamento em relação aos restantes dados com o objetivo de avaliar se se trata de um modelo adequado.

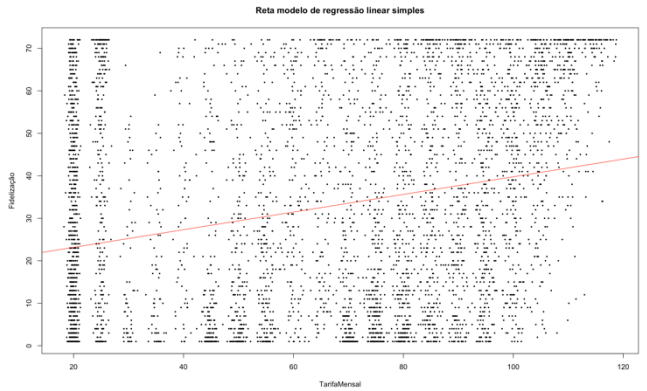


*Figura 2 - Diagrama de Dispersão entre Tarifa Mensal e Fidelização*

Pela observação do diagrama na Figura 2, podemos verificar que os valores se encontram bastante dispersos e que não existe preferência de proximidade dos mesmos em relação à reta de regressão linear.

Recorrendo às métricas de avaliação de desempenho, obtiveram-se valores relativamente baixos, como 20.57 para o Erro Médio Absoluto e 23.50 para a RMSE, o que nos diz que os dados teste utilizados estão próximos do valor absoluto das medições do instrumento de referência.

Tendo em conta o conjunto de dados treino em utilização, pretende-se prever o valor de um outro atributo, TotalTarifas, com o auxílio de todos os atributos. Para isso, começamos por utilizar o Modelo de Regressão Múltipla.

Uma vez que existem pelo menos 40 colunas de atributos, realizou-se em primeiro lugar um estudo de regressão múltipla para todos os atributos e de seguida um estudo apenas para os atributos significativos.

| | Estimate | Std. Error | t value | Pr(>\|t\|) |
|---|---|---|---|---|
| (Intercept) | -1887.098203 | 56.4161765 | -33.4495941 | 3.964576e-221 |
| Genero | 27.539696 | 19.8028971 | 1.3906903 | 1.643825e-01 |
| Maior65 | -21.165349 | 28.8696369 | -0.7331353 | 4.635110e-01 |
| Colaborador | 15.753651 | 24.0036231 | 0.6563031 | 5.116600e-01 |
| Dependentes | -33.335224 | 25.3756647 | -1.3136690 | 1.890191e-01 |
| Fidelização | 61.102343 | 0.6886199 | 88.7316014 | 0.000000e+00 |
| TipoServiço | -114.118936 | 242.3242334 | -0.4709349 | 6.377082e-01 |
| FaturaEletronica | 3.850526 | 22.2818466 | 0.1728100 | 8.628079e-01 |
| TarifaMensal | 38.700452 | 9.5282570 | 4.0616507 | 4.948844e-05 |
| Ativo | -156.337229 | 26.3426210 | -5.9347636 | 3.145217e-09 |
| LinhasMultiplas_Não | -72.238475 | 53.6759618 | -1.3458254 | 1.784211e-01 |
| ServiçoInternet_DSL | 382.406381 | 620.5618250 | 0.6162261 | 5.377740e-01 |
| ServiçoInternet_Fibra_ótica | 328.100778 | 858.2108701 | 0.3823079 | 7.022496e-01 |
| SegurançaOnline_Não | -184.955851 | 54.3899845 | -3.4005498 | 6.778822e-04 |
| CópiadeSegurançaOnline_Não | -238.788279 | 53.3773281 | -4.4735899 | 7.865879e-06 |
| ProteçãoTM_Não | -166.972400 | 54.2303583 | -3.0789470 | 2.088837e-03 |
| SuporteTécnico_Não | -156.169688 | 54.9123397 | -2.8439817 | 4.473827e-03 |
| ServiçoStreamingTV_Não | -73.388880 | 98.3177409 | -0.7464460 | 4.554339e-01 |
| ServiçoStreamingFilmes_Não | -86.938847 | 98.1858621 | -0.8854518 | 3.759565e-01 |
| TipodeContrato_Anual | -37.041449 | 31.2997506 | -1.1834423 | 2.366913e-01 |
| TipodeContrato_BiAnual | -203.447547 | 37.5175123 | -5.4227355 | 6.151293e-08 |
| `MétododePagamento_Cartão_de_Crédito_(automatico)` | 35.088142 | 29.9920045 | 1.1699165 | 2.420915e-01 |
| MétododePagamento_Cheque_Eletrónico | -48.842939 | 29.1793249 | -1.6738886 | 9.421635e-02 |
| MétododePagamento_Cheque_por_email | 249.278200 | 32.1084556 | 7.7636310 | 9.980090e-15 |

*Figura 3 - Coeficientes utilizados na Regressão Linear Múltipla com p-value diferente de NA*

Formularam-se as seguintes hipóteses:

- H0: O preditor não é significativo para o modelo de regressão linear.
- H1: O preditor é significativo para o modelo de regressão linear.

Para um nível de significância igual a 0.05, os valores de *p-value* que se encontram abaixo desse nível são significativos pois rejeita-se a hipótese nula. Assim, com base nos *p-values* dos preditores, na Figura 3, pode-se afirmar que os atributos Fidelização e TarifaMensal são preditores bastante significativos para o modelo em questão, entre outros.

Para os atributos significativos, definidos anteriormente, obteve-se a seguinte equação (2):

$$T_t = 66.03Fid + 36.19\,T_m \quad (2)$$

Seguidamente, aplicando um modelo de regressão diferente, contruiu-se uma Árvore de Regressão com base nos dados dos clientes. Como resultado, foi construída uma árvore que demonstra como diferentes valores de entrada influenciam o resultado de previsão do nó final.

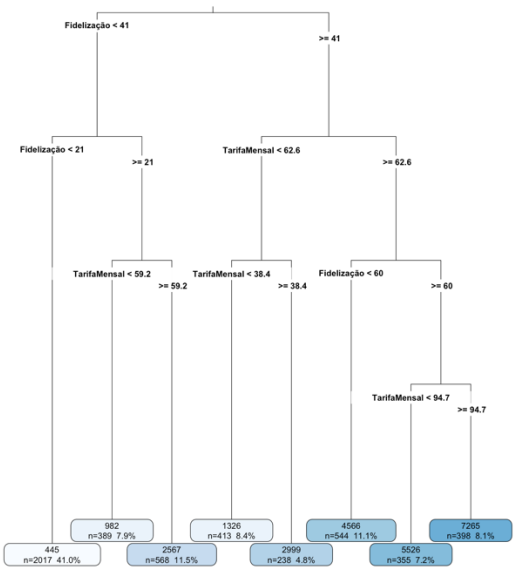


*Figura 4 - Árvore de Regressão para prever o atributo TotalTarifas*

Na construção desta árvore, foram chamados todos os atributos, contudo na Figura 3 apenas se verificam a

"Fidelização" e "TarifaMensal", uma vez que são estes os atributos que influenciam o valor do total de tarifas e com estes se consegue prever o seu valor.

Finalmente, foi aplicado um outro modelo, anteriormente referido, conhecido por Rede Neuronal. Neste modelo, foram definidos diferentes números de nós intermédios como, por exemplo, um nó, três nós e ainda seis seguidos de dois nós intermédios. Os dados de input utilizados para construir a rede neuronal, foram pré-processados através da normalização *min-max* para todos os valores binários. Como exemplo da rede neuronal resultante, na Figura 4, é possível visualizar uma rede com seis e dois nós intermédios.

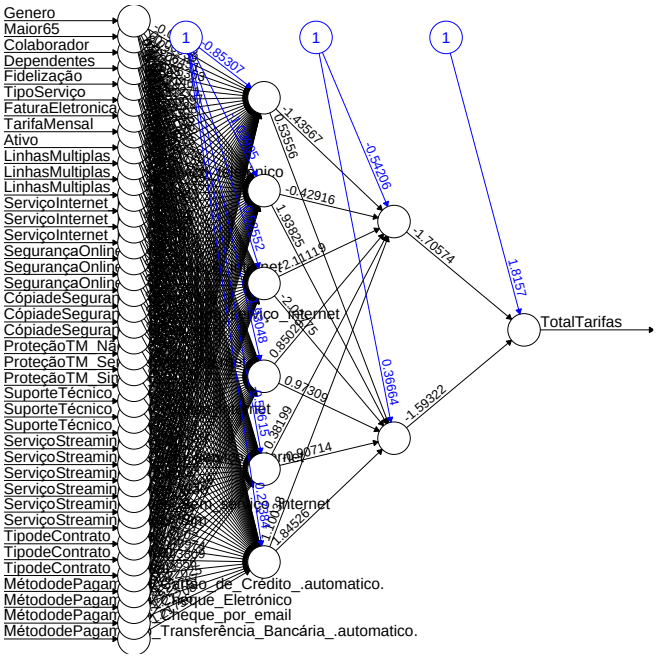


*Figura 5 - Rede Neuronal com seis e dois nós intermédios para prever o atributo TotalTarifas*

De modo a verificar qual o modelo que apresenta o melhor desempenho, foram aplicadas as métricas de avaliação MAE e RMSE aos dados teste (30% dados) em relação aos dados treino (70% dados) usados para construir os modelos, obtendo os resultados representados na Tabela 1.

*Tabela 1 - Comparação entre o MAE e RMSE de diferentes algoritmos*

| | MAE | RMSE |
|---|---|---|
| Regressão Múltipla | 540.89 | 677.49 |
| Árvore de Regressão | 457.77 | 577.37 |
| Rede Neuronal (1) | 223.96 | 273.27 |
| Rede Neuronal (3) | 52.61 | 74.07 |
| Rede Neuronal (6-2) | 51.53 | 74.04 |

Analisando os valores da tabela para cada métrica de performance, podemos verificar que os modelos com o melhor desempenho em termos de precisão na determinação da previsão do resultado para observações não vistas, ou seja, dados de teste, que não foram usados para construir o modelo, são as redes neuronais. Tal se deve ao facto de apresentar o menor valor de erro em relação aos restantes modelos, que quão mais baixo for, melhor é a performance. Adicionalmente, à medida que se adicionam nós intermédios, a performance da rede neuronal também aumenta.

Para verificar se os resultados obtidos para os dois melhores modelos são estatisticamente significativos, foi realizado um *t.test()* com os valores obtidos na Tabela 2 em relação à Rede Neuronal e Árvore de Regressão.

H0: Os resultados obtidos para os dois melhores modelos são estatisticamente significativos.
H1: Os resultados obtidos para os dois melhores modelos não são estatisticamente significativos.

Uma vez que o valor do *p-value* obtido (0.09826) é superior ao nível de significância (0.05), foi comprovado que os resultados são estaticamente significativos pois a hipótese nula não é rejeitada.

## B. *Classificação*

Para o problema de classificação, foram inicialmente selecionados 3 modelos para avaliar a capacidade preditiva do atributo Ativo. Sendo eles: Árvore de decisão, redes neuronais e K-Nearest Neighbor. O primeiro a ser desenvolvido foi o modelo da árvore de decisão. Tendo isso em conta, foi realizado um pré processamento, que consistiu em eliminar o atributo "TotalTarifas" o qual era extremamente correlacionado com o "TarifaMensal", uma vez que não seria necessário prevê-lo. Após isso, os dados, foram aleatoriamente separados em dados de teste e de treino. Subsequentemente, foi obtido como resultado a seguinte árvore de classificação

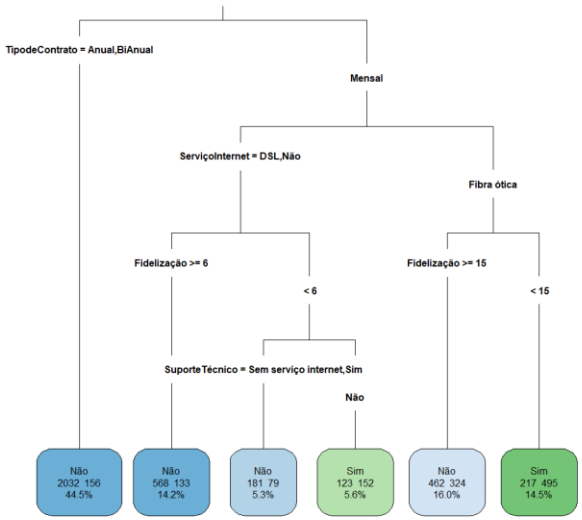


*Figura 6 - Árvore de decisão*

Após o treino, o teste do modelo apresentou 78.15% de taxa de acerto ao ser executado o seu *predict* com os dados de teste.

Com a conclusão do desenvolvimento do modelo supracitado, iniciou-se o desenvolvimento das redes neuronais. Como primeira abordagem, foi necessário recorrer a uma biblioteca externa para substituir os dados em texto para numérico, pois as redes neurais não aceitam *labels*. A primeira rede foi implementada apenas com um nó, e obteve

20.66% de taxa de acerto. A segunda rede foi definida com 4 nós e apenas uma camada, resultado 22.46% de taxa de acerto. Subsequentemente, foi desenvolvida uma rede com 6 nós e 2 camadas, que resultou em 25.55% de taxa de acerto. Por fim, foi desenvolvida uma rede com 10 nós e 3 camadas, resultando 38.82% de taxa de acerto.

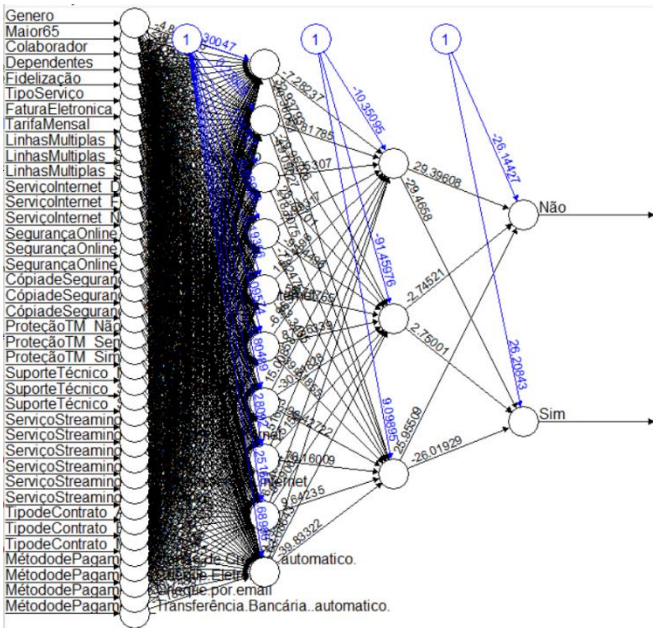


*Figura 7 - Rede neural 10 nós e 3 camadas*

Ao comparar o desempenho das redes, foi possível concluir que conforme a rede apresentava mais nós e camadas, o seu desempenho aumentava, resultando taxas de acertos mais elevadas.

Para concluir a fase de treino dos modelos, foi implementado o algoritmo K-Nearest Neighbor, que assim como as redes neuronais, só aceita valores numéricos, então exigiu também um pré processamento dos dados. Ao finalizar o processamento, foi realizar o treinamento do modelo de 1 até 50 vizinhos com passo igual a 2 e sempre que um modelo era treinado, era testado logo em seguida e seu número de vizinhos e precisão eram guardados juntos em um *array*. Ao finalizar o treinamento, foi observado que o modelo que melhor desempenhou foi o que apresentou 35 vizinhos, registando uma taxa de acerto de 72,70%.

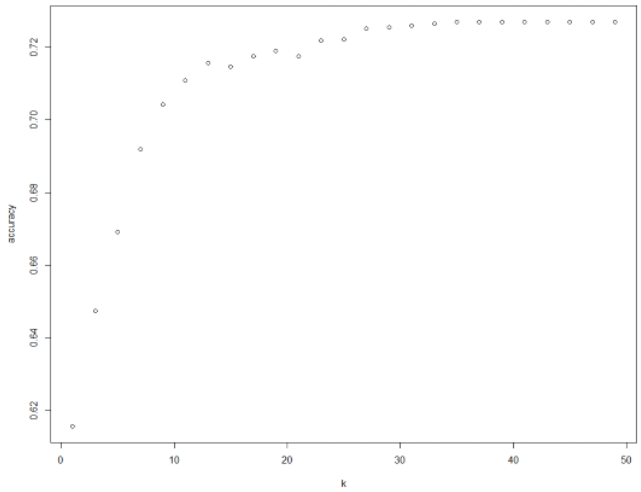


*Figura 8 - KNN*

Ao comparar as taxas de acerto dos modelos, concluiu-se que os que desempenharam melhor dos três modelos citados anteriormente foram a árvore de decisão e o KNN.

Tendo isso em conta, foi realizado o treino e teste dos modelos seguindo o método *k-fold cross validation* para avaliar a taxa de acerto e sua média. Ao finalizar a execução dos testes e treinos, foram obtidos os seguintes resultados.

*Tabela 2 - Valores comparativos KNN e Árvore de decisão*

| KNN | Árvore de decisão |
|---|---|
| 0.7888283 | 0.7901907 |
| 0.7790210 | 0.7972028 |
| 0.8049793 | 0.7980636 |
| 0.8005865 | 0.7815249 |
| 0.7789017 | 0.7890173 |
| 0.7639939 | 0.7685325 |
| 0.7908962 | 0.8022760 |
| 0.7714681 | 0.7686981 |
| 0.8017621 | 0.8105727 |
| 0.7788595 | 0.7760779 |

Com vista a identificar a existência de diferença significativa no desempenho dos modelos, recorreu-se a testes estatísticos de hipótese. Como se trata de duas amostras independentes, as opções possíveis seriam a utilização de um *t-test* ou teste de Wilcoxon. Para utilizar o *t-test*, seria necessário cumprir dois requisitos: as amostras devem ser independentes e devem ser provenientes de uma distribuição normal.

Como o primeiro requisito já havia sido confirmado, para confirmar o segundo foi necessário executar um teste a normalidade. Para isso, foi selecionado o teste de Shapiro para trabalhar sobre as seguintes hipóteses:

H0: São provenientes de uma distribuição normal.

H1: Não apresentam distribuição normal.

O resultado apresentado foi o *p-value* de 0.4 e 0.9 para as amostras. Portanto, como os valores são superiores ao nível de significância, não se rejeita a hipótese nula e assim, pode-se afirmar que as amostras são normalizadas.

Com base nos resultados obtidos, foi executado o t-test para verifica a diferença significativa entre o desempenho com o intuito de refutar umas das seguintes hipóteses:

H0: Não apresenta diferença significativa no seu desempenho.

H1: Apesenta diferença significativa no seu desempenho.

O resultado obtido foi 0.939. Ou seja, não se rejeita a hipótese nula e assim, conclui-se que não há diferença significativa entre o desempenho dos modelos árvore de decisão e KNN.

Finalmente, para avaliar melhor os modelos, foi realizada a execução de 10 instâncias da árvore e do KNN com o intuito de aglomerar os valores médios de precisão, *recall*, *accuracy*, *specificity* e F1, resultando os seguintes valores para cada modelo.

*Tabela 3 - Valores comparativos entre KNN e Árvore de decisão*

| X | Árvore | KNN |
|---|---|---|
| *Accuracy* | 79.07%<br>Desvio = 0.0048 | 72.73%<br>Desvio = 0.0013 |
| *Recall* | 0.922%<br>Desvio = 0.018 | 0.997%<br>Desvio = 0.002 |
| *Specificity* | 0.432 %<br>Desvio = 0.047 | 0.002%<br>Desvio = 0.003 |
| Precisão | 0.817%<br>Desvio = 0.013 | 0.729%<br>Desvio = 0 |
| F1 | 0.035224 | 0.0485101 |

Assim, através dos valores obtidos, pode-se concluir que a Árvore desempenhou melhor sobre o KNN.

## V. Conclusões

Em suma, o estudo descrito neste artigo foca na aplicação de algoritmos de aprendizagem automática com o objetivo de avaliar os dados dos clientes de uma operadora móvel que se encontra com perda de clientes. Esta avaliação foi benéfica na identificação das causas do problema e medidas de prevenção para evitar que tal volte a acontecer.

Para o estudo em foco, a aplicação dos vários e diferentes modelos levou a concluir que o modelo que apresenta o melhor desempenho em termos de previsão de dados, é a árvore de decisão. Porém, na abordagem de um problema de aprendizagem automática, é importante estudar vários modelos, compará-los e concluir qual deles é o melhor a ser utilizado.

No caso da rede reuronal, apesar desta ter dominado no desempenho de previsão, no modelo de regressão, esta consome um grande número de recursos máquina durante o treino dos dados pelo que a sua performance não é considerada a melhor. Recomenda-se o uso de redes neuronais em problemas que possuam um extenso e complexo número de dados pois esta retorna melhores resultados com o tempo.

A regressão linear é um modelo específico que apenas pode ser aplicado em problemas de regressão e não de classificação pois retorna um output contínuo em relação aos dados de entrada que possuem uma relação linear entre si. Por isto, no geral, o seu uso na previsão de dados com vários atributos, não é o mais aconselhável.

Em relação aos resultados obtidos pelo modelo kNN, este teve um desempenho superior à rede neuronal, quando apresentava o valor de 35 vizinhos mais próximos e seu processamento foi extremamente inferior, resultando uma opção considerável em função da rede. No entanto, as métricas comparativas entre o kNN e a árvore de classificação apontaram a árvore como o melhor desempenho, mesmo não apresentando diferenças significativas.

## Referências


[1] S. Raschka, L01: Intro to Machine Learning, 2018.

[2] adminvooo, "Um tutorial completo sobre modelagem baseada em árvores de decisão (códigos R e Python)," Vooo, 8 maio 2019. [Online]. Available: https://www.vooo.pro/insights/um-tutorial-completo-sobre-a-modelagem-baseada-em-tree-arvore-do-zero-em-r-python/. [Accessed junho 2022].

[3] T. Mitchell, Machine Learning, McGraw-Hill, 1997.

[4] [Online]. Available: https://moodle.isep.ipp.pt/pluginfile.php/194086/mod_resource/content/3/Cross-validation.pdf. [Accessed Junho 2022].