Machine Learning

Análise de desempenho das técnicas de Aprendizagem do Machine Learning

João Francisco Junqueira Flores
Departamento de Engenharia Informática
Instituto Superior de Engenharia do Porto
Porto, Portugal
1171409@isep.ipp.pt

Patrick Semedo Timas
Departamento de Engenharia Informática
Instituto Superior de Engenharia do Porto
Porto, Portugal
1171352@isep.ipp.pt

José António Moreira da Mota
Departamento de Engenharia Informática
Instituto Superior de Engenharia do Porto
Porto, Portugal
1161263@isep.ipp.pt

***Resumo* — Atualmente a tecnologia vem-se a desenvolver cada vez mais rapidamente. Este desenvolvimento contribui para alterações em diversas áreas, como, por exemplo, a investigação, ou tomada de decisão. Para tal recorre-se à área da Inteligência Artificial, mais conhecida pela sua sigla, AI.**

Este artigo tem como objetivo analisar quatro algoritmos associados ao Machine Learning (aplicação do AI) e comparar e discutir os seus resultados.

Palavras-chave—Machine Learining,Rrede Neurais, Árvores de decisão,Rregressão Linear, k-vizinhos-mais-próximos

I. Introdução

O presente relatório enquadra-se na unidade curricular de Analise de Dados em Informática, do Curso de Licenciatura em Engenharia Informática do Instituto Superior de Engenharia do Porto (ISEP). Com a evolução constante da ciência e da tecnologia, a maioria das tarefas começam a ser automatizadas, podendo substituir os humanos. Surge, assim, o conceito Inteligência Artificial (AI), que será abordado no capítulo seguinte. Este relatório está dividido em 3 partes: introdução e contextualização do tema através do estado da arte, analise e discussão do desempenho das técnicas de aprendizagem derivadas do AI e, por fim, uma seção dedicada a conclusões e apreciações do estudo realizado.

II. Estado da arte

Como sabemos, estamos num mundo em constante mudança e, com os avanços científicos e tecnológicos, muitos processos/tarefas começam a ser automatizados. A inteligência Artificial irá substituir os humanos em muitas tarefas.

A Inteligência Artificial [1] surge derivada à ciência da computação com foco em resolver tarefas nas quais os seres humanos são bons, através do uso de robots a imitar a inteligência humana. Surge então um conceito denominado Machine Learning. Machine Learing [2] é uma aplicação de inteligência artificial que se dedica ao estudo científico de algoritmos e modelos estatísticos que os sistemas computacionais usam para que uma determinada tarefas em específico seja realizada a partir de padrões e inferência. Utiliza observações para criar padrões nesses mesmos dados e tomar as melhores decisões no futuro.

O Machine Learning possui várias técnicas/ algoritmos que podem ser aplicados, sendo eles:

- Redes Neurais
- Árvores de decisão
- Árvores de regressão/Regressão Linear
- K-vizinhos mais próximos

A. Redes neurais

As redes neurais artificiais são modelos inspirados na estrutura das redes neurais biológicas. O objetivo inicial desta técnica de aprendizagem era resolver problemas da mesma forma que um cérebro humano resolveria. Ao longo tempo, o seu objetivo mudou, sendo que passou a ser a execução de tarefas específicas, com desvios da biologia. As redes neurais foram usadas em variadas tarefas, tais como: tradução automática, filtragem de redes sociais, diagnósticos médicos, visão computacional e em atividades que são consideradas apenas fazíveis pelo

ser humano (casos como a pintura). De uma forma simples, o funcionamento de uma rede neural pode ser compreendido através da figura seguinte:

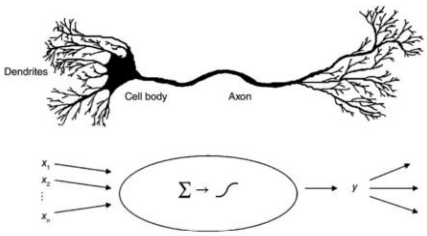

Figura 1-Analogia entre Rede Neural e Rede Neural Artificial [3]

Através da figura 1 conseguimos perceber o fluxo numa rede neural artificial uma vez que se assemelha a uma rede neural do sistema nervoso humano. Existe uma primeira camada (camada de entrada), responsável por receber os sinais de informação, existe uma camada que unifica a informação, produzindo um novo sinal (valor resultado da união de toda a informação recebida) e por fim uma camada de saída, sendo o local por onde é transmitido o novo sinal.
A rede neural mais Conhecida e mais utilizada é a rede FeedFoward [3]. Esta rede foi a primeira e a mais simples a ser criada. O sinal de informação move-se apenas num sentido, começando nos nós de entrada, passando pelos nós ocultos/ intermédios (caso existam) e, por fim, para os nós de saída. Deste modo não existem ciclos na rede.

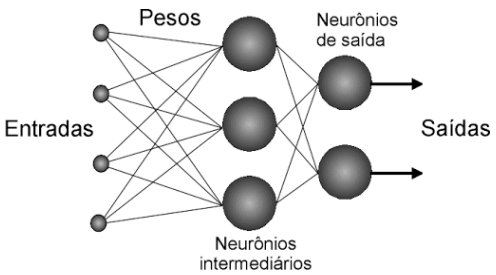

Figura 2- Rede Neural do Tipo FeedForward [4]

As redes FeedFoward podem ser compostas por uma única camada ou então multicamada. Como seria expectável, as redes multicamadas não mais poderosas no sentido em que têm mais aplicações como por exemplo reconhecimento de padrões e criação de modelos matemáticos por analise de regressão.

Existem 3 métodos de aprendizagem utilizados pelas redes neurais [4]. O primeiro é o **Supervised Learning**. Este método é caracterizado pela presença de um "professor" externo, que tem a função de fornecer a resposta desejada durante o processo de aprendizado. A diferença entre a resposta desejada e a resposta observada na saída é denominada sinal de erro, e de acordo com esse erro, os parâmetros da rede são ajustados. Depois, temos o método **Unsuvervised Learning** onde uma rede neural aprende através de um processo iterativo de ajuste de seus pesos sinápticos. O processo de aprendizagem segue a seguinte sequência: estímulo da rede neural pelo ambiente de informação; estrutura interna da rede é alterada como resultado do estímulo; devido a estas alterações na estrutura interna, a rede tem modificada a sua resposta aos estímulos do ambiente. Por fim, temos os **Reinforcement Learning,** método este que se preocupa com a maneira de como os agentes de software devem executar ações num ambiente para maximizar a noção de recompensa cumulativa. Difere do Supervised Learning por não precisar apresentar pares de entrada/saída rotulados e por não precisar que ações subótimas sejam explicitamente corrigidas. Em vez disso, o foco é encontrar um equilíbrio entre a exploração (de território desconhecido) e a exploração (do conhecimento atual).

De todos os algoritmos existentes de aprendizagem, os mais utilizados são: Binary Perceptron e o Backpropagation [4].
O **Binary Perceptron** é um algoritmo de classificadores binários. Faz parte da aprendizagem supervisionada. Só é capaz de solucionar problemas que se são linearmente separáveis caso aplicado em redes de camada simples. Caso seja uma rede multicamadas, já é possível utilizar este algoritmo para resolução de problemas não lineares.
Já o **Backpropagation**, é um algoritmo utilizado no treinamento de redes neurais multicamadas, e consiste em dois passos de computação: o processamento direto e o processamento reverso.
No processamento direto, uma entrada é aplicada à rede neural e seu efeito é propagado pela rede, camada a camada. Durante o processamento direto, os pesos da rede permanecem fixos.
No processamento reverso, um sinal de erro calculado na saída da rede é propagado no sentido reverso, camada a camada, e ao final deste processo os pesos são ajustados de acordo com uma regra de correção de erro.
O algoritmo de Backpropagation segue os seguintes passos: Inicialização onde inicializa os pesos da rede aleatoriamente ou segundo algum método. Em segundo, ocorre o processamento direto, onde apresenta-se um padrão à rede, é computado as ativações de todos os neurónios da rede e é calculado o erro. Depois temos o Passo reverso, onde se calcula os novos pesos para cada neurônio da rede, no sentido retroativo (isto é, da saída para a entrada), camada a camada. Por fim, Teste de parada. Teste o critério de parada adotado. Se satisfeito, termine o algoritmo; Caso contrário volte ao passo 2.

As redes neurais apresentam vantagens, porém também têm as suas desvantagens.

Como vantagens, as redes neurais permitem a aquisição automática de conhecimentos empíricos a partir de uma base de exemplos de aprendizado referente a um problema; a manipulação de dados quantitativos, aproximados e mesmo incorretos com uma degradação gradual das respostas; grande poder de representação de conhecimentos através da criação de relações ponderadas entre as entradas do sistema.
Possui também desvantagens, sendo elas: a dificuldade de configuração das redes em relação à sua estrutura inicial e também no que se refere aos parâmetros dos algoritmos de aprendizado; dificuldade de explicitar os conhecimentos adquiridos pela rede através de uma linguagem compreensível para um ser humano; dificuldade de convergência (bloqueios) e instabilidade, inerentes aos algoritmos de otimização empregados e lentidão do processo de aprendizados/ adaptação.

B. Arvore de decisão

Uma árvore de decisão [5] é uma ferramenta de suporte à tomada de decisão que usa um gráfico no formato de árvore e demonstra visualmente as condições e as probabilidades para se chegar a resultados.
Uma árvore de decisão é uma forma de visualizar as regras de negócio que levam a determinados grupos de indivíduos, construídos com base em uma variável.
Em alguns casos de algoritmos de machine learning, como redes neurais, por exemplo, os modelos podem acabar se tornando tão complexos que se tornam difíceis de serem explicados com facilidade. Já nos modelos de árvores de decisão os resultados são fáceis de explicar e de serem convertidos em decisões práticas de negócio no dia a dia das empresas.

1) Estrutura

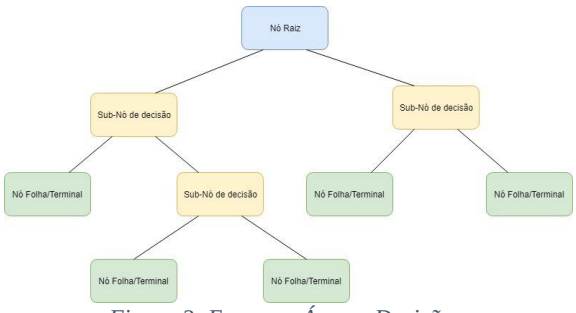

Figura 3- Esquema Árvore Decisão

- **Nó Raiz**: O nó que executa a primeira divisão. Esta divisão pode dividir o nó em dois ou mais sub-nós.
- Quando um **sub-nó** se divide em outros **sub-nós**, denomina-se como **nó de decisão**.
- Quando se remove **sub-nós** de um nó de decisão, esse processo é denominado de **Poda**. O oposto da poda é a **divisão**.
- **Nós / Folhas do Terminal**: nós que preveem o resultado (não se dividem).
- Para todo nó dividido em **sub-nós**, é chamado de **nó pai** de seus **sub-nós**, consequentemente seus **sub-nos** são chamados de **filhos**.
- **Ramos**: setas conectando nós, mostrando o fluxo da pergunta para a resposta.

2) Tipos

a) Árvores de regressão

O algoritmo para modelos de árvore de decisão funciona particionando repetidamente os dados em vários subespaços, de forma que os resultados em cada subespaço final sejam o mais homogêneos possível. Essa abordagem é tecnicamente chamada de particionamento recursivo. O resultado obtido consiste em um conjunto de regras usadas para prever a variável de resultado, continua para árvores de regressão e categórica para árvores de classificação.

b) Árvores de classificação

Uma árvore de classificação é muito semelhante a uma árvore de regressão, exceto que é usada para prever uma resposta qualitativa em vez de quantitativa.
Para uma árvore de classificação, prevê-se que cada observação pertence à classe de observações de treino mais comum na zona à qual pertence. Portanto quando se interpreta os resultados de uma árvore de classificação, muitas vezes estará se interessado não só na previsão de classe correspondente a região de nó terminal, mas também nas proporções de classe entre as observações de treino que pertencem aquela zona.

c) Entropia

Quando o algoritmo toma a decisão de dividir uma árvore, haverá uma definição de pureza em relação ao seu target. A entropia é definida como uma forma de mensurar a pureza de cada subconjunto de uma determinada árvore decisão. Basicamente mede a probabilidade de obter uma ocorrência de evento positivo (0 a 1), a partir de uma seleção aleatória do subconjunto de dados.
A entropia está sempre relacionada ao ganho informacional, que é baseado na redução da entropia. Quando se constrói uma árvore de decisão, quanto mais homogêneo for os ramos da árvore, menor a entropia, que resulta em maior ganho de informação (chance de acontecer algo novo/diferente é muito baixa).

d) Overfitting

Existem casos onde uma árvore de decisão pode ficar com uma quantidade de arestas muito grande, aumentando a sua complexidade. Isso pode gerar um problema conhecido como overfitting. Para resolver este tipo de situação existem métodos de "poda" das árvores de decisão.

3) Vantagens das árvores de decisão [6]

Interpretação.

- Percebe- se a razão da decisão.

Facilidade em lidar com diversos tipos de informação.

- Real, nominal, ordinal, etc.
- Não é necessário definir "importância relativa".

Insensível a fatores de escala.

Escolha automática dos atributos mais relevantes em cada caso.

- Atributos mais relevantes aparecem mais acima na árvore.

Adaptável também a problemas de regressão.

- Modelos locais lineares como folhas.

4) Desvantagens de árvores

Fronteiras lineares e perpendiculares aos eixos.

Sensibilidade a pequenas perturbações no conjunto de treino (geram redes muito diferentes).

C. K-vizinhos-mais-próximos

Método de avaliação de dados não paramétricos baseado em distâncias, usado para classificação e regressão, o mais comum é com a forma euclidiana, para trabalhar sobre este método é necessário avaliar sobre dados numéricos, no caso de ser preciso avaliar dados nominais convertemos para valores numéricos.

O forma como obtêm valores é realizando uma métrica de dois valores, requer dados de input para treino, input para teste do treino. [7]

1) Vantagens

Os dados de treino são robustos e têm muito valor.

O método é bastante eficaz quando os dados de treino são volumosos.

2) Desvantagens

É necessário determinar o valor do parâmetro k, se não for avaliado este pode invalidar a analise.

Não está determinado qual o tipo de distância e quais os atributos a usar, se não forem bem escolhidos a analise poderá ser pobre.

O custo computacional é bastante elevado.

3) Taxa de erro

Com a necessidade de o input de dados ser elevado, vários dados podem ser errados, de tal modo existem vários algoritmos para relacional e avaliar o erro. Para além de erros nos dados também pode haver redundâncias, por exemplo o peso representado em kg ou em libras, para evitar estas imparidades é realizada a extração para o vetor de recursos.

D. Regressão

[8]A regressão linear é a correlação entre duas variáveis, conforme um dado é procurada o valor que se ajusta melhor à linha projetada. A regressão linear tem tendência a ter melhores resultados quando é notório uma evolução constante.

Para determinar se existe correlação entre as variáveis utiliza-se o coeficiente de Pearson, este mede o nível da correlação e a direção. Os valores da correlação variam entre -1 e 1, sendo que quanto mais próximo de zero menor os graus de correlação.

1) Vantagens

- Regressão é bastante percetível e explicável.
- Pode ser regularizada para evitar overfitting.

2) Desvantagens

- Quando a relação não é linear o rendimento não é agradável.
- Não é flexível para avaliar padrões mais complexos.

E. Cross Validation

A validação cruzada é uma técnica que envolve a reserva de uma amostra especifica de um conjunto de dados, com essa amostra os dados são treinados e de seguida testados.

A validação cruzada através do k-folds deve ser realizada com um grande conjunto de dados. A implementação do método é realizada com uma divisão aleatória de folds, realizadas previsões através dos testes, registado o erro, isto é repetido para as diferentes folds, utilizando todos os dados.

1) Vantagens

- Utiliza todos os dados.
- Diferentes métricas
- Funciona com informação dependente

2) Desvantagens

- Não é dos melhores métodos a usar quando os dados são sequencias (ex: datas)

III. Análise de desempenho de técnicas de aprendizagem

Foi realizado um estudo a 3000 trabalhadores do sexto masculino. Foram recolhidos vários dados como idade, estado, raça grau de escolaridade, emprego, salário, entre outros. O principal objetivo

deste estudo era perceber de que forma os vários dados/características influenciavam o salário dos trabalhadores. Para este estudo, foi utilizado os algoritmos explorados no capítulo anterior.

A. Regressão linear entre salário e idade

Para verificar o valor da variável dependente salário e a variável independente foi utilizada a regressão linear.
Servindo-se do "data-frame" criado, é criada uma variável que relaciona o salário com a idade.
Através da variável criada é um diagrama com o salário relacionado com a idade, também é ilustrado uma reta que corresponde à reta de regressão linear.

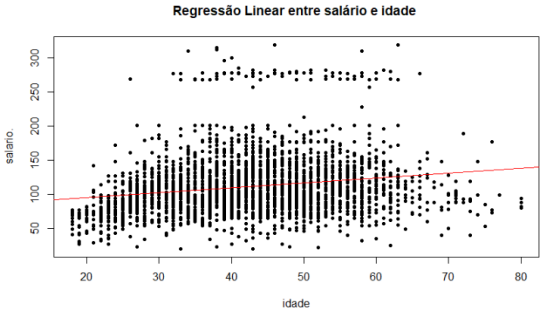

Figura 4 - Diagrama de regressão linear

A reta é do tipo "y=mx+b", através do x é possível obter uma estimativa do y, este corresponde ao salário. Os coeficientes são dois uma para o m: 81.7716 e outro para o b: 0.7069.

Foi necessário verificara vários resíduos da regressão, a normalidade, a homocedasticidade e independência. Quando é referido o nível de significância é utilizado o valor de 5%.
Para a normalidade foi realizado o teste de shapiro, com a hipótese nula verificar a existência de normalidade e a alternativa ser impossível descartar a hipótese. O valor obtido de p-value foi 2.2e-16, bastante inferior ao de significância, por isso os dados não apresentam uma distribuição normal.
Para analise da homocedasticidade criou-se a hipótese nula de se verificar a homocedasticidade. Para se verificar foi calculada a mediana da idade, sendo esta 42, com a mediana é efetuado o teste dividindo em dois conjuntos, os com idade inferior a mediana e ou que têm superior. Para verificar é obtido o p-value de 1.866e-09, sendo este valor novamente superior ao de significância, não se rejeita a hipótese nula e anulando a hipótese de homocedasticidade.
O último resíduo a testar foi a independência, através do teste de DurbinWatson obteve-se novamente o p-value, neste caso foi de 0.208, superior ao nível de significância, não podendo assim refutar a hipótese nula de os dados serem independentes.

B. Regressão linear entre variável dependente "Salario" e as variáveis independentes "Idade", "Grau" e o "Emprego".

Para estimar os valores solicitados na aliena a) e b) usou-se o comando *predict* com o modelo de regressão linear múltipla onde a variável dependente é o salário e as variáveis independentes são *idade*, *grau* e *emprego.* Cria-se um *data frame* com as variáveis independentes com valores requisitados para cada alínea.
Ao fim verificou-se que quanto para a aliena a) quanto para a b) os valores são crescentes de acordo com o grau ou com o emprego.

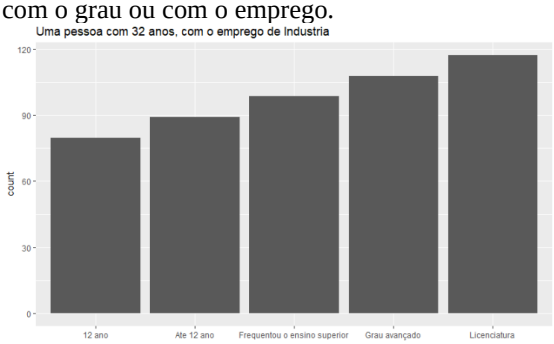

Figura 5 - Diagrama da relação entre o grau e a media de salários

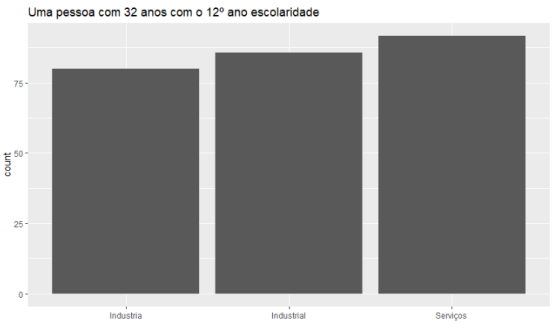

Figura 6 - Diagrama da relação entre o emprego e a media de salários

C. Derival um novo atribuito Nível, atráves da mediana do salário, criando dois niveis o alto e o baixo

Através do salário é calculada a mediana da coluna, realizando um ciclo para verificar se cada valor é superior ou inferior à mediana, apôs obter a coluna esta é adicionada à "data frame".

D. Análise exploratória de dados

No que diz respeito as medidas descritivas de localização, dispersão e forma:
Ano: mínimo=2003, máximo=2009, média=2005.8, mediana=2006, moda=2003, quartil=2004 e 2008, distancia entre quartis=4, kurtosis=1.73, skewness=0.14;
Idade: mínimo=18, máximo=80, média=42.4, mediana=42, moda=40, quartil=33.75 e 51,

distancia entre quartis=17.25, kurtosis=2.55, skewness=0.14;
Estado: a maior parte das pessoas são casadas.
Raça: a maior parte das pessoas são brancas.
Grau: a maioria tem o 12º ano.
Emprego: maioritariamente das pessoas trabalham na indústria.
Saúde: mais do que a metade das pessoas tem uma saúde classificada como muito bom.
Seguro: mais de 2000 trabalhadores têm seguro.
Salário: mínimo=20, máximo=99743383, média=11260567.2, mediana=3836, moda=118, quartil=104 e 8339856, distancia entre quartis=8339752, kurtosis=8.47, skewness=2.55;

E. *Estude a capacidade preditiva de alguns métodos de previsão relativamente ao atributo nível*

Para as analises é criada uma amostra de treino com 70% dos dados e uma amostra de teste com os dados restantes. Figura 7.
Para a arvore os valores obtidos de *accuracy* os valores obtidos foram de 0.98, para o *recall* 0.99, a *precision* 0.99 e o f1 0.99. Figura 8.

Para a rede neuronal os valores obtidos foram: para o *accuracy* 0.96, o *recall* 0.98, a *precision* 0.97 e o f1 0.96.

O desempenho dos diversos algoritmos foi positivo, tendo em consideração um nível de significância de 5%, foram realizados diversos testes. É testada a normalidade dos diversos valores apresentados acima, todos eles para o teste de shapiro apresentam um p-value < 0.05 logo não rejeita a hipótese nula de normalidade, de seguida realizam-se os teste de lillie para todos valores são obtidos p-values > 0.05 podendo assim realizar o t.test, para todos os valores são obtidos valores menores que os de significância

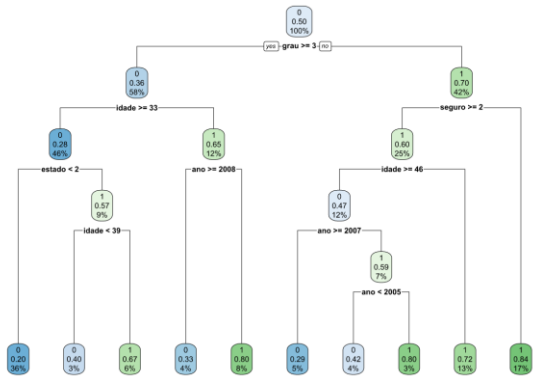

Figura 7 - Árvore de decisão

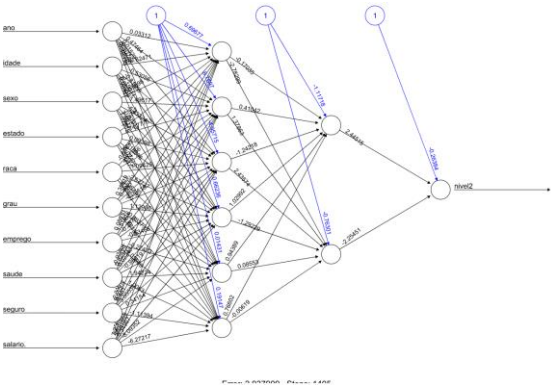

Figura 8 - Rede neuronal

F. *Estudo da capacidade de previsão do atributo salario*

Para este estudo é novamente criado uma matriz de treino e outra de teste, a primeira com 70% e a segunda com 30%.

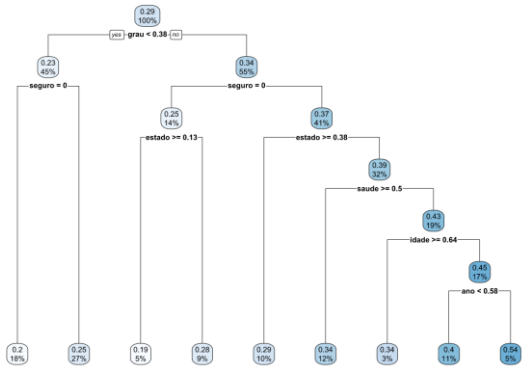

Figura 9 - Árvore de regressão

Com apoio de bibliotecas próprias e com o cross validation obteve se uma matriz de error, onde esse valor depende do fold que está a avaliar.

Foram realizadas médias para ambas os métodos, a média do k vizinhos é 34.43 com desvio padrão de 1.61, a média da árvore foi de 37.69 com desvio padrão de 0.86.

Novamente foram realizados teste para verificar a normalidade e a diferença entre os dados tal como os referidos no ponto E.

IV. CONCLUSÃO

A realização do trabalho permitiu fundamentar os conhecimentos em algoritmos e métodos de avaliação de dados.
É de notar que a normalização de alteração dos dados por vezes pode ser necessária de modo a obter dados mais precisos.
A regressão provou que com o aumento da idade é notável um aumento progressivo do salário. No que diz respeito às previsões salariais, utilizando uma

regressão linear generalizada, tendo como variáveis independentes a idade, o grau de escolaridade e o tipo de emprego, conclui-se que:

Para um homem com 32 anos e emprego indústria, o salário tende a aumentar com o aumento do nível de escolaridade. Para um homem com 32 anos e 12º ano de escolaridade, o salário é ligeiramente superior nos serviços do que na indústria.

Relativamente às avaliações realizadas, é notório os valores obtidos. A previsão tem um valor quase perfeito, tal pode ser por overfitting ou por uma adaptação dos algoritmos com uma granularidade alta aumentando a sua complexidade, que funciona bem para os nossos dados, mas caos algo seja alterado poderá surgir problemas. É de salientar também a falta de provas dos resultados através dos testes, pois estes não obtêm valores de significância para comprovar diferenças significativas entre os quatros algoritmos analisados.

De acordo com os valores obtidos podemos concluir que os diversos algoritmos têm performances altas, as performances de previsão estão dependentes das amostras e dos métodos utilizados.

V. Referências

[1] iberdrola, "O QUE É A INTELIGÊNCIA ARTIFICIAL?," [Online]. Available: https://www.iberdrola.com/inovacao/o-que-e-inteligencia-artificial. [Accessed 8 Junho 2020].

[2] Wikipedia, "Machine learning," [Online]. Available: https://en.wikipedia.org/wiki/Machine_learning. [Accessed 8 Junho 2020].

[3] R. Fonseca , F. Pereira and E. Didoné, "Modelos de predição da redução," 2012. [Online]. Available: https://www.scielo.br/pdf/ac/v12n1/v12n1a11. [Accessed 10 Junho 2020].

[4] A. M. Madureira, "NEURAL NETWORKS," 27 Maio 2020. [Online]. Available: https://moodle.isep.ipp.pt/pluginfile.php/330787/mod_resource/content/1/NeuralNetworks%20slides.pdf. [Accessed 11 Junho 2020].

[5] G. Stankevix, "Arvore de Decisão em R," 15 Outubro 2019. [Online]. Available: https://medium.com/@gabriel.stankevix/arvore-de-decis%C3%A3o-em-r-85a449b296b2. [Accessed 1 Junho 2020].

[6] V. Lobo, "Sistemas de Apoio à Decisão– Árvores de decisão," 2005. [Online]. Available: https://www.novaims.unl.pt/docentes/vlobo/isegi_SAD/SAD_5_arvores_6.pdf. [Accessed 1 Junho 2020].

[7] O. Harrison, "Machine Learning Basics with the K-Nearest Neighbors Algorithm," 10 Setembro 2018. [Online]. Available: https://towardsdatascience.com/machine-learning-basics-with-the-k-nearest-neighbors-algorithm-6a6e71d01761. [Accessed 8 Junho 2020].

[8] Stat, "Linear Regression," [Online]. Available: http://www.stat.yale.edu/Courses/1997-98/101/linreg.htm. [Accessed 6 Junho 2020].