1 - Identificação de uma base de dados de interesse:

Em primeiro lugar, é notório que a educação é essencial para formação do cidadão e transformação da sociedade. Ela é responsável é a responsável pela multiplicação do conhecimento e pelo desenvolvimento de habilidades úteis para a atuação do indivíduo em sua comunidade. Dessa forma, levando em consideração a importância desse tema, foi escolhida um banco de dados relacionada a esse tema.

A base de dados escolhida aborda o desempenho dos alunos no ensino secundário de duas escolas portuguesas. Os atributos de dados incluem notas dos alunos, características demográficas, sociais e relacionadas à escola e foram coletados por meio de relatórios e questionários escolares. São fornecidos dois conjuntos de dados relativos ao desempenho em duas disciplinas distintas: Matemática (mat) e Língua Portuguesa (por), entretanto, visando uma possível simplificação do estudo, apenas a primeira foi considerada para o treinamento, sendo que a segunda foi usada novo para testes dos modelos de classificação em uma base de dados nova, tudo isso com auxilio do software Orange Mining.

**(TABELA 1 – DICIONARIO DE DADOS)**

**(TABELA 2 – AMOSTRA DE DADOS)**

2 - Preparação/pré-processamento dos dados:

É importante lembrar que existe uma fase intermediaria entre a identificação dos dados e a classificação do mesmo, denominada de preparação ou pré-processamento dos dados. Essa etapa é crucial para obter resultados de análise de melhor qualidade e veracidade. Para uma melhor analise nesse caso, foram realizadas etapas para o processamento dos dados, por meio do Orange Mining. Antes de tudo, foi realizado a limpeza dos dados, removendo registros com atributos nulos. Além disso, visando o aumento da eficiência e diminuição dos custos foi feito a redução em 50% dos registros de toda base de dados, sendo que o restante foi utilizado apenas para testes de qualidade e confiança dos modelos criados. Por fim, foi efetuada uma seleção das colunas e do objetivo que seria utilizado para classificação, assim foi escolhido as seguintes colunas, por serem consideradas as mais relacionadas com o alvo: Studytime, Famrel, Romantic, Fedu, Medu, Traveltime, Freetime e Absences para o objeto de classificar se o aluno foi ou não aprovado, através da coluna “Aprovado”.

3 - Tarefa de Aprendizado de Máquina: Classificação

Para esse estudo, pode-se dizer que foram criados dois tipos de modelos de classificação de dados, que tem a finalidade de determinar se um aluno foi aprovado (nota final maior ou igual a 12) ou reprovado na disciplina de matemática. Sendo levado em consideração o tempo de estudo, a qualidade familiar, o nível de educação dos pais, o tempo livre e de viagem, envolvimento em um relacionamento, e as faltas na disciplina.

Como dito anteriormente, para tal tarefa, foi utilizado o Orange, que é um kit de ferramentas de visualização de dados de código aberto, aprendizado de máquina e mineração de dados. Dessa forma, após conclusão foi disponibilizado o fluxograma da ferramenta.

**FOTO DO FLUXOGRAMA**

3.1 – Árvore de Decisão

3.1.1 DEFINICAO

A Árvore de Decisão é uma técnica de aprendizado supervisionado que pode ser usada tanto para problemas de classificação quanto de regressão, mas é preferida principalmente para resolver problemas de classificação. É um classificador estruturado em árvore, onde os nós internos representam as características de um conjunto de dados, os ramos representam as regras de decisão e cada nó folha representa o resultado.

3.1.2 MODELO CRIADO

O modelo criado para esse trabalho levou em consideração 50% da base de dados para treinamento, ou seja, 199 registros e o restante para validação e testes. Além disso, é notório revelar que todas as opões de parâmetros foram deixadas do mesmo jeito que a plataforma recomenda, ou seja, houve uma indução para criação de uma arvore binaria, tive um mínimo de duas instâncias nas folhas, conjuntos menores que cinco não foram divididos, e por fim a profundidade máxima da árvore foi definida para cem.

3.1.3 PRECISAO DO MODELO

É significante analisar e compreender as métricas do modela, para realiza-se uma avaliação do mesmo. No caso, dessa arvore de decisão nos dados treinados previamente foi obtido uma acurácia de 87,9% , uma precisão de 91,2% e uma revocação de 77,5%. Entretanto, nos dados considerados como novos, ou seja , de uma disciplina diferente, em uma tentativa de prever os aprovados, esse mesmo modelo teve uma acurácia de 58,3%, uma precisão de 61,2% e uma revocação de 58,3%.

3.2 – Rede Neural

3.2.1 DEFINICAO

Uma rede neural é uma série de algoritmos que se esforça para reconhecer relacionamentos subjacentes em um conjunto de dados por meio de um processo que imita a maneira como o cérebro humano opera. Nesse sentido, as redes neurais referem-se a sistemas de neurônios, sejam de natureza orgânica ou artificial.

3.2.2 MODELO CRIADO

O modelo criado para esse trabalho levou em consideração 50% da base de dados para treinamento, ou seja, 199 registros e o restante para validação e testes. Além disso, é notório revelar que todas as opões de parâmetros foram alteradas, ou seja, houve a utilização de 1500 neurônios nas camadas da rede, foi utilizada a ativação denominada de ReLu e o solucionador Adam, a regularização foi definida como zero, e finalmente, o número de passos máximos do modelo foi ajustada para 2000.

3.3.3 PRECISAO DO MODELO

Neste caso, foi obtida métricas que indicam uma maior eficiência do que o outro modelo. Em um conjunto de dados que foi utilizado para o treinamento, foi alcançado uma acurácia de 99,5% , uma precisão de 91,2% e uma revocação de 100%. Porém, é notório que existe uma grande diferença quando analisado com dados novos, assim conquistado uma acurácia de 58%, uma precisão de 60,2% e uma revocação de 58%.

**(FOTO TABELA DO ORANGE MINING – PREDICTIONS)**

**(FOTO TABELA DO ORANGE MINING – TEST AND SCORE)**

4- Análise crítica e conclusão

Em primeiro lugar, é importante ressaltar a diferença das métricas analisadas entre os modelos citados anteriormente, quando analise os dados da disciplina de matemática que foram usados no treinamento. Em relação a acurácia da classificação, precisão e revocação possui uma distinção de 11,6%, 7,6% e 22,5%, respectivamente, sendo que a rede neural teve melhores resultados em todas métricas. Todavia, quando essa análise é feita em uma base de dados nova, no caso da disciplina de português, essa diferença modifica-se a 0,03%, 1,1%,0,03%, respectivamente, sendo que a arvore de decisão teve melhores resultados em todas métrica, diferentemente do outro cenário. Portanto, nota-se que o modelo de rede neural foi muito mais eficiente quando analisado com uma base de dados previamente treinada, e um pouco menos eficiente com dados novos, porém devido a grande diferença no primeiro cenário, conclui-se que este foi o melhor modelo de classificação considerando o contexto e o desempenho.