Trabalho – Avaliação A1 (nota máxima 2,5) (primeira semana outubro).

Fazer um gráfico para verificar os melhores pontos de corte, este gráfico deve conter as seguintes especificações:

1. Equipes de até 4 pessoas.
2. O gráfico deve generalizar para qualquer modelo de classificação binária, independentemente da base de dados utilizada.
3. O gráfico será baseado em histogramas, baseado na saída de probabilidade do modelo de classificação binária para a classe positiva (ex: é um spam? é um sentimento positivo? Etc..).
4. No eixo X do gráfico, deve ser utilizado a escala da saída probabilidades do modelo (de 0% a 100% ). O eixo Y deve ser utilizado para quantidade percentual populacional.
5. Deve ser estabelecida uma escala baseada em steps(passos) para as probabilidades (ex: de 10% em 10% [0%, 10%, 20%, 30% ... 100%]).
6. Usar a legenda azul para plotar os histogramas da população em geral.
7. Usar a legenda vermelha para plotar os histogramas da população onde o target é positivo.
8. Estabelecer 3 pontos de corte para o máximo aproveitamento do modelo: Um para a classe negativa, um para a classe positiva e um para mesa ou análise manual. Justificar a escolha dos pontos de corte através do acúmulo populacional e erro do modelo.

Ex:

Lousa branca com texto preto sobre fundo branco

Descrição gerada automaticamente com confiança média

Prova – Avaliação A1 ( nota máxima 2,5) (segunda semana outubro):

Será aplicada após a conclusão da avaliação do trabalho do gráfico.

Análise e construção de um modelo de Machine Learning (classificação binária), utilizando aprendizagem de máquina supervisionada, com avaliação de desempenho, a prova terá as seguintes especificações:

1. Equipes de até 4 pessoas, cada equipe receberá uma base de dados diferente. A base será um corpus de classificação binária.
2. Será fornecido um arquivo TXT que poderá ser transformado em CSV para ser carregado no google Colab.
3. Verificar o balanceamento percentual da base de dados.
4. Utilizar a técnica de extração de características mais adequada.
5. Executar pelo menos 2 modelos de machine learning e comparar a performance entre eles. Se possível, otimizar os hiper parâmetros.
6. Avaliar os dois modelos com as seguintes métricas: Acurácia, Precisão, Recall, F1-score, AUC-ROC, AUC-PR, além de plotar visualmente a curva ROC e a curva PR. Justificar qual é métrica é mais adequada para avaliar a performance de cada modelo e qual é o melhor modelo. Explicar como as curvas são plotadas.
7. Utilizar o programa que plota o gráfico para pontos de corte (trabalho - feito em grupo), estabelecendo 3 pontos de corte para o máximo aproveitamento do modelo: Um para a classe negativa, um para a classe positiva e um para mesa ou análise manual. Justificar a escolha dos pontos de corte através do acúmulo populacional e erro do modelo.

Considerações:

\*Em ambas avaliações, o trabalho/prova feitos em grupo, caso estejam corretos, valerão 1 ponto cada. Em seguida serão feitas 3 perguntas individuais para cada aluno (3 para a o trabalho e 3 para a prova), valendo 0.5 pontos cada resposta correta.

Importante!!

\*\*A prova e o trabalho foram fornecidos com bastante antecedência, tirem suas dúvidas antes!!!