Template para entrega do projeto da disciplina
Ciência de Dados e Inteligência Artificial
Fase 2

Nome do estudante	**Eduarda Pereira Blanco**

Desenvolva um processo de ciência de dados no Orange Data Mining, cobrindo os elementos abaixo. **Para cada um dos itens solicitados é necessário inserir imagens que evidenciem o trabalho realizado**.

Exploração dos dados

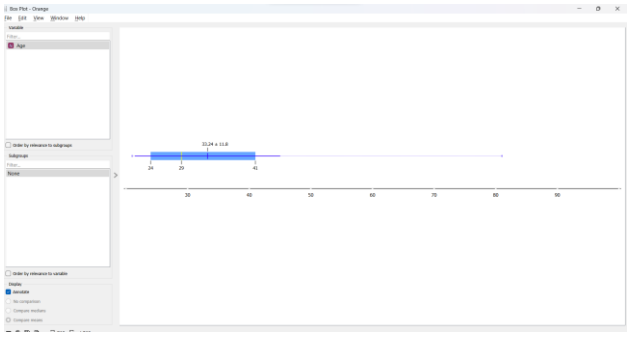
Que tipo de experimentos você fez na exploração dos dados (verificação de outliers, cálculos de médias, dados inválidos etc.).

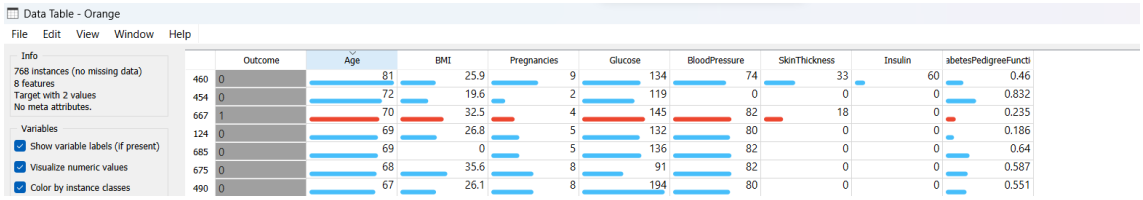
O conjunto de dados escolhidos compreende um dataset de diabetes, que é originário do National Institute of Diabetes and Digestive and Kidney, disponível em: https://www.kaggle.com/datasets/akshaydattatraykhare/diabetes-dataset e considerado o padrão gold de dados.

O objetivo dos dados é prever diagnosticamente se o paciente tem diabetes, com base em medidas de diagnósticos incluídas no conjunto de dados. Todos os pacientes são mulheres com idade superior a 21 anos.

Alguns conceitos foram levantados durante a exploração de dados:

- Resultados: 500 com resultado de “não diabetes” e 268 com resultado de “diabetes”
- Idade:
  - Média: 33,2 anos
  - Moda: 22 anos
  - Encontrado o valor máximo de 81 anos, podendo ser considerado um único outlier como apresentado pelo box plot





- Quantidade de vezes grávida:
    - Média: 3,8
    - Moda: 1
    - Máximo: 17 porém não consideraria outlier visto que é um valor muito próximo de demais

- BMI (Body Mass Index):
- O equivalente ao IMC aqui no Brasil, quanto maior o valor, maior a relação com sobrepeso e obesidade.
- Média: 31 – considerado obesidade
- Moda: 32 – considerado de obesidade
- Máximo: 67.1

- Tabela referência:

Escolha de, ao menos, três algoritmos de aprendizado para a modelagem

Apresente os algoritmos utilizados e justifique a escolha.

1. KNN

A utilização do modelo K-Nearest Neighbors (KNN) foi uma escolha para a classificação de diabetes em mulheres devido à sua simplicidade e à falta de suposições sobre a distribuição dos dados. Ele classifica novos casos com base na proximidade aos K vizinhos mais próximos no conjunto de treinamento, o que é intuitivo e eficaz para detectar padrões não lineares. Também, optei devido a sua pré prepação de dados realizada através do sistema Orange.

2. Logistc Regression

A regressão logística foi escolhida devido a eficácia para a classificação de diabetes devido à sua simplicidade e capacidade de fornecer probabilidades interpretáveis sobre a presença da doença. Ela assume uma relação linear entre as características e a probabilidade de uma classe, o que facilita a modelagem e a interpretação dos. A regressão logística também permite uma avaliação clara do desempenho por meio de métricas como acurácia – AUC.

3. Tree

As árvores de decisão são adequadas para a classificação de diabetes por sua capacidade de modelar relações complexas e não lineares de maneira visual e intuitiva, facilitando a interpretação dos critérios de decisão. Elas lidam bem com dados categóricos e contínuos, não exigem normalização dos dados e são robustas a outliers como os apresentados acima.

Preparação dos dados de acordo com as características dos algoritmos de aprendizado escolhidos

Descreva o processo realizado para essa etapa.

Retirei os dados nulos/vazios e realizei os primeiros experimentos com os dados não tratados, considerando que por padrão o modelo KNN, realiza um pré-processamento deste dados.

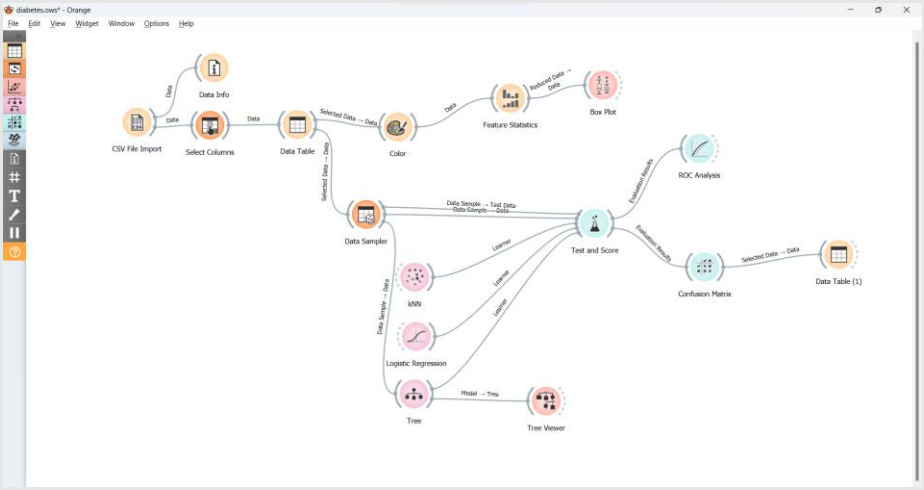
Já para a aplicação dos algoritmos Tree e Logistc Regression realizei um equilíbrio de casos positivos e negativos para acusação de diabetes, utilizando os métodos apresentados na aula 07, parte 05, evitando vieses. Assim normalizando os números entre pacientes com e sem diabetes, tornando os dados equilibrados para aplicação do algoritmo.

Apesar da normalização dos dados, realizei mais uma vez Tree e Logistic Regression nos nados não tratados para comparar seus resultados.

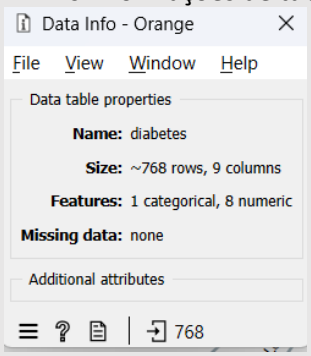
Execução dos experimentos de aprendizado e coleta das métricas

Descreva o processo e os resultados obtidos.

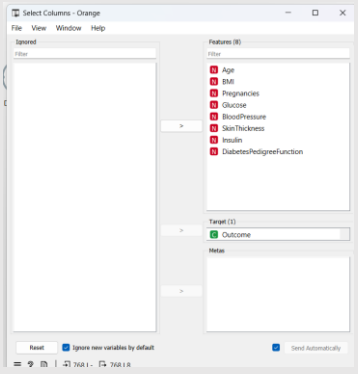
Parte 1 (dados não normalizados):



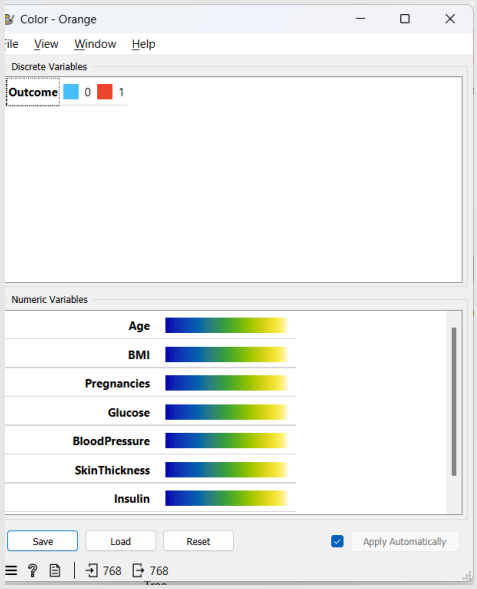
1. Arquivo CSV importado
2. As informações de tabela apresentadas foram:



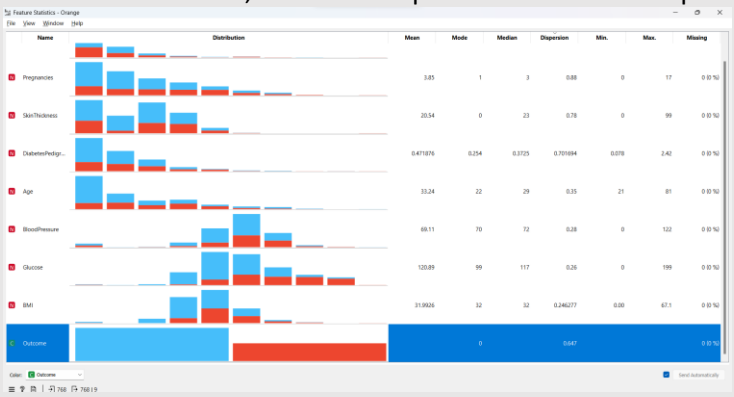
3. A seleção de colunas com o parâmetro de alvo a coluna resultado, considerando que para a coluna de resultado os valores atribuídos foram 0 para “não diabetes” e 1 para “diabetes”.

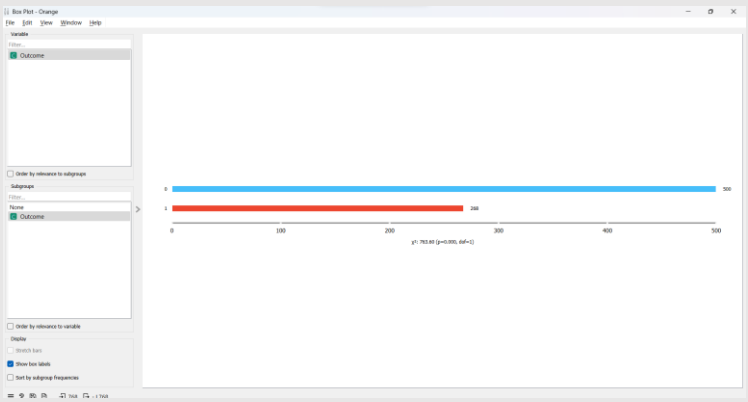


4. Para iniciar a exploração de dados coloquei a opção de coloração por resultado: azul para “não diabetes” e vermelho para “diabetes”.

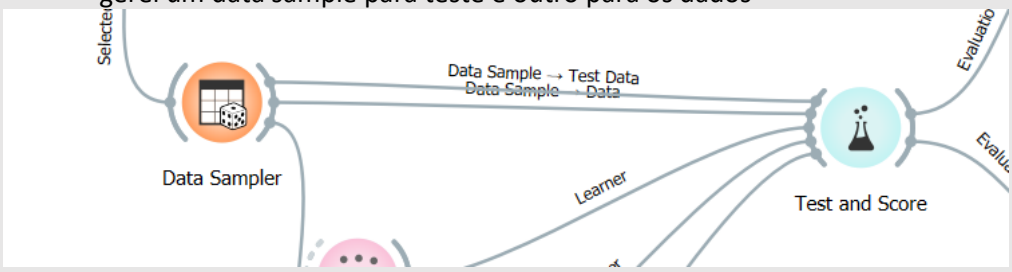


4.1. Seguindo no âmbito da exploração de dados foi incluído o visualização de feauture statistics e box plot, onde consegui visualizar os conceitos de média, moda, máxima e os outliers, além de compreender o valor total por resultado.

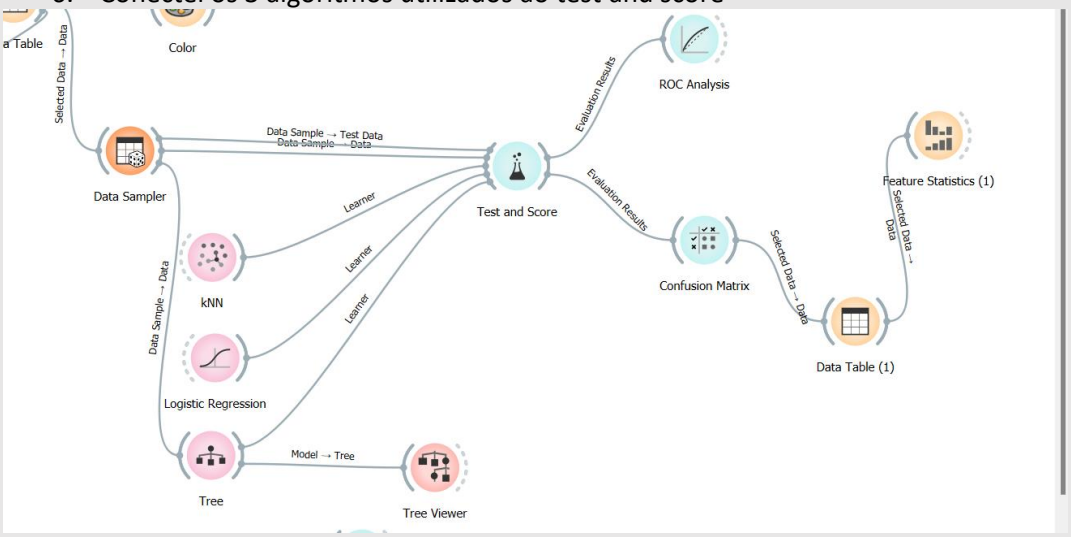




5. Através da seleção de dados utilizei data sample que conectado ao test and score, gerei um data sample para teste e outro para os dados

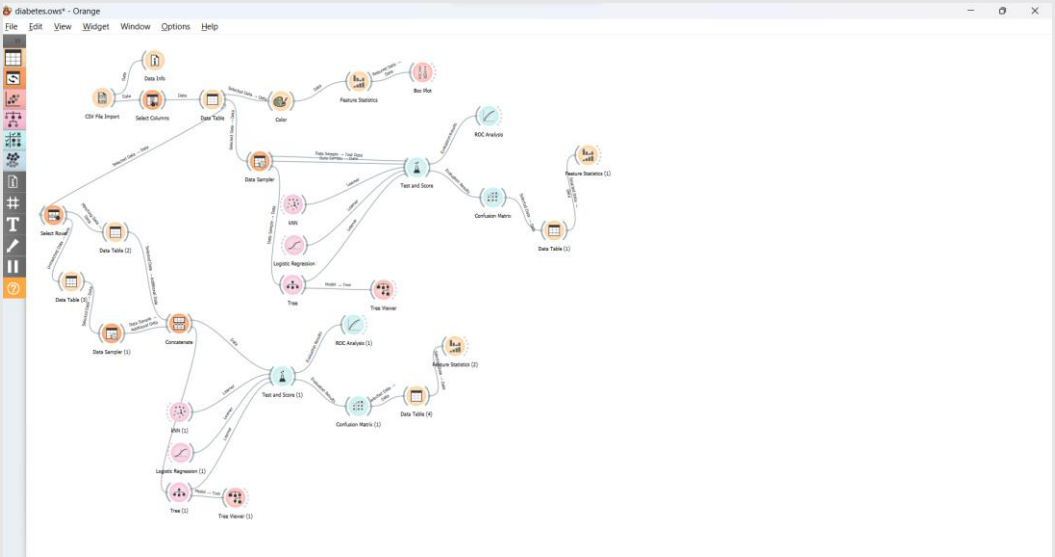


6. Conectei os 3 algoritmos utilizados ao test and score



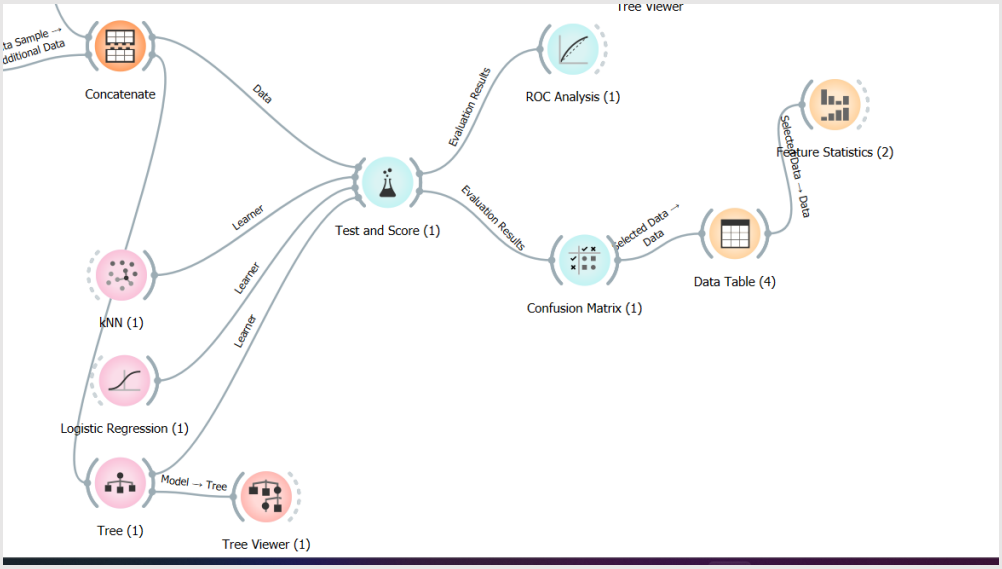
6.1. Conectei a visualização da árvore a própria arvore para análises
7. Conectei a visualização da confusion matrix e ROC analysis
8. Por fim, conectei uma visualização de tabelas na confusion matrix para identificar os dados que obtiveram erro durante o experimento.

Parte 2 (dados normalizados):



9. Extrai uma seleção de linhas que possuíam o resultado de 1, no caso diabetes. Totalizando os 268 casos.

10. Conectei uma tabela de dados com esses 268 casos
11. Conectei uma tabela com o restante dos casos de resultado 0, não diabetes.
12. Conectei uma amostragem a tabela de dados com resultado 0, deixando apenas 268 casos.
13. Realizei a concatenação de ambas as tabelas gerando uma de 536 casos e rodei os mesmos modelos novamente.



Relato dos experimentos e lições aprendidas

Apresente uma reflexão acerca dos resultados obtidos com este projeto.

- KNN

Considerando os dados não normalizados o modelo apresentou a consistência entre os valores de acurácia (0.7089 e 0.6942) está mostrando um desempenho relativamente estável em diferentes conjuntos de validação ou diferentes execuções. A acurácia média (em torno de 0.70) indica que o modelo tem um desempenho razoável, mas há espaço para melhorias.

Considerados os dados normalizados a acurácia média dos valores fornecidos é bastante alta, especialmente o valor de 0.7876, que sugere que o modelo KNN(1) está alcançando um bom desempenho geral na maioria dos casos.

- Tree

Considerando os dados não normalizados o modelo de árvore de decisão apresenta uma acurácia média que varia entre 63% e 70% em diferentes execuções ou validações, com algum nível de consistência em torno de 70% em algumas validações. No entanto, o valor baixo de 0.3486 em uma métrica de erro sugere que o modelo pode ter problemas de desempenho em aspectos específicos. Em comparação com outros modelos, a árvores de decisão não está performando tão bem.

Considerando os dados normalizados a comparação entre os dois conjuntos de scores para o modelo de árvore de decisão mostra uma melhoria significativa na versão mais recente (Tree (1)). A acurácia aumentou de uma faixa de 63%-70% para 74.9%-78.8%, indicando um desempenho muito melhor e mais consistente. A métrica de erro também mostrou uma

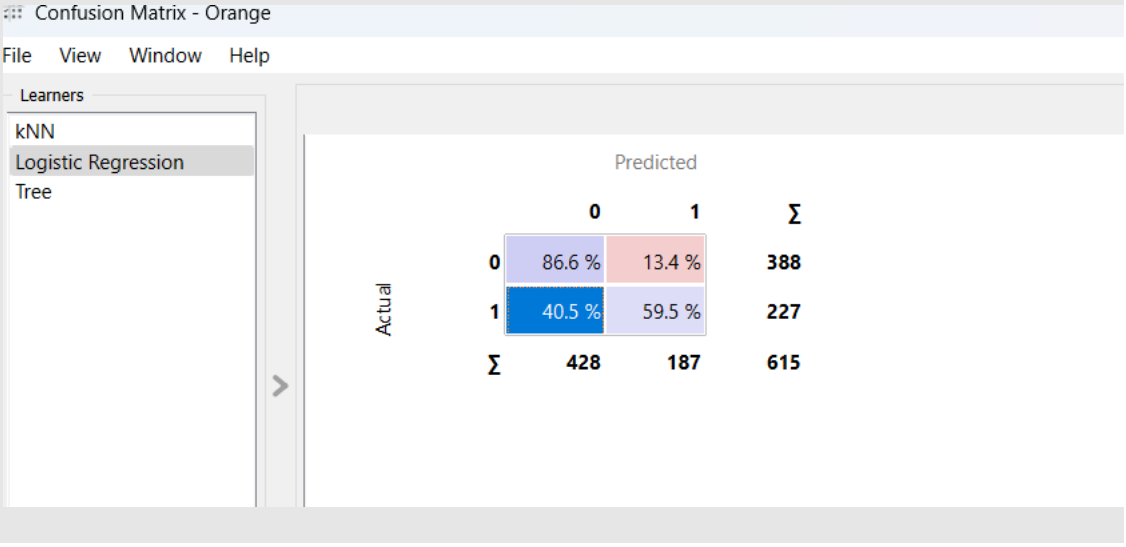
variação, mas a versão mais recente do modelo demonstra um avanço em termos de desempenho geral.

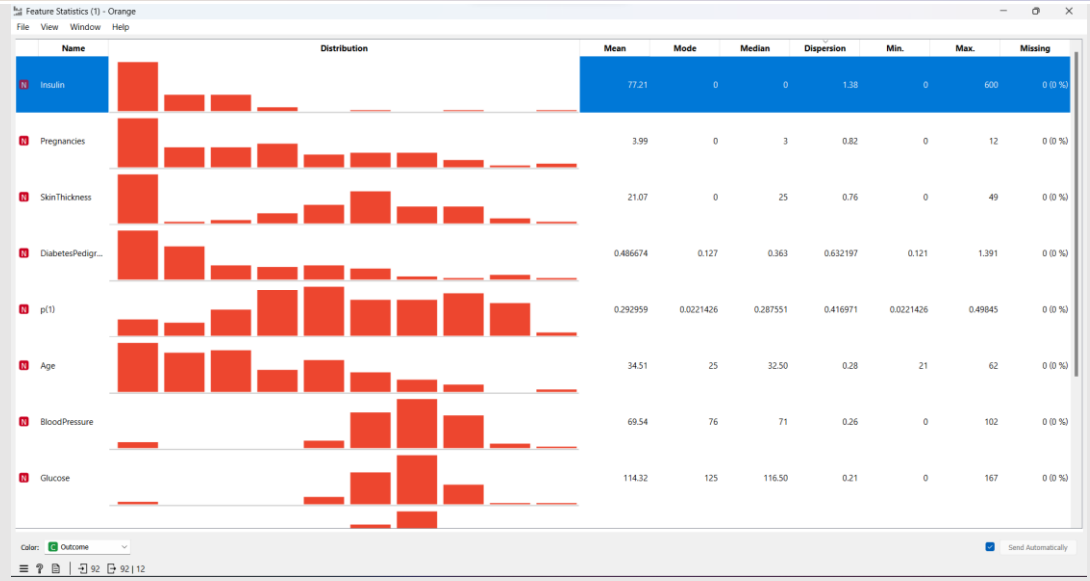
- Logistc Regression

O modelo de Regressão Logística tem um desempenho robusto, com acurácia média alta e consistente na versão original, variando de 76.6% a 81.8%. A versão mais recente (Logistic Regression (1)) mostra uma acurácia máxima melhor (84.0%), mas com uma média um pouco menor e uma métrica de erro mais alta (0.4965). Apesar de a melhoria na acurácia máxima seja notável, a métrica de erro sugere que ainda há espaço para ajustes. No geral, ambos os modelos têm um desempenho competitivo, mas a versão (1) pode indicar que a acurácia foi aprimorada em alguns aspectos, enquanto o erro precisa ser melhorado.

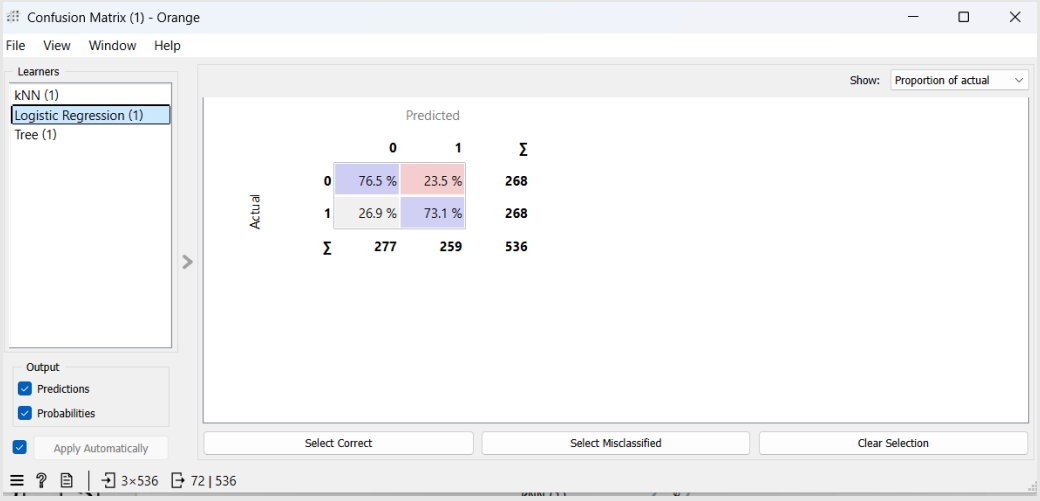
Caso:

Optei por normalizar os dados visto que a logistic regression apresentou um número alto de falso negativo, levando em consideração o contexto dos dados, um falso negativo é muito mais prejudicial neste caso, imagine o seguinte cenário: paciente recebe o resultado de não ter diabetes, já considerando os dados vistos anteriormente que possui agravantes como sobrepeso e alto valor de insulina, quando olhamos para os casos que apresentaram falsos negativos, temos uma média de insulina de 77, considerando valores coletados em jejum o normal previsto é entre 5 a 29,0 µIU/mL. Sabendo as complicações da diabetes a médio e longo prazo um erro de 40,5% de pacientes apresentaria até mesmo a morte desses pacientes.





Já ao realizar a normalização dos dados o valor de falso negativo cai mais de 10 p.p proporcional a amostragem.



Considero importante para os modelos a normalização e amostragem de dados com valores muito próximos de classificação, ainda mais quando pensando no contexto proposto de saúde/doença, mesmo que os modelos sejam treináveis, é compreensível um olhar humano no impacto dessa margem de erro, aqui, poderia significar 10% a menos de possíveis mortes ocasionadas pela diabete com diagnóstico errado.

LINK PARA O ARQUIVO DO PROJETO DO ORANGE E DOS DADOS UTILIZADOS

Insira os links para os arquivos.

Dados: https://www.kaggle.com/datasets/akshaydattatraykhare/diabetes-dataset

Orange:https://drive.google.com/file/d/1sMGi-thQMKbHCz9TrgMnxhwVeF-K-3Eh/view?usp=sharing