Trabalho Prático 2

André Góis

Luís Santos

Rafael Balinha

Especificação do trabalho

- Este projeto tem como função fazer com que consigamos abordar um caso real da Ciência de Dados, usando um conjunto de dados do Carcinoma Hepatocelular recolhido no Centro Hospitalar e Universitário de Coimbra.
- Iremos desenvolver várias técnicas desde a limpeza de dados e o seu pré-processamento até o uso de aprendizagem supervisionada e a comunicação de resultados.
- Desta forma, a principal finalidade deste projeto é desenvolver um pipeline de aprendizado de máquina capaz de determinar a capacidade de sobrevivência dos pacientes em 1 ano após o diagnóstico (por exemplo, "vive" ou "morre").
- A métrica de sucesso para este projeto é a exatidão da previsão. Queremos minimizar tanto os falsos positivos (prever que um paciente sobreviverá quando não sobreviverá) quanto os falsos negativos (prever que um paciente morrerá quando sobreviverá). Ambos os tipos de erros têm implicações significativas na vida real e devem ser minimizados.

Bibliografia

Para desenvolver este projeto, teremos como base os seguintes recursos:

- Material disponível no moodle, por exemplo, na pasta “iris”, o Jupyter Notebook ‘Exercise5_IART1’, como uma referência a nível de estrutura e organização do pensamento, ou o “auto” Notebook para ajudar na implementação dos modelos.
- Inteligência Artificial, por exemplo o ChatGPT, para auxiliar em métodos mais complexos da Ciência de Dados.
- Youtube, assistindo vídeos educativos que possam dar uma ajuda significativa no desenvolvimento do trabalho (exemplo, How to Do Data Exploration (step-by-step tutorial on real-life dataset) (youtube.com)).

Descrição das ferramentas e algoritmos

Ao longo da análise dos dados fornecidos, iremos seguir alguns passos importantes:

1. Exploração dos Dados
2. Pré-processamento dos Dados
3. Modelagem dos Dados
4. Avaliação dos Dados
5. Interpretação dos Resultados

- Para isto iremos utilizar o Jupyter Notebook para o desenvolvimento do projeto, com a distribuição Anaconda, usufruindo das bibliotecas que nesta já estão disponíveis, tais como, o numpy, matplotlib, seaborn, sklearn, os e pandas.
- Desta forma, adotamos os algoritmos necessários para chegar da forma mais eficaz e precisa ao objetivo final.

Pré-processamento

- Testamos inicialmente várias formas que afetaram a precisão dos modelos.
- No que diz respeito ao valores ausentes:

- > 45%: Eliminamos estes valores pois tinham uma grande percentagem de valores em falta;
- < 15%: Substituímos pela média (em variáveis numéricas) e pela moda (em variáveis categóricas) pois estes tinham uma percentagem reduzida;
- 15% a 45%: Como tem uma percentagem de valores ausentes até certo ponto significativa, usamos um método de imputação avançado, chamado imputação iterativa que é um processo que utiliza informações das outras características para estimar os valores ausentes de forma mais precisa.

Isto foi eficaz para a precisão dos resultados, visto que testamos várias percentagens e vários métodos, como substituir até aos 45%, unicamente, pela média ou fazer a média dos sobreviventes e substituir nos valores correspondentes, tal como nos não sobreviventes. Contudo não trouxe melhorias nenhumas, pelo contrário, piorou a acurácia do modelo.

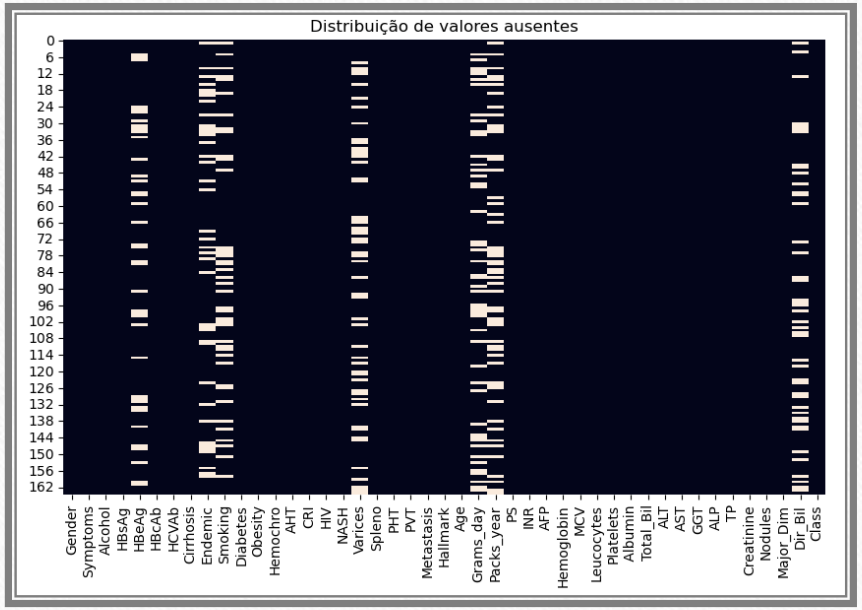
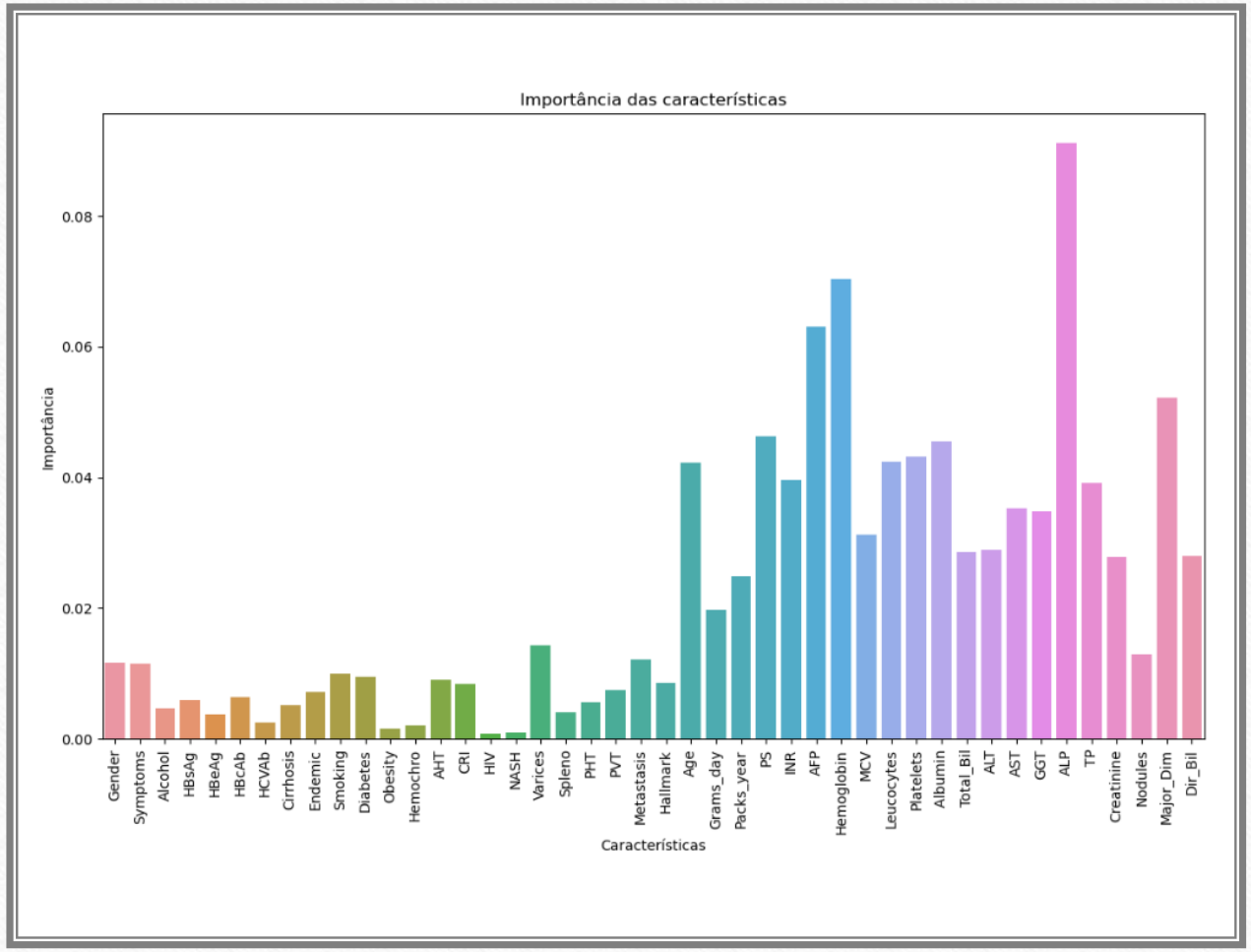


Gráfico após a remoção das variáveis >45% e a imputação da média e da moda das <15%

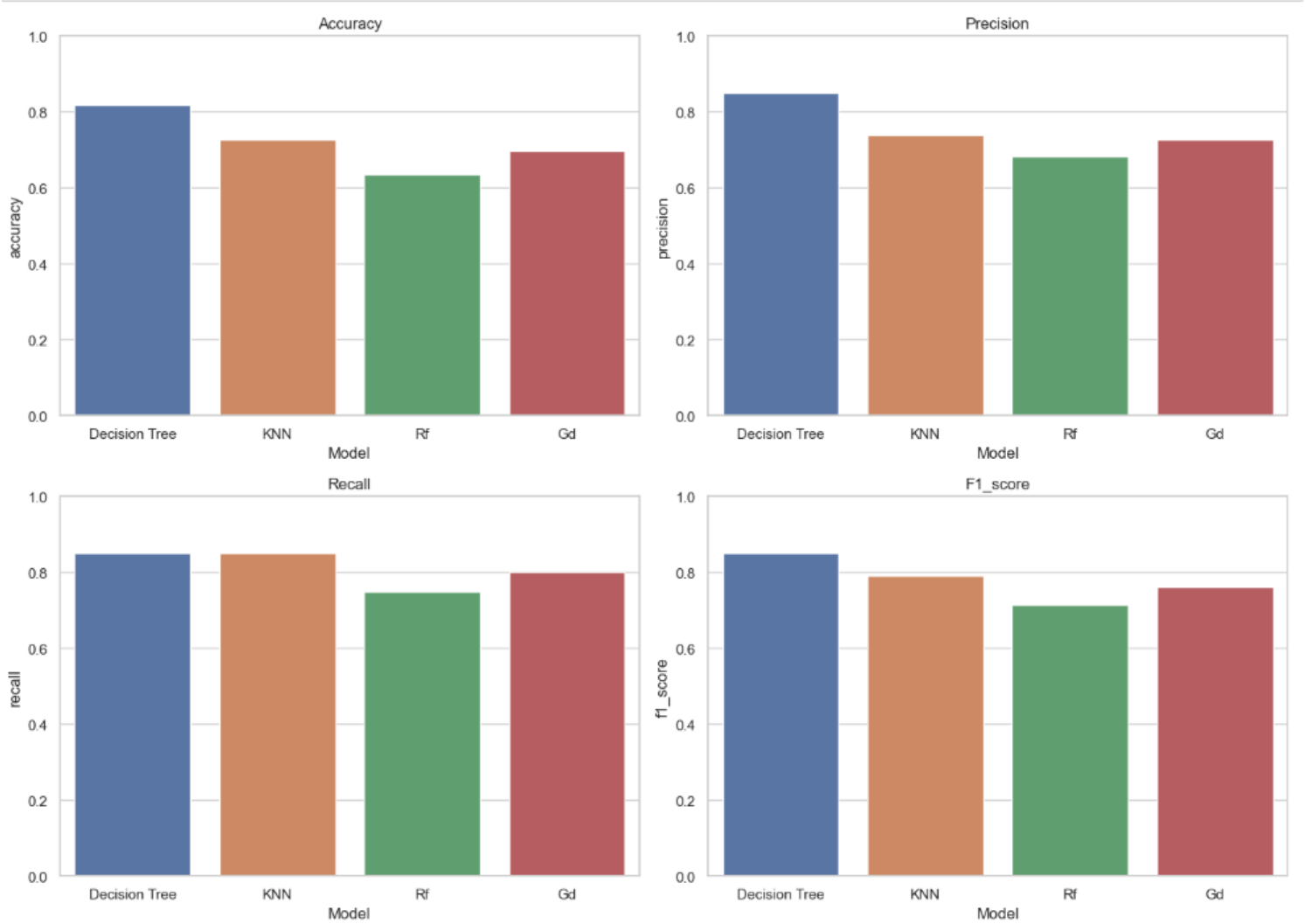


Pré-processamento

- Também foi feito uma avaliação da importância de cada variável com vários gráficos, e posteriormente foram feitos testes analisando como os modelos se comportavam usando todas as variáveis da mesma forma, ou atribuindo um peso a elas de acordo com a sua importância.
- Podemos, então, concluir que a atribuição destes pesos foi imprescindível para uma melhor precisão e acurácia.

Interpretação dos Resultados

Aqui podemos comparar as diferentes métricas em relação a cada modelo implementado, tirando várias conclusões que serão apresentadas no diapositivo seguinte.

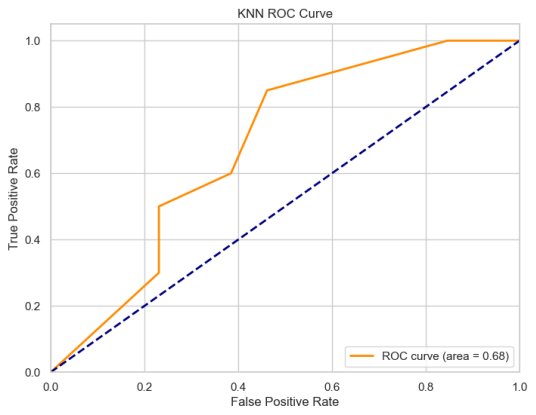


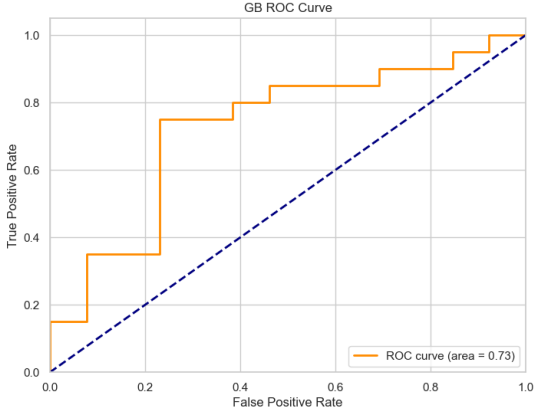
Interpretação dos Resultados

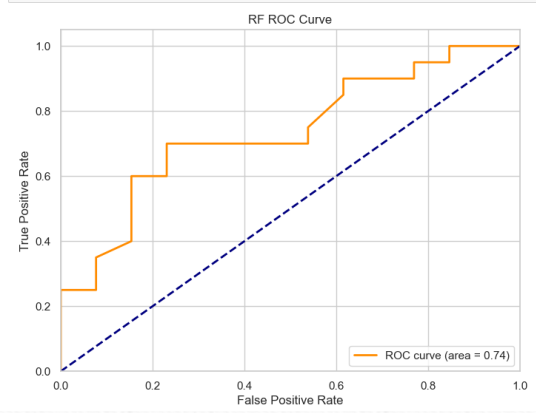
- Árvore de decisão:
- Este foi o melhor método, com índices de sucesso mais elevados, tendo assim uma acurácia de 82% e precisão de 85%. Usamos dois algoritmos para calcular a melhor profundidade, um que está num exemplo do Notebook 'auto' dado pelo professor no moodle, e outro chamado GridSearch (testa todas as combinações possíveis para encontrar os melhores hiperparâmetros). Podemos concluir que o melhor foi o do professor, dando uma sugestão de profundidade máxima de 50, o que trouxe a maior e melhor acurácia do projeto.
- O algoritmo começa com o nó raiz, que usa um dos atributos do conjunto de dados para fazer uma decisão. Seguidamente, o conjunto de dados é dividido em subconjuntos. Cada subconjunto então passa por um novo nó de decisão, onde outro atributo é usado para tomar uma decisão e dividir o subconjunto ainda mais. Este processo é repetido até que os nós folha sejam alcançados, no qual vão representar as previsões ou saídas do modelo.
- KNN:
- Este foi o segundo melhor método obtendo-se uma acurácia de 73%. Tal como a árvore de decisão, este modelo esteve sujeito a um algoritmo de pesquisa Grid search para encontrar o melhor número de vizinhos a usar, dando-nos uma resposta muito agradável, e que corresponde com as expectativas, trazendo uma boa acurácia e uma precisão de 74%.
- Este faz uma "pesquisa pelos vizinhos" mais próximos para fazer suas previsões e, para problemas de classificação, atribui ao ponto não classificado a classe que é mais comum entre seus k vizinhos mais próximos.
- GradientBoosting:
- Cria modelos que fazem poucas suposições sobre os dados, árvores de decisão simples, e junta-os num modelo forte de maneira iterativa, trazendo-nos uma acurácia de 70%, sendo o terceiro melhor método usado.
- RandomForest:
- Este foi o pior método utilizado, trazendo uma acurácia de apenas 64%. Como este método combina várias árvores de decisão, seria expectável uma boa eficácia contudo, esta discrepância em relação à árvore de decisão pode ser devida à profundidade não ser a mais correta e à possibilidade de as árvores estarem sobreajustadas.

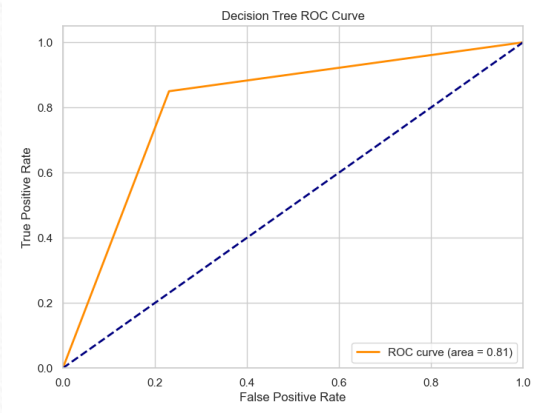
Interpretação de Resultados

- Por outro lado, um dos objetivos do projeto era minimizar os falsos negativos e os falsos positivos, o que foi algo concluído com sucesso pois, por exemplo, na árvore de decisão tivemos 3 casos de falsos positivos e falsos negativos contra 10 e 17 verdadeiros positivos e negativos. Os outros modelos, como também possuem uma acurácia menor, é de esperar que possuem uma maior discrepância nestes valores.
- Isto pode ser visível com recuso aos gráficos ROC e área AUC, no qual as linhas de todos os gráficos estão acima da linha aleatória, refletindo um bom resultado. A nível individual, a árvore de decisão possui uma maior área e mais próxima de 1, e o seu gráfico é o mais próximo ao cantor superior esquerdo em relação aos restantes, sendo a mais precisa.









Conclusão

Para análises futuras, com base neste projeto, iremos ter vários aspetos em conta. Por exemplo, começando por fazer uma boa exploração do problema em si e dos dados, para refletir e perceber bem o que é suposto fazer e como fazê-lo, como também, quais os dados que possuímos e como o computador está a lhes tratar.

Por fim, será interessante refletir que teremos de possuir vários modelos para analisar as suas diferenças, como também métodos de pesquisa que nos permitam maximizar a acurácia destes modelos.

Outro aspeto, talvez o mais importante é o Pré-processamento. Teremos sempre em consideração fazer um bom Pré-processamento, tratando bem dos dados ausentes, eliminando os desnecessários e avaliando as importâncias das variáveis para que, desta forma, possamos criar um modelo de classificação eficaz e com elevada precisão.

Não desmerecer também que uma boa reflexão sobre os resultados é importante, tanto para melhorar o projeto e os erros que possam existir, como também para uma melhor interpretação do que foi feito.