Andreia Araújo nº62912 (16 horas)

Carolina Rodrigues nº62910 (16 horas)

Filipa Branco nº62762 (16 horas)

INTRODUÇÃO

O presente trabalho tem como objetivo explorar e comparar o desempenho de três modelos de regressão distintos – ***Random Forest Regression***, ***Linear Regression*** e ***Decision Tree Regression*** – no contexto da inferência da atividade molecular sobre os recetores de Dopamina2.

Ao utilizar diferentes modelos de regressão, procurou-se identificar a abordagem que melhor se adaptaria aos dados específicos e, consequentemente, proporcionaria previsões mais meticulosas. A escolha dos modelos - *Random Forest, Linear Regression* e *Decision Tree* - foi baseada na necessidade de explorar diferentes paradigmas de modelagem, desde métodos baseados em árvores até abordagens mais lineares.

Além disso, a abordagem usada inclui uma análise do processo de pré-processamento dos dados, a avaliação do desempenho dos modelos e uma discussão sobre os resultados obtidos. Ir-se-á, por fim, justificar qual o modelo mais adequado ao *dataset*, dos três utilizados.

O repositório online do *Github* deste trabalho, contendo os ficheiros pertinentes, pode ser acessado a partir deste link: https://github.com/Princesacorderosa/AAut.

OBJETIVO

O objetivo é, através da exploração de diversos modelos de regressão, entender qual o modelo mais indicado para a previsão da atividade molecular dos recetores da Dopamina2.

OS MODELOS

No início deste projeto, analisaram-se diferentes modelos de regressão, tendo-se optado por usar os três modelos que se encontram na *Tabela 1*, por todos os pontos a favor referidos na mesma.

Foram então treinados estes modelos de forma a entender qual destes se adequaria melhor ao *dataset* em questão.

Tabela 1: Modelos de Regressão utilizados.

Modelo	Pontos a Favor
Random Forest Regression	Aleatoriedade , menos sensível a *overfitting*, requer menor ajuste de hiperparâmetros
Linear Regression	Modelo simples, grande interpretabilidade, rápido de treinar
Decision Tree Regression	Versátil, não necessita de normalização de *features*, menor ajuste de hiperparâmetros

Andreia Araújo nº62912 (16 horas)

Carolina Rodrigues nº62910 (16 horas)

Filipa Branco nº62762 (16 horas)

DISCUSSÃO DE RESULTADOS

No âmbito desta análise, uma etapa crítica envolveu a identificação de *missing values* nos conjuntos de dados ***X_train***, ***X_ivs*** e ***y_train***. Uma vez que não existiam, não foi necessário removê-los, pudemos assim prosseguir para os próximos passos.

Na fase de seleção de características (*feature selection*) , utilizou-se um *Random Forest Regressor* para avaliar a importância de cada característica. O limiar de importância (*threshold*) foi definido em 0,005. Características com importância superior a este limiar são consideradas relevantes para os modelos de regressão. Este processo permitiu concentrar a análise num conjunto mais restrito que contribuíram significativamente para a modelagem preditiva.

A Análise de Componentes Principais (PCA) com ***n_components = 3***, resultou na extração de três componentes principais que retêm a maior parte da variabilidade dos dados originais. A avaliação da eficácia dessa redução de dimensionalidade foi crucial para compreender se a variância foi mantida.

A validação cruzada foi implementada com *KFold* (***n_splits = 5***) para garantir uma avaliação robusta e evitar enviesamento nos resultados.

O IVS foi utilizado como uma métrica abrangente para entender e comparar aspetos diversos relacionados ao desenvolvimento e qualidade do modelo, neste caso o ***Random Forest Regression.***

Desta forma, para uma melhor compreensão, obteve-se uma visualização gráfica de cada modelo de regressão, que oferecem insights valiosos sobre o desempenho dos modelos.

Através dos gráficos e das métricas utilizadas, como ***RVE***, ***RMSE***, ***Correlation Score***, ***Maximum Error*** e ***Mean Absolute Error***, será possível orientar a escolha da seleção do modelo mais adequado.

A. PCA

PCA é uma técnica de pré-processamento que transforma as suas características originais em combinações lineares (componentes principais **PC0**, **PC1**, **PC2**), da qual são frequentemente usadas para reduzir a dimensionalidade de um determinado conjunto de dados. A escolha de como se usa essas componentes pode influenciar de como os modelos de regressão interpretam e preveem os dados.

Tabela 2: Variação da Variância pela análise de PCA

	Variância explicada	**Variância Total**
PC0	0.2048	0.2048
PC1	0.1716	0.3764
PC2	0.1068	0.4832

Andreia Araújo nº62912 (16 horas)

Carolina Rodrigues nº62910 (16 horas)

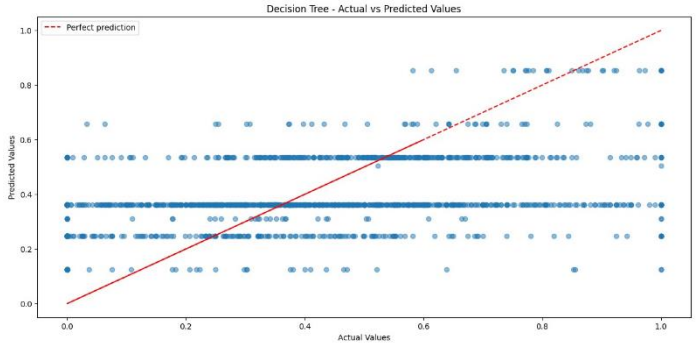
Filipa Branco nº62762 (16 horas)

Esses resultados (*Tabela 2*) refletem a influência da escolha das componentes principais dos modelos. A técnica do PCA ajudou a melhorar o desempenho, reduzindo a dimensionalidade dos dados e destacando as informações mais importantes.

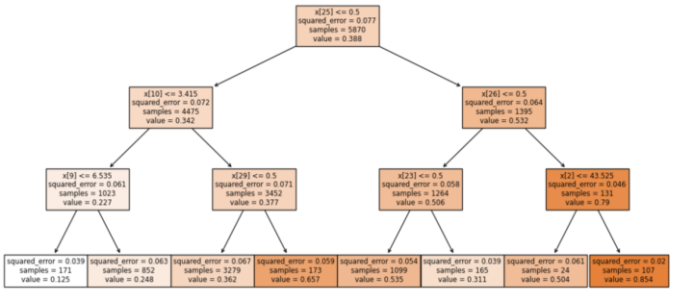
Com base na análise, retira-se que:

- **PC0 - *Variance explained: 0.2048 - Total Variance: 0.2048***
  Percentagem relativamente baixa (20.48%). PC0 não consegue capturar completamente as informações mais importantes nos dados originais.
- **PC1 - *Variance explained: 0.1716 - Total Variance: 0.3764***
  PC0 e PC1 explicam 37.64% da variância total. Essa percentagem sugere que essas duas componentes principais, em conjunto, representam uma parte significativa das variações nos dados originais.
- **PC2 - *Variance explained: 0.1068 - Total Variance: 0.4832***
  Com PC0, PC1 e PC2 juntos, apresentam 48.32% da variância total.

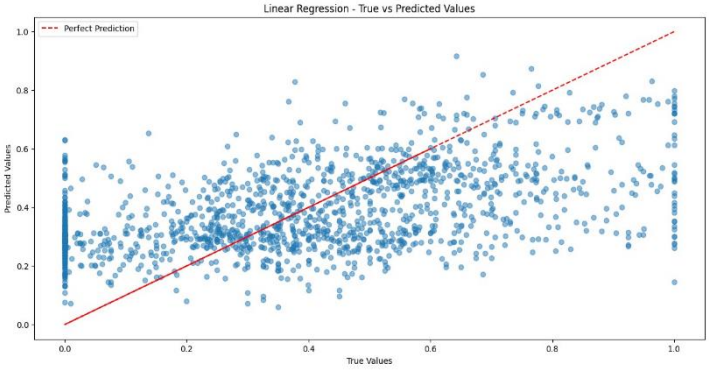
B. REGRESSION MODELS



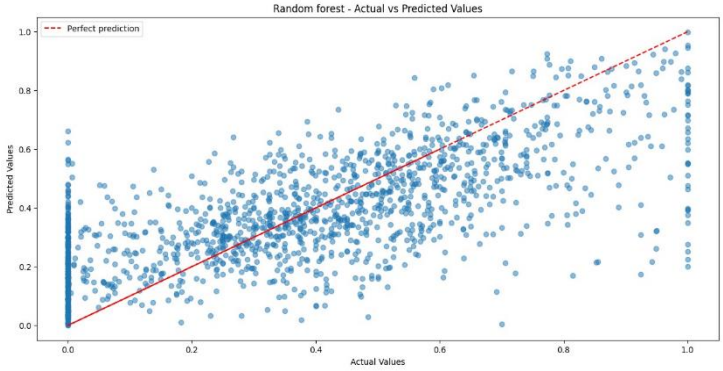
a)



b)



c)



d)

Fig. 1. Representação gráfica dos modelos de Regressão de Árvore de Decisão a) e b), Regressão Linear c), e Random Forest (d).

Andreia Araújo nº62912 (16 horas)

Carolina Rodrigues nº62910 (16 horas)

Filipa Branco nº62762 (16 horas)

Tabela 3: Parâmetros dos Modelos de Regressão utilizados.

	RVE	***RMSE***	***Correlation Score***	***Maximum Error***	***Mean Absolute Error***
Regressão Linear	0.258	0.236	0.508	0.855	0.189
Árvore de Decisão	0.170	0.249	0.416	0.875	0.198
Random Forest	**0.498**	0.194	**0.705**	0.800	0.145

Com base nas métricas apresentadas, o modelo de ***Random Forest*** parece ser o melhor entre os modelos utilizados.

No caso da *Random Forest*, um valor alto de ***RVE*** sugere que o modelo está capturando efetivamente a variabilidade nos dados, porém, um ***RMSE***, ***Maximum Error*** e ***MAE***, baixos indicam que o modelo apresenta previsões próximas aos valores reais. Desta forma, este modelo é o mais promissor fornecendo uma relação linear eficaz e previsões precisas.

Desta forma, com base nas informações fornecidas, a escolha de incluir PC0 na análise com o modelo ***Random Forest*** foi orientada pelo objetivo de reter a maior quantidade possível de informação individual.

PC0, que explica 20.48% da variância total, foi considerado crucial para preservar as características individuais dos dados durante a previsão com o modelo *Random Forest*. Essa estratégia visa maximizar a capacidade do modelo em capturar padrões complexos e fornecer previsões mais precisas.