Blue text on a white background

Description automatically generated

**APRENDIZAGEM AUTOMÁTICA 2023/2024**

**SECOND HOME ASSIGNMENT**

|  |  |
| --- | --- |
| **Grupo N.º 03** | **Horas Dedicadas** |
|  |  |
| Diogo Mataloto, n.º 62747 | 25h |
| Rita Morais, n.º 38844 | 25h |
| Paulo Correia, n.º 22303 | 25h |

**INTRODUÇÃO**

O presente relatório visa explorar a atividade molecular de compostos sobre os Receptores de Dopamina D2 por meio de técnicas de Machine Learning. O enfoque principal reside na utilização de modelos de regressão para inferir e compreender a influência molecular sobre esses receptores, com especial atenção aos padrões, que acreditamos serem, de interação e resposta biológica.

**ACERCA DOS DADOS:**

O conjunto de dados utilizado neste estudo é caracterizado por uma extensa coleção de variáveis, totalizando 2132, com um número significativo de observações, ultrapassando 7000. Durante a análise exploratória, foi identificado que a grande maioria dessas variáveis é binária, assumindo valores de 0 ou 1. Além disso, uma parcela menor das variáveis é composta por dados contínuos, enquanto algumas são discretas, especificamente aquelas identificadas como D11 até D21, D23, e D26.

**Uma imagem com captura de ecrã, Saturação de cores, file, Retângulo

Descrição gerada automaticamenteUma imagem com diagrama, Gráfico, captura de ecrã, file

Descrição gerada automaticamenteUma imagem com captura de ecrã, file, Azul elétrico, Gráfico

Descrição gerada automaticamenteFig 1.** Principais observações**: A)** Em relação às variáveis binárias, destacou-se uma alta prevalência de valores 0; **B)** As variáveis contínuas, por outro lado, revelaram uma forte correlação entre si durante a análise exploratória (expecto a variavel D38); **C)** Distribuição dos dados y\_test.

**C**

**B**

**A**

**METODOLOGIA**

A metodologia implementada procedeu à avaliação de diversos modelos de regressão (KNeighborsRegressor, DecisionTreeRegressor, RandomForestRegressor e LinearRegression) através da aplicação de uma procura aleatória de hiperparâmetros em conjunto com a utilização de validação cruzada. Esta estratégia permitiu a comparação entre os modelos, identificando as combinações de hiperparâmetros que otimizam a variância explicada durante a procura. A adoção desta estratégia contribuiu para uma estimativa robusta do desempenho dos modelos e prevenindo possíveis vazamentos de informações provenientes do conjunto de teste.

Inicialmente, os dados foram avaliados sem qualquer transformação. Em uma segunda tentativa, implementou-se a seleção de características utilizando Random Forests (RFs), o que resultou na escolha de 355 variáveis. Finalmente, optou-se pela técnica de Análise de Componentes Principais (PCA) dos dados previamente escalados.

Uma imagem com texto, captura de ecrã, Tipo de letra

Descrição gerada automaticamente**RESULTADOS**

Uma imagem com texto, captura de ecrã, Tipo de letra

Descrição gerada automaticamente**Fig 2.** Resultados dos dados sem transformação. O RandomForestRegressor apresentou o melhor resultado.

**Fig 3.** Resultados após seleção das variáveis com maior importância (cerca de 355 variáveis). Outra vez o RandomForestRegressor foi o melhor.

Uma imagem com texto, captura de ecrã, Tipo de letra

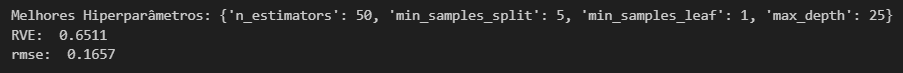
Descrição gerada automaticamente

**Fig 4.** Resultados do PCA para n\_components=10. O kNeighborsRegressor aqui apresentou melhores resultados.

**FINE-TUNING OF HYPERPARAMETERS**

Após várias tentativas, muitas das quais não foram apresentadas aqui, o modelo que proporcionou os melhores resultados foi o RandomForestRegressor, utilizando apenas os dados escalados das 355 variáveis selecionadas através do SelectFromModel.

Desta forma procurou-se os melhores hyperparamentos recorrendo ao RandomizedSearchCV.

Neste caso o melhor resultado de RVE que se obteu foi **0.6511**

**Fig 4.** Resultados do PCA para n\_components=10. O kNeighborsRegressor aqui apresentou melhores resultados.

**CONSIDERAÇÕES FINAIS**

Em considerações finais, a metodologia adotada para a avaliação de modelos de regressão foi abrangente e sistemática, visando uma compreensão aprofundada do desempenho dos diferentes algoritmos. Também conseguimos perceber o impacto que os diferentes processamentos dos dados tiveram no desempenho dos modelos.

O refinamento dos hiperparâmetros, realizado através do RandomizedSearchCV, permitiu obter o modelo RandomForestRegressor mais eficiente, alcançando um RVE de 0.6511.