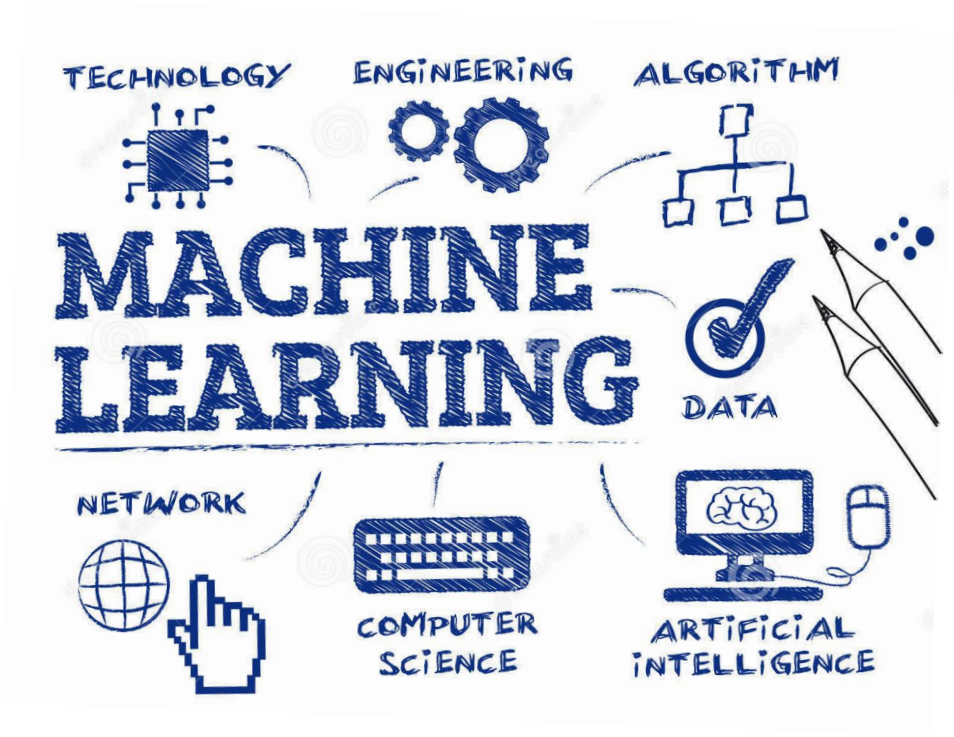
Engenharia do Conhecimento

Projeto 2



Grupo: 41

Francisco Henriques – 56348

Guilherme Marques – 55472

Miguel Seabra – 56344

Docentes: Sofia Teixeira, Cátia Pesquita

Data: 17/05/2023

Ciências ULisboa

- **Registo de horas de trabalho:**

Aluno	Horas de Trabalho
Francisco Henriques	25 Horas
Guilherme Marques	7 Horas
Miguel Seabra	5 Horas

- **Introdução e Objetivos:**

O objetivo deste projeto baseia-se em desenvolver o melhor modelo de classificação possível para prever a variável *Biodegradable*, utilizando uma versão aumentada e editada do conjunto de dados: *QSAR biodegradation Data Set*. Disponibilizado pelos docentes da UC. A tarefa consistiu em 4 passos gerais como: Processar os dados disponibilizados; Selecionar as features mais importantes; Examinar diferentes híper parâmetros para cada modelo; Selecionar o melhor modelo;
Para atingir este objetivo, foi utilizado um JupyterNotebook de modo a ser mais fácil a execução do código necessário e da organização do output do mesmo. Esse ficheiro está em anexo junto deste relatório.

- **Processamento de dados:**

Tendo em conta que o conjunto de dados era relativamente grande (cerca de 42 colunas) e que não estava normalizado, existindo células em branco nos dados, começámos por carregar o csv disponibilizado para um panda, sendo depois utilizado o método KNNInputer para **preencher as células vazias**, este método tem em conta um K numero de vizinhos, em que neste caso foi escolhido K=5, de modo a não termos overfitting nem underfitting, com estes 5 vizinhos mais próximos o KNNInputer atribui um valor médio desses vizinhos à célula vazia.

Tendo o conjunto de dados completo e tendo em conta que a variável alvo (Biodegradable) era uma variável binária (RB ou NRB), procedemos à sua conversão em número binário 1 para RB e 0 para NRB.

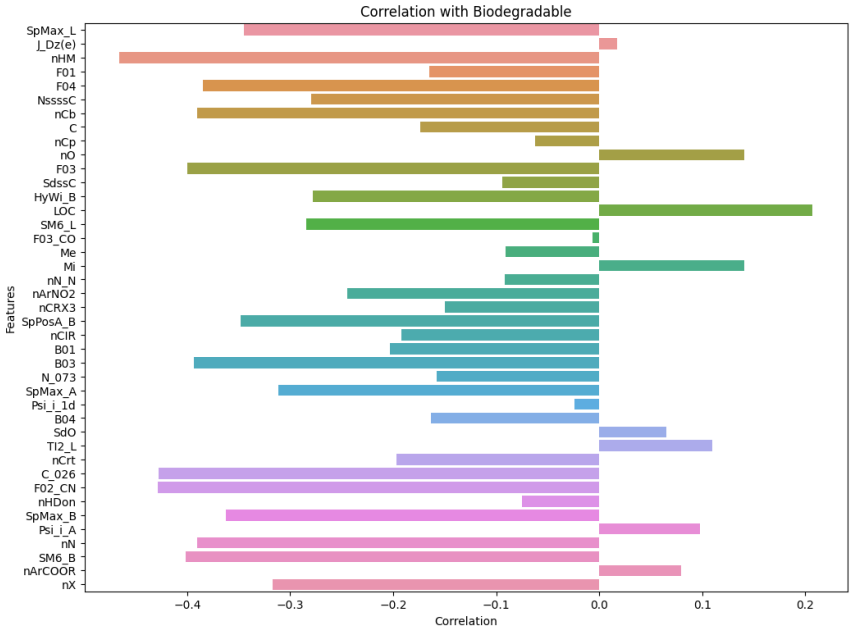
Depois foram feitos testes para determinar qual o scaler que melhor se adaptaria aos dados, para isso foi utlizada uma função do notebook da TP04, que imprimia estatísticas para diferentes scalers um modelo, e um certo conjunto de dados. Nos testes observou-se que para a maioria dos modelos, o StandartScaler era o que trazia melhores resultados tanto a nível de accuracy como de precision, recall, etc. Os resultados dos testes foram os seguintes:

Model	Scaler	Accuracy	Precision	Recall	F1 Score	MCC
Logistic Regression	StandartScaler	0.968	0.9634	0.9914	0.9772	0.8441
	RobustScaler	0.9575	0.9611	0.9898	0.9753	0.8304
	MinMaxScaler	0.9416	0.9425	0.9914	0.9663	0.7612
Decision Trees	StandartScaler	0.9628	0.9736	0.9827	0.9781	0.8552
	RobustScaler	0.9582	0.9734	0.9772	0.9753	0.8387
	MinMaxScaler	0.9575	0.9741	0.9757	0.9749	0.837
Random Forests	StandartScaler	0.9735	0.9805	0.9882	0.9844	0.897
	RobustScaler	0.9735	0.9813	0.9874	0.9844	0.8972
	MinMaxScaler	0.9735	0.982	0.9867	0.9843	0.8975
SVM	StandartScaler	0.9688	0.9716	0.9922	0.9817	0.877
	RobustScaler	0.9608	0.962	0.9929	0.9772	0.844
	MinMaxScaler	0.9502	0.9531	0.9898	0.9711	0.7992

Como podemos ver em 3 dos 4 modelos testados o StandartScaler obteve os melhores resultados, apenas no modelo RandomForests obtivemos um resultado muito idêntico entre o RobustScaler e o StandartScaler, sendo o Recall um pouco melhor no StandartScaler e a precision um pouco pior, face às pequenas diferenças nos valores consideramos que não vale a pena ser utilizado o RobustScaler.

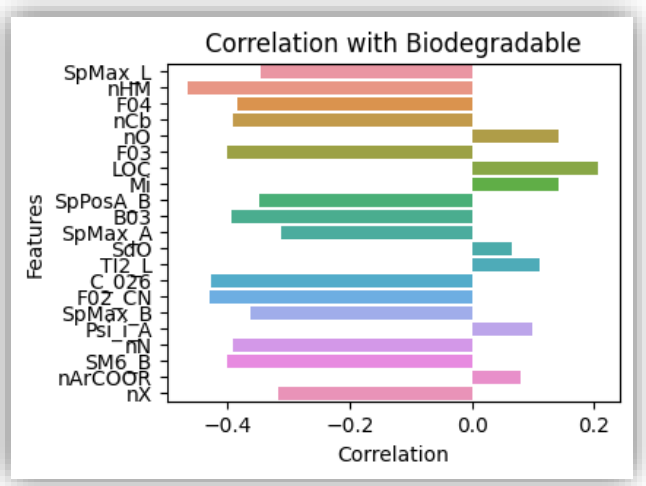
- **Seleção de Variáveis:**

Na hora de selecionar as variáveis, o primeiro objetivo foi remover as variáveis irrelevantes. Para isso usamos o correlation selection method. Para encontrarmos as variáveis irrelevantes temos de procurar as variáveis que têm uma correlação perto de 0 pois perto de 1 significa que apoiam fortemente a feature alvo (Biodegradable) e perto de -1 rejeitam fortemente essa mesma feature, daí as que têm uma correlação com valor perto de 0 serem irrelevantes pois nem apoiam nem rejeitam a feature alvo, não trazendo assim vantagens suficientes para continuarem no dataSet. Tendo em conta a correlação das features em relação à variável alvo obtivemos o seguinte gráfico:



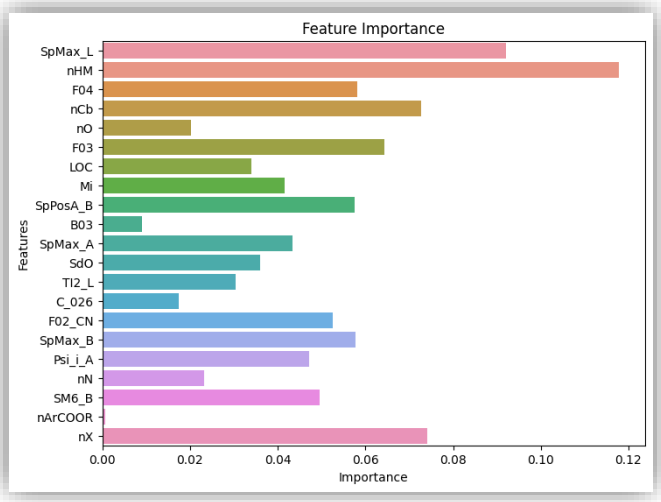
Como podemos ver pelo gráfico anterior existem muitas features que praticamente não são relevantes para a previsão da variável Biodegradable (a nossa variável alvo) por este motivo é que devemos reduzir o número de colunas para obter somente as que são relevantes , reduzindo a dimensionalidade, o que vai ser muito importante para a performance dos nossos modelos e até logo a seguir quando formos avaliar quais são as features mais importantes.

Ao remover as variáveis irrelevantes (cerca de 20 colunas), usando o modelo RandomForests e escalando os dados com StandartScaler, obtivemos o seguinte gráfico:



Neste gráfico facilmente percebe-se que não existem features com performance muito perto de 0 logo apenas estão incluídas as relevantes.

Além de remover as features irrelevantes fizemos também uma análise das features mais importantes no que toca à classificação da variável alvo, essa análise foi efetuada com um modelo RandomForests utilizando o método Stepwise Feature Selection (SFS), este método selecionou 6 features importantes(escolhemos 6 pois seria cerca de 30% das features restantes após a remoção das irrelevantes) sendo elas : ['nHM' 'nCb' 'nO' 'F03' 'Mi' 'F02_CN']. Tal como no gráfico a seguir consegue-se ver bem o motivo destas terem sido as escolhidas;



Neste gráfico podemos observar também que as features: nHM, nCb,F03 e F02_CN tem valores muito elevados de importância, mas que curiosamente as features: nO e Mi não apresentam uma importância tão elevada, quanto muitas outras, nós pensamos que isto deve-se ao facto de o gráfico da correlação apresentar para estas duas variáveis uma correlação positiva bastante elevada face às outras daí o método SFS as ter selecionado tendo outras hipóteses com maior valor de importância.

- **Afinação dos modelos com os híper parâmetros:**

Para iniciar a afinação começámos por obter uma tabela com estatísticas para cada modelo considerado, sendo estes: LR; Árvores de Decisão; RandomForests e SVM. Obtendo a seguinte tabela com os modelos sem qualquer tipo de afinação de híper parâmetros:

Model	Logistic Regression	Decision Trees	SVM	Random Forests
Accuracy	0.943597	0.975448	0.942933	0.956868
Precision	0.948679	0.978345	0.965517	0.956916
Recall	0.986656	0.992936	0.967033	0.993721
F1 Score	0.967295	0.985586	0.966275	0.974971
MCC	0.770878	0.903969	0.780941	0.82719

Tendo, portanto, os valores estatísticos base, começámos por testar a ferramenta GridSearchCV utilizada nas TPs com o SVC. Com esta ferramenta conseguimos "afinar" o modelo SVM com os melhores hiperparametros, que neste caso são: C = 10 e gamma = 0.1

Depois tentámos afinar o modelo RandomForest (o nosso melhor modelo até agora), com o método RandomizedSearchCV (utilizando uma grid de parâmetros possíveis) com alguns exemplos encontrados aqui[1] . Depois de encontramos os melhores hiperparametros para afinar o modelo conseguimos perceber uma pequena melhoria no modelo, em comparação com o antigo RandomForest sem os hiperparametros e testado com os mesmos dados. Depois de termos conseguido realizara a afinação com o RandomForests procedemos à afinação e testes dos 4 modelos sempre testados com os mesmos dados e o StandartScaler.

[1] https://towardsdatascience.com/hyperparameter-tuning-the-random-forest-in-python-using-scikit-learn-28d2aa77dd74

- **Testes dos modelos:**

À medida que íamos realizando a afinação dos modelos para cada um deles, procedemos também com um teste de avaliação, deste modo afinávamos e testávamos cada um dos modelos. Nestes testes foram obtidos os seguintes resultados:

STATS	Logistic Regression		Decision Trees		SVM		Random Forests	
	Base	Afinado	Base	Afinado	Base	Afinado	Base	Afinado
Accuracy	0.943597	0.943597	0.975448	0.945587	0.942933	0.971466	0.956868	0.974784
Precision	0.948679	0.948679	0.978345	0.9613	0.965517	0.976025	0.956916	0.977589
Recall	0.986656	0.986656	0.992936	0.974882	0.967033	0.990581	0.993721	0.992936
F1 Score	0.967295	0.967295	0.985586	0.968044	0.966275	0.983249	0.974971	0.985202
MCC	0.770878	0.770878	0.903969	0.785944	0.780941	0.888194	0.82719	0.901284

Como podemos observar pela tabela, tanto o modelo SVM como o RandomForest conseguiram melhores resultados em todas as métricas. Comparando então os dois modelos:

STATS	SVM	RF	RF – SVM
Accuracy	0.971466	0.974784	0.003318
Precision	0.976025	0.977589	0.001564
Recall	0.990581	0.992936	0.002355
F1 Score	0.983249	0.985202	0.001954
MCC	0.888194	0.901284	0.01309

Como podemos ver o modelo RandomForest é o melhor em todas as métricas sendo que no F1 Score temos uma pequena diferença, mas no MCC acabamos por ter um décimo de percentagem a mais.

Ficou assim escolhido o modelo **RandomForests** com os seguintes hiperparametros:

n_estimators	min_samples_split	min_samples_leaf	max_features	max_depth	bootstrap
100	2	1	sqrt	None	FALSE

Descrição dos parâmetros: **n_estimators:** Nº de árvores de decisão independentes que serão criadas; **min_samples_split:** define o nº mínimo de amostras necessárias num nó para que ocorra a criação de novos nós; **min_samples_leaf:** nº mínimo de amostras exigidas para formar uma folha numa árvore; **max_features:** nº máximo de características que o algoritmo considera enquanto procura a melhor divisão num nó - sendo sqrt: raiz quadrada do nº de features existentes; **max_depth:** profundidade máxima da árvore; **bootstrap:** o modelo é treinado sem reposição (ie. cada arvore é treinada num subconjunto diferente dos dados de treino) no caso False e vice-versa para True;

- **Discussão e conclusões sobre os resultados obtidos:**

Neste projeto de Engenharia do Conhecimento, o objetivo foi desenvolver um modelo de classificação para prever a variável "Biodegradable" utilizando o conjunto de dados QSAR Biodegradation. O processo envolveu várias etapas, incluindo o processamento dos dados, seleção de variáveis, afinação dos modelos e testes.

No processamento dos dados, carregámos o arquivo CSV fornecido, num Pandas DataFrame e utilizámos o método KNNInputer para preencher os valores ausentes. Em seguida, realizámos testes com diferentes scalers e observámos que o StandardScaler obteve os melhores resultados.

Na seleção de variáveis, utilizamos o método de correlation selection para remover as variáveis irrelevantes, Em seguida, reduzimos o conjunto de dados para incluir apenas as features relevantes. Também aplicámos o método Stepwise Feature Selection para identificar as features mais importantes

Na afinação dos modelos, utilizámos técnicas como GridSearchCV e RandomizedSearchCV para encontrar os melhores hiperparametros principalmente para o modelo SVM e RandomForest, pois eram os modelos com melhores score. Essa afinação resultou, no geral, em melhorias nos modelos em comparação com os modelos sem os híper parâmetros.

Realizámos testes dos modelos afinados e comparando os resultados, observámos que o modelo RandomForest afinado com os hiperparametros encontrados, obteve o melhor desempenho em todas as métricas avaliadas. Portanto, foi escolhido esse mesmo modelo (RandomForests) com os híper parâmetros específicos definidos, como o melhor modelo para prever a variável "Biodegradable".