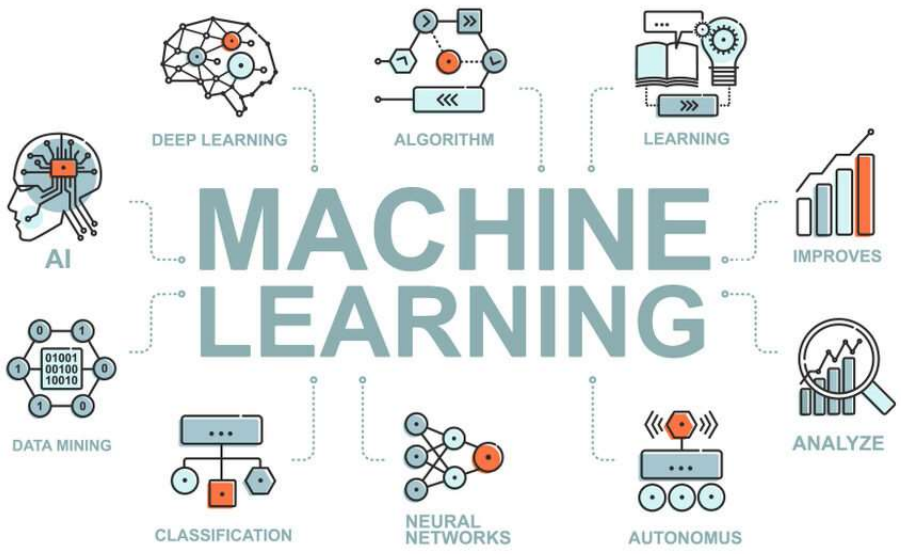
PROJETO FINAL DE APRENDIZAGEM AUTOMÁTICA

Online Shoppers Purchasing Intention Data Set



Instituto Politécnico de Setúbal – Escola Superior de Tecnologia do Barreiro

Unidade curricular Aprendizagem Automática

Docente Jacinto Estima

2º Ano 2º Semestre

João Yanga nº2020000167

Sérgio Pinto nº202000087

Alexandre Duarte nº202000198

Índice de Figuras

Índice

Introdução

Na elaboração deste projeto, o Sr. Professor Jacinto Estima, propôs *data sets* dos quais os grupos tiveram que escolher. Escolheu-se o dataset "*Online Shoppers Purchasing Intention Data Set*"

O *data set* contém cerca de 12000 sessões onde temos informações dos acessos num site que terminaram ou não em transações.

O objetivo deste trabalho é então conseguir retirar conhecimento a partir deste algomerado de dados, usando as técnicas de *Machine Learning* aprendidas.

Neste projeto utilizaram-se três modelos diferentes.

Sendo eles o, KNN (*K-nearest Neighbour*) e *logistic regression* que são algoritmos de classificação e regressão (aprendizagem supervisionada), K-means que é um algoritmo de *clustering* (aprendizagem não-supervisionada).

Procurou-se comparar resultados e responder a certas perguntas elaboradas pelo o grupo. Para a aplicação dos mesmos, o *data set* teve que passar por um processo de limpeza e tratamento dos dados.

Limpeza de dados e análise exploratória

No caso do nosso *dataset* foi preciso este passo, pois havia muito *missing data*, ou seja, os dados foram adaptados de modo a que estivessem preparados para entrar nos diferentes modelos.

Portanto, começou-se por substituir os valores nulos que estão nas colunas pelo valor da média das respetivas colunas. Depois algumas colunas que tinham valores *float* foram passadas para valores inteiros, como por exemplo a coluna *Administrative* entre outros. Em seguida todas as colunas que tinham N/A passaram a ter a média dos valores das colunas.

Métodos aplicados de Machine Learning

Como foi referido na introdução (pág.2) foram usados 3 modelos diferentes, sendo eles:

- K-means;
- KNN (*K-nearest Neighbour*);
- *Logistic regression*.

K-means

Clustering é a tarefa de agrupar um conjunto de objetos de tal forma que os objetos do mesmo grupo sejam mais semelhantes (em algum sentido) entre si do que com os de outros grupos (clusters). (Cluster_analysis, 2022)

Utilizámos este algoritmo de classificação pois adequava-se mais ao nosso *dataset* e também porque uma das suas vantagens é de usar princípios simples podem ser explicados por termos que não são estatísticos. (Estima, 2022)

Comecou-se por reduzir o *data set* a duas colunas: "*Bounce Rates* e "*Exit Rates*", pois decidiu-se focar nestes dois atributos para agregar os dados. Após a criação do modelo fez-se o treino do mesmo modelo.

Quanto à definição do número de clusters primeiramente permitiu que o modelo decidisse por defeito. Foram escolhidos 8 clusters. Numa segunda abordagem reduziu-se o número de clusters a 3. (Robinson, 2022)

KNN

O algoritmo KNN (K-*Nearest Neighbour*) é um dos mais utilizados em *Machine Learning* e também um dos mais simplistas, analisando seu processo de cálculo.

Tem a possibilidade de utilização do mesmo tanto para classificação quanto para regressão, neste caso foi usado para classificação. (Luz, 2019)

Também é conhecido como o método de aprendizagem "*lazy*" para não construir um modelo. O custo deste algoritmo é durante a classificação e depende muito da dimensão dos dados de treinamento.

A pergunta que foi feita para este modelo foi da possibilidade da página dos produtos da loja é mais visitada durante o fim de semana.

No desenvolvimento do modelo do KNN, começou-se por escolher quais as variáveis que se queria utilizar no modelo. Em seguida foram separados os dados treino e os dados teste, dando um ratio de 0.30 de *test size*, o que significa que 70% dos dados vão ser usados para treinar o modelo e que 30% é para testar usando *train_test_split*.

Depois fez-se o *import* no *KNeighborsClassfier* do *sklearn*, e em seguida treinou-se o modelo com os *training sets*. Quando terminámos isso, decidimos testar a *accuracy* dos dados de teste e de treino. Depois de obtermos a *accuracy* dos nossos dados, usamos o *cross-validation* para encontrar o melhor valor para k. (Dwivedi, 2021)

Logistic Regression

A regressão logística é um dos algoritmos básicos e populares para resolver um problema de classificação. Fornece um número/valor, por exemplo o valor de mercado de uma determinada casa que irá ser colocada a venda. Este algoritmo é, e foi usado para analisar diversas *features* (colunas com dados, as quais geraram informação). (Estima, 2022)

Neste modelo foi-se questionado se as *BounceRates* e *ExitRates* ocorriam mais aos fins de semana ou não.

Deste modo, foi preciso fazer o *import* do *linear_model* dos pacotes do *sklearn*. Foram selecionadas as colunas acima referidas e também a coluna *Weekend*. Depois da seleção foi aplicado o algoritmo foi feito o *fit* dos dados ao modelo. Definimos os valores precisos para as *rates* e foi feita a previsão.

Discussão de resultados

Foi feita uma matriz de correlação (Figura 1) para determinar as variáveis que tinham maior correlação com outras ajudando assim a descobrir as melhores maneiras de aplicar os modelos e a que dados os aplicar.

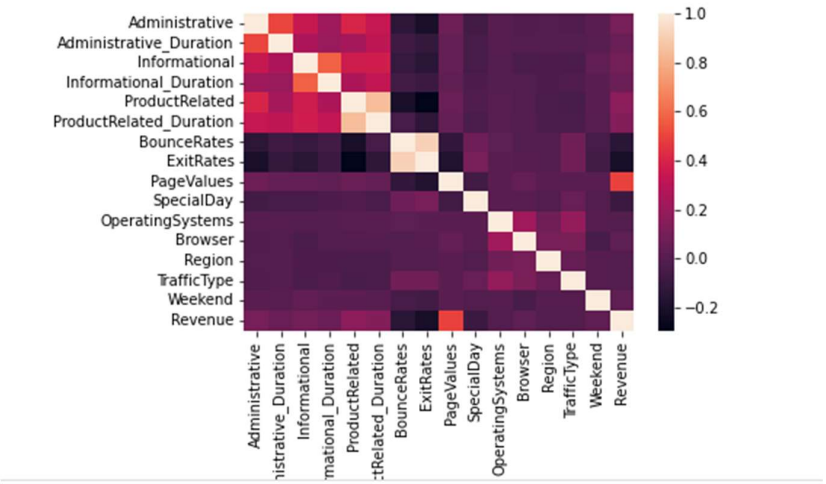


Figura 1 - Matriz de correlação do dataset

De acordo, os resultados do modelo K-NN chegamos à mesma conclusão que em que mais clientes visitam os sites sem comprar nada durante a semana do que ao fim de semana. Suportanto assim a conclusão do modelo de *logistic regression*

Com base nos resultados, não se conseguiu chegar a grandes conclusões.

Provavelmente as variáveis escolhidas não tenham sido as melhores ou deveria-se abranger mais variáveis, para que a homegeneidade fosse maior.

Decidiu-se fazer mais que uma vez o teste com um menor número de *clusters*, pois por defeito o modelo selecionava 8 *clusters*. Ora apesar dos grupos estarem demasiado descriminados isso podia facilmente resultar num problema de *overfiting*.

Apreciação Global do Projeto

Neste projeto cada elemento do grupo foi posto à prova e testamos as nossas capacidades e conhecimento no que toca a aprendizagem automática e algoritmos de *Machine Learning*.

Devido ao *missing data* nas colunas e à baixa normalização do *dataset* sentiu-se alguma dificuldade ao trabalhar com este documento. Partilhou-se da opinião que o *dataset* podia estar mais completo e mais detalhado.

Por outro lado, obrigou-nos a procurar/relembrar e aplicar todo o conhecimento e que nos foi trasmitido nas aulas e até mesmo conhecimento de outras unidades curriculares conseguimos aplicar.

Referências

Cluster_analysis. (10 de junho de 2022). Obtido de Wikipedia: https://en.wikipedia.org/wiki/Cluster_analysis

Dwivedi, R. (5 de julho de 2021). *How Does K-nearest Neighbor Works In Machine Learning Classification Problem?* Obtido de analyticsteps: https://www.analyticssteps.com/blogs/how-does-k-nearest-neighbor-works-machine-learning-classification-problem

Estima, J. (2022). *Aprendizagem Automática*. Obtido de Moodle: https://moodle.ips.pt/2122

Luz, F. (21 de fevereiro de 2019). *ALGORITMO KNN PARA CLASSIFICAÇÃO*. Obtido de inferir: https://inferir.com.br/artigos/algoritimo-knn-para-classificacao/

Robinson, S. (2022). *K-Means Clustering with Scikit-Learn*. Obtido de StackAbuse: https://stackabuse.com/k-means-clustering-with-scikit-learn/