Comparação de Algoritmos de Categorização

Categorizar um veículo a partir das suas informações

Comparation of Categorization Algorithms

Categorize a vehicle based on its information

Martinho Caeiro - 23917 —— Paulo Abade - 23919
Instituto Politécnico de Beja
Escola Superior de Tecnologia e Gestão
Beja, Portugal
23917@stu.ipbeja.pt —— 23919@stu.ipbeja.pt

***Resumo*—Este artigo apresenta um estudo para a comparação entre algoritmos de categorização de veículos. O objetivo de cada um dos algoritmos é categorizar um veículo a partir das suas informações, sendo que este veículo será categorizado consoante o seu país de origem. Foi escolhido este tema para facilitar a nossa compreensão sobre o assunto e tornar mais agradável o estudo destes algoritmos. O estudo foi realizado com base em algoritmos de aprendizagem supervisionada, nomeadamente o algoritmo *Binary Tree*, algoritmo *Random Forest* TODO: Escolher mais algoritmos para fazer a comparação entre eles. Para fazer a comparação destes, foi utilizado um dataset, com cerca de 400 entradas, onde possui informações de veículos de diferentes países e este foi utilizado nos diferentes algoritmos para treinar e testar os mesmos. Este dataset, já foi alterado no primeiro trabalho da Disciplina de *Sistemas de Informação*, sendo ligeiramente diferente, por já ter sido tratado. Os resultados obtidos foram comparados e analisados para perceber qual o algoritmo que melhor categoriza um veículo a partir das suas informações. A comparação dos algoritmos foi feita com base na sua precisão, sensibilidade e especificidade. TODO: Adicionar mais informações sobre o estudo realizado. Gráficos ROC?**

***Index Terms*—algoritmos; veículos; categorização; aprendizagem supervisionada; árvore binária; random forest; precisão; sensibilidade; especificidade; orange; datamining; machine learning; kaggle.**

I. Introdução

Neste artigo temos como objetivo comparar diferentes algoritmos de categorização de modo a entender qual o algoritmo mais eficiente para a questão que está a ser feita. Para isso, foi utilizado um dataset com informações de veículos de diferentes países, onde o objetivo é determinar o pais de origem dado as informações do veículo. Este dataset foi utilizado para treinar e testar os diferentes algoritmos de categorização, nomeadamente os algoritmos *Tree*, *Random Forest*, *Logistic Regression* e *Neural Network*.

II. Dataset

O dataset "Car information dataset"[1] utilizado neste estudo foi retirado do site *Kaggle* e contém informações de veículos de diferentes países. Estas informações incluem a marca/modelo, a economia de combustivel, o número de cilindros, a cilindrada, a potência, o peso, a aceleração, o ano de fabrico e o país de origem. Este dataset possui cerca de 400 entradas em cada uma das colunas.

III. Algoritmos de Decisão

Nesta secção, vamos apresentar os diferentes algoritmos de decisão utilizados para a categorização dos veículos.

A. Tree

É um modelo baseado numa estrutura hierárquica em forma de árvore. Cada nó representa uma condição ou regra (geralmente um atributo do conjunto de dados), e os ramos dividem os dados com base nessa regra. O objetivo é chegar a uma decisão ou classificação no final de cada ramo (folha). É simples, interpretável e útil para problemas de classificação e regressão.

Atual/Previsão	Europeu	Japão	EUA	Total
Europeu	50	12	6	68
Japão	16	57	6	79
EUA	12	12	221	245
Total	78	81	233	392

Tabela I. **Matriz de Confusão do Algoritmo Tree**

B. Random Forest

Este algoritmo é um conjunto de árvores de decisão. Cria várias árvores independentes, cada uma treinada com um subconjunto dos dados e das features (atributos) selecionados aleatoriamente. No final, combina os resultados (por votação, na classificação, ou pela média, na regressão) para melhorar a precisão e reduzir o risco de overfitting, comparado a uma única árvore.

Atual/Previsão	Europeu	Japão	EUA	Total
Europeu	39	15	14	68
Japão	11	57	11	79
EUA	10	12	223	245
Total	60	84	248	392

Tabela II. **Matriz de Confusão do Algoritmo Tree**

C. Logistic Regression

Apesar do nome, é um método usado principalmente para classificação. Modela a probabilidade de um resultado pertencente a uma classe específica, usando uma função logística. É simples, rápido e eficaz em problemas de classificação binária, embora também possa ser estendido para múltiplas classes.

Atual/Previsão	Europeu	Japão	EUA	Total
Europeu	32	27	9	68
Japão	15	51	13	79
EUA	6	19	220	245
Total	53	97	242	392

Tabela III. **Matriz de Confusão do Algoritmo Logistic Regression**

D. Neural Network

Inspiradas pelo cérebro humano, consistem em camadas de ”neurónios”interligados. Cada neurónio recebe entradas, aplica uma ponderação e uma função de ativação, e passa o resultado para os neurónios da camada seguinte. São altamente versáteis e podem lidar com problemas complexos, como reconhecimento de imagens ou processamento de linguagem natural, mas requerem mais dados e poder computacional.

Atual/Previsão	Europeu	Japão	EUA	Total
Europeu	34	18	16	68
Japão	17	45	17	79
EUA	10	16	219	245
Total	61	16	219	392

Tabela IV. **Matriz de Confusão do Algoritmo Neural Network**

IV. Comparações finais

Como podemos visualizar na tabela e figuras abaixo, o algoritmo *Random Forest* obteve os melhores resultados AUC, e o algoritmo Tree obteve os melhores resultados em termos de precisão, sensibilidade e especificidade.

Algoritmo	AUC	CA	F1	Precision	Recall	MCC
Tree	0.882	0.837	0.840	0.846	0.837	0.706
Random Forest	0.934	0.814	0.812	0.812	0.814	0.652
Logistic Regression	0.914	0.773	0.773	0.779	0.773	0.582
Neural Network	0.912	0.760	0.757	0.755	0.760	0.548

Tabela V. **Comparação de Resultados dos Algoritmos**

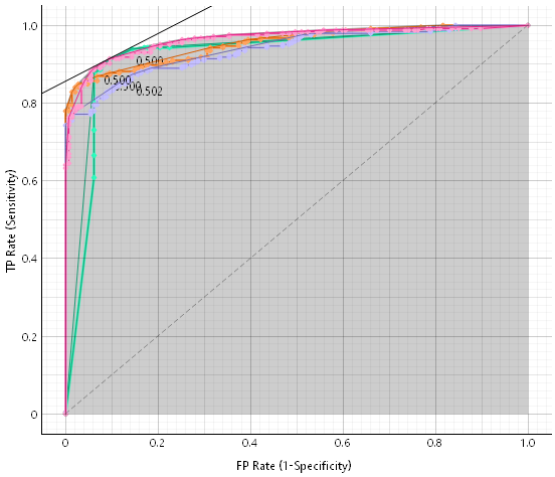

Figura 1: Curva ROC dos Algoritmos - EUA

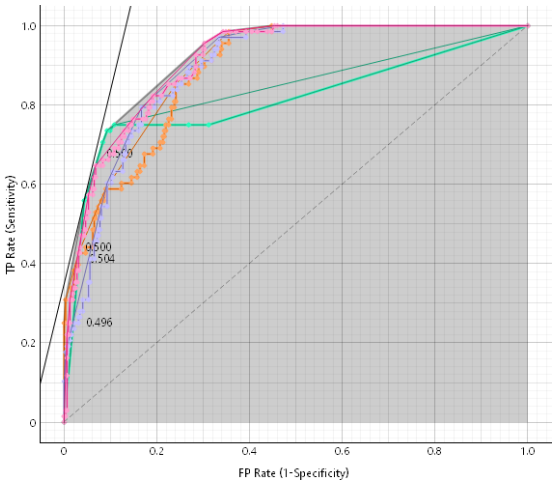

Figura 2: Curva ROC dos Algoritmos - Europa

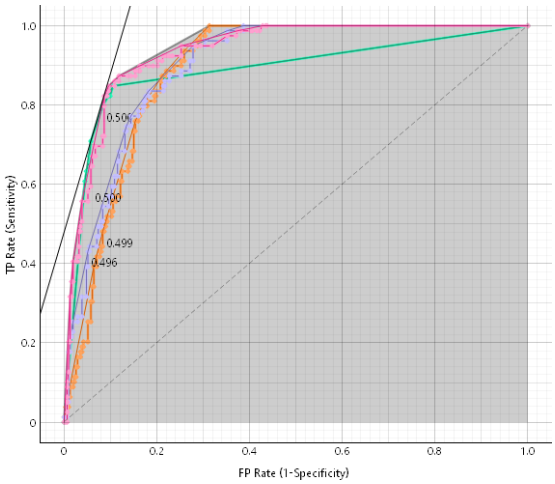

Figura 3: Curva ROC dos Algoritmos - Japão

V. Conclusões

Concluimos que a escolha do algoritmo correto é de extrema importância dado que a precisão, sensibilidade e especificidade deve ser garantida e que a sua eficácia em grande escala tem um grande impacto no resultado final.

VI. Trabalhos Relacionados

Os trabalhos ”Automobile EDA”[2] e ”EDA CAR INFORMATION DATA”[3] são exemplos de trabalhos relacionados com este dataset. Estes trabalhos tem outras abordagens e análises ao dataset, que tem como objetivo analisar as diferentes caracteristicas dos veículos. em vez de diferentes algoritmos, mesmo assim, são trabalhos foram úteis para uma compreensão mais aprofundada do dataset.

Referências

[1] T. Elmetwally, “Car information dataset”, *Kaggle*, Maio 2023.
[2] A. Aboraida, “Automobile EDA”, *Kaggle*, Setembro 2024.
[3] V. Salodkar, “EDA CAR INFORMATION DATA”, *Kaggle*, Junho 2023.