

TCC - CIÊNCIA DE DADOS E BIG DATA

CATEGORIZAÇÃO DE PRODUTOS EM MARKETPLACES

1

ENTENDENDO O PROBLEMA

O intuito deste trabalho foi criar um modelo para **categorizar produtos vendidos em marketplaces, com modelagem de algoritmos de Machine Learning (ML)** e, dessa forma, automatizar o processo de categorização de itens vendidos pela empresa *Olist Store*, **sem depender de ações manuais**.

2

COLETA DE DADOS

As informações de produtos vendidos na *Olist Store* foram coletadas da plataforma *Kaggle* (comunidade voltada para ciência de dados). O conjunto de dados utilizados contém informações como, **peso**, **dimensões** e **preço** de produtos diversos, bem como sua **categoria**.

3

TRATAMENTO DE DADOS

Antes de iniciar a modelagem dos algoritmos de ML foi **importante para obtenção de bons resultados**:

- concatenar bases de dados;
- eliminar colunas insignificantes;
- eliminar produtos duplicados;
- eliminar dados faltantes;
- alterar unidades;
- renomear dados e colunas;
- remover outliers;
- redimensionar os dados.

4

ANÁLISE DE DADOS

Após o tratamento dos dados foi construída uma **base balanceada e robusta** com **2.035** produtos da categoria **"cama, mesa e banho"** e **2.027** produtos da categoria **"beleza e saúde"**.

Atributo	categoria 0 cama_mesa_banho		categoria 1 beleza_saude	
	De	Até	De	Até
peso_kg	0,31	2,29	-0,07	1,59
comprimento_cm	22,29	42,27	15,7	27,58
altura_cm	5,4	16,68	7,21	19,15
largura_cm	18,61	34,13	12,14	21,04
preco	29,3	134,1	26,21	142,57

5

MODELAGEM DE ML

O *XGBoost* apresentou a melhor acurácia, seguido por **Árvores de Decisão**. A performance destes modelos foi acentuada com otimização Bayesiana. O **melhor resultado obtido foi com o XGBoost ajustado**.

Algoritmo	Classificador	Acurácia	Erro
Árvores de Decisão	cif_dt=tree.DecisionTreeClassifier()	88,85	11,15
Árvores de Decisão	dt_bayes_search.best_estimator_	89,13	10,87
Regressão Logística	cif_lr = LogisticRegression()	80,44	19,56
Naive Bayes	cif_nb = GaussianNB()	78,72	21,28
XGBoost	cif_xgb = xgb.XGBClassifier()	92,72	7,28
XGBoost	xgb_bayes_search.best_estimator_	92,75	7,25

6

AVALIAÇÃO DOS RESULTADOS

O **modelo campeão** deste projeto foi o XGBoost ajustado, que **conseguiu prever corretamente a categoria de 93,52% dos produtos da base de teste**. Isso representa uma boa acurácia.

Matriz de confusão

