Classificação de Cogumelos (Mushroom UCI) usando modelos de aprendizado de máquina: proposta experimental e comparação com a literatura

Matheus Henrique, João Guilherme, Renan Aprígio, Arthur Tomé

Curso Integrado de Desenvolvimento de Sistemas – Instituto Federal de Educação, Ciência e Tecnologia de Pernambuco (IFPE - Campus Jaboatão dos Guararapes)

jglf@discente.ifpe.edu.br, atx@discente.ifpe.edu.br, radm@discente.ifpe.edu.br, mham4@discente.ifpe.edu.br

Orientador: Luciano de Souza

Resumo.

Este trabalho apresenta uma solução experimental para a tarefa de classificação binária (comestível vs. venenoso) sobre o dataset Mushroom (UCI). Implementamos um pipeline reprodutível de pré-processamento (One-Hot Encoding, escalonamento quando necessário) e avaliamos sete modelos: GaussianNB, Logistic Regression, Decision Tree, KNN, SVM, RandomForest e XGBoost. As medições foram obtidas via validação estratificada (Stratified K-Fold) e as métricas comparadas com resultados reportados na literatura recente. Observou-se que modelos baseados em árvore/ensemble (Decision Tree, RandomForest, XGBoost) atingiram precisão perfeita (1.000), enquanto modelos probabilísticos e lineares apresentaram desempenho inferior (GaussianNB ≈ 0.922; LogisticRegression ≈ 0.948). Discutimos por que esses resultados concordam com trabalhos anteriores, implicações sobre viés/overfitting e sugestões para validação adicional.

Abstract.

This paper reports an experimental study on the Mushroom (UCI) dataset. We implemented a reproducible pipeline and evaluated seven classifiers (GaussianNB, Logistic Regression, Decision Tree, KNN, SVM, RandomForest, XGBoost) using stratified cross-validation. Tree-based and ensemble methods reached perfect precisão (1.000). We compare our findings to the literature and discuss dataset characteristics that explain the results and propose further robustness checks.

1. Introdução

A identificação automática de cogumelos como comestíveis ou venenosos tem sido um problema padrão em estudos introdutórios e comparativos de aprendizado de máquina devido à disponibilidade do Mushroom dataset (UCI). Diversos trabalhos relatam desempenhos muito altos, especialmente para classificadores baseados em árvores e ensembles. Este trabalho apresenta uma solução experimental completa (pré-processamento, pipeline, sete modelos), reporta métricas relevantes e compara os resultados empíricos obtidos com os achados da literatura recente.

2. Trabalhos Relacionados

Estudos comparativos e relatórios aplicados sobre o Mushroom frequentemente mostram que modelos de árvore e ensembles alcançam acurácias em torno de 98–100%; modelos lineares e probabilísticos normalmente ficam abaixo desses valores (90–96%); KNN e SVM costumam apresentar resultados próximos aos ensembles quando o pré-processamento é adequado.

3. Materiais e Métodos

3.1. Dataset

Usou-se o dataset Mushroom da UCI (8.124 instâncias, 22 atributos categóricos + rótulo class), com classes edible e poisonous. Atributos notáveis incluem odor, spore-print-color e gill-size.

3.2. Pré-processamento

Leitura do CSV bruto; tratamento de valores faltantes (dataset padrão não tem valores ausentes significativos); One-Hot Encoding para todas as variáveis categóricas; escalonamento aplicado quando o classificador exige (SVM, LogisticRegression); validação com StratifiedKFold (n_splits=5, shuffle=True, random_state=42).

3.3. Modelos avaliados

Gaussian Naive Bayes; Logistic Regression; Decision Tree; K-Nearest Neighbors; Support Vector Machine; Random Forest (n_estimators=200); XGBoost.

3.4. Métricas

precisão, Precisão (macro), Recall (macro) e F1-Score (macro). Métricas calculadas a partir das pontuações médias em validação cruzada estratificada (5 folds).

4. Resultados

Os resultados médios obtidos nos experimentos foram os seguintes (ordenados pelos valores fornecidos):

Tabela 1. Resultados médios por modelo (Precisão, Recall e F1-Score)

Modelo	precisão	Precisão	Recall	F1-Score
DecisionTree	1.000000	1.000000	1.000000	1.000000
RandomForest	1.000000	1.000000	1.000000	1.000000
XGBoost	1.000000	1.000000	1.000000	1.000000
KNN	0.996308	0.992386	1.000000	0.996178
SVM	0.992615	0.998705	0.985934	0.992278
LogisticRegression	0.947692	0.943949	0.947570	0.945756
GaussianNB	0.921846	0.909887	0.929668	0.919671

5. Discussão — Comparação com a Literatura

As principais concordâncias com a literatura: árvores e ensembles alcançaram precisão perfeita, como relatado em diversos estudos; KNN e SVM obtiveram desempenho muito alto (≥99%) quando o pré-processamento foi adequado; logistic regression e gaussianNB apresentaram desempenho inferior (≈92–95%).

Interpretação crítica: o dataset é altamente solucionável, o que explica os resultados perfeitos; recomenda-se cautela quanto a overfitting e verificação de data leakage. Sugestões para robustez: validação externa, remoção de atributos altamente informativos (ex.: odor), análise de importância de features (SHAP), experimentos com ruído sintético.

6. Conclusões

Implementamos um pipeline reprodutível e avaliamos sete classificadores. Decision Tree, RandomForest e XGBoost atingiram 100% de precisão, corroborando a literatura. KNN e SVM performaram >99% e modelos lineares/probabilísticos ficaram em 92–95%. Recomendamos análises adicionais para testar robustez e generalização.

7. Referências (seleção)

UCI Machine Learning Repository — Mushroom Dataset.
Wagner, D.; Heider, D.; Hattab, G. (2021). Mushroom data creation, curation and simulation to support classification tasks.
Tarawneh, O.; Tarawneh, M.; Sharrab, Y.; Husni, M. (2022). Mushroom classification using machine-learning techniques.
Paudel, S.; Bhatta, R. (2022). Comparative study: Random Forest and other classifiers on Mushroom dataset.