RELATÓRIO DE ANÁLISE - MINERAÇÃO DE DADOS

Seleção de Atributos para Classificação e Regressão

RESUMO EXECUTIVO

Este relatório apresenta a aplicação de técnicas de seleção de atributos no Wine Quality Dataset para tarefas de classificação e regressão. Foram implementados e comparados quatro métodos de seleção: ANOVA, Mutual Information, RFE e Feature Importance, utilizando Random Forest como algoritmo base. Os resultados demonstram que a seleção adequada de atributos mantém ou melhora a performance dos modelos enquanto proporciona ganhos significativos em interpretabilidade e eficiência computacional.

INTRODUÇÃO

Contexto do Problema

A seleção de atributos é uma etapa crucial no processo de mineração de dados, permitindo identificar as variáveis mais relevantes para a construção de modelos preditivos. Este trabalho aborda duas tarefas distintas:

Classificação: Prever se um vinho é de alta qualidade (≥7) ou não

Regressão: Prever o valor de pH do vinho

Objetivos

Aplicar e comparar métodos de seleção de atributos

Avaliar impacto na performance preditiva

Analisar ganhos em interpretabilidade

Verificar eficiência computacional

METODOLOGIA

Dataset e Pré-processamento

Características do Dataset:

Fonte: UCI Machine Learning Repository

Amostras: 1.599 instâncias iniciais

Atributos: 11 features físico-químicas

Duplicatas removidas: 240 instâncias

Dataset final: 1.359 instâncias

Pré-processamento aplicado:

Remoção de duplicatas

Normalização com StandardScaler

Balanceamento com SMOTE (apenas classificação)

Criação de variável target binária para classificação

Métodos de Seleção Implementados

Método	Descrição	Parâmetros
ANOVA F-value	Seleção baseada em teste estatístico	k=5 features
Mutual Information	Seleção baseada em teoria da informação	k=5 features
RFE	Eliminação recursiva de features	n_features=5
Feature Importance	Baseado na importância do Random Forest	threshold=percentil 70

Configuração Experimental

Divisão treino/teste: 70%/30%

Validação cruzada: 5-fold

Algoritmo base: Random Forest (100 árvores)

Métricas:

Classificação: Acurácia, Precision, Recall, F1-Score

Regressão: MSE, R^2

ANÁLISE DE CLASSIFICAÇÃO

Distribuição das Classes

Classe	Descrição	Quantidade	Proporção
0	Vinho ruim (qualidade < 7)	1.175	86.46%
1	Vinho bom (qualidade ≥ 7)	184	13.54%

Observação: Dataset significativamente desbalanceado, necessitando aplicação de SMOTE.

Resultados dos Métodos de Seleção

Método	Acurácia	Nº Features	Precision (Classe 1)	Recall (Classe 1)
Baseline	87.75%	11	0.58	0.35
Mutual Information	*88.97%*	*5*	*0.62*	*0.42*
Feature Importance	87.25%	5	0.54	0.38
ANOVA	87.25%	5	0.54	0.38
RFE	87.25%	5	0.54	0.38

Análise Detalhada do Melhor Modelo

Método vencedor: Mutual Information

Relatório de Classificação Detalhado:

	PRECISION	RECALL	F1-SCORE	SUPPORT
0	**0.91**	**0.96**	**0.93**	**353**
1	**0.62**	**0.42**	**0.50**	**55**
ACCURACY			**0.89**	**408**
MACRO AVG	**0.76**	**0.69**	**0.72**	**408**
WEIGHTED AVG	**0.87**	**0.89**	**0.88**	**408**

Features Selecionadas:

1. Álcool
2. Sulfatos
3. Acidez Volátil
4. Densidade
5. Ácido Cítrico

Matriz de Confusão - Modelo Final

	Predito 0	Predito 1
Real 0	**338**	**15**
Real 1	**32**	**23**

Interpretação:

95.8% dos vinhos ruins corretamente classificados

41.8% dos vinhos bons corretamente identificados

Melhoria significativa na classe minoritária comparado ao baseline

ANÁLISE DE REGRESSÃO

Estatísticas da Variável Target (pH)

Estatística	Valor
Média	**3.311**
Desvio Padrão	**0.154**
Mínimo	**2.740**
Máximo	**4.010**
Mediana	**3.310**

Resultados dos Métodos de Seleção

Método	R^2	MSE	Nº Features
Baseline	98.81%	0.0003	10
Feature Importance	**99.46%**	**0.0001**	**4**
Mutual Information	98.92%	0.0002	5
RFE	98.85%	0.0002	5
F-Regression	62.57%	0.0089	5

Análise do Melhor Modelo

Método vencedor: Feature Importance

Features Selecionadas:

Acidez Fixa (99.87% de importância)

Densidade

Álcool

Acidez Volátil

Performance Detalhada:

R^2: 99.46% (melhoria de 0.65% sobre baseline)

MSE: 0.0001 (redução de 66.7% no erro)

Validação Cruzada: 99.32% ± 0.08%

Análise de Resíduos

Distribuição normal dos resíduos

Homocedasticidade mantida

Sem padrões aparentes nos resíduos vs preditos

Modelo demonstra excelente ajuste aos dados

ANÁLISE COMPARATIVA

Performance Geral

Métrica	*Baseline*	*Com Seleção*	*Variação*
Acurácia (Classificação)	*87.75%*	*88.97%*	***+1.22%***
R^2 (Regressão)	*98.81%*	*99.46%*	***+0.65%***
Nº Features Médio	*10.5*	*4.8*	***-54.3%***

Eficiência Computacional

Tarefa	*Tempo Baseline*	*Tempo com Seleção*	*Redução*
Classificação	*2.34s*	*1.41s*	***39.7%***
Regressão	*2.18s*	*1.25s*	***42.7%***

Features Mais Relevantes por Tarefa

Classificação (Qualidade):

Álcool (17.6%) - Correlação positiva com qualidade

Sulfatos (13.0%) - Relacionado a preservação

Acidez Volátil (10.9%) - Correlação negativa com qualidade

Regressão (pH):

Acidez Fixa (99.9%) - Determinante principal do pH

Densidade - Relacionada à composição

Álcool - Influencia propriedades químicas

DISCUSSÃO E INSIGHTS

Impacto da Seleção de Atributos

Vantagens Identificadas:

Manutenção da Performance: Redução de ~55% nas features com manutenção ou melhoria das métricas

Interpretabilidade: Identificação clara dos fatores mais relevantes

Eficiência: Redução de ~40% no tempo de treinamento

Robustez: Menor propensão a overfitting

Limitações:

Performance na classe minoritária ainda desafiadora

Métodos diferentes selecionam features diferentes

Dependência do algoritmo base para alguns métodos

Insights do Domínio

Para Produção de Vinhos de Alta Qualidade:

Focar em teor alcoólico adequado

Controlar níveis de sulfatos

Monitorar acidez volátil (impacto negativo)

Para Controle de pH:

Acidez fixa como principal controlador

Densidade como indicador secundário

Teor alcoólico com influência moderada

CONCLUSÕES

Principais Conclusões

Seleção Efetiva: Técnicas de seleção permitiram reduzir significativamente o número de features mantendo a performance

Métodos Específicos por Tarefa:

Classificação: Mutual Information mais eficaz

Regressão: Feature Importance superior

Ganhos Tangíveis: 40% de redução no tempo de treinamento com manutenção da acurácia

Interpretabilidade: Identificação clara das features mais impactantes

REFERÊNCIAS

UCI Machine Learning Repository - Wine Quality Dataset

Scikit-learn Documentation

IMBLearn Documentation - SMOTE

Hastie, T., et al. - "The Elements of Statistical Learning"

RELATÓRIO GERADO POR: Samir Lopes Rosa

DISCIPLINA: Mineração de Dados