Relatório Técnico - IFSP: Classificação de Nomes Brasileiros via Machine Learning

Autor: Augusto Dos Santos Souza - GU3020207

Professor: Rodrigo Bortoletto

Disciplina: Tópicos Especiais 2

Data: 07/12/2024

Introdução

Este relatório técnico descreve um estudo sobre a classificação de gêneros de nomes brasileiros utilizando técnicas de Machine Learning. O objetivo é automatizar a identificação do gênero de um nome, tarefa útil em diversas áreas, como marketing, sociologia e linguística. O projeto foi inspirado pela disponibilidade de dados relevantes no Brasil, obtidos do [Brasil.io], e pelo desejo de aplicar conhecimentos teóricos em um problema prático.

Base de Dados

A base de dados utilizada neste estudo é composta por dois arquivos CSV que foram combinados em um único dataframe:

- grupos.csv: Contém informações sobre grupos de nomes com características comuns, como frequência de uso e proporção de nomes masculinos e femininos.
- nomes.csv: Contém uma lista de nomes brasileiros, com informações sobre sua frequência de uso por gênero e a classificação do nome (masculino (M) ou feminino (F)).

A base de dados foi obtida do [Brasil.io] e contém um total de 100.787 registros. Uma análise estatística preliminar revelou que 45.533 nomes são classificados como masculinos e 55.254 como femininos. Portanto, a base de dados apresenta um 45,1% de nomes masculinos e 54,9% de nomes femininos.

Descrição das Variáveis

Variável	Tipo	Descrição	Independência
nome	Qualitativa	Nome próprio	Independente

classificacao	Qualitativa	Classificação do nome por gênero (M - Masculino, F - Feminino)	Dependente
frequencia_feminina	Quantitativa	Frequência de uso do nome por mulheres	Independente
frequencia_masculina	Quantitativa	Frequência de uso do nome por homens	Independente
frequencia_total	Quantitativa	Frequência de uso total do nome	Independente
nome_grupo	Qualitativa	Grupo de nomes com características comuns	Independente
nomes_alternativos	Qualitativa	Nomes que podem ser considerados sinônimos	Independente
porcentagem_feminina	Quantitativa	Porcentagem de uso do nome por mulheres	Independente
porcentagem_masculina	Quantitativa	Porcentagem de uso do nome por homens	Independente

proporcao	Quantitativa	Proporção de uso do nome por gênero (1.0 indica que o nome é usado apenas por um gênero)	Independente

Pré-processamento dos Dados

O pré-processamento dos dados envolveu as seguintes etapas:

1. Limpeza de dados nulos: Os dados nulos (NaN) foram preenchidos com 0, considerando que a ausência de informação sobre a frequência de uso de um gênero em um nome indica que este nome não é usado para aquele gênero.
2. Remoção de duplicatas: Duplicatas foram removidas dos datasets para evitar o enviesamento dos modelos.
3. Criação de novas colunas: Foram criadas colunas binárias para cada posição do nome (LETRA_1_A, LETRA_2_B, etc.), representando a presença de cada letra do alfabeto em cada posição. Essa etapa visou representar os nomes como vetores numéricos para entrada nos modelos de Machine Learning.
4. Label Encoding: A coluna de classificação ("classificacao") foi codificada utilizando o LabelEncoder, transformando as categorias (masculino (M) e feminino (F)) em números. Essa transformação é necessária para o treinamento de modelos de Machine Learning.

Modelos de Machine Learning

Foram utilizados diversos modelos de Machine Learning para a classificação de nomes:

- Regressão Logística: É um modelo linear que calcula a probabilidade de um nome ser masculino ou feminino.
- KNN (k-Nearest Neighbors): É um modelo baseado na distância entre os nomes. O modelo prevê o gênero de um nome com base nos gêneros dos k nomes mais próximos.
- Naive Bayes (GaussianNB): É um modelo probabilístico que calcula a probabilidade de um nome ser masculino ou feminino com base na frequência das letras do nome.
- Random Forest: É um modelo baseado em árvores de decisão. O modelo agrupa diversas árvores, cada uma treinada com um subconjunto aleatório dos dados, para obter uma previsão mais robusta.
- Redes Neurais: É um modelo inspirado no cérebro humano, com diversas camadas de neurônios interconectados. O modelo aprende padrões complexos nos dados e é capaz de realizar previsões precisas.
- Bagging de Redes Neurais: Utilizando a técnica de bagging, múltiplos modelos de redes neurais foram treinados com amostras parte das dados, e suas previsões são combinadas para gerar uma previsão final.

- Boosting (AdaBoost): É um modelo que combina diversos modelos fracos (por exemplo, modelos de Regressão Logística) para formar um modelo forte. O modelo corrige os erros dos modelos anteriores e foca mais nos dados que foram classificados incorretamente.
- Stacking: É um modelo que combina diversos modelos de Machine Learning. O modelo final usa as previsões dos modelos base para fazer a previsão final.

Devido ao grande volume de dados e ao baixo poder de processamento disponível, não foi possível testar modelos como SVM.

Validação Cruzada

A validação cruzada (k-fold) foi utilizada para avaliar a generalização dos modelos de Machine Learning. Essa técnica consiste em dividir o conjunto de dados em k partes (folds) e treinar o modelo em k-1 folds, utilizando o fold restante para avaliação. O processo é repetido k vezes, utilizando cada fold como conjunto de teste. A média dos resultados obtidos em cada fold é utilizada para avaliar a performance geral do modelo.

Resultados

Os modelos foram avaliados utilizando as seguintes métricas:

- F1-score: Métricas que combinam a precisão e o recall do modelo.
- Acurácia: Proporção de previsões corretas.
- Matriz de Confusão: Mostra a quantidade de previsões corretas e incorretas para cada gênero.
- Relatório de Classificação: Fornece informações detalhadas sobre a precisão, recall e F1-score para cada gênero.

A tabela abaixo apresenta um resumo dos resultados obtidos com os diferentes modelos, utilizando a informação que você forneceu:

Modelo	F1-Score	Acurácia	Falsos Positivos	Falsos Negativos
Regressão Logística	0.865	0.87	1097	1489
KNN	0.875	0.88	1046	1371
Naive Bayes	0.655	0.70	141	5877
Random Forest	0.938	0.94	479	694
Redes Neurais	0.99	0.95	694	575

Bagging de Redes Neurais	0.953	0.95	692	722
Boosting (AdaBoost)	0.805	0.81	1710	4002
Stacking	0.940	0.94	853	934

Os resultados indicam que o modelo de Redes Neurais obteve o melhor desempenho, com um F1-score de 0.99 e uma acurácia de 0.95. A análise da matriz de confusão e do relatório de classificação revela que o modelo apresenta um baixo número de falsos negativos, indicando que ele é capaz de identificar a maioria dos nomes femininos corretamente.

Considerações Finais

O estudo demonstrou que técnicas de Machine Learning podem ser utilizadas para classificar o gênero de nomes brasileiros com precisão. O modelo de Redes Neurais apresentou os melhores resultados e pode ser utilizado para automatizar a identificação do gênero de nomes em diversas aplicações.

Limitações:

- A base de dados utilizada é limitada a nomes brasileiros e pode não ser representativa de outras culturas.
- A base de dados é baseada em estatísticas de frequência de uso de nomes, que podem mudar ao longo do tempo.
- O modelo não é capaz de identificar nomes neutros ou de gêneros não binários.

Sugestões para trabalhos futuros:

- Expandir a base de dados para incluir nomes de diferentes culturas e idiomas.
- Investigar o uso de técnicas de Processamento de Linguagem Natural (PNL) para melhorar a precisão do modelo.
- Explorar a utilização de modelos de aprendizado profundo (Deep Learning) para a classificação de nomes.
- Implementar o modelo em uma aplicação web ou API, tornando a solução mais acessível para usuários.