Previsão de probabilidade de diabetes em estágio inicial usando técnicas de mineração de dados

Jonas v. Moreira

TP555 - Inteligência Artificial e Aprendizado de Máquina
Inatel - Instituto Nacional de Telecomunicações

Introdução

Problema Global da Diabetes

- **Aumento da Incidência da Doença**
  - Mais de **422 milhões de pessoas** no mundo têm diabetes.
  - Principalmente **diabetes tipo 2**, ligada ao estilo de vida (dieta e sedentarismo).
- **Consequências da Diabetes Não Controlada**
  - **Doenças cardiovasculares**: aumento do risco de infarto e AVC.
  - **Insuficiência renal**: danos nos rins, podendo levar à diálise.
  - **Retinopatia diabética**: perda de visão, podendo causar cegueira.
  - **Neuropatia**: danos nervosos, resultando em amputações.
- **Importância do Diagnóstico Precoce**
  - Permite intervenções mais eficazes e controle da doença.
  - Melhora a qualidade de vida dos pacientes e reduz complicações graves.
- **Desafios no Diagnóstico**
  - Métodos tradicionais: exames de glicemia em jejum, com limitações.
  - Dificuldade de levar os exames a areas remotas e/ou rurais.

Objetivos

Objetivos da pesquisa

- Estudar o artigo **Likelihood Prediction of Diabetes at Early Stage Using Data Mining Techniques**
- **Reproduzir os resultados do artigo** e validar a eficácia do Random Forest como o algoritmo mais preciso.
- **Explorar outras técnicas de aprendizado de máquina** utilizando o **AutoGluon** [7], a fim de identificar algoritmos que possam superar o desempenho do Random Forest.
- **Desenvolver uma aplicação prática** que permita predições em tempo real, utilizando uma interface de programação no estilo arquitetural de transferência de estado representacional (API REST) e uma interface gráfica simples.

Estudo do Artigo

Likelihood Prediction of Diabetes at Early Stage Using Data Mining Techniques

O artigo aborda a previsão do risco de diabetes utilizando algoritmos de mineração de dados.

Algoritmos utilizados:

- Random Forest
- Naive Bayes
- Regressão Logística

Conjunto de dados: Coletado em Bangladesh.
Resultados:

- **Random Forest** teve o melhor desempenho:
  - 97,4% de precisão com validação cruzada.
  - 99% de precisão com divisão percentual 80:20.

Explorando outros algoritmos

Objetivo

Explorar algoritmos diferentes do **Random Forest** para encontrar uma alternativa mais eficiente.

Explorando outros algoritmos

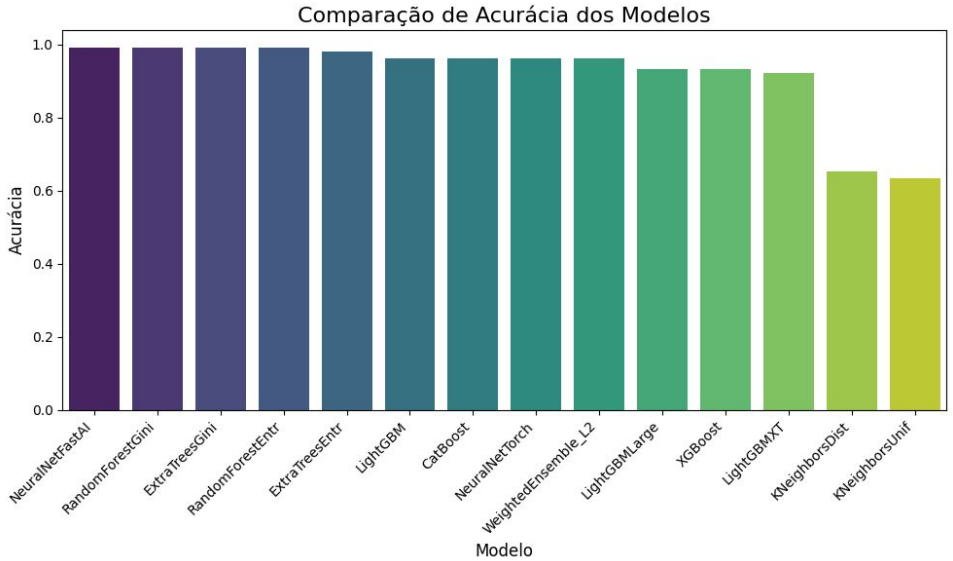
Processo

1. **Carregamento e Pré-processamento dos Dados**:
   - Conjunto de dados de diabetes carregado e convertido em **DataFrame**.
   - Divisão dos dados: 80% para treino e 20% para teste.
2. **Uso do AutoGluon**:
   - **TabularPredictor** configurado com a variável alvo **class** e métrica de **acurácia**.
3. **Treinamento Automatizado**:
   - AutoGluon realiza **seleção automática de modelos** e **otimização de hiperparâmetros**.
4. **Avaliação e Ranking**:
   - Comparação de modelos com base em acurácia, utilizando dados de teste e **ranking detalhado**.

Explorando outros algoritmos

Resultado

NeuralNetFastAI obteve acurácia de **99,04%**, ligeiramente superior ao **Random Forest** (98,85%).



Treinando o Modelo

Objetivo

Treinar um modelo de **Random Forest** e exportá-lo para uso em outras aplicações.

Treinando o Modelo

Processo

1. **Pré-processamento dos Dados**:
   - Conversão de valores categóricos binários ('Yes'/'No') e 'Gender' ('Male'/'Female') para valores numéricos (1 e 0).
2. **Separação dos Dados**:
   - Divisão dos dados em variáveis independentes (**X**) e dependentes (**y**).
   - Dados divididos em 75% para **treinamento** e 25% para **teste**.
3. **Treinamento do Modelo**:
   - Modelo de **Random Forest** com 100 árvores de decisão.
   - O algoritmo é treinado para prever a variável dependente com base nas variáveis independentes.
4. **Exportação do Modelo**:
   - Modelo treinado salvo em um arquivo `.pkl` utilizando **joblib**.
   - Arquivo salvo como **'diabetes_rf_model.pkl'** para uso posterior.

Aplicação Web

Objetivo

Validar o modelo de machine learning com uma aplicação web interativa de diagnóstico de diabetes.

Aplicação Web

Arquitetura do Sistema

1. **Backend**:
   - Usando **Flask** para expor uma API de previsão via requisição HTTP **POST**.
   - Modelo carregado com **joblib** a partir de um arquivo serializado (`diabetes_rf_model.pkl`).
   - Dados recebidos no formato **JSON**, estruturados em um **DataFrame** com **Pandas**.
2. **Frontend**:
   - Interface HTML com **CSS** e **JavaScript** para interação com o usuário.
   - Formulário coleta dados (idade, gênero, sintomas) e envia via requisição assíncrona.
   - Exibição do resultado de previsão: diagnóstico positivo ou negativo para diabetes.

Funcionalidades

- **CORS**: Permite acesso ao backend por clientes externos.
- **Predição**: O modelo retorna a probabilidade de diabetes com base nos sintomas informados.

Resultados

Comparação dos Algoritmos

- **Algoritmos Avaliados**: Naive Bayes, Logistic Regression e Random Forest.
- **Resultados**:
    - **Random Forest**:
        - **97,4% de acurácia** na validação cruzada.
        - **99% de acurácia** na divisão percentual.

Explorando outros algorigmos com o AutoGluon

- Redes neurais, como o **NeuralNetFastAI**
        - **99,04% de acurácia**.

Resultados

Escolha do Modelo

- **Random Forest** foi selecionado para a implementação prática devido à:
  - **Simplicidade**

Validação Prática

- Integração do modelo em **API REST** e **Interface Gráfica** foi eficiente.
- Testes simulados confirmaram a **precisão em tempo real** das previsões.

Discussões

Resultados e Contribuições

- **Robustez do Random Forest**:
  - Confirmada sua eficácia na previsão do risco de diabetes.
- **Exploração do AutoGluon**:
  - Destacou a importância de testar **múltiplas abordagens** para otimizar a precisão do modelo.
- **Aplicabilidade Prática**:
  - **API REST** e **interface gráfica** ampliaram o uso do modelo em triagens clínicas iniciais.

Limitações Identificadas

- **Tamanho e Especificidade do Conjunto de Dados**:
  - Dados pequenos e limitados a uma região, impactando a **generalização** do modelo.
- **Dependência de Sintomas Específicos**:
  - Modelo altamente dependente de **poliúria** e **polidipsia**, o que pode limitar sua eficácia em **casos sem esses sintomas**.

Conclusões

Contribuições do Trabalho

- **Validação dos Resultados**:
  - Reproduziu os resultados do artigo original e expandiu a análise com novas abordagens.
- **Integração Prática**:
  - Demonstra que algoritmos de aprendizado de máquina podem ser usados em **triagem de diabetes** em estágio inicial por meio de **API** e **interface gráfica**.

Pesquisas Futuras

- **Expansão do Conjunto de Dados**:
  - Ampliar a base de dados para melhorar a **generalização** e robustez do modelo.
- **Modelos Híbridos**:
  - Explorar combinações de algoritmos para otimizar resultados.
- **Validação em Ambientes Clínicos**:
  - Validar a aplicabilidade dos modelos em cenários **clínicos reais**.

Referências

1. M. M. F. Islam, R. Ferdousi, S. Rahman, and H. Y. Bushra, "Likelihood Prediction of Diabetes at Early Stage Using Data Mining Techniques," in *Proceedings of the 2020 International Conference on Data Mining and Big Data*, 2020. Available at: https://doi.org/10.1007/978-981-15-1910-9 4.
2. Organização Mundial da Saúde (OMS), "Diabetes," Available at: https://www.who.int/news-room/fact-sheets/detail/diabetes.
3. AutoGluon Team, "AutoGluon: AutoML for Text, Image, and Tabular Data," Available at: https://auto.gluon.ai, 2023.
4. F. Pedregosa, et al., "Scikit-learn: Machine Learning in Python," *Journal of Machine Learning Research*, vol. 12, pp. 2825-2830, 2011. Available at: https://scikit-learn.org.
5. Flask Documentation, "Flask: Web Application Framework," Available at: https://flask.palletsprojects.com, 2023.
6. Joblib Team, "Joblib: Lightweight Pipelines in Python," Available at: https://joblib, 2023.
7. AutoGluon Team, "AutoGluon: AutoML for Text, Image, and Tabular Data," Available at: https://auto.gluon.ai, 2023.