

Trabalho 2 - Data Science

Introdução

O objetivo deste trabalho é construir um modelo preditivo utilizando as técnicas de classificação ou regressão. O modelo deve ser capaz de prever a gravidade dos acidentes ou identificar padrões de risco, a fim de possibilitar ações preventivas.

Metodologia

A metodologia para o desenvolvimento do modelo preditivo foi dividida nas seguintes etapas:

Pré-processamento dos Dados:

Limpeza de Dados: Dados ausentes ou inconsistentes foram tratados (drop).

Transformação das Variáveis: Variáveis categóricas, como "Tipo de Acidente" e "Condições Climáticas", foram convertidas para variáveis numéricas utilizando técnicas como one-hot encoding, também foi usado o label encoding quando precisou de um valor ordinal.

Seleção de Modelos

Foram testados os seguintes tipos de modelos: modelos de classificação e modelos de regressão.

- **Classificação:** O objetivo aqui foi prever a probabilidade de um acidente ser grave ou não, com base nas variáveis fornecidas. Os modelos testados foram:
 - Regressão Logística
 - K-Nearest Neighbors (KNN)
 - Random Forest Classifier
- **Regressão:** O objetivo foi prever o número de vítimas em um acidente. O modelo utilizado foi o Regressão Linear. Embora simples, ele forneceu uma boa linha de base para a análise.

Resultados Obtidos

Modelo de Classificação: Regressão Logística

- **Acurácia: 77.84%**
- **Classification Report:**
 - **Precision: 0.78**
 - **Recall: 1.00**
 - **F1-Score: 0.88**
 - **Classe 0 (não grave):**
 - **Precision: 0.78**
 - **Recall: 1.00**
 - **F1-Score: 0.88**
 - **Support: 5699**
 - **Classe 1 (grave):**
 - **Precision: 0.42**
 - **Recall: 0.01**
 - **F1-Score: 0.01**
 - **Support: 1617**
 - **Acurácia total: 78%**
 - **Macro avg: Precision: 0.60, Recall: 0.50, F1-Score: 0.44**
 - **Weighted avg: Precision: 0.70, Recall: 0.78, F1-Score: 0.68**

Modelo de Classificação: K-Nearest Neighbors (KNN)

- **Acurácia: 73.33%**

- **Classification Report:**

- **Precision: 0.79**
- **Recall: 0.90**
- **F1-Score: 0.84**
- **Classe 0 (não grave):**
 - **Precision: 0.79**
 - **Recall: 0.90**
 - **F1-Score: 0.84**
 - **Support: 5699**
- **Classe 1 (grave):**
 - **Precision: 0.28**
 - **Recall: 0.13**
 - **F1-Score: 0.18**
 - **Support: 1617**
- **Acurácia total: 73%**
- **Macro avg: Precision: 0.53, Recall: 0.52, F1-Score: 0.51**
- **Weighted avg: Precision: 0.67, Recall: 0.73, F1-Score: 0.69**

Modelo de Classificação: Random Forest Classifier

- **Acurácia: 74.85%**
- **Classification Report:**
 - **Precision: 0.79**

- **Recall: 0.93**
- **F1-Score: 0.85**
- **Classe 0 (não grave):**
 - **Precision: 0.79**
 - **Recall: 0.93**
 - **F1-Score: 0.85**
 - **Support: 5699**
- **Classe 1 (grave):**
 - **Precision: 0.31**
 - **Recall: 0.11**
 - **F1-Score: 0.17**
 - **Support: 1617**
- **Acurácia total: 75%**
- **Macro avg: Precision: 0.55, Recall: 0.52, F1-Score: 0.51**
- **Weighted avg: Precision: 0.68, Recall: 0.75, F1-Score: 0.70**

Modelo de Regressão: Regressão Linear

- **Erro Quadrático Médio (MSE): 0.170**
- **Erro Médio Absoluto (MAE): 0.341**
- **Erro Médio Percentual Absoluto (MAPE): 780.37%**

Tentativa com SVM (Support Vector Machine)

O modelo de SVM foi tentado para reduzir a dimensionalidade, mas não obteve sucesso devido à alta lentidão no treinamento.

Discussão dos Resultados

Os modelos de classificação apresentaram resultados variados. O modelo de Regressão Logística, apesar de ter uma boa acurácia global de 78%, apresentou problemas em prever a classe 1 (acidentes graves), com uma taxa de recall muito baixa.

O KNN também mostrou um bom desempenho em termos de recall para a classe 0, mas com baixa precisão para a classe 1.

O Random Forest teve um desempenho equilibrado, mas também não foi eficiente em prever a classe 1.

Pontos Fortes

- Random Forest e Regressão Logística apresentaram boa capacidade de prever acidentes não graves (classe 0).
- A Regressão Linear foi eficaz em prever a variabilidade no número de vítimas, com um MSE relativamente baixo.

Pontos Fracos

- Todos os modelos apresentaram dificuldades em prever a classe 1, com baixa precisão e recall.
- O modelo de SVM não foi viável devido à lentidão no processamento, o que inviabilizou sua aplicação no tempo necessário para o treinamento.

Viabilidade Prática

Apesar das limitações de previsão da classe 1, os modelos têm grande viabilidade prática para prever a gravidade de acidentes em casos não graves, o que pode ser útil para a alocação de recursos em áreas com maior risco de acidentes.

Conclusão

Este trabalho teve como objetivo a construção de um modelo preditivo para identificar padrões e prever a gravidade dos acidentes de trânsito com base em dados históricos.

O modelo de Random Forest Classifier apresentou os melhores resultados gerais, mas há margem para melhorias, especialmente na previsão de acidentes graves. A implementação de modelos preditivos como este tem um grande potencial para aumentar a segurança no trânsito e fornecer insights valiosos para políticas públicas voltadas à redução de acidentes.

A implementação de modelos preditivos para acidentes de trânsito possui um grande potencial de mercado. Com o aumento da demanda por soluções de segurança no trânsito, os modelos baseados em inteligência artificial podem ajudar a prever e mitigar acidentes, direcionando recursos de forma eficiente para áreas de risco.

Enfim, esse tipo de tecnologia pode ser integrado a sistemas de monitoramento de tráfego e seguros, reduzindo custos e aumentando a segurança viária.