Relatório “Modelos de Classificação” Titanic Dataset

[1]Alvaro Cristian da Silva Botelho

[1]MBA Data Science e Big Data – Universidade do Vale dos Sinos (UNISINOS)
Caixa Postal 15.064 – 91.501-970 – Porto Alegre – RS – Brazil

Cristian.ucpel@gmail.com

https://pandas.pydata.org/pandas-docs/stable/groupby.html

1. INTRODUÇÃO

Criminalidade é uma constante em nossas vidas e a utilização de ferramentas de data mining e machine learning pode ser a chave para o entendimento e talvez antecipação de ações criminosas.

O objetivo do Criminal Data Mining é entender os padrões do comportamento buscando prevenir a atividade criminal. Existe uma tendência na aplicação da lei que tenta identificar as áreas de crime potencial fazendo assim com que as abordagens policiais sejam mais assertivas, este movimento é denominado Policiamento Preditivo. Várias ações vêm sendo tomadas em território nacional tentando incorporar Policiamento Preditivo ao cotidiano da população. O Sistema de predição desenvolvido pela Microsoft em parceria com a Prefeitura de New York, *PREDICTA*, já vem sendo utilizado pela Prefeitura de São Paulo com intuito de combater possíveis ameaças e terrorismo.

2. PROBLEMA

Com o passar dos anos a onda de crimes vem aumentando tornando assim um problema alarmante para a sociedade. A necessidade de identificar padrões que possam ajudar na inibição de atividades criminais é uma tendência e será abordada neste projeto.

3. OBJETIVO

I. Como o crime mudou com o passar dos anos;

II. E possível prever onde e quando o crime será cometido;

III. Avaliar a utilização de inteligência artificial para predizer se uma abordagem culminara em uma prisão.

4. HIPÓTESES

I. Os pontos de interesse (e.g. shoppings, mercados) possuem relação com a valorização do imóvel?

II. Imóveis próximos a parques possuem maior valorização que os demais?

III. O perfil de venda dos imóveis possuí relação sazonal?

5. RESULTADOS ESPERADOS

I. Estimar com 80% de certeza a valorização dos imóveis;

II. Identificar perfis de imóveis e seus potenciais locais de valorização.

6. RESTRIÇÕES DO PROJETO

Se houver.

7. PROJETO

Neste trabalho será utilizado o dataset que reflete os incidentes reportados na cidade de Chicago. Incidentes envolvendo assassinatos onde existem dados das vítimas não foram disponibilizados, portanto não fazem parte deste estudo. Os dados foram extraídos do sistema do Departamento de Polícia de Chicago CLEAR(*Citzen Law Enforcement Analysys and Report*). Para proteger a privacidade das vítimas não são utilizados os endereços exatos. A fonte de dados foi disponibilizada no Kaggle (https://www.kaggle.com/currie32/crimes-in-chicago).

O dataset utilizado é dividido em 4 arquivos CSV sendo eles Chicago_Crimes_2001_to_2004.csv, Chicago_Crimes_2005_to_2007.csv, Chicago_Crimes_2008_to_2011.csv, Chicago_Crimes_2012_to_2017.csv. Cada um dos arquivos e dividido em 23 atributos e aproximadamente 1.900.000 registros. Algumas colunas como Latitude e Longitude contém o maior número de dados faltantes. Explorando os dados foi um dos tipos de crimes

mais comuns listados é o de roubo de carros, ***MOTOR VEHICLE THEFT***. Pode-se avaliar se existe alguma relação com os horários dos roubos ou até mesmo se a presença de viaturas na região. Utilizando machine Learning Supervisionado e testando os algoritmos de classificação para procurar a relação dos fatores que determinam a ocorrência de roubo de carro.