CENTRO UNIVERSITÁRIO EURÍPIDES DE MARÍLIA

CURSO DE CIÊNCIA DE DADOS

STEPHANIE BERLINI RUFINO

RELATÓRIO CIENTÍFICO SOBRE O PROJETO DE CIÊNCIA DE DADOS "SOBREVIVENDO AO TITANIC"

MARÍLIA

2020

1. Objetivo

O presente relatório cientifico tem como objetivo relatar o experimento de ciência de dados com o projeto Titanic, o qual baseado na análise de dados e *machine learning* irá predizer se uma pessoa sobreviveria ao naufrágio do Titanic.

2. Introdução

Baseado no desafio do site *Kaggle "Titanic: machine learning from disaster"*, o projeto integrador do 1º semestre de 2020 da Turma I de Ciência de Dados do Centro Universitário Eurípedes de Marilia (UNIVEM), teve como proposta inicial replicar o *notebook* do *GitHub* do usuário Carlos FAB, com o intuito de prever se um indivíduo sobreviveria ao naufrágio. Idealizado pela discente Stephanie Berlini Rufino, obteve-se êxito em sua execução e conclusão, os quais serão relatados neste documento.

Para sua execução, foi utilizada como fonte dos dados o repositório do Titanic no *Kaggle*, com arquivos em formato CSV. A criação do notebook foi em linguagem Python, utilizando-se a ferramenta *Google Colab*. Todo o projeto está localizado no repositório da aluna no site *GitHub*.

3. Experimentos

Utilizando a metodologia CRISP-DM, todos os experimentos feitos no projeto serão relatados por etapas tais quais o do método citado.
Inicialmente, houve uma tentativa de criar o código com um *notebook* próprio do *Kaggle*, porém surgiram problemas em relação ao código e optou-se por construir um *notebook* na ferramenta *Google Colab*, o qual obteve êxito.

3.1 *Business Understanding*

Na primeira etapa do método, foi essencial compreender as necessidades do projeto. Girando em torno da questão "Quais problemas serão resolvidos?", tal etapa foi relativamente simples, visto que o problema a ser solucionado era se uma pessoa seria sobrevivente do já citado naufrágio utilizando os dados disponíveis para análise.

3.2 *Data Understanding*

Nessa etapa, o objetivo principal é coletar, organizar e documentar os dados disponíveis.

A coleta de dados foi feita no repositório do desafio, o qual disponibiliza arquivos em formato CSV. O primeiro, o qual serve como base para o treino da máquina, traz dados para nos auxiliar na previsão da sobrevivência (mais adiante, perceberemos que alguns dados não são tão relevantes para essa previsão e outros são cruciais). São eles: número de identificação do passageiro (*PassengerId*), se houve sobrevivência (*Survived*), a classe do bilhete (*Pclass*), nome do passageiro (*Name*), sexo (*Sex*), idade (*Age*), se possuía cônjuges ou irmãos a bordo (*SibSp*), se possuía filhos ou pais a bordo (*Parch*), número da passagem (*Ticket*), preço (*Fare*), número da passagem (*Cabin*) e em qual porto foi feito o embarque (*Embarked*). Já o segundo arquivo, traz dados para que o modelo possa ser testado; aqui, o dado se o passageiro sobreviveu ou não, não está incluído, uma vez que o intuito desse arquivo é usá-lo como teste.

Um fator interessante a ser notado com a plotagem dos gráficos e a análise percentual de sobrevivência por cada atributo é que passageiros do sexo feminino, da 1ª classe, mais novos e com poucos acompanhantes (pais, filhos, cônjuge e irmãos), tinham maior probabilidade de sobrevivência. Mulheres, por exemplo, tinham uma chance de 74,20% e homens, 18,89%. Por sua vez, viajantes da 1ª classe tinham percentual de 62,96%; 2ª, 47,28% e 3ª classe, 24,23%. Passageiros com cônjuges ou irmãos a bordo com número entre zero e dois, tinham 34,53%, 53,58% e 46,42% chances, respectivamente; já com três cônjuges ou irmãos, 25%, quatro, 16,66% e acima disso, 0% de chances. Ademais, passageiros com pais ou filhos teriam as probabilidades de 34,36% desacompanhados, 55,08% com apenas um deles, 50%, dois, 60%, três e acima disso, 0% de chances.

3.3 *Data Preparation*

Nessa etapa, os dados são tratados. Fez-se um código para saber qual variável mais tinha valores faltantes e chegou-se à conclusão de que "*Cabin*" possuía 77% valores faltantes; "*Age*", 20% e "*Embarked*", 2%.

Por conseguinte, houve uma minuciosa observação sobre quais *features* (variáveis) seriam relevantes para o modelo ou não: classe do bilhete, nome do passageiro, número da passagem e número da cabine foram desconsideradas por não influenciarem no resultado final.

A seguir, as *features* que restaram foram preparadas: *age* e *fare*, a título de exemplo, possuíam alguns valores faltantes que foram substituídos pela mediana (valor mais frequente); *embarked* com valores faltantes foi substituído pelo porto de maior frequência de embarque, o qual foi Southampton, além disso, essa *feature*, que era apenas uma coluna, foi substituída por outras com o nome dos portos de embarque, para, assim, haver uma atribuição de valores numéricos (caso o embarque foi naquele porto, teremos 0 como "não" e 1 como "sim"); por fim, em *sex*, os valores que eram do formato categórico, careciam conversão para números, portanto, *male* (masculino) foi comutado por 0 e *female* (feminino) por 1.

3.4 *Modeling*

Essa etapa consiste em definir os modelos de teste. Inicialmente, reproduzindo o código do repositório de Carlos FAB, foram usados dois modelos de *machine learning*: regressão logística e árvore de decisão com acurácia (precisão de acerto) de 80,13%, o primeiro e de 82,72% para o segundo.

Após pesquisas com o intuito de aumentar a acurácia do projeto, decidiu-se pelo uso do método KNN, o qual teve uma acurácia bem mais alta: 97,64%.

3.5 *Evaluating*

Aqui, obtivemos a avaliação dos resultados. Dessa forma, optou-se por usar os três modelos como forma de comparar e gerar resultados positivos ou negativos quanto à sobrevivência ou não ao naufrágio.

Com o intuito de responder a questão principal, tivemos como passageiros, a aluna e seus acompanhantes para avaliar o que ocorreria a cada um: Elaine, a mãe, José, o pai e Bruno, o namorado. Supondo-se que todos viajariam na 2 ª classe, pagando o preço médio do *ticket* e embarcando no porto de Southampton, auferimos resultados interessantes a serem notados: utilizando os modelos de regressão logística e árvore de decisão, apenas Stephanie e sua mãe sobreviveriam, já com o modelo KNN - de maior acurácia -, sua mãe não sobreviria, mas seu namorado, Bruno, sim.

4. Conclusão

O presente projeto de ciência de dados com análise preditiva da sobrevivência ao naufrágio do navio Titanic conseguiu atingir uma acurácia de mais de 97% utilizando o modelo KNN de *machine learning*, além de correlacionar variáveis que, aparentemente, não possuíam concordância.

Por fim, a questão principal a ser respondida no início do projeto obteve sua conclusão: a discente Stephanie teria sobrevivido ao naufrágio do Titanic.