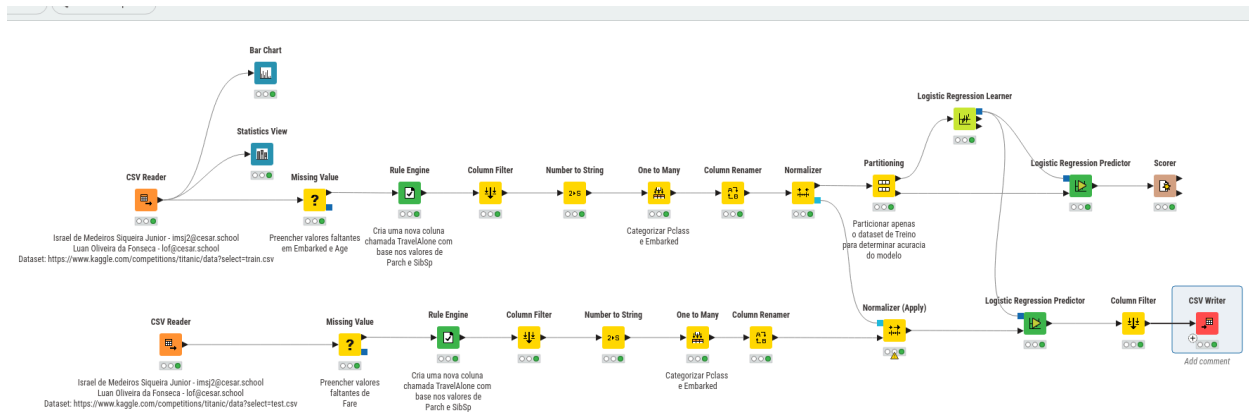


Israel de Medeiros Siqueira Junior: imsj2@cesar.school

Luan Oliveira da Fonseca: lof@cesar.school

Dataset: <https://www.kaggle.com/datasets/heptapod/titanic>

Workflow



Análise de performance

O nosso dataset escolhido estava dividido em duas partes: uma para treino e outra para teste. As duas partes possuíam desafios diferentes para o tratamento de dados.

Nos dois datasets haviam valores faltantes. No de teste, apenas 1 valor da coluna de Fare estava faltando e para consertar isso nos utilizamos do nó “Missing Value” do Knime para preencher esse valor com a Média dos valores de Fare. Nos dados de treino, havia bastante valores faltantes nas colunas de Embarked, Age e Cabin. Veja abaixo quantos valores estavam faltando para cada coluna:

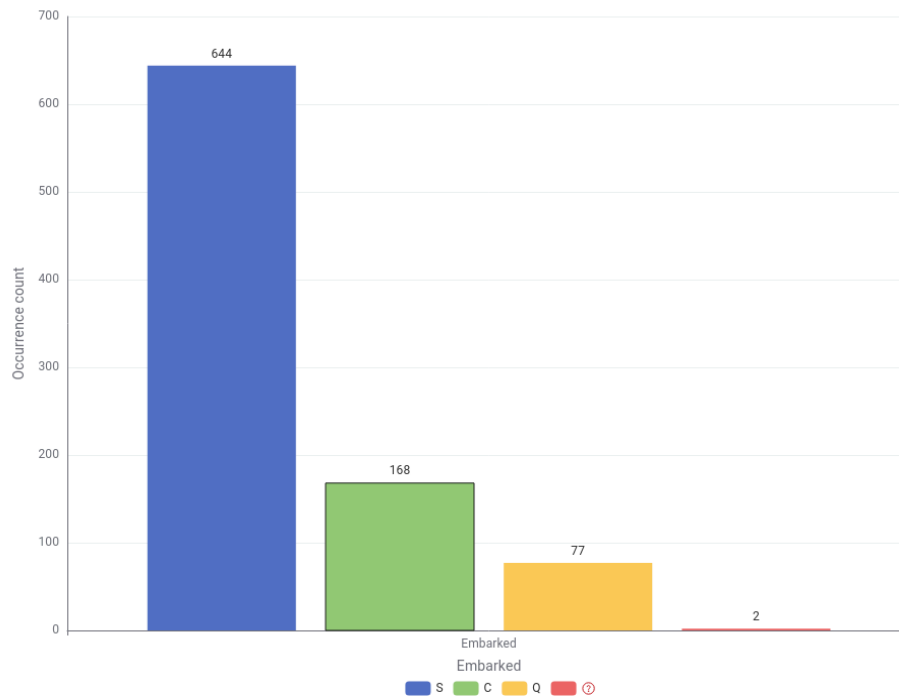
Statistics

Rows: 3 | Columns: 6

Name	Type	# Missing values	Minimum	Maximum	Mean ↓
Age	Number (double)	177	0.42	80	29.699
Cabin	String	687	?	?	?
Embarked	String	2	?	?	?

Para a coluna de Age, utilizamos a média para preencher os valores faltantes. Optamos por remover a coluna da Cabin do treinamento devido a quantidade de dados ausentes. Já para a coluna de Embarked, foi utilizado o valor mais comum (“S”) como substituto para não alterar a distribuição dos dados.

Passageiros em cada categoria de Embarked



O dataset possuía duas colunas (Parch e SibSp) para identificar a quantidade de familiares que estavam acompanhando o passageiro na viagem. De acordo com a descrição do dataset:

- SibSp: número de irmãos/cônjuges a bordo do Titanic
- Parch: de pais/filhos a bordo do Titanic

Optamos por combinar essas informações em uma nova coluna chamada TravelAlone com o objetivo de identificar se o passageiro estava viajando sozinho. Isso foi feito com o auxílio do nó RuleEngine do Knime.


Filtramos as seguintes colunas para removê-las do treinamento:

- PassengerId, Name e Ticket: são atributos pessoais e devem ser removidos
- SibSp e Parch: foram combinados na nova coluna TravelAlone
- Cabin: muitos valores ausentes

O próximo passo foi categorizar as colunas de Pclass e Embarked utilizando a ferramenta One to Many do Knime.

Por fim, as colunas de Age e Fare foram normalizadas por se tratarem de valores numéricos. Com a finalidade de analisar a acurácia do nosso modelo o dataset de treino foi particionado. Uma particularidade destes datasets é que o de teste não possui a coluna Survived portanto não podemos depender dele para analisar a acurácia localmente. Para isso, ao final do treinamento e execução dos testes submetemos o resultado no Kaggle para que eles retornassem a verdadeira acurácia.

Para o treinamento do modelo utilizamos uma Regressão Logística pois ela se mostrou eficaz em categorizar os dados, alcançando uma acurácia de 85.4%, com base nos valores particionados do dataset de treinamento. Após a submissão no Kaggle a acurácia foi de 74.6%.

 KAGGLE · GETTING STARTED PREDICTION COMPETITION · ONGOING

Submit Prediction ...

Titanic - Machine Learning from Disaster


Start here! Predict survival on the Titanic and get familiar with ML basics

Overview Data Code Models Discussion Leaderboard Rules Team Submissions

Submissions

All Successful Errors

Recent ▼

Submission and Description	Public Score ⓘ
<div> tiitanic_knime_logistic_reg.csv Complete · 2m ago</div>	0.74641