Titanic Kaggle Competition Report

Advanced Data Analytics Assignment 1

Naudé Conradie – 19673418

# Introduction

A dataset of the passengers on the Titanic is provided, with the intention of creating a machine learning model capable of predicting whether or not individual passengers survived its sinking. The dataset contains various attributes, which may be used as predictors for the model. The goal is to obtain a model that predicts the survival rate of passengers the most accurately.

# Attributes

The attributes are listed in Table 1 below, along with a brief description and discussion on their relevance and use, and an example value.

Table 1 - Attributes

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| **Attribute** | **Description** | **Relevance** | **Use** | **Example** |
| PassengerId | Unique identifier for each passenger, starting at 1 and incrementing by 1 until the last passenger | Not relevant for training | Used for identification of each passenger | 1 |
| Survived | Binary value of 0 or 1, representing the survival of the passenger, with 0 meaning they did not survive and 1 they did | The key attribute, as the model will be tested to determine its accuracy in predicting this value correctly |  | 0 |
| Pclass | Value ranging from 1 to 3, representing the class of the passenger | Higher class passengers, i.e. 1st class, were more likely to survive than lower class passengers, i.e. 3rd class | Used as is to train the model | 3 |
| Name | A string containing the passenger’s full name and title | Not relevant for training | Not used | Braund, Mr. Owen Harris |
| Sex | A string indicating the passenger’s sex, i.e. male or female | Female passengers were more likely to survive than male passengers | Converted to binary value similar to the Survived attribute, with 0 meaning male and 1 female | male |
| Age | A value indicating the passenger’s age | Younger passengers were more likely to survive than older passengers | Used as is to train the model | 22.0 |
| SibSp | An integer indicating the amount of siblings and/or spouses the passenger had on board | People with close relationships to other passengers were more likely to survive than lone passengers, as they would likely attempt to ensure that the entire group survived | Used as is to train the model | 1 |
| Parch | An integer indicating the amount of parents and/or children the passenger had on board | Used as is to train the model | 0 |
| Ticket | A string with the passenger’s unique ticket number | Not relevant for training | Not used | A/5 21171 |
| Fare | A value indicating the fare the passenger paid | Similar to the Pclass attribute | Used as is to train the model | 7.2500 |
| Cabin | A string indicating the passengers cabin number | Potentially relevant, as the location of a passenger’s cabin is related to their class and/or may place them closer to lifeboats. However, no information about the layout of the Titanic is provided, and, upon inspection, the majority of the attributes values are empty, and the rest are often erroneous. | Not used | NaN |
| Embarked | One of three characters indicating where the passenger embarked the ship, i.e. C for Cherbourg, Q for Queenstown and S for Southampton | Relevance unclear but easy to incorporate into the model | Converted into three new attributes, each with the name of the town and a binary value with 0 meaning no embarking from the respective town, and 1 meaning embarking from that town | S |