Intern Name: Muskaan
Intern ID: 170007

PROJECT ACTIVITY - 3

Predicting the Survival of Titanic Passengers

Problem Statement:

In this project you will use real data from the Titanic Tragedy to take a closer look at the people who survived this disaster and also predict the likelihood of survival.

On April 15, 1912, the largest passenger liner ever made collided with an iceberg during her maiden voyage. When the Titanic sank it killed 1502 out of 2224 passengers and crew. This sensational tragedy shocked the international community and led to better safety regulations for ships. One of the reasons that the shipwreck resulted in such loss of life was that there were not enough lifeboats for the passengers and crew. Although there was some element of luck involved in surviving the sinking, some groups of people were more likely to survive than others.

The titanic.csv file contains data of the real Titanic passengers. Each row represents one person. The column heading variables have the following meanings:

·   **survival:** Survival (0 = no; 1 = yes)

·   **class:** Passenger class (1 = first; 2 = second; 3 = third)

·   **name:** Name

·   **sex:** Sex

·   **age:** Age

·   **sibsp:** Number of siblings/spouses aboard

·   **parch:** Number of parents/children aboard

·   **ticket:** Ticket number

·   **fare:** Passenger fare

·   **cabin:** Cabin

·   **embarked:** Port of embarkation (C = Cherbourg; Q = Queenstown; S = Southampton)

·   **boat:** Lifeboat (if survived)

·   **body:** Body number (if did not survive and body was recovered)

Intern Name: Muskaan
Intern ID: 170007

Section: A (Exploratory Data Analysis)

1. Find out the overall chance of survival for a Titanic passenger.
2. Find out the chance of survival for a Titanic passenger based on their sex and plot it.
3. Find out the chance of survival for a Titanic passenger by traveling class wise and plot it.
4. Find out the average age for a Titanic passenger who survived by passenger class and sex.
5. Find out the chance of survival for a Titanic passenger based on number of siblings the passenger had on the ship and plot it.
6. Find out the chance of survival for a Titanic passenger based on number of parents/children the passenger had on the ship and plot it.
7. Plot out the variation of survival and death amongst passengers of different age.
8. Plot out the variation of survival and death with age amongst passengers of different passenger classes.
9. Find out the survival probability for a Titanic passenger based on title from the name of passenger.
10. What conclusions are you derived from the analysis?

Section-B (Classification)

We'll now predict if the person onboard has survived or not.   In this section, three different machine learning algorithms are implemented in Python using Scikit-learn library (i.e., **k-nearest neighbors, naive Bayes**, and **decision tree**) to study Titanic Tragedy dataset and deduce useful information to know the knowledge of the reasons for the survival of some travelers and sinking the rest. Finally, the results have been analyzed and compared with to evaluate the performance of each classifier. The following steps should be followed during building a classifier:

i.   Import the libraries required to build a decision tree in Python.

ii.   Load the dataset using the read_csv () function in pandas.

iii.   **Data Cleaning**: Preprecessing of the dataset.

      a)   **Missing Values:** Check where there are missing values and fix them appropriately.

iv.   **Feature Selection:** Separate the independent and dependent variables using the slicing method.

v.   **Encoding to numeric data:** Convert each of the categorical variables in to numeric data for modeling. For handling categorical variables, there are many methods like One Hot Encoding or Dummies.

vi.   **Splitting Data:** Split the data into training and testing sets.

vii.   **Building a Model:** Train the model using the classifier.

viii.   **Evaluating Model:** Predict the test data set values using the model above.

Intern Name: Muskaan
Intern ID: 170007

    ix.   Calculate the accuracy of the model using the accuracy score function.

Conclusions:

1. Which is the best model?

Solution:

https://github.com/MuskaanSehgal/BDDA2_ProjectActivity3.git