*Session9: Assignment 1*

1. Introduction

This assignment will help you to consolidate the concepts learnt in the session.

1. Problem Statement

Predicting Survival in the Titanic Data Set:

We will be using a decision tree to make predictions about the Titanic data set from

Kaggle. This data set provides information on the Titanic passengers and can be used to

predict whether a passenger survived or not.

Here is the code to load the data

import numpy as np

import pandas as pd

import seaborn as sb

import matplotlib.pyplot as plt

import sklearn

from pandas import Series, DataFrame

from pylab import rcParams

from sklearn import preprocessing

from sklearn.linear\_model import LogisticRegression

from sklearn.cross\_validation import train\_test\_split

from sklearn import metrics

from sklearn.metrics import classification\_report

**Url=**

https://raw.githubusercontent.com/BigDataGal/Python-for-Data-Science/master/titanic

-train.csv

titanic = pd.read\_csv(url)

titanic.columns =['PassengerId','Survived','Pclass','Name','Sex','Age','SibSp','Parch','Ticket','Fare','Cabin','E

mbarked']

You use only Pclass, Sex, Age, SibSp (Siblings aboard), Parch (Parents/children aboard),

and Fare to predict whether a passenger survived.

**Note: The solution shared through Github should contain the source code used**

**and the screenshot of the output.**

***3. Solution:***

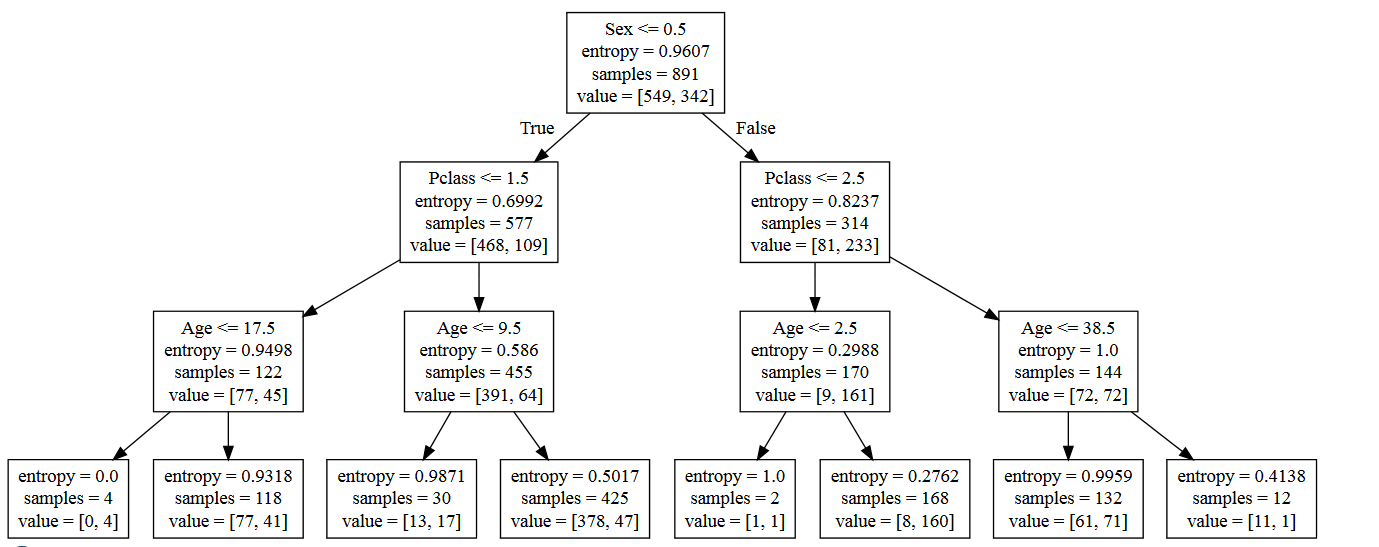
**Please fine the solution in the attached HTML file, I am unable to upload it directly to Github link.**

**Kindly enable the word file before clicking on the object so that it opens in a browser.**

**There are other files in the zipped folder used for input test csv file and output- decision tree dot file,** **DecisionTree png file and predicted survival based on gender -gender\_submission1 csv file**

** **

**DECISION TREE:**



**Predicting Survival in the Titanic Data Set:**

**Decision Tree Score: 87%** and Random Forest Score: 90%.

I just used Random Forest Algorithm to see the difference in prediction %. There is improvement score using Random Forest Algorithm.

Here I completed the analysis of what sorts of people were likely to survive. In particular, applied the tools of machine learning to predict which passengers survived the titanic tragedy.

**Data Description:**

The training set should be used to build machine learning models. For the training set, I will provide the outcome (also known as the “ground truth”) for each passenger. The model will be based on “features” like passengers’ gender, age, sex, family size and class.

The test set should be used to see how well my model is performing on unseen data. For the test set, we do not provide the ground truth for each passenger. It has to predict the outcomes. For each passenger in the test set,we have to use the model which we trained to predict whether or not they survived the sinking of the Titanic.

We also include gender\_submission.csv, a set of predictions that assume all and only female passengers survive, as an example of what a submission file should look like.

**Data Dictionary Variable Definition Key**

* survival Survival 0 = No, 1 = Yes
* pclass Ticket class 1 = 1st, 2 = 2nd, 3 = 3rd
* sex Sex =Male or Female
* Age Age in years
* sibsp #of siblings/spouses aboard the Titanic
* parch #of parents / children aboard the Titanic
* ticket Ticket number
* fare Passenger fare
* cabin Cabin number
* embarked Port of Embarkation C = Cherbourg, Q = Queenstown, S = Southampton

**Variable Notes**

* pclass: A proxy for socio-economic status (SES)=> 1st = Upper, 2nd = Middle and 3rd = Lower
* age: Age is fractional if less than 1. If the age is estimated, is it in the form of xx.5
* sibsp: The dataset defines family relations in this way=> Sibling = brother, sister, stepbrother, stepsister and Spouse = husband, wife (mistresses and fiancés were ignored)
* parch: The dataset defines family relations in this way=> Parent = mother, father and Child = daughter, son, stepdaughter, stepson. Some children travelled only with a nanny, therefore parch=0 for them.