**Session 23: Assignment 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.

Loading Data and modules

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.

Code:

**​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.tree import DecisionTreeClassifier**

**from sklearn.model\_selection import train\_test\_split**

**from sklearn import metrics**

**from sklearn.metrics import confusion\_matrix**

**from sklearn.metrics import classification\_report**

**# Url= https://raw.githubusercontent.com/BigDataGal/Python-for-Data-Science/master/titanic-train.csv**

**titanic = pd.read\_csv('https://raw.githubusercontent.com/BigDataGal/Python-for-Data-Science/master/titanic-train.csv')**

**titanic.columns**

**titanic.columns.values**

**titanic.info()**

**titanic.describe()**

**titanic.head()**

**total = titanic.isnull().sum().sort\_values(ascending=False)**

**percent\_1 = titanic.isnull().sum()/titanic.isnull().count()\*100**

**percent\_2 = (round(percent\_1, 1)).sort\_values(ascending=False)**

**missing\_data = pd.concat([total, percent\_2], axis=1, keys=['Total', '%'])**

**missing\_data.head()**

**# Age**

**survived = 'survived'**

**not\_survived = 'not survived'**

**sb.set\_style("darkgrid")**

**fig, axes = plt.subplots(nrows=1, ncols=2,figsize=(20, 8))**

**women = titanic[titanic['Sex']=='female']**

**men = titanic[titanic['Sex']=='male']**

**ax = sb.distplot(women[women['Survived']==1].Age.dropna(), bins=18, label = survived, ax = axes[0], kde =False, color='g')**

**ax = sb.distplot(women[women['Survived']==0].Age.dropna(), bins=40, label = not\_survived, ax = axes[0], kde =False, color='r')**

**ax.legend()**

**ax.set\_title('Female')**

**ax = sb.distplot(men[men['Survived']==1].Age.dropna(), bins=18, label = survived, ax = axes[1], kde = False)**

**ax = sb.distplot(men[men['Survived']==0].Age.dropna(), bins=40, label = not\_survived, ax = axes[1], kde = False)**

**ax.legend()**

**\_ = ax.set\_title('Male')**

**# Pclass and Sex**

**FacetGrid = sb.FacetGrid(titanic, row='Embarked', size=4.5, aspect=1.6)**

**FacetGrid.map(sb.pointplot, 'Pclass', 'Survived', 'Sex', palette=None, order=None, hue\_order=None )**

**FacetGrid.add\_legend()**

**# Pclass**

**sb.barplot(x='Pclass', y='Survived', data=titanic)**

**grid = sb.FacetGrid(titanic, col='Survived', row='Pclass', size=4.2, aspect=1.6)**

**grid.map(plt.hist, 'Age', alpha=.5, bins=20, color="M")**

**grid.add\_legend();**

**# SibSp and Parch**

**data = [titanic]**

**for dataset in data:**

**dataset['relatives'] = dataset['SibSp'] + dataset['Parch']**

**dataset.loc[dataset['relatives'] > 0, 'not\_alone'] = 0**

**dataset.loc[dataset['relatives'] == 0, 'not\_alone'] = 1**

**dataset['not\_alone'] = dataset['not\_alone'].astype(int)**

**titanic['not\_alone'].value\_counts()**

**axes = sb.factorplot('relatives','Survived', data=titanic, aspect = 3.3, color="C")**

**Output**











