BAHIRDAR UNIVERSITY INSTITUTE OF TECHNOLOGY

DEPARTMENT OF SOFTWARE ENGINEERING

MACHINE LEARNING INDIVIDUAL ASSIGNMENT 1

TITANIC DATASET

NAME :ADERAW MOLLA

ID:BDU 1200811

submitted to :Ass.Prof.Assefa M

Submitted date:02/11/2014 E

1. Briefly describe what the dataset is about and size of the dataset (e.g. number of tables, number of instances and attributes, etc.

Overview of titanic dataset

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 for 891 of the real Titanic passengers. Each row represents one person. The columns describe different attributes about the person including whether they survived (SS), their age (AA), their passenger-class (CC), their sex (GG) and the fare they paid (XX).

Data Dictionary

test.csv

- Data Source - [http://www.kaggle.com/c/titanic-gettingStarted/data]
- Data Information - Test data for Kaggle Titanic introductory comp.

train.csv

- Data Source - [http://www.kaggle.com/c/titanic-gettingStarted/data]
- Data Information - Training data for Kaggle Titanic introductory comp.

Data Variables (Test / Train)

Describes the variables in the test / train .csv files. This data dictionary and subsequent info was obtained from Kaggle.

Variable	Description	Details
survival	Survival	0 = No; 1 = Yes
pclass	Passenger Class	1 = 1st; 2 = 2nd; 3 = 3rd

Variable	Description	Details
name	First and Last 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

SPECIAL NOTES:
PClass can be a proxy for socio-economic status (SES)

- 1st ~ Upper;
- 2nd ~ Middle;
- 3rd ~ Lower

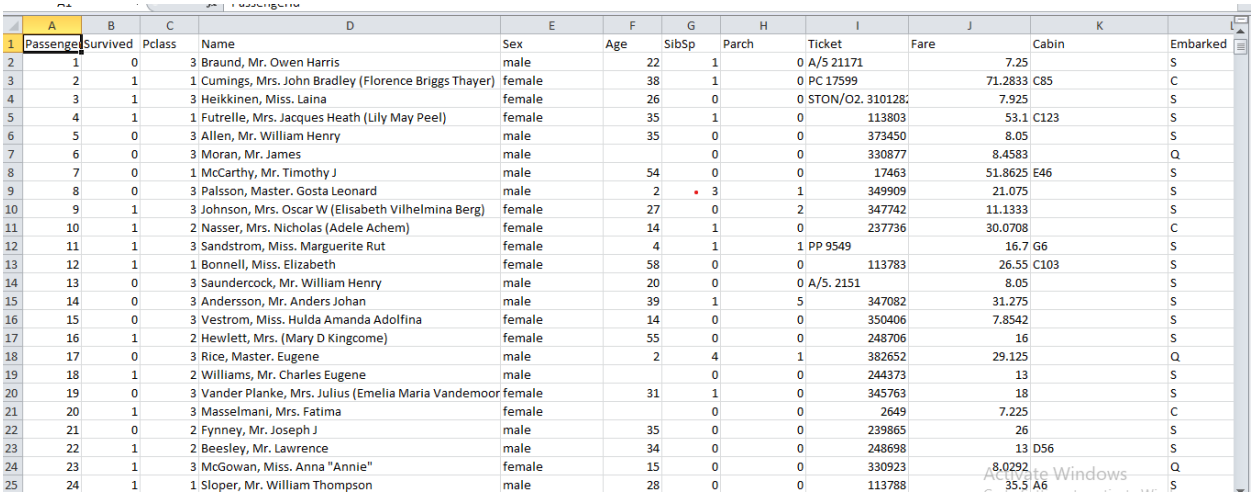
Age is in Years; - Fractional if Age less than One (1) - If the Age is Estimated, it is in the form xx.5

With respect to the family relation variables (i.e. sibsp and parch) some relations were ignored. The following are the definitions used for sibsp and parch.

Sibling: Brother, Sister, Stepbrother, or Stepsister of Passenger Aboard Titanic Spouse: Husband or Wife of Passenger Aboard Titanic (Mistresses and Fiances Ignored) Parent: Mother or Father of Passenger Aboard Titanic Child: Son, Daughter, Stepson, or Stepdaughter of Passenger Aboard Titanic

Other family relatives excluded from this study include cousins, nephews/nieces, aunts/uncles, and in-laws. Some children travelled only with a nanny, therefore parch=0 for them. As well, some travelled with very close friends or neighbors in a village, however, the definitions do not support such relations.

Sample excel screenshot of the dataset



	A	B	C	D	E	F	G	H	I	J	K	L
1	PassengerId	Survived	Pclass	Name	Sex	Age	SibSp	Parch	Ticket	Fare	Cabin	Embarked
2	1	0	3	Braund, Mr. Owen Harris	male	22	1	0	A/5 21171	7.25		S
3	2	1	1	Cumings, Mrs. John Bradley (Florence Briggs Thayer)	female	38	1	0	PC 17599	71.2833	C85	C
4	3	1	3	Heikkinen, Miss. Laina	female	26	0	0	STON/O2. 310128	7.925		S
5	4	1	1	Futrelle, Mrs. Jacques Heath (Lily May Peel)	female	35	1	0	113803	53.1	C123	S
6	5	0	3	Allen, Mr. William Henry	male	35	0	0	373450	8.05		S
7	6	0	3	Moran, Mr. James	male		0	0	330877	8.4583		Q
8	7	0	1	McCarthy, Mr. Timothy J	male	54	0	0	17463	51.8625	E46	S
9	8	0	3	Palsson, Master. Gosta Leonard	male	2	3	1	349909	21.075		S
10	9	1	3	Johnson, Mrs. Oscar W (Elisabeth Vilhelmina Berg)	female	27	0	2	347742	11.1333		S
11	10	1	2	Nasser, Mrs. Nicholas (Adele Achem)	female	14	1	0	237736	30.0708		C
12	11	1	3	Sandstrom, Miss. Marguerite Rut	female	4	1	1	PP 9549	16.7	G6	S
13	12	1	1	Bonnell, Miss. Elizabeth	female	58	0	0	113783	26.55	C103	S
14	13	0	3	Saundercock, Mr. William Henry	male	20	0	0	A/5. 2151	8.05		S
15	14	0	3	Andersson, Mr. Anders Johan	male	39	1	5	347082	31.275		S
16	15	0	3	Vestrom, Miss. Hulda Amanda Adolfina	female	14	0	0	350406	7.8542		S
17	16	1	2	Hewlett, Mrs. (Mary D Kingcome)	female	55	0	0	248706	16		S
18	17	0	3	Rice, Master. Eugene	male	2	4	1	382652	29.125		Q
19	18	1	2	Williams, Mr. Charles Eugene	male		0	0	244373	13		S
20	19	0	3	Vander Planke, Mrs. Julius (Emelia Maria Vandemoor	female	31	1	0	345763	18		S
21	20	1	3	Masselmani, Mrs. Fatima	female		0	0	2649	7.225		C
22	21	0	2	Fynney, Mr. Joseph J	male	35	0	0	239865	26		S
23	22	1	2	Beesley, Mr. Lawrence	male	34	0	0	248698	13	D56	S
24	23	1	3	McGowan, Miss. Anna "Annie"	female	15	0	0	330923	8.0292		Q
25	24	1	1	Sloper, Mr. William Thompson	male	28	0	0	113788	35.5	A6	S

To analyse titanic dataset let us import some liberaries and see thire result

1.import pandas for reading the dataset

```
import pandas as pd #in this project to use read_csv
from pandas import read_csv #to read a file as acommon data datatype in pandas i.e dataframe
```

path=("C:/Users/WIN- 10/Desktop/Titanic-Dataset.csv")

data=read_csv(path)

```
In [5]: data.shape
Out[5]: (891, 12)
```

This indicates that the selected dataset contains 12 columns or attributes and 891 rows or instances

Attributes /inputs are

passangerId ……..the id of each passanger

pclass……the class of passanger at theat time there were three classes 1,2,and 3

name……the name of each passanger

sex……male/female

age……..age of passangers

sibsp……number of sibling and spouse each passanger have

parch……parent or children have

ticket ……..the ticket number for each passanger

fare……..the payment payed by each passaanger

cabin……..cabin

embarked…… place of passangers

survived……indicates weather the passanger die or live after titanic incident

2. Discuss potential machine learning applications for the dataset. Name one or two types of machine learning application (classification,Regression, clustering, etc.) you think would be relevant and discuss the potential results

Regression algorithms fall under the family of Supervised Machine Learning algorithms which is a subset of machine learning algorithms. One of the main features of supervised learning algorithms is that they model dependencies and relationships between the target output and input features to predict the value for new data. Regression algorithms predict the output values based on input features from the data fed in the system. The go-to methodology is the algorithm builds a model on the features of training data and using the model to predict the value for new data.
Logistic regression one form of supervised learning regression with two discrete outputs .in this case survived or not.

For this dataset I prefer logistic regression algorithm to others because the model try to predict weather the passanger is survived or not i.e two outcome based on input variables age, pclass,sex,sibsp,parch,fare and embarked conditions.

3. Discuss appropriate measures of the central tendency and dispersion for the attributes. For numerical attributes of the dataset ,compute the mean, median, mode, range, Inter quartile Range, variance and Standard Devation for the attributes

```
data.describe()
```

	PassengerId	Survived	Pclass	Age	SibSp	Parch	Fare
count	891.000000	891.000000	891.000000	714.000000	891.000000	891.000000	891.000000
mean	446.000000	0.383838	2.308642	29.699118	0.523008	0.381594	32.204208
std	257.353842	0.486592	0.836071	14.526497	1.102743	0.806057	49.693429
min	1.000000	0.000000	1.000000	0.420000	0.000000	0.000000	0.000000
25%	223.500000	0.000000	2.000000	20.125000	0.000000	0.000000	7.910400
50%	446.000000	0.000000	3.000000	28.000000	0.000000	0.000000	14.454200
75%	668.500000	1.000000	3.000000	38.000000	1.000000	0.000000	31.000000
max	891.000000	1.000000	3.000000	80.000000	8.000000	6.000000	512.329200

Interpretation of each measure of centeral tendency

Count……indicates the number of record available for each attribute

This indicates that for n column dataset if n=count the dataset column have no null value for that attribute.else indicates there there are some null values and further processing is required.

Example.

from titanic dataset we can see that age=714 n=891 therefore age have 177 null values.so,further processing is required

mean……indicates the average sum of all values for a given column.

Example

For passangerid all the sum of passanger id divide by 891 is 446

Std……indicates the avarge deviation of agiven column data from thire mean.

Example

For passanger the average devation from the mean 446 is 257

Min and max……these indicates the minimum and maxiumum from the column

Example

Passanger Id have aminimum of 1 and maximum of 891

25%,50%,75%……indicates 1 st ,2 nd ,3 rd quartile

25% or 1 st quartile……indicates that 25 % of data for that column is lessthan or equal to value of quartile

Example

25 % of passangers are under the age of 20

50% or 2 st quartile……indicates that 50 % of data for that column is lessthan or equal to value of quartile

Example

50 % of passangers are age 28 and below age 28

75% or 3 st quartile……indicates that 75 % of data for that column is lessthan or equal to value of quartile

Example

75 % of passangers are age 38 and below age 38

Range ……found by the difference of maxiumum and nminimum value

i.e max- min
variance ……calculated as the squre of variance std^2
IQR……the difference of 3 quartile and 1 st quartile

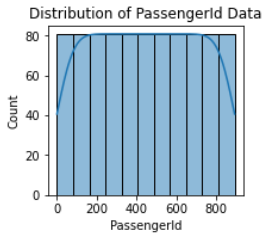
q3-q1

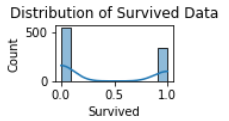
4. Plot shapes using the possible data visualization technique for the attributes and comment on the technique you selected

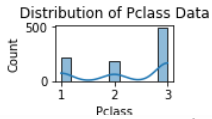
visualization techniqes I use

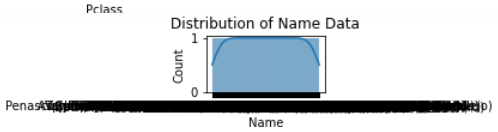
1.number of counts versus attribute

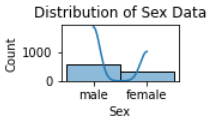
```
import seaborn as sns #in this assignment for plotting
import matplotlib.pyplot as plt #to plot shapes
plt.figure(figsize=(10,12))
for i,col in enumerate(data.columns,1):
    plt.subplot(4,3,i)
    plt.title(f"Distribution of {col} Data")
    sns.histplot(data[col],kde=True)

    plt.plot()
    plt.show()
print('\n')
```

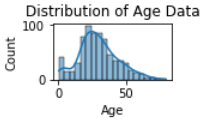


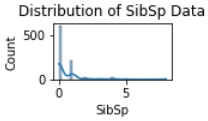












This hist indicates the number of counts with thire frequency of occurrence

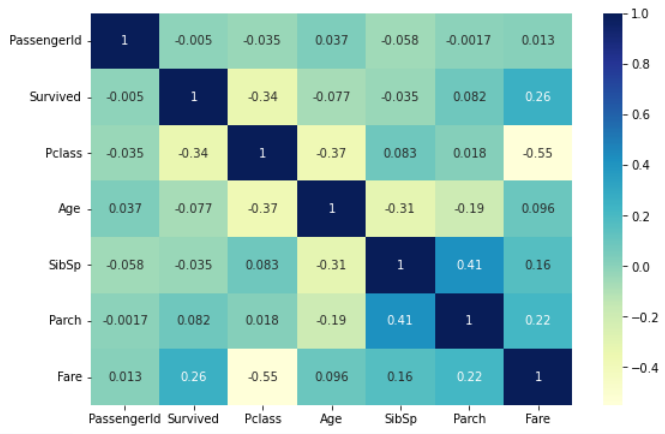
Example

For the attribute of pclass nearly 500 passangers are in the 3 rd class

The other visualization way is correlation factor which indicates interdependency of each attributs. Example screenshot of the code and result

```
corr = data.corr()
print(corr)
plt.figure(figsize=(9,6))
sns.heatmap(corr, annot=True,cmap='YlGnBu')
print("Correlation Plot of the sales Prediction")
print('\n')
```

```
             PassengerId  Survived    Pclass       Age     SibSp     Parch  \
PassengerId     1.000000 -0.005007 -0.035144  0.036847 -0.057527 -0.001652
Survived       -0.005007  1.000000 -0.338481 -0.077221 -0.035322  0.081629
Pclass         -0.035144 -0.338481  1.000000 -0.369226  0.083081  0.018443
Age             0.036847 -0.077221 -0.369226  1.000000 -0.308247 -0.189119
SibSp          -0.057527 -0.035322  0.083081 -0.308247  1.000000  0.414838
Parch          -0.001652  0.081629  0.018443 -0.189119  0.414838  1.000000
Fare            0.012658  0.257307 -0.549500  0.096067  0.159651  0.216225

                 Fare
PassengerId  0.012658
Survived     0.257307
Pclass      -0.549500
Age          0.096067
SibSp        0.159651
Parch        0.216225
Fare         1.000000
Correlation Plot of the sales Prediction
```



Example

We can intrprete as parch(parent or child) and survived have higest correlation than the others which is 0.082 that indicates the increase and decreasment of dependenyt variable directly affects the independent variable.

We can also interprete as passangerId have almost no factor for survival

Pclass higly affects survival inversely than other attributes i.e as pclass increases or go from first class to third class probability of survival decrease

5.Discuss data quality issues of the dataset. Are there (potential) problems with certain data attributes? What would be appropriate responses to these quality issues? Discuss the data preprocessing techniques that are likely required for the dataset. E.g. Is data reduction required and what would be an appropriate technique. Select one attribute and normalize the attribute based on min-max

Problem of the dataset

```
data.info()
```

```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 891 entries, 0 to 890
Data columns (total 12 columns):
 #   Column       Non-Null Count  Dtype
---  ------       --------------  -----
 0   PassengerId  891 non-null    int64
 1   Survived     891 non-null    int64
 2   Pclass       891 non-null    int64
 3   Name         891 non-null    object
 4   Sex          891 non-null    object
 5   Age          714 non-null    float64
 6   SibSp        891 non-null    int64
 7   Parch        891 non-null    int64
 8   Ticket       891 non-null    object
 9   Fare         891 non-null    float64
 10  Cabin        204 non-null    object
 11  Embarked     889 non-null    object
dtypes: float64(2), int64(5), object(5)
memory usage: 83.7+ KB
```

From this the dataset have problems

The age column have only 714 non-null values.so,the other values are null so either filling appropriate measure or discard it preferable for further processing to train our model for the future.

But,in this dataset discarding is not preferable because it affects the model I will develop.so.filling with mean or median is appropriate

Embareked column have 2 empty values .in this case either leave as it is or fill with appropriate measure is preferable

cabin column have only 294 null.so discarding the attribute is preferable as it does not affect our model .The dataset have columns which have empty columnds.so,filling these columns is appropriate techniques for preprocessing .
1.remove unnessary columns

```
from sklearn import preprocessing #for data preprocessing mainly in this project cleaning
def clean(data):
    data=data.drop(["Ticket","Cabin","Name","PassengerId"], axis=1)
    cols=["SibSp","Parch","Fare","Age"]
    for col in cols:
        data[col].fillna(data[col].median(),inplace=True)
    data.Embarked.fillna("U" ,inplace=True)
    return data
data=clean(data)
le=preprocessing.LabelEncoder()
cols=["Embarked","Sex"]
for col in cols:
    data[col]=le.fit_transform(data[col])
print(data)
```

```
     Survived  Pclass  Sex   Age  SibSp  Parch     Fare  Embarked
0           0       3    1  22.0      1      0   7.2500         2
1           1       1    0  38.0      1      0  71.2833         0
2           1       3    0  26.0      0      0   7.9250         2
3           1       1    0  35.0      1      0  53.1000         2
4           0       3    1  35.0      0      0   8.0500         2
..        ...     ...  ...   ...    ...    ...      ...       ...
886         0       2    1  27.0      0      0  13.0000         2
887         1       1    0  19.0      0      0  30.0000         2
888         0       3    0  28.0      1      2  23.4500         2
889         1       1    1  26.0      0      0  30.0000         0
890         0       3    1  32.0      0      0   7.7500         1
```


In this dataset column reduction is nessary because they have no any factor on our model
The above code removes columns Ticket,Cabin,Name,PassangerId anf fill numerical null values with thire median.
It also convert non-numbered attributes into numbered attributes this simplifies next step of data processing .
Example Embarked attribute converted to value of either 0,1,2 without affecting the result.

To normalize let for example the age attribute using min-max normalization

```
from sklearn import preprocessing
scaler = preprocessing.MinMaxScaler()
attribute = ["Age"]
d = scaler.fit_transform(data[attribute])
scaled_df = pd.DataFrame(d, columns=attribute)
scaled_df.head(10)
```

	Age
0	0.271174
1	0.472229
2	0.321438
3	0.434531
4	0.434531
5	0.346569
6	0.673285
7	0.019854
8	0.334004
9	0.170646

Summary of titanic dataset

The data set is taken back the incident of titanic in 1912.the dataset contains attributes for each passanger.

Depending on this the dataset identifies how many of them survived and died based on thire attributes such as pclass,age,sex,sibsp,parch,fare and embarked.

Sources:

https://www.kaggle.com/datasets/brendan45774/test-file

https://youtu.be/pUSi5xexT4Q