Name : Abhishek Jani

Department : Information Technology

Experiment No.1
Collect, Clean, Integrate and Transform Healthcare Data based on specific disease
Date of Performance: 4/8/2023
Date of Submission: 11/8/2023

Aim: Collect, Clean, Integrate and Transform Healthcare Data based on Heart disease.

Objective: The objective of this experiment is to perform basic pre processing on healthcare data set using python libraries

Theory:

Data Collection- Data collection is the process of gathering and measuring information from countless different sources. In order to use the data we collect to develop practical artificial intelligence (AI) and machine learning solutions, it must be collected and stored in a way that makes sense for the business problem at hand.

Data Cleaning: Cleaning data refers to the way of deleting wrong, corrupted, wrongly formatted, duplicate information, or incomplete information from a dataset. The possibility of duplicating or mislabelling data increases when two or more data sources are combined.

Data Integration: Data integration is the practice of consolidating data from disparate sources into a single dataset with the ultimate goal of providing users with consistent access and delivery of data across the spectrum of subjects and structure types, and to meet the information needs of all applications and business processes.

Data transformation: Data transformation is the process of converting, cleansing, and structuring data into a usable format that can be analyzed to support decision making processes, and to propel the growth of an organization. Data transformation is used when data needs to be converted to match that of the destination system.

Code: -

AIML EXP 1

File Edit View Insert Runtime Tools Help All changes saved

Comment Share

+ Code + Text

RAM Disk

```
# read csv
df = pd.read_csv("heart.csv")
```

```
[4] # To print no. of samples and attributes
    print(df.shape)
```

```
(303, 14)
```

```
[5] # getting the columns of the dataset
    columns = list(df.columns)
    print(columns)
```

```
['age', 'sex', 'cp', 'trestbps', 'chol', 'fbs', 'restecg', 'thalach', 'exang', 'oldpeak', 'slope', 'ca', 'thal', 'target']
```

```
[6] # To print first five samples
    print(df.head())
```

```
   age  sex  cp  trestbps  chol  fbs  restecg  thalach  exang  oldpeak  slope  \
0   63    1   3       145   233    1        0      150      0      2.3      0
1   37    1   2       130   250    0        1      187      0      3.5      0
2   41    0   1       130   204    0        0      172      0      1.4      2
3   56    1   1       120   236    0        1      178      0      0.8      2
4   57    0   0       120   354    0        1      163      1      0.6      2

   ca  thal  target
0   0     1       1
1   0     2       1
2   0     2       1
3   0     2       1
```

0s completed at 02:01

AIML EXP 1

File Edit View Insert Runtime Tools Help All changes saved

Comment Share

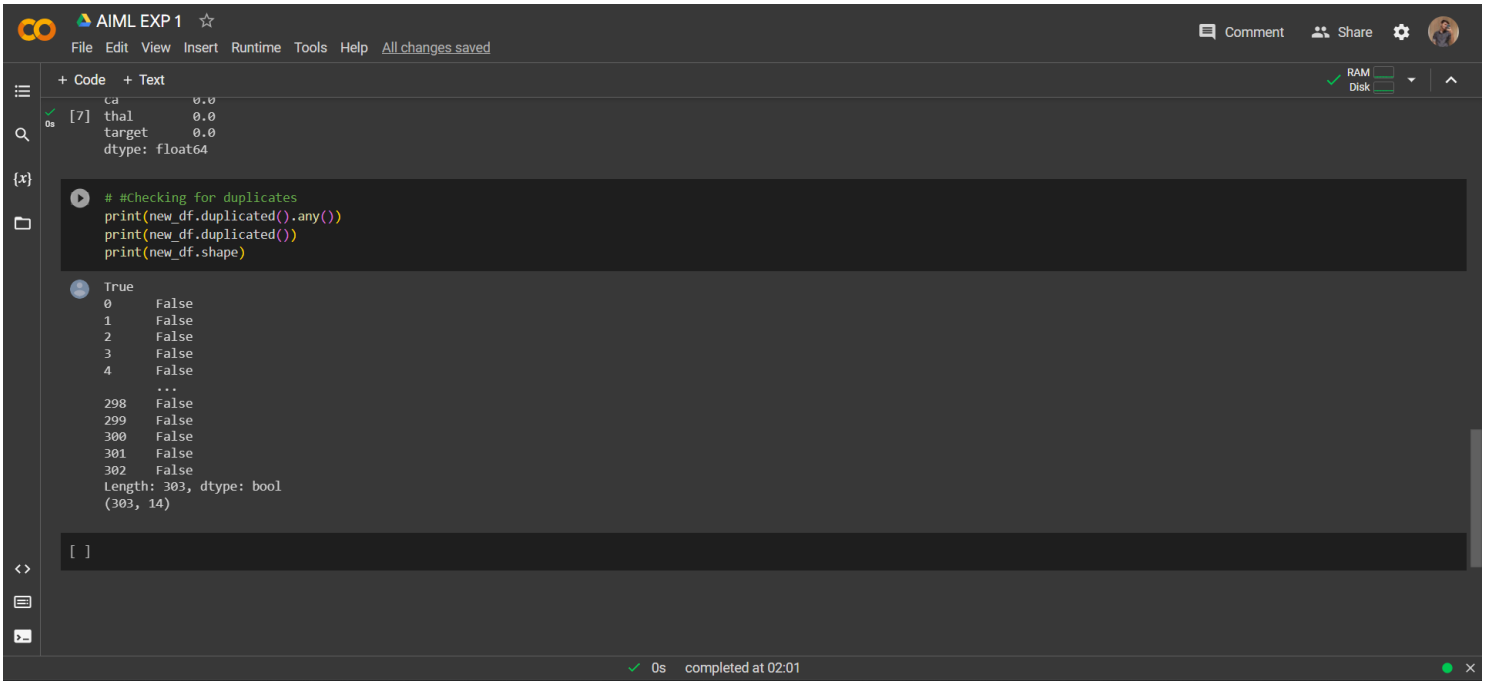
+ Code + Text

RAM Disk

```
#New dataframe
new_df = df
#new_df.isnull()
#Checking for null values
print(new_df.isnull().sum())
print("Missing values distribution: ")
print(new_df.isnull().mean())
#print(new_df.shape)
#new_df.duplicated()
```

```
age         0
sex         0
cp          0
trestbps    0
chol        0
fbs         0
restecg     0
thalach     0
exang       0
oldpeak     0
slope       0
ca          0
thal        0
target      0
dtype: int64
Missing values distribution:
age         0.0
sex         0.0
cp          0.0
trestbps    0.0
chol        0.0
```

0s completed at 02:01



Google Collaboratory Link: -

https://colab.research.google.com/drive/1eWmsGfryXAfUEqOHx35JA856gwL2n55w?usp=sharing

Conclusion: - Thus, we have successfully Collected, Cleaned, Integrated and Transformed our healthcare data.