Statistical Machine Learning Approaches to Liver Disease Prediction

TEAM ID: PNT2022TMID48226

Team Leader : P.M.Priyadharshini

Team Member : N.N.Deepika

Team Member : T.G.Yamini

Team Member : S.S.Megha

Data Pre-Processing

Data Pre-processing includes the following main tasks

1. Import the Libraries.

2. Reading the dataset.

3. Analyse the data.

4. Taking Care of Missing data.

5. Data Visualization.

6. Splitting the Dataset into Dependent and Independent variables.

7. Splitting Data into Train and Test

Importing The Libraries

The first step is usually importing the libraries that will be needed in the program.

The required libraries to be imported to Python script are:

```python
import pandas as pd
import numpy as np
import matplotlib.pyplot as plt
import seaborn as sns
import pickle
```

Numpy:

It is an open-source numerical Python library. It contains a multi-dimensional array and matrix data structures. It can be used to perform mathematical operations on arrays such as trigonometric, statistical, and algebraic routines.

Example:

```python
import numpy as np
a = np.array([0, 1, 2, 3])        # Create a rank 1 array
print(a)                          #print array a
print(type(a))                    #type of array a
print(a.ndim)                     #dimension of array a
print(a.shape)                    #shape(row,column) of array a
print(len(a))                     #length of array a
```

```
[0 1 2 3]
<class 'numpy.ndarray'>
1
(4,)
4
```

Pandas:

It is a fast, powerful, flexible and easy to use open source data analysis and manipulation tool, built on top of the Python programming language.
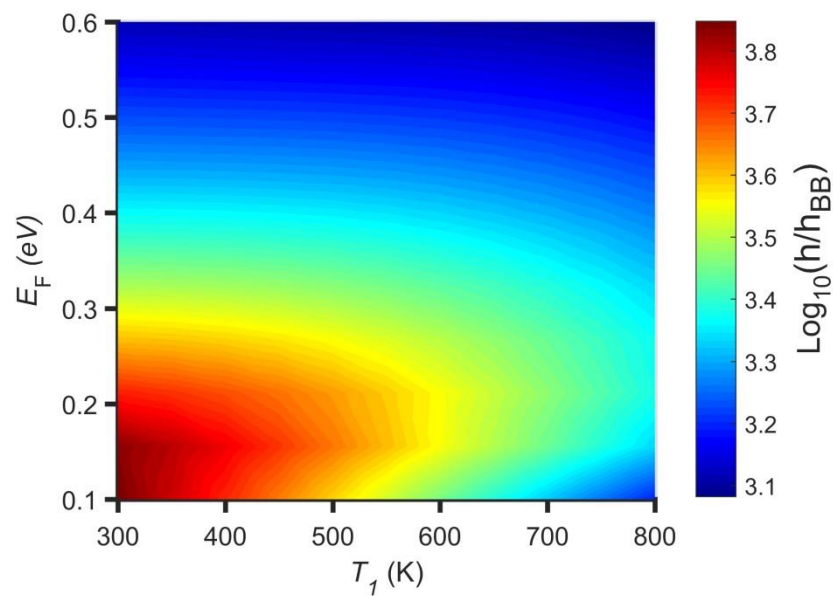
Example:

```
label = ['a', 'b', 'c']
my_data = [10, 20, 30]
pd.Series(data = my_data, index = label)
```

```
a    10
b    20
c    30
dtype: int64
```

Matplotlib:

Visualisation with python. It is a comprehensive library for creating static, animated, and interactive visualizations in Python.
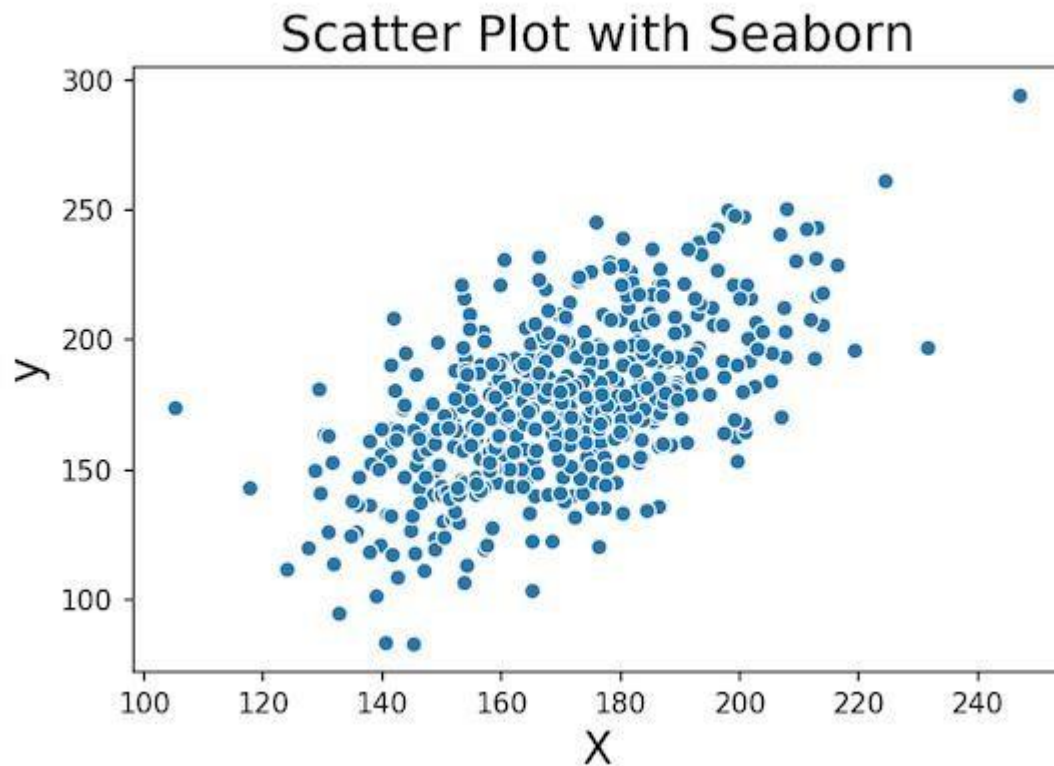
Example:

Seaborn:

Seaborn is a library for making statistical graphics in Python. Seaborn helps you explore and understand your data. Its plotting functions operate on dataFrames and arrays containing whole datasets and internally perform the necessary semantic mapping and statistical aggregation to produce informative plots.

Example:

Pickle:

The pickle module implements serialization protocol, which provides an ability to save and later load Python objects using special binary format.