MACHINE LEARNING PARADIGMS IN BIOINFORMATICS

Presented By

- Farooq Ansari (10800116102)
- Rohit Kumar Majee (10800116061)
- Arkadeep Bagal (10800116108)
- Rohan Kumar Singh (10800116062)

Introduction

In this short presentation, we would like to talk about and explore the various ways popular machine learning algorithms can help us extract useful and actionable data from seemingly ordinary datasets with its implementation.
We would also explore some popular machine learning algorithms and how they work.

Outline

- Introduction
- What is Machine Learning?
- Types of Machine Learning Algorithms
- Implementation of a Working Model
- Planned Future Work
- Conclusion
- Acknowledgement
- References

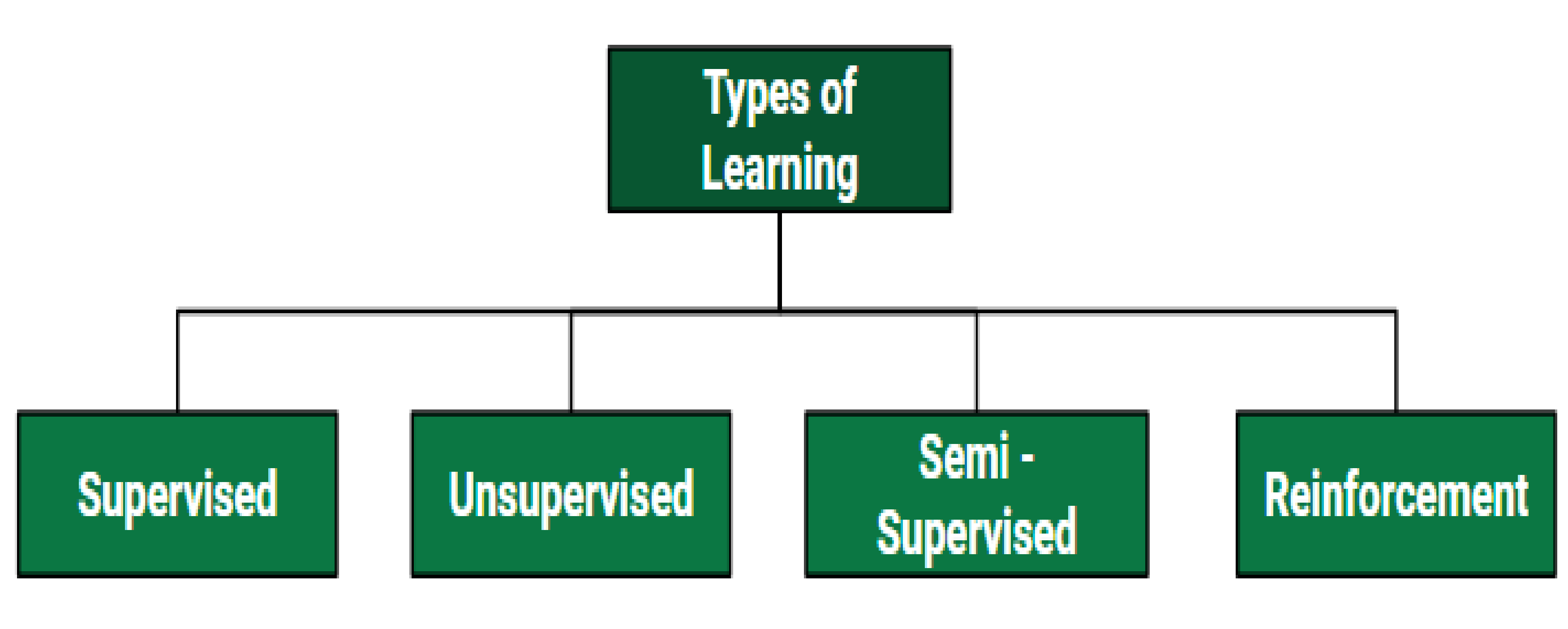
What is Machine Learning?

Some popular definitions are as follows

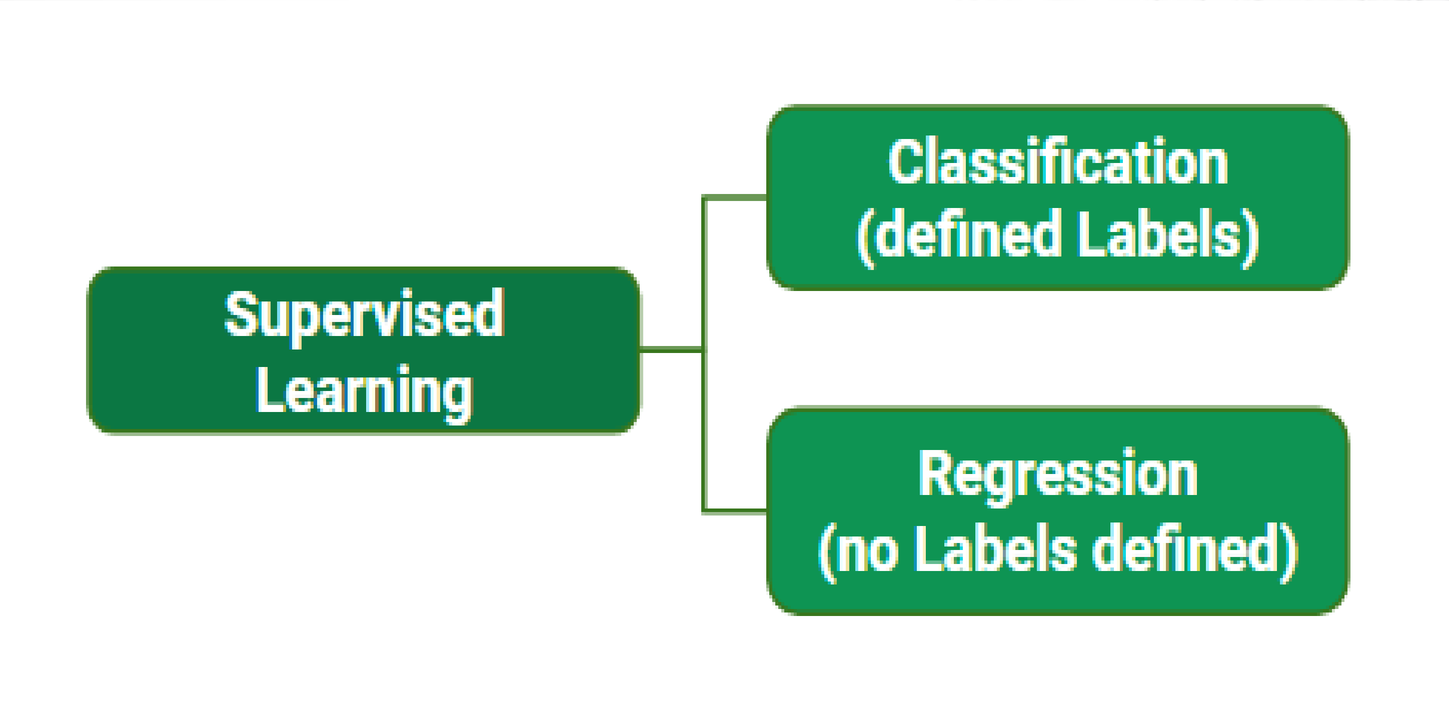
- **Arthur Samuel(1959):** Field of study that gives computers the ability to learn without being explicitly programmed.
- **Tom Mitchell(1998):** A computer program is said to learn from experience**(E)** with respect to some task **(T)** and some performance measure **(**P) , if its performance on **T** , as measured by **P** , improves with experience **E.**

Types of
Machine Learning Algorithms

Machine Learning Algorithms



Supervised Learning



Supervised Learning

Classification (defined Labels)

Regression (no Labels defined)

Supervised Learning

- Supervised learning is when the model is getting trained on a labelled dataset.
- **Labelled** dataset is one which have both input and output parameters. In this type of learning both training and validation datasets are labelled
- Example:

  Linear Regression

  Logistic Regression

Types of Supervised Learning

1. Classification Algorithms

These are the kinds of algorithms with produces defined labels or discrete values as output.

Ex. - Logistic Regression, Neural Networks

Types of Supervised Learning

Logistic Regression is a Machine Learning classification algorithm that is used to predict the probability of a categorical dependent variable.

In logistic regression, the dependent output variable is a binary variable that contains data coded as 1 (yes, success, etc.) or 0 (no, failure, etc.).

Types of Supervised Learning

1. Regression Algorithms

Here, we fit a curve / line to the data points, in such a manner that the differences between the distances of data points from the curve or line is minimized. We obtain a continuos output variable.

Ex. - Linear Regression, Support Vector Regression

Types of Supervised Learning

Linear Regression is a supervised machine learning algorithm where the predicted output is continuous and has a constant slope. It’s used to predict values within a continuous range, (e.g. sales, price) rather than trying to classify them into categories (e.g. cat, dog).

Unsupervised Learning

It is the training of machine using information that is neither classified nor labeled and allowing the algorithm to act on that information without guidance. Here the task of machine is to group unsorted information according to similarities, patterns and differences without any prior training of data.

Ex. K-Means Clustering, Principle Component Analysis

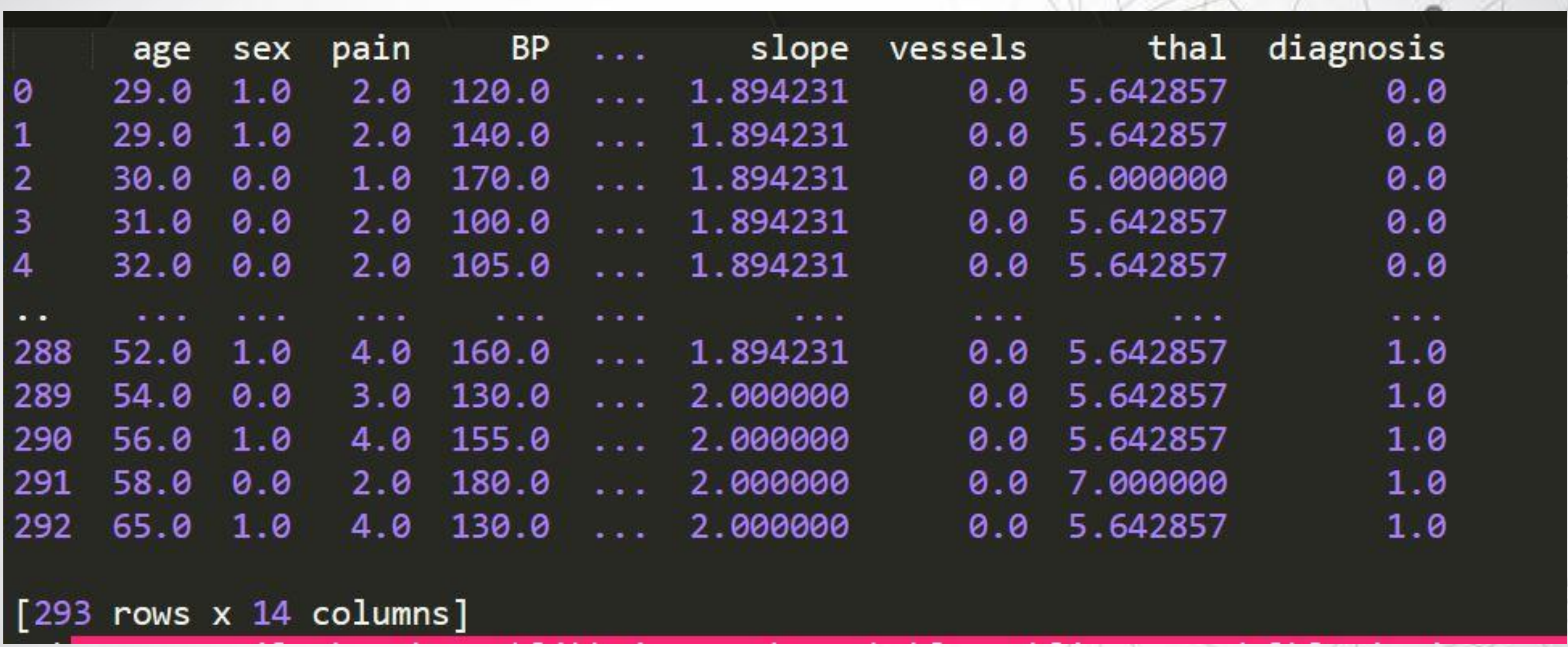
Types of Unsupervised Learning

K-Means Clustering is an iterative clustering algorithm which helps you to find the highest value for every iteration. Initially, the desired number of clusters are selected. In this clustering method, you need to cluster the data points into k groups.

The output of the algorithm is a group of "labels." It assigns data point to one of the k groups. In k-means clustering, each group is defined by creating a centroid for each group, they capture the points closest to them and adds them to the cluster.

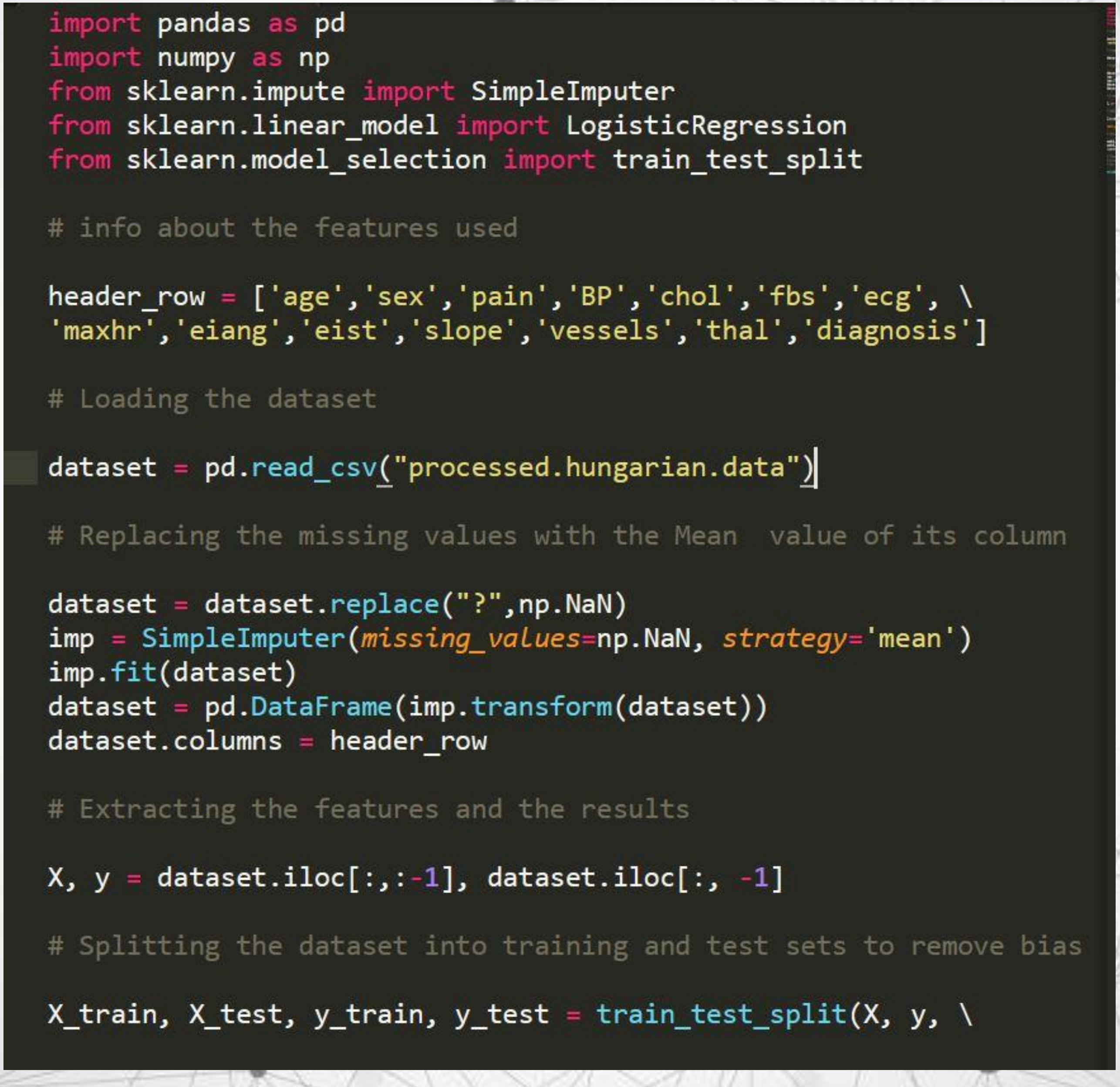
Implementation of Machine Learning for Bioinformatics

The Hungarian Heart Disease Dataset



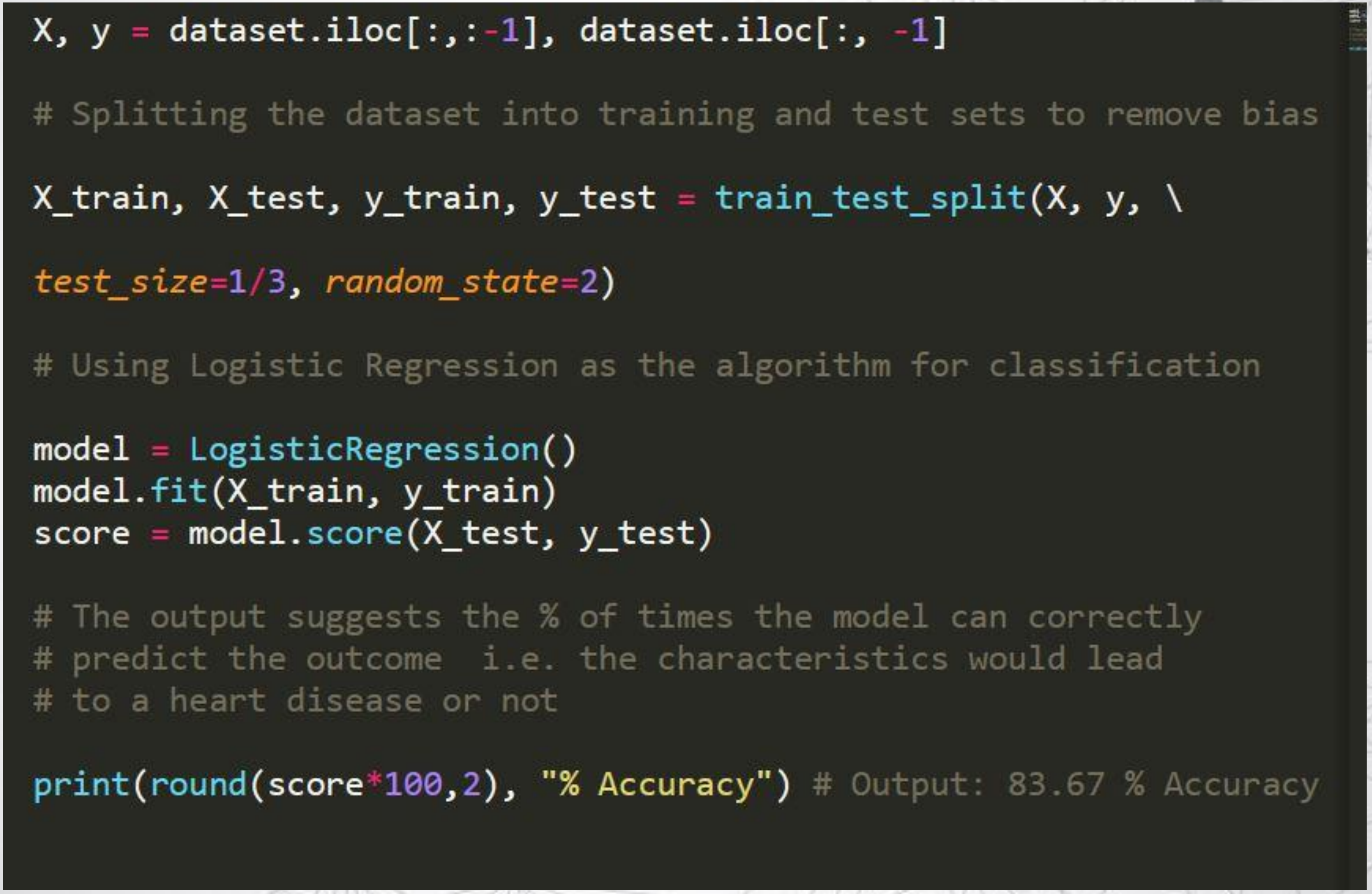
```
     age  sex  pain     BP  ...     slope  vessels      thal  diagnosis
0    29.0  1.0   2.0  120.0  ...  1.894231      0.0  5.642857        0.0
1    29.0  1.0   2.0  140.0  ...  1.894231      0.0  5.642857        0.0
2    30.0  0.0   1.0  170.0  ...  1.894231      0.0  6.000000        0.0
3    31.0  0.0   2.0  100.0  ...  1.894231      0.0  5.642857        0.0
4    32.0  0.0   2.0  105.0  ...  1.894231      0.0  5.642857        0.0
..    ...  ...   ...    ...  ...       ...      ...       ...        ...
288  52.0  1.0   4.0  160.0  ...  1.894231      0.0  5.642857        1.0
289  54.0  0.0   3.0  130.0  ...  2.000000      0.0  5.642857        1.0
290  56.0  1.0   4.0  155.0  ...  2.000000      0.0  5.642857        1.0
291  58.0  0.0   2.0  180.0  ...  2.000000      0.0  7.000000        1.0
292  65.0  1.0   4.0  130.0  ...  2.000000      0.0  5.642857        1.0

[293 rows x 14 columns]
```

Source Code



```
import pandas as pd
import numpy as np
from sklearn.impute import SimpleImputer
from sklearn.linear_model import LogisticRegression
from sklearn.model_selection import train_test_split

# info about the features used

header_row = ['age','sex','pain','BP','chol','fbs','ecg', \
'maxhr','eiang','eist','slope','vessels','thal','diagnosis']

# Loading the dataset

dataset = pd.read_csv("processed.hungarian.data")

# Replacing the missing values with the Mean  value of its column

dataset = dataset.replace("?",np.NaN)
imp = SimpleImputer(missing_values=np.NaN, strategy='mean')
imp.fit(dataset)
dataset = pd.DataFrame(imp.transform(dataset))
dataset.columns = header_row

# Extracting the features and the results

X, y = dataset.iloc[:,:-1], dataset.iloc[:, -1]

# Splitting the dataset into training and test sets to remove bias

X_train, X_test, y_train, y_test = train_test_split(X, y, \
```

Source Code



```
X, y = dataset.iloc[:,:-1], dataset.iloc[:, -1]

# Splitting the dataset into training and test sets to remove bias

X_train, X_test, y_train, y_test = train_test_split(X, y, \

test_size=1/3, random_state=2)

# Using Logistic Regression as the algorithm for classification

model = LogisticRegression()
model.fit(X_train, y_train)
score = model.score(X_test, y_test)

# The output suggests the % of times the model can correctly
# predict the outcome  i.e. the characteristics would lead
# to a heart disease or not

print(round(score*100,2), "% Accuracy") # Output: 83.67 % Accuracy
```

Future Work

1. More visually descriptive
2. Improving the accuracy
3. Cleaning the data
4. Performance measures.
5. Analyzing the original 1987 Dataset

Conclusion

As we can observe, there is plenty of scope in the field of Bio-Medical Informatics for us, as budding Computer Science students to explore and implement ground-breaking algorithms & machine learning pipelines to extract meaningful inferences that can directly help people.

It is our conviction that we would contribute to this ever-growing field and improve our initial understandings.

Acknowledgement

We would like to thank our mentor Mr. Sabyasachi Mukherjee Sir, for his continuous guidance during the preparation for this topic.

We would like to thank the Head of Computer Science & Engineering department, Dr. Monish Chatterjee for providing us with the opportunity to explore such technological fields.

References

1. Dataset was taken from the popular UCI Machine Learning Repository

http://archive.ics.uci.edu/ml/machine-learning-databases/heart-disease/

2. The Python in-built machine learning module "Scikit-Learn"

https://scikit-learn.org/stable/

3. Understanding of the Algorithms and various Data Science costructs from the popular MOOC by Andrew Ng

https://www.coursera.org/learn/machine-learning-with-python