Configuration Manual

MSc Research Project
MSc in Data Analytics

Sahil Pathak
Student ID: 23215852

School of Computing
National College of Ireland

Supervisor: Athanasios Staikopoulos

National College of Ireland

MSc Project Submission Sheet

School of Computing

Student Name: …Sahil Pathak……

Student ID: 23215852……..……

Programme: …Msc in Data Analytics…… **Year:** …2024-2025..

Module: …Msc in Data Analytics…….……

Lecturer: …… Athanasios Staikopoulos …….……

Submission Due Date: …12th Dec 2024…….……

Project Title: … Prediction of patient adherence to medication using random forest, decision tree and neural networks techniques …….……

Word Count: …118…… **Page Count:** …8…….……..……

I hereby certify that the information contained in this (my submission) is information pertaining to research I conducted for this project. All information other than my own contribution will be fully referenced and listed in the relevant bibliography section at the rear of the project.
ALL internet material must be referenced in the bibliography section. Students are required to use the Referencing Standard specified in the report template. To use other author's written or electronic work is illegal (plagiarism) and may result in disciplinary action.

Signature: ……Sahil Pathak……

Date: …12th Dec 2024……

PLEASE READ THE FOLLOWING INSTRUCTIONS AND CHECKLIST

Attach a completed copy of this sheet to each project (including multiple copies)	☐
Attach a Moodle submission receipt of the online project submission, to each project (including multiple copies).	☐
You must ensure that you retain a HARD COPY of the project, both for your own reference and in case a project is lost or mislaid. It is not sufficient to keep a copy on computer.	☐

Assignments that are submitted to the Programme Coordinator Office must be placed into the assignment box located outside the office.

Office Use Only	
Signature:	
Date:	
Penalty Applied (if applicable):	

Configuration Manual

Sahil Pathak
Student ID: 23215852

1 Introduction

This particular research project is about patient adherence to prescribed medications. Following sections will outline the tools and technologies needed and were installed. Also the code execution step by step will be explained further.

2 System Configuration

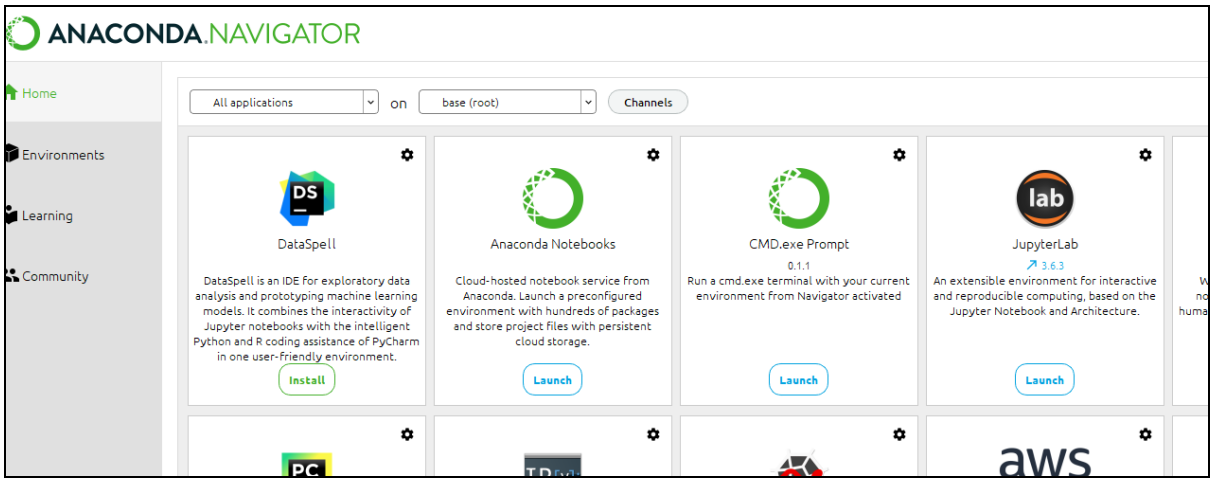
2.1. HARDWARE SPECIFICATION

- **OPERATING SYSTEM :** Windows 10
- **PROCESSOR:** x64 bit i7 processor
- **HARD DRIVE:** 1 TB
- **RAM:** 16 GB

2.2. SOFTWARE SPECIFICATION

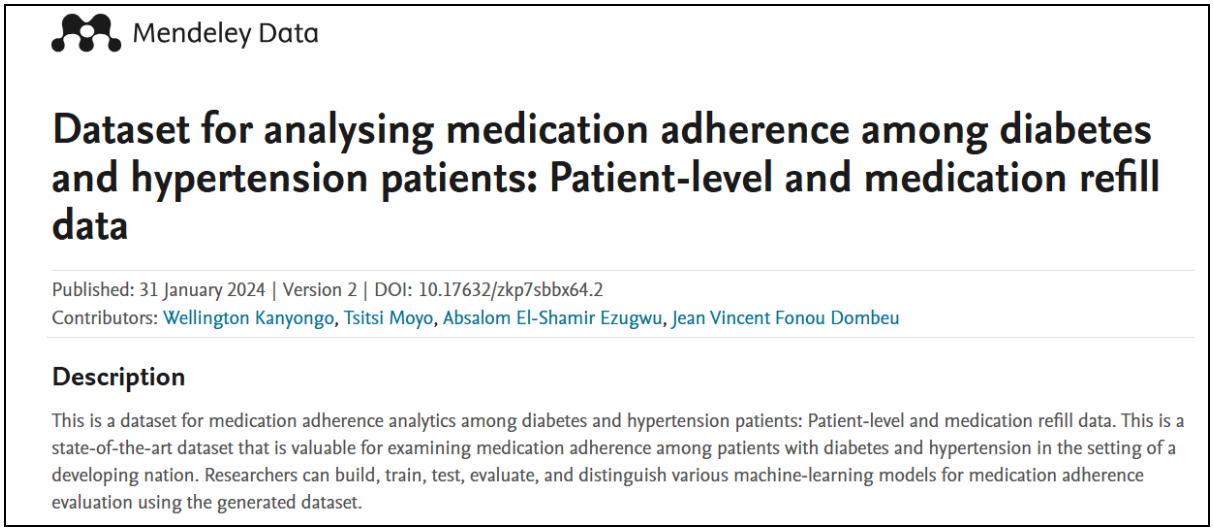
- PYTHON
- JUPYTER NOTEBOOK

3 Installation and Environment Setup



4 Data Collection

The data was collected from an open source website. Total columns it contains is around 41 columns. Below is the screenshot for the same:



5 Implementation

5.1 Libraries used in the research project

- Numpy
- Pandas
- Seaborn
- Plotly
- Matplotlib
- Scipy

```python
# Here I am importing the necessary packages :

import numpy as np
import pandas as pd
import seaborn as sns
import plotly.express as px
import matplotlib.pyplot as plt
import datetime
from matplotlib import style
%matplotlib inline
import warnings
warnings.filterwarnings('ignore')

import scipy
from sklearn.preprocessing import LabelEncoder
from sklearn.preprocessing import StandardScaler
from sklearn.preprocessing import OneHotEncoder
from sklearn.model_selection import train_test_split , cross_val
from sklearn.metrics import accuracy_score, log_loss
from sklearn.metrics import accuracy_score, precision_score, rec
from sklearn.ensemble import RandomForestClassifier
from sklearn.tree import DecisionTreeClassifier
from sklearn.neural_network import MLPClassifier
import xgboost as xgb
from sklearn.linear_model import LogisticRegression
```

5.2 Flow of implementation

Reading the source files for diabetes and hypertension

```python
diabetes = pd.read_csv('Diabetes_Adherence_Data.csv')
```

```python
diabetes
```

	MEMBER	INO	PZTIQNT NZXQ	ADHERENCE	GENDER	PRACTICE NO	DIS	PROVIDER	AS AT NETWORKS	REFERRING PRACTICE	...	DL	CLAIM NO	CLAIM LINE NO	DUPLICATE CLAIM
0	98788489	0	XS TZFZDZWZ NYZPIXBI	2	F	860662	60	81302 PLUS TWO PHARMACY	NaN	847321.0	...	N	3285628	7248261	NaN
1	98795328	0	XR TSITSIKZYI XUDOTI	7	M	877344	60	92398 CIMAS HCC MUTARE PHARMACY	CIMAS PHARMACIES (2123)	866776.0	...	N	2760001	6299632	NaN
2	98795328	0	XR TSITSIKZYI	7	M	877344	60	92398 CIMAS HCC MUTARE	CIMAS PHARMACIES	866776.0	...	N	2760001	6299633	NaN

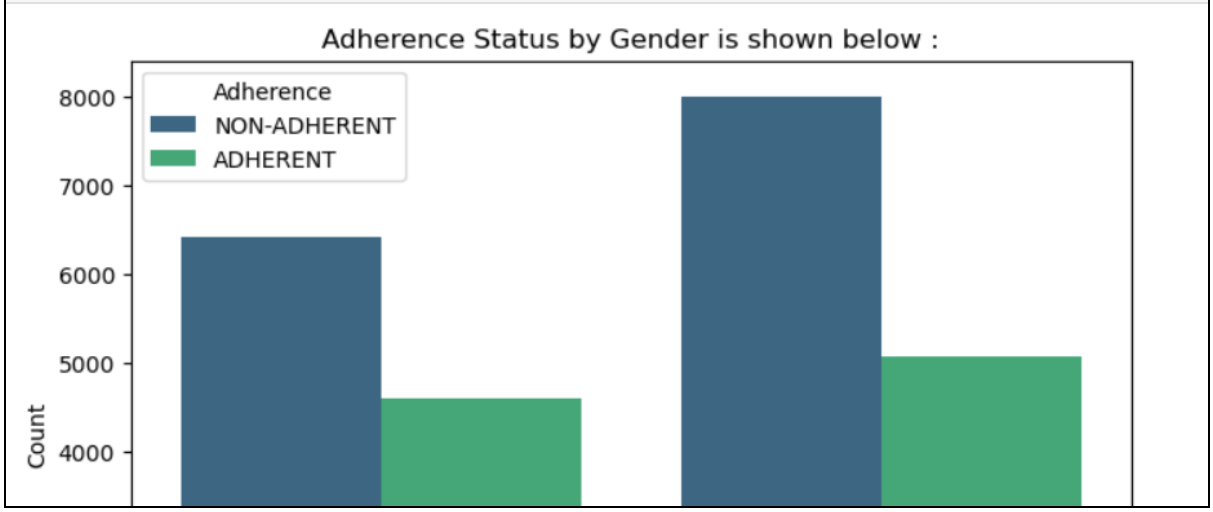
Data Preprocessing

```python
# Removing the unwanted columns or variables from the dataset as shown in the below code:
diabetes_new = diabetes[['BENEFICIARY', 'ADHERENCE', 'GENDER', 'UNITS','AMOUNT CLAIMED', 'OPTION NAME
htn_new = htn[['BENEFICIARY', 'ADHERENCE', 'GENDER', 'UNITS','AMOUNT CLAIMED', 'OPTION NAME', 'CURREN
```

```python
diabetes_new
```

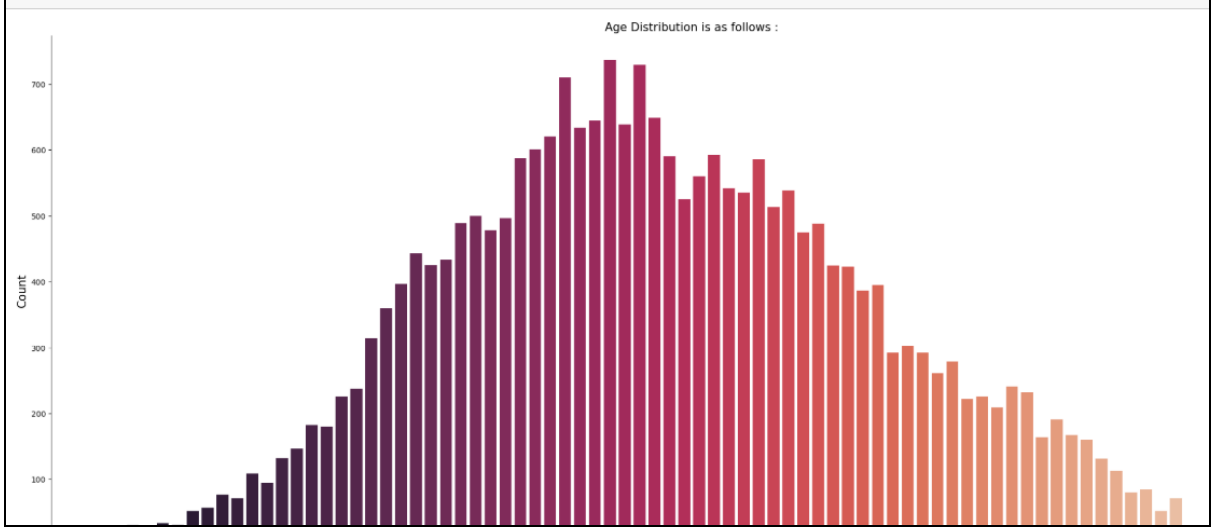
	BENEFICIARY	ADHERENCE	GENDER	UNITS	AMOUNT CLAIMED	OPTION NAME	CURRENT AGE

```python
# Grouped bar plot of adherence by gender is shown as below :
plt.figure(figsize=(8, 6))
sns.countplot(data=df_new, x='GENDER', hue='ADHERENCE', palette='viridis')
plt.title('Adherence Status by Gender is shown below :')
plt.xlabel('Gender')
plt.ylabel('Count')
plt.legend(title='Adherence')
plt.show()
```




```python
# Plotting the visualizations based on age :

fig_var, j = plt.subplots()
fig_var.set_size_inches(30, 12)
sns.countplot(x = 'AGE',  palette="rocket", data = df_new)
j.set_xlabel('Age', fontsize=15)
j.set_ylabel('Count', fontsize=15)
j.set_title('Age Distribution is as follows :', fontsize=15)
sns.despine()
```



Model Building

Model Building

```
# Independent Variables
X=df_new[['AGE', 'ANNUALCONTRIBUTION', 'ANNUALCLAIMAMOUNT', 'UNITSTOTAL', 'GENDER_M', 'SCHEMETYPE_MEDIUM', 'SCH
        'DIAGNOSIS_HYPERTENSION', 'COVERTYPE_STANDARD', 'COMORBIDITY_NO_COMORBIDITY', 'COMPLICATIONDEVELOPMENT_N
X
```

	AGE	ANNUALCONTRIBUTION	ANNUALCLAIMAMOUNT	UNITSTOTAL	GENDER_M	SCHEMETYPE_MEDIUM	SCHEMETYPE_PREMIUM
0	24	690384.0	10400.00	60.0	False	False	False
1	53	2718696.0	306366.53	1380.0	True	True	False
2	46	808284.0	77480.00	600.0	True	False	False
3	76	808284.0	198800.58	840.0	True	False	False
4	69	808284.0	116872.04	720.0	False	False	False
...	...	...	...	...	...	...	...
19274	50	9724602.0	3500.00	30.0	True	False	True
19275	71	9724602.0	4763.96	30.0	False	False	True
19276	55	6889900.0	3899.19	30.0	True	False	True
19277	43	6889900.0	4.13	30.0	False	False	True

```
rf_model = RandomForestClassifier(n_estimators=100, random_state=42)
rf_model.fit(X_train, y_train)
```

```
RandomForestClassifier
RandomForestClassifier(random_state=42)
```

```
# Making the prediction using the testing data , code is shown below :

y_pred = rf_model.predict(X_test)
```

```
# Metrics calculations :

accuracy = accuracy_score(y_test, y_pred)
precision = precision_score(y_test, y_pred, pos_label='ADHERENT')
recall = recall_score(y_test, y_pred, pos_label='ADHERENT')
f1 = f1_score(y_test, y_pred, pos_label='ADHERENT')


print(f"\n The Accuracy of Random Forest in scikit learn library is as follows : {accuracy}")
print(f" The Precision of Random Forest in scikit learn library is as follows : {precision}")
print(f" The Recall of Random Forest in scikit learn library is as follows : {recall}")
print(f" The F1 Score of Random Forest in scikit learn library is as follows : {f1}")
```

```
 The Accuracy of Random Forest in scikit learn library is as follows : 0.806947135344589
 The Precision of Random Forest in scikit learn library is as follows : 0.73210378243201
```

Decision Tree

```
4]: dt_model = DecisionTreeClassifier(random_state=42)

5]: dt_model.fit(X_train, y_train)

5]: DecisionTreeClassifier
    DecisionTreeClassifier(random_state=42)

6]: y_pred = dt_model.predict(X_test)

7]: # Metrics calculations :

    accuracy = accuracy_score(y_test, y_pred)
```

Neural Networks

```
149]: nn_model = MLPClassifier(hidden_layer_sizes=(100,), max_iter=

150]: nn_model.fit(X_train, y_train)

150]: MLPClassifier
      MLPClassifier(max_iter=300, random_state=42)

151]: y_pred = nn_model.predict(X_test)

152]: # Metrics calculations :

      accuracy = accuracy_score(y_test, y_pred)
```