Configuration Manual

MSc Research Project
MSc in Data Analytics

Kaviprasad Shanmugasundaram
Student ID: x22241850

School of Computing
National College of Ireland

Supervisor: Prof: Anu Shani

National College of Ireland

MSc Project Submission Sheet

School of Computing

Student Name:	Kaviprasad Shanmugasundaram
Student ID:	22241850
Programme:	Msc in Data Analytics **Year:** 2025
Module:	Msc Research Project
Supervisor:	Anu Shani
Submission Due Date:	28-01-2025
Project Title:	Comparative Analysis of Machine Learning Models for Mental Health Assessment Using Music Therapy
Word Count:	662 words **Page Count**: 9

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.

<u>ALL</u> 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:	Kaviprasad Shanmugasundaram
Date:	12-12-2024

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

Kaviprasad Shanmugasundaram
Student ID: x22241850

1 Introduction

This manual provides a comprehensive guide to set up, use, which employs machine learning models to predict depression percentages. Follow the steps outlined below to configure the environment, execute the code, and interpret the results.

2 System Requirements

The specific needs for hardware as well as software to put the research into use are detailed in this section.

2.1. HARDWARE SPECIFICATION

- **OPERATING SYSTEM : MacOS(Ventura13.4)**
- **PROCESSOR: M2Processor with build in 10 CPU and GPU**
- **HARD DRIVE: SSD(256GB)**
- **RAM:16GB**

2.2 Software Requirements

- PYTHON
- JUPYTER NOTEBOOK
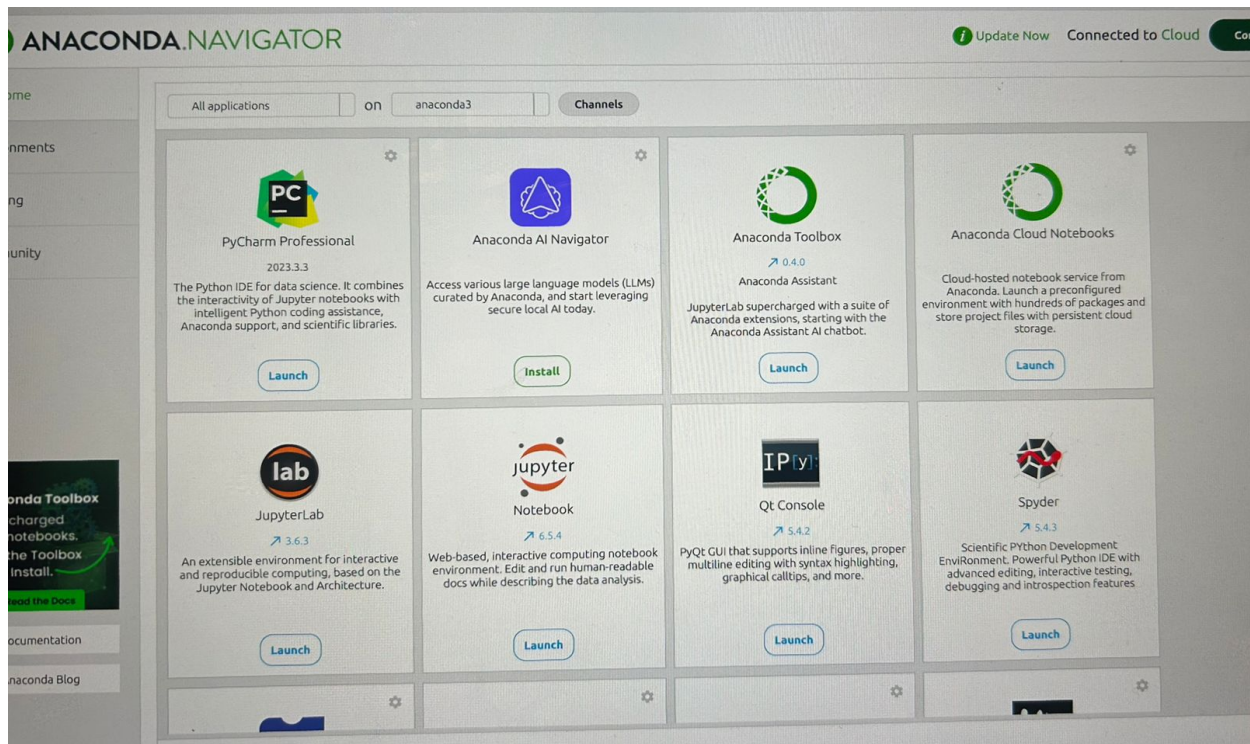
2.3 Installation and Environment Setup

Fig 2 : Anaconda Dashboard

3. Data Collection and Preparation

The dataset is taken from Kaggle public repository from the link

Data Collection

The analysis utilized two datasets:

Dataset1: https://www.kaggle.com/datasets/catherinerasgaitis/mxmh-survey-results

Dataset2: https://www.kaggle.com/datasets/thedevastator/uncover-global-trends-in-mental-health-disorder

Dataset Preparation
Ensure you have the following datasets saved in the project directory:
Mental health Depression disorder Data.csv
mxmh_survey_results.csv

Loading Datasets
Use the following code snippet to load the datasets into the environment:
```python
import pandas as pd

df1 = pd.read_csv('Mental health Depression disorder Data.csv')
df2 = pd.read_csv('mxmh_survey_results.csv')
```

User Merged Dataset: used in the code
/Users/kaviprasad/Documents/Kavi_Projects_sem3/THESIS/Datasets/x22241850_Final
merged dataset.csv

4. Libraries

- pandas
- numpy
- matplotlib
- seaborn
- scikit-learn
- joblib

Install these libraries using the following command:
#pip install pandas numpy matplotlib seaborn scikit-learn joblib

```
In [1]:    1  # Import necessary libraries
           2  import pandas as pd
           3  import numpy as np
           4  from sklearn.model_selection import train_test_split
           5  from sklearn.preprocessing import LabelEncoder
           6  from sklearn.tree import DecisionTreeRegressor
           7  from sklearn.ensemble import RandomForestRegressor, GradientBoostingRegressor
           8  from sklearn.linear_model import LinearRegression
           9  from sklearn.svm import SVR
          10  from sklearn.neighbors import KNeighborsRegressor
          11  from sklearn.metrics import mean_squared_error, r2_score
          12  import matplotlib.pyplot as plt
          13  import seaborn as sns
          14  import joblib
```

Fig 3: Libraries required for research.

5. Exploratory Data Analysis

- This Figure Visualize the pie chart for distribution of primary streaming services. Streaming services are like Pandora, Apple music, Spotify, YouTube music, etc.
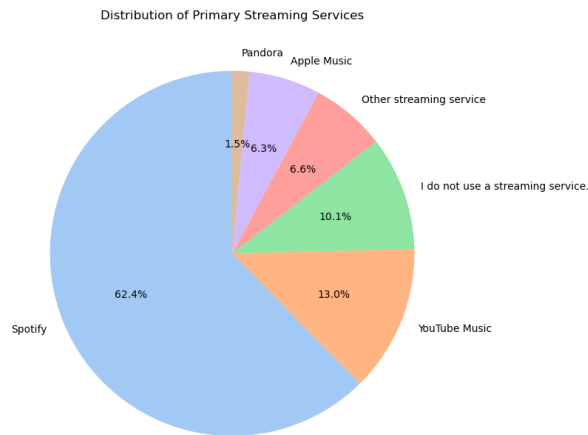
Figure 4: Distribution of primary streaming services.

- The Figure 5, Visualize the pie chart of effects of music distribution in which place it is improved and any kind of changes.
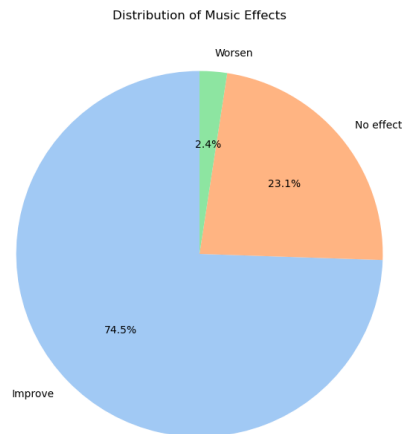

Figure 5: Distribution of music effects

- This Figure 6, represents the comparison level of OCD, Anxiety, Depression while working

```
In [13]:  1  fig, axes = plt.subplots(nrows=3, ncols=1, figsize=(6, 10), sharex=True)
          2
          3  emotional_states = ['OCD', 'Anxiety', 'Depression']
          4
          5  for i, state in enumerate(emotional_states):
          6      sns.barplot(x='While working', y=state, data=df, ax=axes[i], palette='viridis')
          7      axes[i].set_ylabel(f'{state} Level')
          8      axes[i].set_xlabel('Listening While Working')
          9
         10  plt.suptitle('Comparison of OCD, Anxiety, and Depression by Listening While Working', y=1.02)
         11  plt.show()
```

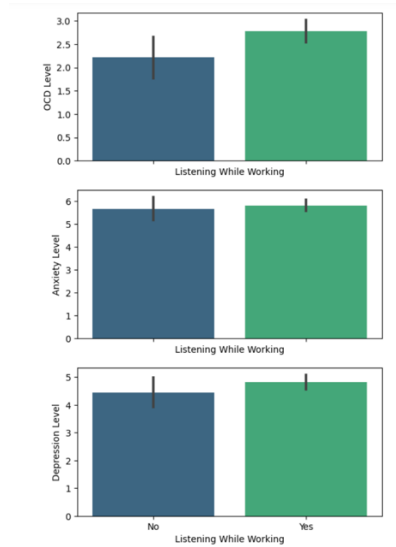Comparison of OCD, Anxiety, and Depression by Listening While Working

Figure 6: OCD, Anxiety, Depression level while listening

6. Model Development and Evaluation

The system uses six machine learning models:
1. K-Nearest Neighbors (KNN)
2. Neural Network
3. Decision Tree Regressor
4. Random Forest Regressor
5. Gradient Boosting Regressor
6. Linear Regression

Key Steps:
- Preprocessing: Handles missing values and encodes categorical data.
- Feature Selection: Extracts features (`X`) and target (`y`).
- Data Splitting: Splits data into training and testing subsets (80%-20%).
- Model Training: Trains models using `X_train` and `y_train`.
- Model Evaluation: Assesses performance using Mean Squared Error (MSE) and R^2 Score.
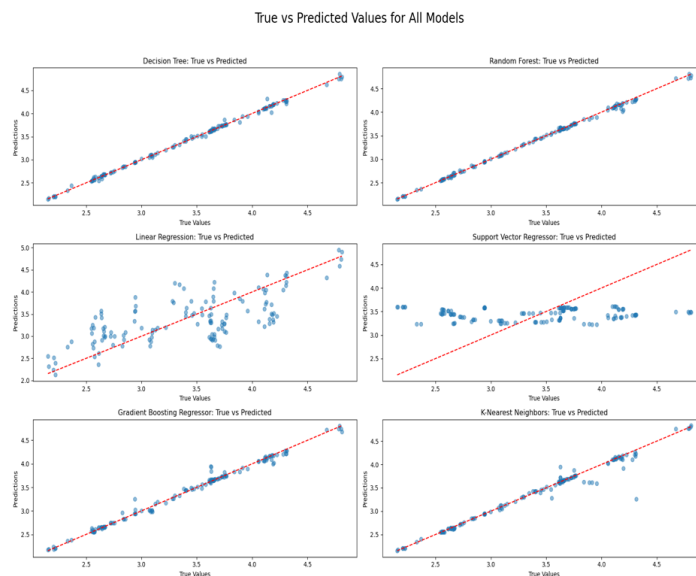Model Evaluation:

Figure7: True vs predicted value for all models
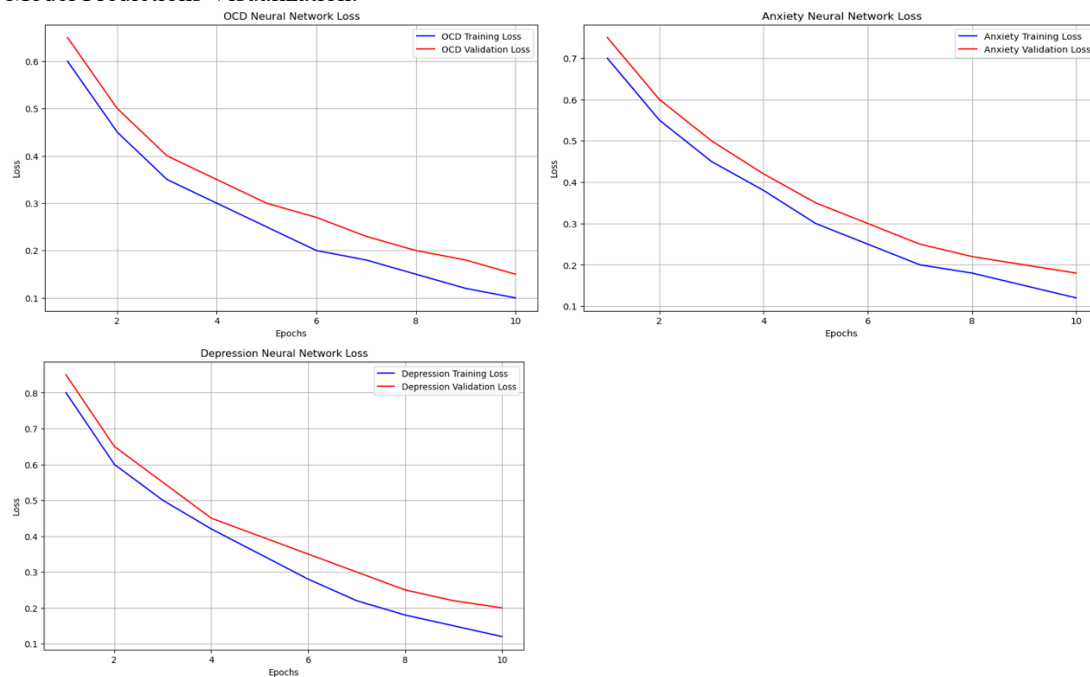
Model Predictions Visualization:

Figure 8: Neural network loss for OCD, Depression, Anxiety.

7. Result and Analysis

This section explains the results and analysis:

The system outputs the same MSE and R^2 Score for each model in a comparative table. The KNN and Neural Network models typically performs best with:

- MSE: 0.0140

- R^2 Score: 0.96

Visualization includes scatter plots for predicted vs. actual values for each model, allowing for quick analysis of accuracy.

8. Implementation Notes

- Execution: Code will be executed in Jupyter Notebook.

9. Model Saving and Loading

Saving Trained Models

The trained models are saved in `.joblib` format for future use.

Example:

import joblib

joblib.dump(knn_model, 'knn_model.joblib')

Loading Trained Models

To load and use a saved model:

knn_model = joblib.load('knn_model.joblib')

predictions = knn_model.predict(new_data)

10. Maintenance and Updates

Dataset Updates

- Replace old datasets with updated files in `.csv` format.

- Re-run the preprocessing and training pipeline.

Library Updates

Keep all Python libraries up-to-date:

pip install --upgrade pandas numpy matplotlib seaborn scikit-learn joblib

11. Troubleshooting

- Error: `FileNotFoundError`

  - Solution: Ensure the datasets are in the specified directory.

- Error: `ModuleNotFoundError`

  - Solution: Install the missing library using `pip install <library_name>`.

12. Future Enhancements

- To enhance generalization of the model by testing more complex architectures of a neural network, including convolutional or recurrent neural networks that might reveal more subtle structures in the data.
- To increase the variety and size of datasets and potentially including conditions other than depression or anxiety could either improve the stability and performance of such models.

13. References

Dataset1: https://www.kaggle.com/datasets/catherinerasgaitis/mxmh-survey-results

Dataset2: https://www.kaggle.com/datasets/thedevastator/uncover-global-trends-in-mental-health-disorder

OpenCV: OpenCV modules

Keras Applications