**Machine Learning using Python**

**Project Report – [Mind-Scope:Mental health prediction using machine learning]**

|  |  |
| --- | --- |
| **College Name:** | **Government College of Engineering , Bodinayakanur** |
| **Team Name :** | **Predictive Pirates** |

|  |  |  |
| --- | --- | --- |
| **Team Members :** | 1. **21CS128-** 2. **21CS123-** 3. **21CS124-** 4. **21CS147-** | **Maheswari V J**  **Kiruba R**  **Kokila M**  **Supria R** |

**1.Abstract**

This study explores the application of multiple machine learning algorithms to predict mental health conditions using a random set of physiological and psychological indicators. The data utilized in this study include heartbeat rate, sleep hours, number of steps walked, mental crisis occurrences, anger levels, anxiety levels, excitement levels, and current emotional states . We employed various machine learning algorithms, including Random Forest, Support Vector Machine (SVM), K-Nearest Neighbors (KNN), and Neural Networks, to analyze these multifaceted data points. The primary objective was to identify patterns and correlations that could serve as early indicators of mental health issues, such as depression . This study highlights the potential of machine learning in mental health prediction, providing a foundation for the development of real-time monitoring tools that can offer personalized mental health insights and interventions.

**2.Model**

Supervised learning is a machine learning technique that is widely used in various fields such as finance, healthcare, marketing, and more. It is a form of machine learning in which the algorithm is trained on labeled data to make predictions or decisions based on the data inputs. The algorithm tries to learn the relationship between the input and output data so that it can make accurate predictions on new, unseen data. In this type of learning both training and validation, datasets are labelled.

Supervised learning is typically divided into two main categories:

Regression: Regression is a supervised learning technique used to predict continuous numerical values based on input features. It aims to establish a functional relationship between independent variables and a dependent variable.

Classification: Classification is a type of supervised learning that categorizes input data into predefined labels. It involves training a model on labeled examples to learn patterns between input features and output classes. In classification, the target variable is a categorical value.

**3.Algorithm Implementation**

1.Naive Bayes (GaussianNB):

* Purpose: Classifies data based on Bayes' theorem, assuming independence between features.
* Implementation: GaussianNB() fits the model to the training data and predicts the target for the test data. The model assumes that features are normally distributed.

2.Support Vector Machine (SVC):

* Purpose: Finds the hyperplane that best separates different classes in the feature space.
* Implementation: SVC() fits the model to the training data and predicts class labels for the test data. It aims to maximize the margin between different classes.

3. Linear Regression (LinearRegression):

* Purpose: Predicts a continuous outcome based on input features. Here, it's used for binary classification by converting predictions to binary values.
* Implementation: LinearRegression() fits the model to the training data, predicts continuous values for the test data, and then rounds these predictions to binary values (0 or1).

4. Logistic Regression (LogisticRegression):

* Purpose: A classification algorithm that estimates the probability of a binary outcome using a logistic function.
* Implementation: LogisticRegression(max\_iter=1000) fits the model to the training data and predicts class probabilities, which are then converted to binary predictions.

5. Random Forest (Random Forest Classifier):

* Purpose: An ensemble method that uses multiple decision trees to make predictions. It aggregates the results of individual trees to improve accuracy and robustness.
* Implementation: Random Forest Classifier() fits multiple decision trees to the training data and combines their predictions to classify the test data.

6. Decision Tree (Decision Tree Classifier):

* Purpose: A tree-based model that splits the data into branches based on feature values to make predictions.
* Implementation: Decision Tree Classifier() fits a decision tree to the training data, where each node represents a decision based on a feature, and predictions are made by traversing the tree.

7. K-Nearest Neighbors (K Neighbors Classifier):

* Purpose: Classifies data based on the majority class among the k-nearest neighbors in the feature space.
* Implementation: K Neighbors Classifier() fits the model to the training data and predicts the class of the test data based on the majority vote from the nearest neighbors.

8. Neural Network (MLP Classifier):

* Purpose: A type of artificial neural network that uses multiple layer of neurons to model complex patterns in the data.
* Implementation: MLP Classifier( max\_iter=1000 ) fits the model to the training data, using multiple layers and activation functions to learn complex patterns and make predictions on the test data.

**4.Predication Comparison Report**

Highest Accuracy: Random Forest (0.6000)

Performance: Random Forest performs the best with an accuracy of 60%. It has the highest number of true negatives (21) and a relatively balanced number of true positives (15). It also maintains the lowest number of false positives and false negatives, making it the most reliable model on this dataset.

Second-Best Accuracy: Logistic Regression & Neural Network (0.5833)

Performance: Both Logistic Regression and Neural Network achieve an accuracy of 58.33%. The Neural Network has slightly more true negatives than Logistic Regression, but Logistic Regression has fewer false positives, making it slightly more conservative in predicting a crisis.

Third-Best Accuracy: Linear Regression (0.5667)

Performance: Linear Regression follows with an accuracy of 56.67%. It has balanced true positives and negatives but struggles slightly with false positives compared to Random Forest.

Moderate Performers: SVM, Decision Tree, KNN (0.5333 - 0.5500)

Performance: These models have comparable accuracies. They demonstrate a moderate balance between true positives and true negatives but have higher rates of false positives and negatives, indicating that they might not be as reliable as Random Forest or Logistic Regression for this dataset.

Lowest Accuracy: Naive Bayes (0.4833)

Performance: Naive Bayes has the lowest accuracy at 48.33%. It struggles with a high number of false positives and negatives, making it the least effective model for this particular dataset.

**5.Final Prediction**

Random Forest emerges as the top-performing model in terms of accuracy and balanced performance between predicting mental crises and non-crises.

**6.Conclusion**

Future generations face increasing mental health challenges due to factor such as social media pressures, economic uncertainty, and global crises, the need for innovative solutions like these becomes even more critical. The ability to predict and address mental health issues at an early stage could play a crucial role in mitigating the rising mental health crisis among younger populations, helping to foster resilience and well-being in the face of unprecedented challenges. Therefore, these predictive tools not only offer hope for timely intervention but also provide a crucial foundation for supporting the mental health of younger generations as they navigate these complex and evolving landscapes.