Name: Shalu Mary Abraham

Register No: 2241156

Class: 6BCA B

**ML Programs Report**

**Dataset** - **Remote Work and Mental Health**

**Description of Impact of Remote Work on Mental Health Prediction Dataset**

As remote work becomes the new norm, it's essential to understand its impact on employees' mental well-being. This dataset dives into how working remotely affects stress levels, work-life balance, and mental health conditions across various industries and regions.

This dataset provides valuable insights into key areas like work location (remote, hybrid, onsite), stress levels, access to mental health resources, and job satisfaction. It’s designed to help researchers, HR professionals, and businesses assess the growing influence of remote work on productivity and well-being. It includes key metrics such as job satisfaction, anxiety levels, and work-life balance.

**Key Features:**

**Diverse Respondents:** Data from different age groups and professions.

**Mental Health Insights:** Explore the effects of remote work conditions.

**Column Descriptions:**

1. **Employee\_ID:** Unique identifier for each employee.
2. **Age:** Age of the employee.
3. **Gender:** Gender of the employee.
4. **Job\_Role:** Current role of the employee.
5. **Industry:** Industry they work in.
6. **Work\_Location:** Whether they work remotely, hybrid, or onsite.
7. **Stress\_Level:** Their self-reported level of stress.
8. **Mental\_Health\_Condition**: Any mental health condition reported (Anxiety, Depression, etc.).
9. **Social\_Isolation\_Rating:** A self-reported rating (1-5) on how isolated they feel.
10. **Satisfaction\_with\_Remote\_Work:** How satisfied they are with remote work arrangements (Satisfied, Neutral, Unsatisfied).

**Overall description of what is done in all programs**

**P1 - Demonstrate data Exploration (Numeric Data)**

\* Comparison of columns - column distribution

\* Correlation Heatmap

**P2 - EDA on Categorical Values**

\* Have done interpretations on the distribution of numeric variables with histogram, distribution of nominal variables using count plot, and different other plotting graphs such as bar graph, scatterplot, box plot and violin plot

**P3 - KNN Classifier**

\* Detailed EDA

\* ⁠Train Test split

\* ⁠Trained the KNN model

\* ⁠Made predictions and evaluated the model

\* ⁠Found the best K value by using cross-validation

\* ⁠Retrained the model with Optimal K

**P4 - Naive Bayes Algorithm**

\* Detailed EDA

\* ⁠Trained Gaussian Naive Bayes Classifier (have got a perfect accuracy result) based on the data

\* ⁠Trained Multinominal Naive Bayes Classifier (have got a perfect accuracy result) based on the data

\* ⁠Done confusion matrix and found Gaussian naive bayes cross-validation scores

\* ⁠Checked unique values after encoding

\* ⁠ROC and AUC curve

\* ⁠Interpretation and Results

**P5 - Decision Tree**

\* Decision tree classifier with criterion Gini-index which includes the gini index, precision, recall and f1-score accuracy, macro average and the weighted average

\* ⁠Decision tree classifier with criterion Entropy which includes the entropy, precision, recall and f1-score accuracy, macro average and the weighted average

\* ⁠Analysis of precision and recall

\* ⁠ROC curve

\* ⁠Comparative Analysis

\* ⁠Visualisation of tree

\* ⁠Confusion matrix of both gini-index and entropy

As a results and conclusion, got a pretty good accuracy score for both gini-index and entropy as both of them are at the range of 0.33, 0.34. This means that the lower the value indicates higher purity (less disorder)

**P5 (continuation)** -

\* As program 5 (continuation), have done pre-pruning and post-pruning and have found the gini-index and entropy’s accuracy and f1 score and based on this I concluded that as a comparison from the previous scores before pruning the tree, the classifier with higher F1 Score and ROC AUC can be considered better for this dataset

\* ⁠ROC curve for both pre-pruned and post-pruned tree and again done visualisation of the tree

**P6 - Simple Linear Regression**

\* Checked the dependent and independent variables and visualized the change in the variables

\* ⁠Divided the data into independent and dependent variables, split the data into train and test sets

\* ⁠Train the algorithm and retrieved both the intercept and the slope

\* ⁠Found the predicted and the actual value

**P7 - Neural Network - Multi-Layer Perceptron**

\* Split the dataset and build the classification model

\* ⁠Making prediction and evaluated the model

**P8 - Clustering**

\* Checked the missing values, plotted it and displayed them then dropped the missing values

\* ⁠Done univariate analysis: visualizing the distributions of numerical variables

\* ⁠Done bivariate analysis and multivariate analysis

\* ⁠Performed clustering using K-Means

\* ⁠Used the Elbow method for optimal number of clusters

\* ⁠Found the silhouette score

\* ⁠Displayed the cluster centers and visualized the distribution of data points across clusters

\* ⁠Interpreted the clusters

\* ⁠Again visualized the clusters by using scatter plot