**Name:** Areshay Asad

**Section:** BS- AI (3A)

**Roll no:** 052

**AI PROJECT REPORT**

**Stress Prediction Project: Final Report**

**1. Project Title**

Stress Prediction System using Machine Learning

**1. Introduction**

Stress has become a significant concern in modern society, affecting both mental and physical health. Accurate stress detection systems can help individuals take timely actions to manage their well-being. This project leverages machine learning to create a Stress Prediction System based on health and lifestyle data. The system predicts stress levels and determines whether a user is likely to experience stress based on user-provided inputs.

**2. Objective**

The primary goals of this project are:

* To develop a machine learning model that predicts stress levels based on user inputs and health metrics.
* To implement an interactive system that determines whether a user is likely stressed based on predefined thresholds and machine learning predictions.
* To provide both quantitative (stress level) and qualitative (stress likelihood) outputs.

**3. Dataset Overview**

The project uses a Sleep Health and Lifestyle Dataset.

**Key Features:**

* **Health Metrics:**
  + Blood Pressure (split into Systolic BP and Diastolic BP)
  + Heart Rate
  + BMI Category
* **Sleep-Related Data:**

Sleep Duration

Quality of Sleep

Sleep Disorder

* **Lifestyle Attributes:**
  + Age
  + Daily Steps
  + Physical Activity Level
* **Target Variable:**
  + Stress Level: Represents the severity of stress on a numerical scale (e.g., 3 to 8).

**Dataset Challenges:**

* Missing values in both numeric and categorical features.
* Mixed data types (numerical, categorical, and text-based).
* Class imbalance in the Stress Level target variable.

**4. Project Workflow**

**4.1 Data Preprocessing**

1. **Splitting Columns:**

* The Blood Pressure column was split into Systolic BP and Diastolic BP for better feature handling.

1. **Handling Missing Values:**

* Numeric columns: Missing values were filled with the median.
* Categorical columns: Missing values were filled with the placeholder "Unknown".

1. **Categorical Encoding:**

* Used LabelEncoder to encode BMI Category and Sleep Disorder.

1. **Feature Selection:**

* Relevant features for prediction:
  + Age, Sleep Duration, Quality of Sleep, Physical Activity Level, BMI Category, Systolic BP, Diastolic BP, Heart Rate, Daily Steps, Sleep Disorder.
* Target variable: Stress Level.

**4.2 Feature Engineering**

**Scaling:**

Used StandardScaler to normalize numeric features for better model performance.

**Threshold-Based Logic:**

Added manual thresholds for features like sleep quality, physical activity, and blood pressure to support predictions.

**4.3** **Model Development**

1. **Pipeline Design:**

* A preprocessing pipeline was created using ColumnTransformer to standardize features before feeding them into the model.
* The pipeline integrates both data preprocessing and the machine learning model.

1. **Model Selection:**

* Random Forest Classifier: Chosen for its robustness and ability to handle mixed data types.

1. **Hyperparameter Tuning:**

Used GridSearchCV to optimize parameters such as:

* n\_estimators: Number of trees in the forest.
* max\_depth: Maximum depth of the trees.
* min\_samples\_split: Minimum samples required to split a node.

**4.4 Evaluation**

**Metrics:**

* Accuracy: Proportion of correctly classified instances.
* Classification Report: Includes precision, recall, and F1-score.
* Confusion Matrix: Shows the distribution of correct and incorrect predictions.

**5. Results**

**5.1 Model Performance**

* **Best Parameters:**
  + max\_depth: 10
  + min\_samples\_split: 5
  + n\_estimators: 100
* **Accuracy: 100.00%**

**Classification Report:**

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| Stress Level | Precision | Recall | F1-Score | Support |
| 3 | 1.00 | 1.00 | 1.00 | 12 |
| 4 | 1.00 | 1.00 | 1.00 | 10 |
| 5 | 1.00 | 1.00 | 1.00 | 14 |
| 6 | 1.00 | 1.00 | 1.00 | 10 |
| 7 | 1.00 | 1.00 | 1.00 | 12 |
| 8 | 1.00 | 1.00 | 1.00 | 17 |

* **Confusion Matrix**:

#### **[[12 0 0 0 0 0]**

#### **[ 0 10 0 0 0 0]**

#### **[ 0 0 14 0 0 0]**

#### **[ 0 0 0 10 0 0]**

#### **[ 0 0 0 0 12 0]**

#### **[ 0 0 0 0 0 17]]**

#### ****5.2 Interactive Prediction****

The system predicts:

1. **Stress Likelihood**:
   * Based on thresholding logic.
2. **Stress Level**:
   * Numerical level (e.g., 6) with a descriptive label (e.g., "High Stress").

**7. Tools and Technologies**

1. **Programming Language**: Python
2. **Libraries**:
   * **pandas**: Data manipulation.
   * **numpy**: Numerical computations.
   * **scikit-learn**: Machine learning and evaluation.
3. **Machine Learning Algorithm**: Random Forest Classifier.
4. **Evaluation Tools**:
   * Accuracy score, classification report, and confusion matrix.

**8. Challenges and Solutions**

1. **Handling Missing Values**:
   * Solution: Used median for numeric columns and "Unknown" for categorical columns.
2. **Categorical Encoding**:
   * Solution: Used LabelEncoder for categorical variables.
3. **Class Imbalance**:
   * Solution: Applied class\_weight='balanced' in the Random Forest Classifier.