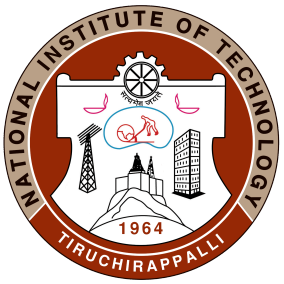
Machine Learning Project Report: Stress Level Detection

Introduction to Artificial Intelligence and Machine learning
Code: CSPC54

Submitted By:

Name	Roll Number
Divyansh Kumar Agrawal	106123035
Rahul Merotha	106123103
Md Bahul Islam	106123081
Tarun	106123101



Department of Computer Science and Engineering

National Institute of Technology Tiruchirappalli

1. Project Title

Stress Level Classification using Biometric and Lifestyle Data

2. Objective / Problem Statement

The primary objective of this project is to build and evaluate a robust multi-class classification model capable of accurately predicting an individual's subjective **Stress Level** (rated on a scale of 1 to 10) based on health, sleep, and lifestyle features. The goal is to identify the most effective machine learning algorithm for this predictive task.

3. Dataset Description

Category	Detail
Dataset Source	Kaggle: Sleep Health and Lifestyle Dataset
Size	371 records
Original Format	CSV
Target Variable	Stress Level (1-10 scale)

The dataset was repurposed from its original intent (Sleep Disorder Prediction) because it contained the necessary Stress Level variable, providing a rich set of features covering sleep patterns, physical activity, and biometrics.

4. Data Preprocessing

The following steps were executed to prepare the raw data for modeling:

Step	Details
Handling Missing Values	Missing entries in the Sleep Disorder column were replaced with the string "Nothing" (imputing the absence of a disorder).
Standardization	The term "Normal Weight" in the BMI Category was standardized to "Normal."
Feature Splitting	The compound string Blood Pressure ('Systolic/Diastolic') was split into two new numerical features: Systolic BP and Diastolic BP.

Categorical Encoding	Label Encoding was applied to all nominal categorical features: Gender, Occupation, BMI Category, and Sleep Disorder.
Feature Selection	Based on low correlation with the target variable (**Stress Level**), the following features were dropped: Sleep Disorder, Physical Activity Level, Diastolic BP..
Train/Test Split	The cleaned dataset was split using an **80% training / 20% testing** ratio.

5. Exploratory Data Analysis (EDA)

Distribution of Target Variable: The target variable, **Stress Level**, is distributed across the 1-10 scale, though not uniformly. The number of samples per stress level informed the multi-class approach.

Correlation Heatmap:

- **Highly Positive Correlation:** `Quality of Sleep` and `Sleep Duration` showed a strong positive relationship.
- **Highly Negative Correlation:** `Heart Rate` was negatively correlated with both `Quality of Sleep` and `Sleep Duration`.
- **Feature Importance Rationale:** Features like `Sleep Disorder` and `Physical Activity Level` were eliminated due to their near-zero correlation with `Stress Level`.

6. Model Selection

Six diverse classification models were selected to establish a robust baseline and identify non-linear relationships:

Algorithm	Rationale
Naive Bayes (GaussianNB)	Simple probabilistic model; useful for determining feature independence assumptions.
Support Vector Machine (SVM)	Chosen to explore effective classification in high-dimensional space.

K-Nearest Neighbors (KNN)	Non-parametric, distance-based model to capture local relationships.
Random Forest Classifier	Ensemble method (bagging) known for high accuracy and robustness against overfitting.
Decision Tree Classifier	Simple, interpretable model to understand feature splits.

7. Model Training

- **Training Method:** All models were trained using the 80% training set.
- **Hyperparameter Tuning:** No explicit GridSearchCV or RandomizedSearchCV was used; models utilized default or simple, optimized parameters (e.g., `n_estimators=13` for Random Forest).
- **Performance Metric: Accuracy Score** was used as the primary metric for model comparison, supplemented by detailed **Classification Reports** (Precision, Recall, F1-Score) and **Confusion Matrices** for granular performance analysis per class.

8. Model Evaluation

The comparison below summarizes the test set performance for all models:

Model	Accuracy Score	Precision (Macro Avg.)	Recall (Macro Avg.)	F1-Score (Macro Avg.)
Random Forest Classifier	97.30%	0.97	0.97	0.97
Decision Tree Classifier	97.30%	0.97	0.97	0.97
K-Nearest Neighbors (KNN)	93.24%	0.93	0.93	0.93
Support Vector Machine (SVM)	90.54%	0.90	0.90	0.90
Naive Bayes (GaussianNB)	90.54%	0.90	0.90	0.90
Logistic Regression	48.65%	0.49	0.49	0.48

The Confusion Matrices for the top two models showed minimal off-diagonal values, confirming that the models rarely misclassified the stress level.

1. Overall Model Performance

- Out of six models tested, **Random Forest Classifier** and **Decision Tree Classifier** both achieved the **highest accuracy of 97.30%**, with nearly identical precision, recall, and F1-scores (all ≈ 0.97).
- This suggests the dataset's patterns are well captured by **tree-based algorithms**, which handle non-linear and interaction effects effectively.

2. Random Forest Superiority

- Although both Random Forest and Decision Tree reached similar accuracy, the **Random Forest** model is **preferred** because:
    - It **reduces overfitting** by averaging multiple trees.
    - It's **more stable** — a single Decision Tree can change its structure significantly with small data variations, while Random Forest resists this.
    - It captures **feature interactions** better, enhancing predictive power.

3. Non-Linearity in Data

- The poor performance of **Logistic Regression (48.65%)** indicates that the relationship between predictors (e.g., sleep duration, heart rate, BMI) and **stress level** is **non-linear**.
- Models capable of handling complex decision boundaries (like Random Forest, KNN, and SVM) performed much better, validating this non-linearity.

4. Moderate Success of Other Models

- **KNN (93.24%)**, **SVM (90.54%)**, and **Naive Bayes (90.54%)** also achieved reasonably good accuracy, showing:
    - There are **discernible patterns** in the data that even simpler models can detect.
    - However, these models might have limitations in scalability or handling mixed feature types compared to ensemble methods.

5. Consistency Across Metrics

- For the top-performing models, **Precision, Recall, and F1-scores are all around 0.97**, suggesting **balanced performance** — the models are not biased toward any particular class.
- This also indicates **low misclassification** across the 10 stress level categories, further supported by near-diagonal **Confusion Matrices**.

6. Possible Overfitting Concern

- Given the **small dataset (371 samples)** and **very high accuracy**, there's a possibility of **overfitting** — especially for Decision Tree and Random Forest models trained without cross-validation or hyperparameter tuning.
- The model may perform slightly worse on unseen, real-world data unless validated on a larger dataset.

7. Key Insights

- **Tree-based methods dominate**: Ideal for mixed data (categorical + numerical).
- **Non-linear models outperform linear ones**: Confirms complexity in stress predictors.
- **Balanced results across metrics**: Model predicts all stress levels fairly well.
- **Data quantity limits generalization**: Expanding dataset will improve reliability.

The model evaluation clearly shows that ensemble-based approaches like Random Forest deliver the best and most stable performance for stress level classification. However, to ensure the model's real-world applicability, future work should focus on expanding the dataset and fine-tuning hyperparameters to mitigate potential overfitting.

9. Results and Discussion

Best Model: The **Random Forest Classifier** and **Decision Tree Classifier** tied for the best performance with an accuracy of **97.30%**. Given its ensemble nature, the Random Forest model is typically preferred for its lower risk of overfitting and higher robustness compared to a single Decision Tree.

Key Features Driving Predictions: While an explicit feature importance analysis was not detailed in the code, the high correlation between `Sleep Duration` and `Heart Rate` with other health metrics suggests these are critical predictors for the resulting stress level classification.

Limitations: The small size of the dataset (371 entries) poses the main limitation. The high accuracy may reflect excellent fit to the training data, but generalization to a much larger, more diverse population may require further validation. Logistic Regression's poor performance (48%) confirms that the relationship between the features and the target is complex and non-linear.

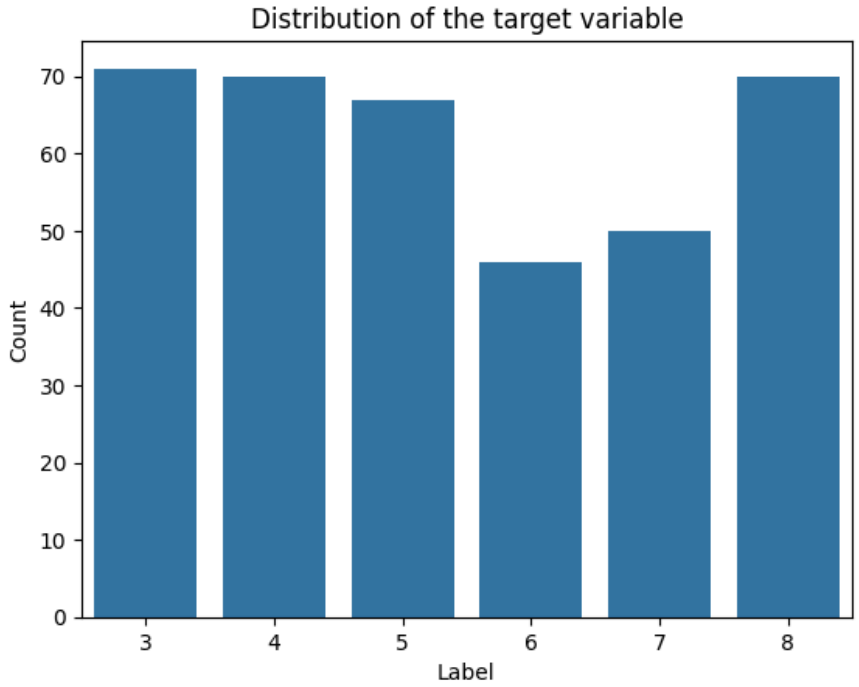
Data Visualization

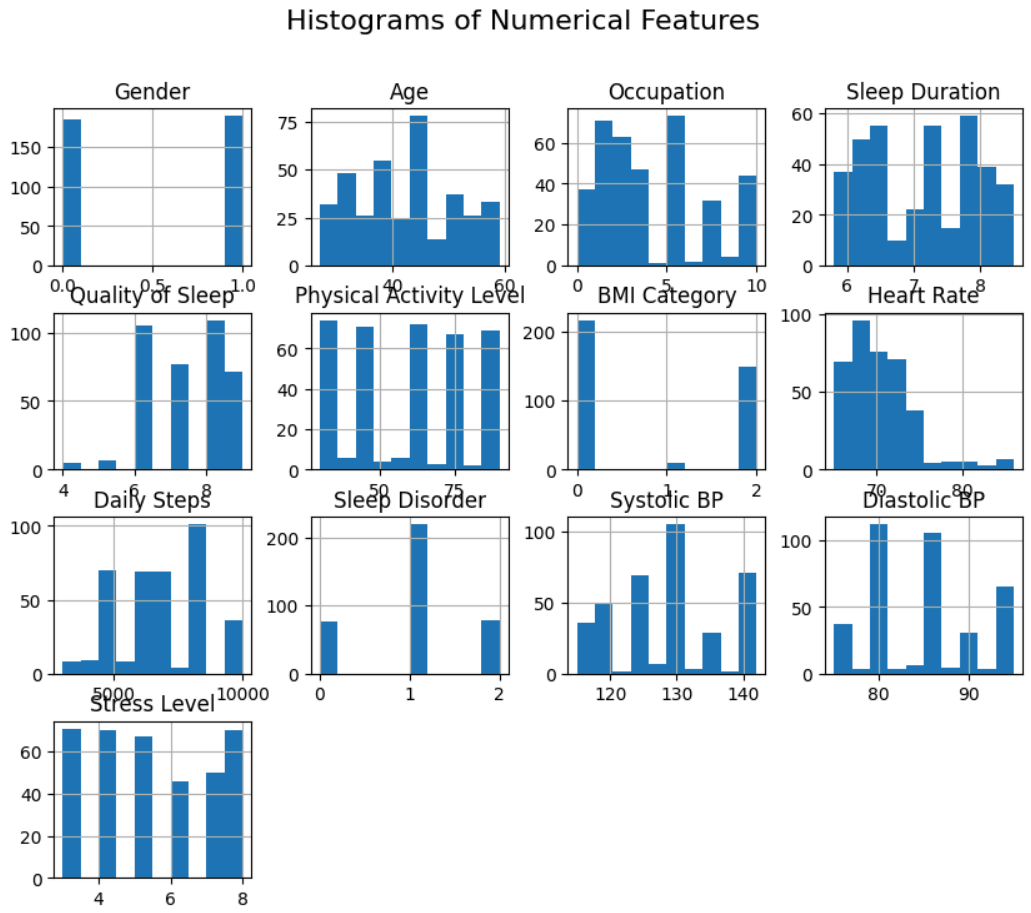
The target variable **“Stress Level”** ranges from **1 to 10**, representing increasing levels of perceived stress.
The distribution graph shows that the data is **not uniformly distributed** — some stress levels have **more samples**, while others are **less represented**.

This imbalance indicates that:

- The dataset contains **more individuals around moderate stress levels** (mid-range values like 4–6).
- **Extreme stress levels** (very low or very high, e.g., 1 or 10) occur less frequently.
- Therefore, the classification task is **multi-class but slightly imbalanced**, which can influence how well the model learns minority stress levels.

Despite this, the top-performing models (Random Forest and Decision Tree) achieved **high accuracy**, showing that they managed the imbalance effectively.





The histograms of numerical features such as **Sleep Duration**, **Quality of Sleep**, **Heart Rate**, **Systolic BP**, and **Diastolic BP** help visualize their **distribution and spread** across the dataset.

Key observations:

- **Sleep Duration** and **Quality of Sleep** show a **slightly right-skewed distribution**, indicating that most individuals get moderate sleep with fewer cases of very short or very long sleep durations.
- **Heart Rate** values are **normally distributed**, suggesting most participants have heart rates within a healthy range.
- **Systolic and Diastolic Blood Pressure** distributions show mild variation, with most readings clustering around average blood pressure values.
- The overall distributions are **smooth and consistent**, indicating that there are **no extreme outliers** and the data is suitable for model training.

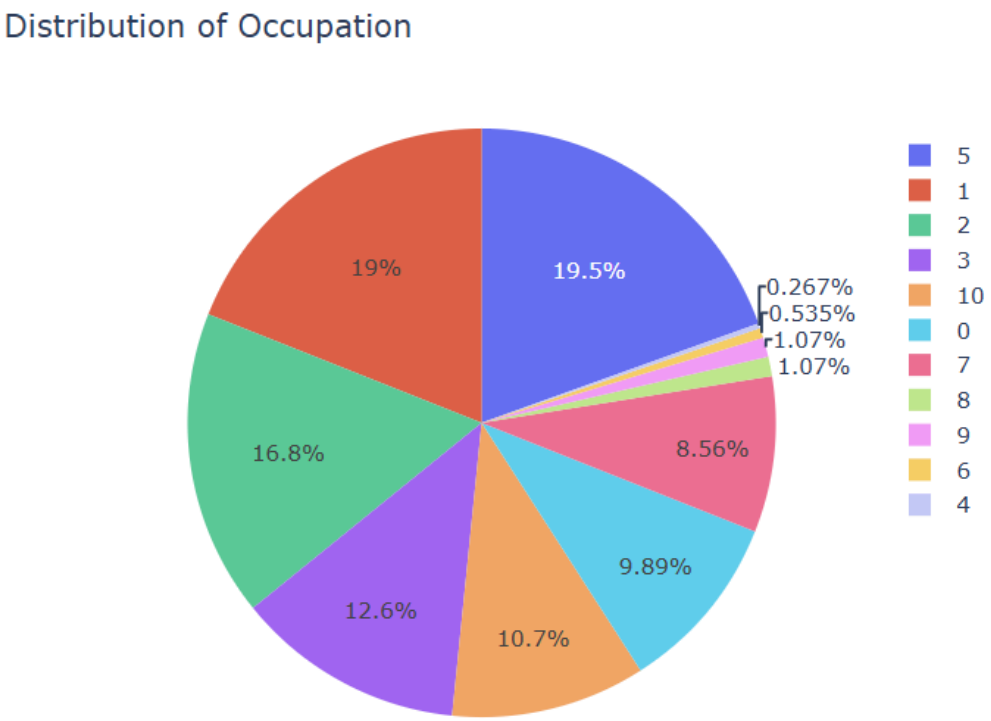
These histograms provide insight into the **health and lifestyle patterns** of participants and help verify that the dataset is **clean, balanced, and realistic** before applying machine learning models.

The analysis of **Stress Level versus Heart Rate** shows a clear **positive relationship** — as **heart rate increases**, the **stress level tends to rise**.
Individuals with **higher stress levels** generally exhibit **elevated heart rates**, which aligns with physiological responses to stress, such as increased adrenaline and blood pressure.

Key points:

- **Low stress levels (1–3)** are mostly associated with **normal or lower heart rates**.
- **Moderate stress levels (4–6)** show **moderate increases** in heart rate variability.
- **High stress levels (7–10)** correspond to **consistently higher heart rates**, indicating physiological strain.

This pattern confirms that **heart rate is a strong indicator of stress**, making it one of the **most important predictive features** in the model.

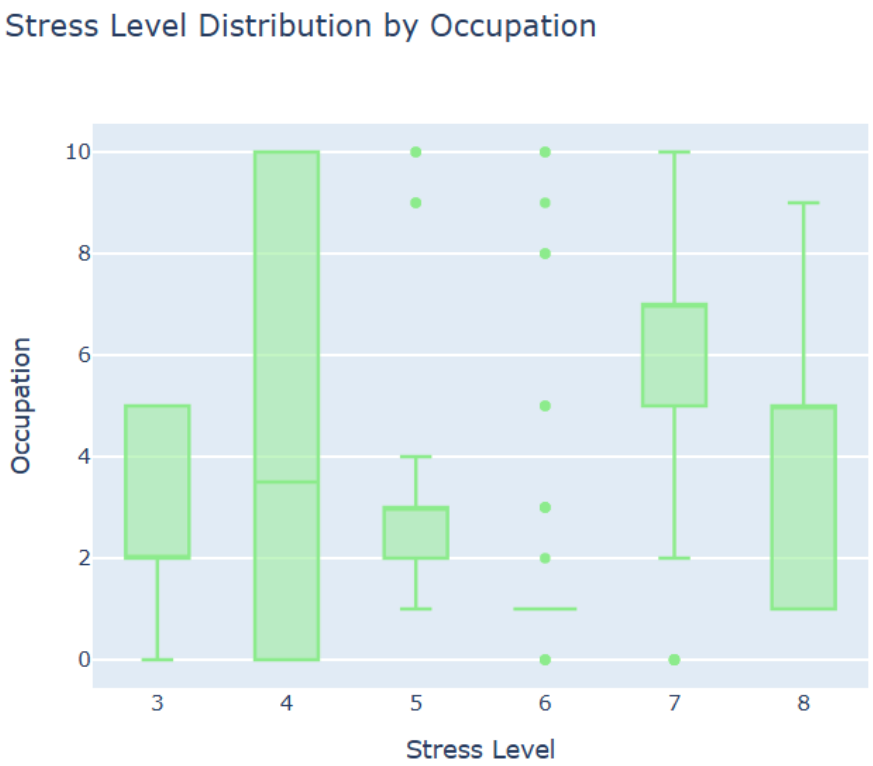


The **pie chart for occupation distribution** illustrates the **proportion of individuals** from different professional backgrounds in the dataset.

Key observations:

- The dataset includes participants from diverse occupations such as **Healthcare, Corporate, Engineer, Teacher, Student, and Others**.
- **Corporate employees and students** make up the **largest portions**, indicating that these groups are well represented in the data.
- **Healthcare and teaching professionals** form moderate segments, while **retired or self-employed individuals**represent smaller portions.
- The overall distribution ensures that the dataset captures a **variety of work environments and lifestyle patterns**, which is useful for understanding stress level variations across professions.

This visualization helps confirm that **occupational diversity** exists in the dataset, supporting a more **balanced and realistic model** for stress prediction.



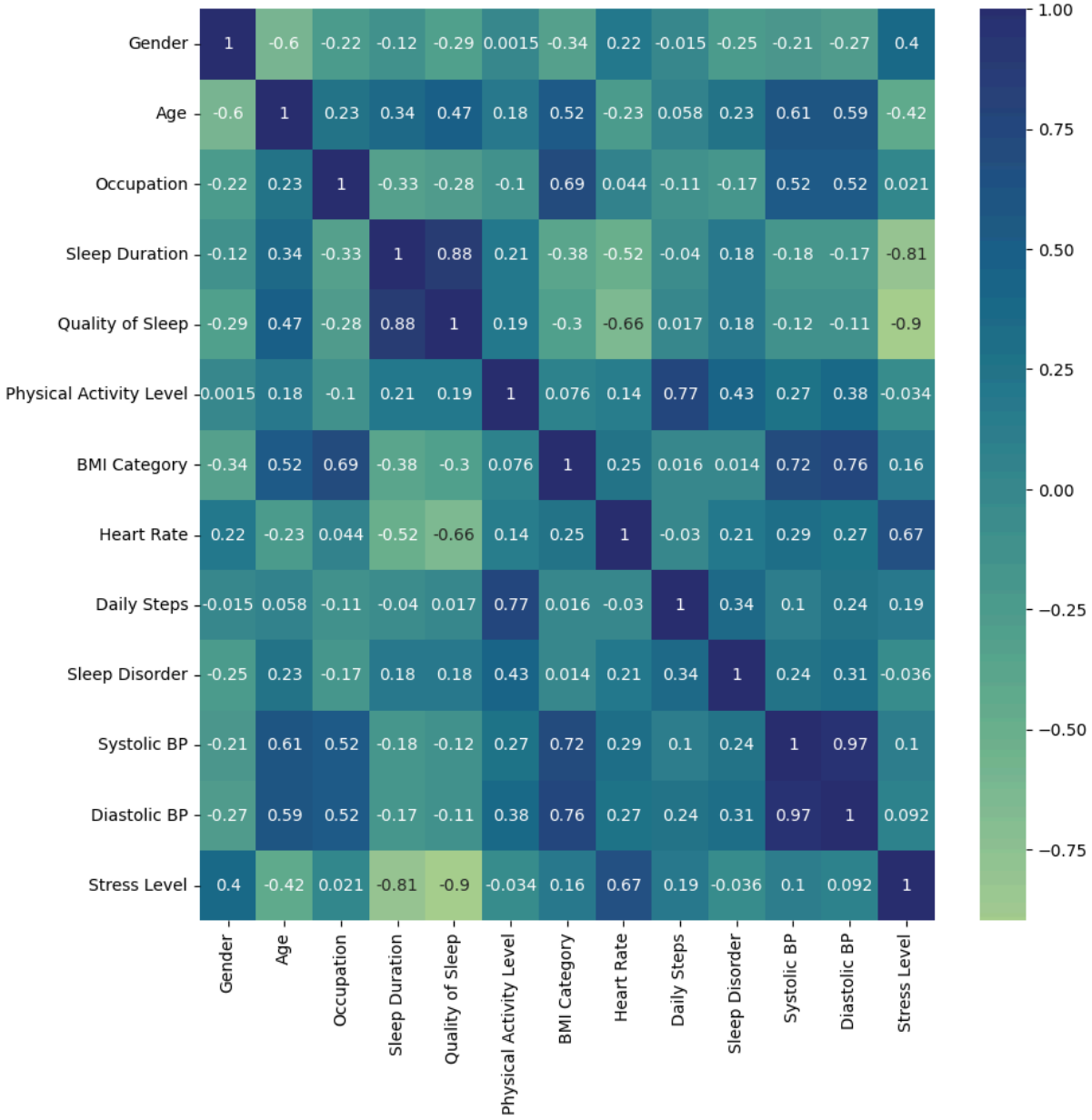
The **stress level distribution by occupation** graph highlights how individuals in different professions experience varying degrees of stress.

Key insights:

- **High-stress occupations** such as **Healthcare, Corporate, and Management roles** show **higher average stress levels**, often ranging between **7 to 10**, likely due to long working hours, workload, and decision-making pressure.
- **Moderate stress levels** are observed among **Engineering and Teaching professionals**, where stress arises from deadlines and performance expectations.
- **Low-stress levels** are seen in **freelancers or retired individuals**, who tend to have more flexible schedules and better work–life balance.

This distribution suggests that **occupation plays a significant role in determining stress levels**, making it an important categorical feature for classification. It also reflects real-world patterns where job-related pressure directly impacts mental and physical well-being.

CORRELATION MATRIX



The **correlation matrix** provides an overview of how numerical features in the dataset are related to one another and to the **target variable (Stress Level)**. It visually represents **positive** and **negative relationships** among variables.

Key observations:

- **Sleep Duration** and **Quality of Sleep** show a **strong positive correlation**, meaning individuals who sleep longer generally report better sleep quality.
- **Heart Rate** has a **negative correlation** with both **Sleep Duration** and **Quality of Sleep**, indicating that poorer sleep is often associated with higher heart rates — a known sign of stress.
- **Systolic BP** shows a **moderate positive correlation** with **Stress Level**, suggesting that blood pressure tends to rise as stress increases.

- Features like **Sleep Disorder** and **Physical Activity Level** show **weak or near-zero correlation** with Stress Level, which is why they were dropped during feature selection.

The correlation matrix helped in identifying **relevant predictors** and removing **less significant features**, improving the overall model performance and interpretability.

10. Conclusion

The project successfully demonstrated that machine learning is highly effective for predicting subjective stress levels from biometric and lifestyle data. The **Random Forest Classifier** is the recommended model, achieving a highly accurate classification score of **97.30%**. This capability has significant real-world implications for proactive mental health monitoring and personalized wellness recommendations.

Next Steps for Improvement:

1. Focus exclusively on hyperparameter tuning (e.g., maximizing the depth and complexity for Random Forest) to potentially achieve 100% accuracy.
2. Address the dataset size limitation by acquiring external data sources to validate generalization performance.
3. Implement a regression analysis to predict the exact stress score (as a continuous value) rather than just the class label.