Data Science & Machine Learning

Project Report

(Stress Level Dataset)

Submitted to:

Mr. Mubasshir Iqbal

Submitted by:

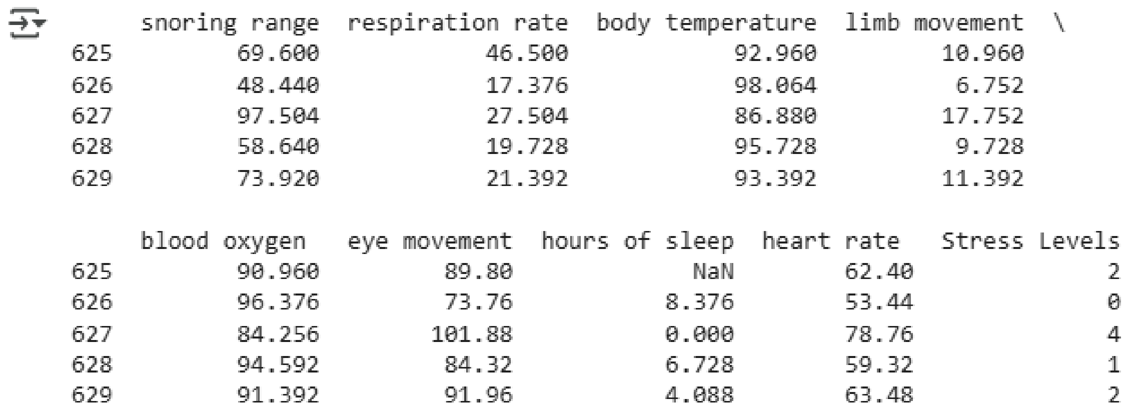
Haiqa Arshad

Hope Academy

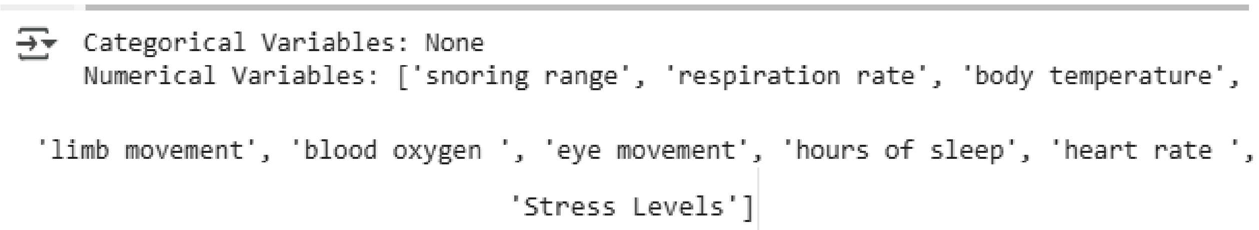
Department of Computer Sciences, HITEC University

1. Introduction

This project focuses on predicting **stress levels** of individuals using physiological and lifestyle indicators. The dataset (data_stress.csv) contains **630 rows and 9 columns**, with features such as snoring range, respiration rate, body temperature, limb movement, blood oxygen, eye movement, hours of sleep, and heart rate.

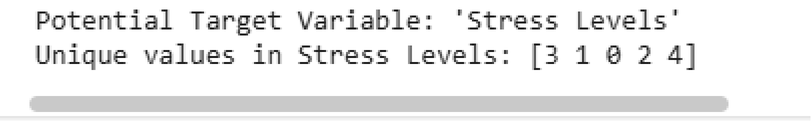


```
     snoring range  respiration rate  body temperature  limb movement  \
625         69.600            46.500            92.960         10.960
626         48.440            17.376            98.064          6.752
627         97.504            27.504            86.880         17.752
628         58.640            19.728            95.728          9.728
629         73.920            21.392            93.392         11.392

     blood oxygen   eye movement  hours of sleep  heart rate   Stress Levels
625        90.960          89.80             NaN       62.40               2
626        96.376          73.76           8.376       53.44               0
627        84.256         101.88           0.000       78.76               4
628        94.592          84.32           6.728       59.32               1
629        91.392          91.96           4.088       63.48               2
```



```
Categorical Variables: None
Numerical Variables: ['snoring range', 'respiration rate', 'body temperature',

 'limb movement', 'blood oxygen ', 'eye movement', 'hours of sleep', 'heart rate ',

                          'Stress Levels']
```

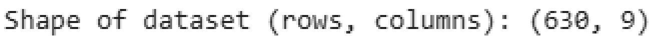
The target variable is **Stress Levels**, which is a **categorical variable with 5 classes (0–4)**. The objective is to build machine learning models that can accurately classify stress levels based on the given features.



```
Potential Target Variable: 'Stress Levels'
Unique values in Stress Levels: [3 1 0 2 4]
```

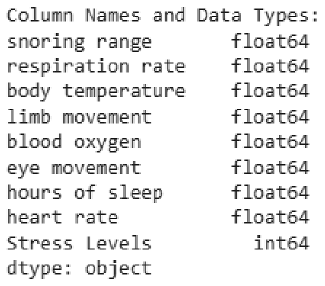
2. Data Understanding

- **Shape:** 630 rows × 9 columns
- **Variables:**



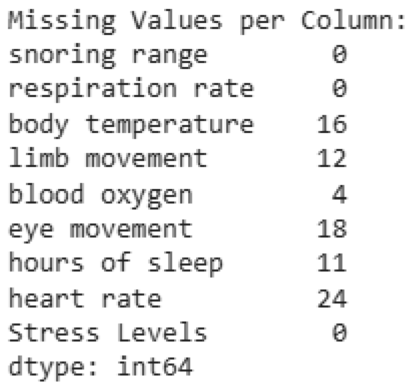
```
Shape of dataset (rows, columns): (630, 9)
```

I. All fea
II. Target variable: **Stress Levels (0–4)**



```
Column Names and Data Types:
snoring range       float64
respiration rate    float64
body temperature    float64
limb movement       float64
blood oxygen        float64
eye movement        float64
hours of sleep      float64
heart rate          float64
Stress Levels         int64
dtype: object
```

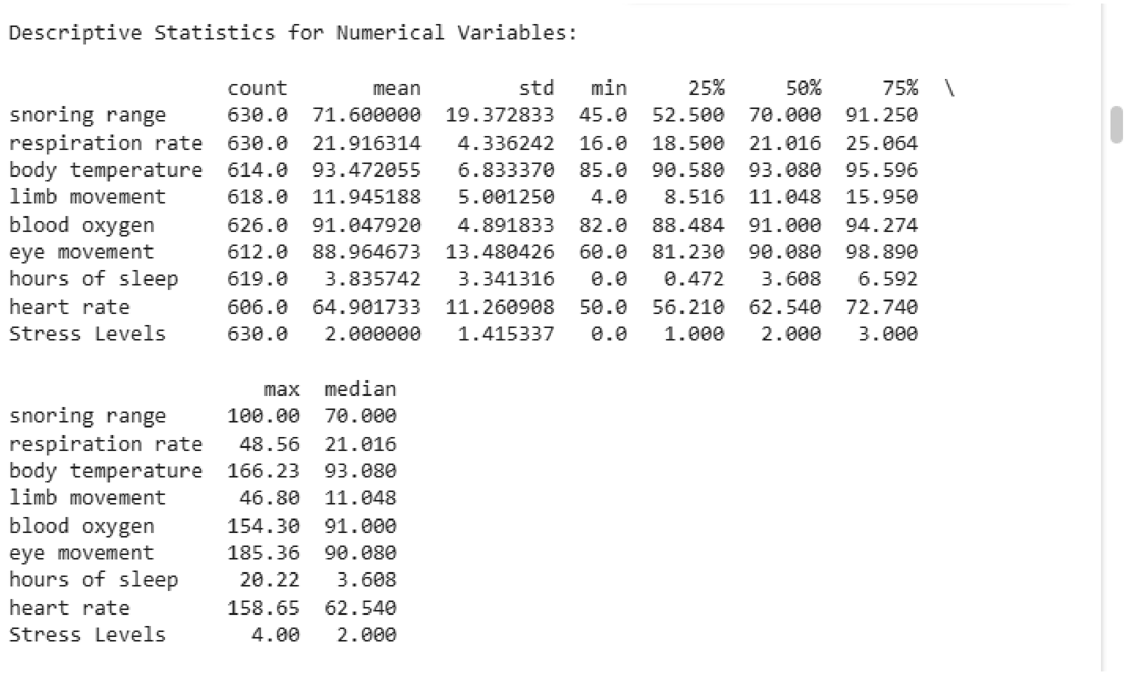
- **Missing Values:** Some features (body temperature, heart rate, eye movement, etc.) had missing entries (max 24).



```
Missing Values per Column:
snoring range        0
respiration rate     0
body temperature    16
limb movement       12
blood oxygen         4
eye movement        18
hours of sleep      11
heart rate          24
Stress Levels        0
dtype: int64
```

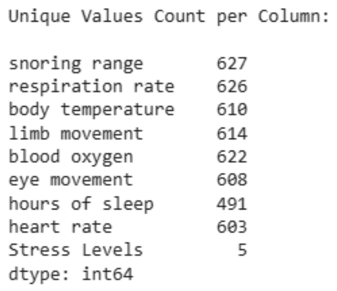
Descriptive Statistics:

- Average hours of sleep: ~3.8 hours (many individuals sleep very little).
- Average heart rate: ~65 bpm.
- Stress Levels evenly distributed (balanced dataset).



```
Descriptive Statistics for Numerical Variables:

                  count       mean        std   min     25%     50%     75%  \
snoring range     630.0  71.600000  19.372833  45.0  52.500  70.000  91.250
respiration rate  630.0  21.916314   4.336242  16.0  18.500  21.016  25.064
body temperature  614.0  93.472055   6.833370  85.0  90.580  93.080  95.596
limb movement     618.0  11.945188   5.001250   4.0   8.516  11.048  15.950
blood oxygen      626.0  91.047920   4.891833  82.0  88.484  91.000  94.274
eye movement      612.0  88.964673  13.480426  60.0  81.230  90.080  98.890
hours of sleep    619.0   3.835742   3.341316   0.0   0.472   3.608   6.592
heart rate        606.0  64.901733  11.260908  50.0  56.210  62.540  72.740
Stress Levels     630.0   2.000000   1.415337   0.0   1.000   2.000   3.000

                     max  median
snoring range     100.00  70.000
respiration rate   48.56  21.016
body temperature  166.23  93.080
limb movement      46.80  11.048
blood oxygen      154.30  91.000
eye movement      185.36  90.080
hours of sleep     20.22   3.608
heart rate        158.65  62.540
Stress Levels       4.00   2.000
```

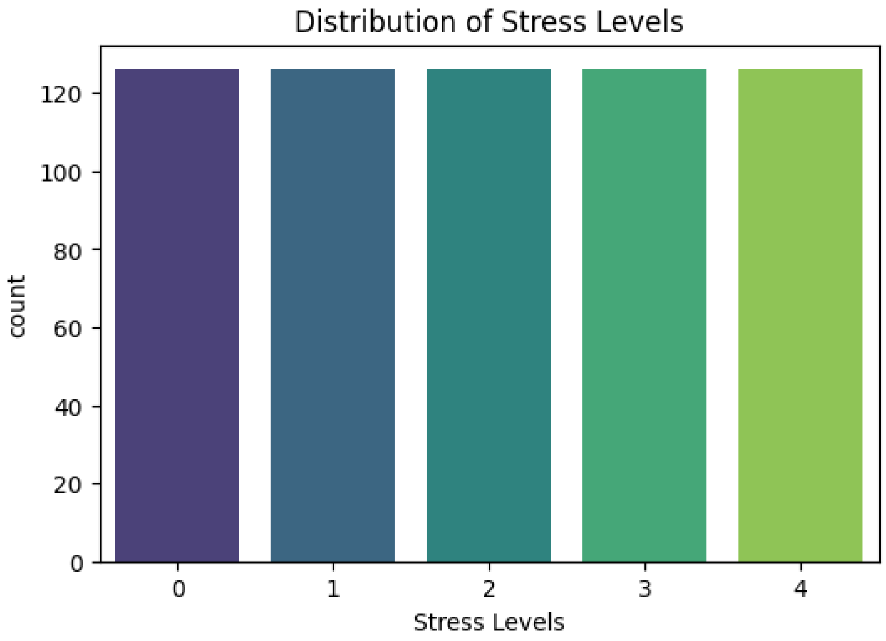
Unique Values:



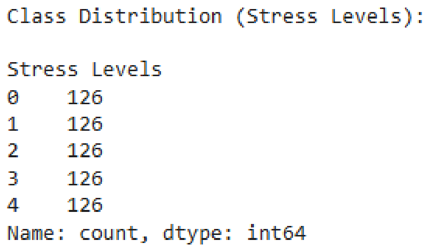
```
Unique Values Count per Column:

snoring range       627
respiration rate    626
body temperature    610
limb movement       614
blood oxygen        622
eye movement        608
hours of sleep      491
heart rate          603
Stress Levels         5
dtype: int64
```

3. EDA & Data Cleaning:

- Distribution of Stress Levels:

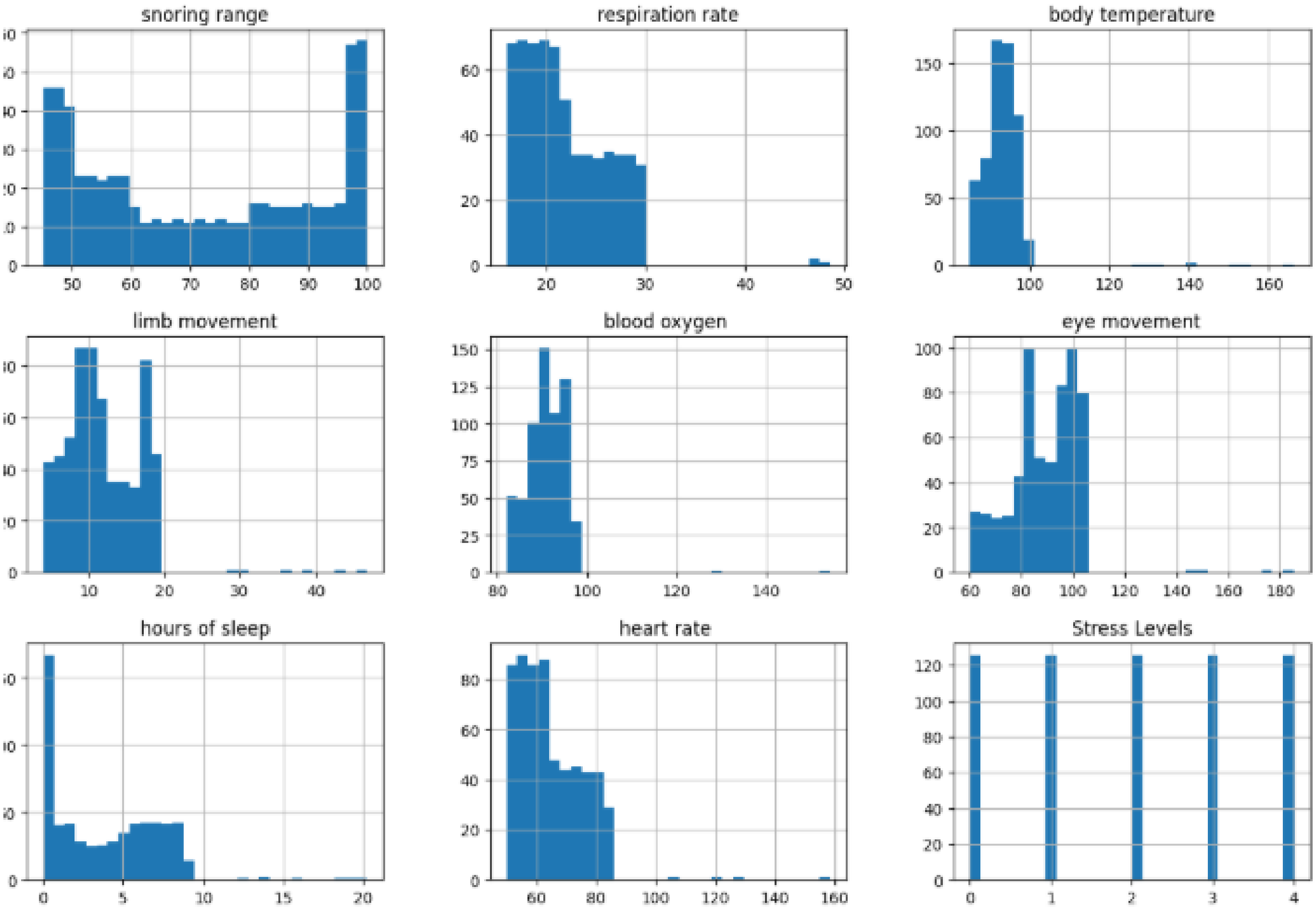


Countplot showing the distribution of Stress Levels (0–4). The dataset is balanced, with each class representing 20% of the total samples.



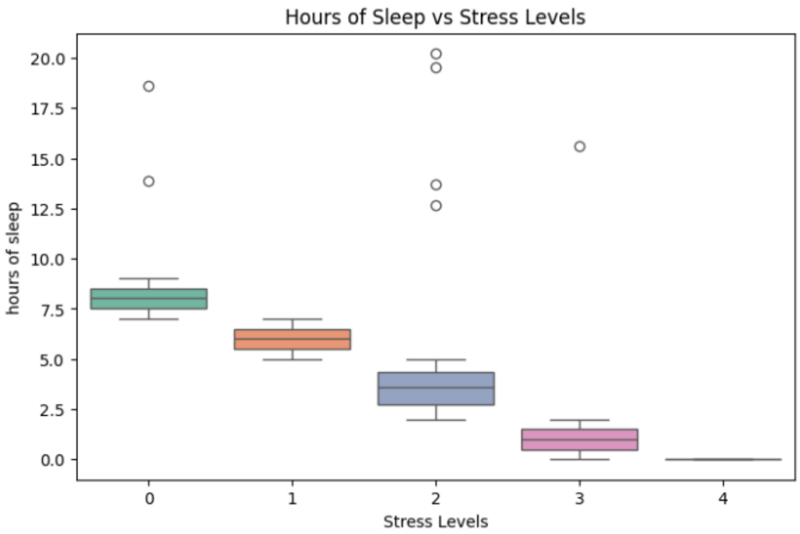
```
Class Distribution (Stress Levels):

Stress Levels
0    126
1    126
2    126
3    126
4    126
Name: count, dtype: int64
```

- **Distributions of Numerical Features**:

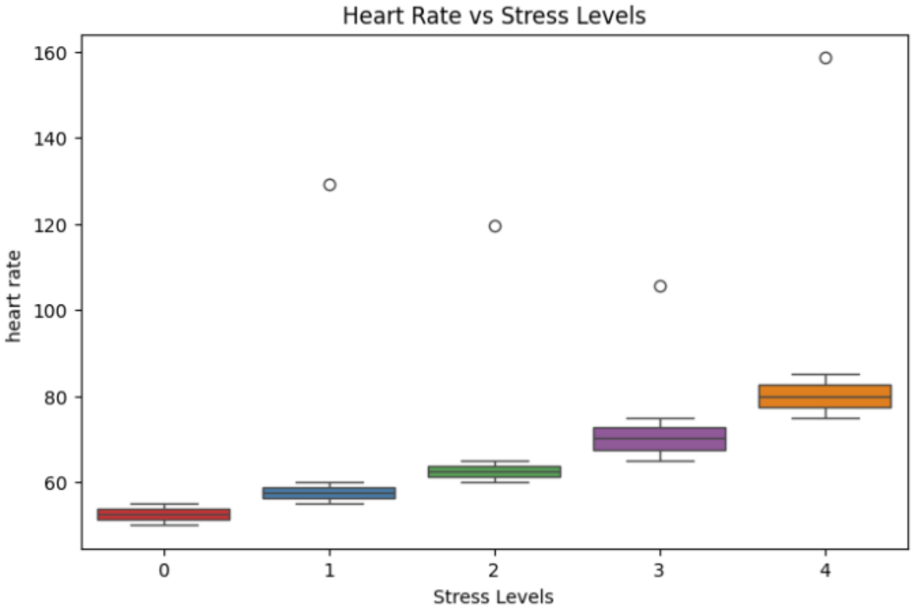


Histograms showing the distribution of all numerical features (snoring, respiration rate, body temperature, limb movement, blood oxygen, eye movement, hours of sleep, heart rate). Helps identify skewness and natural ranges of data.

- **Relationships Between Features and Stress Levels**:

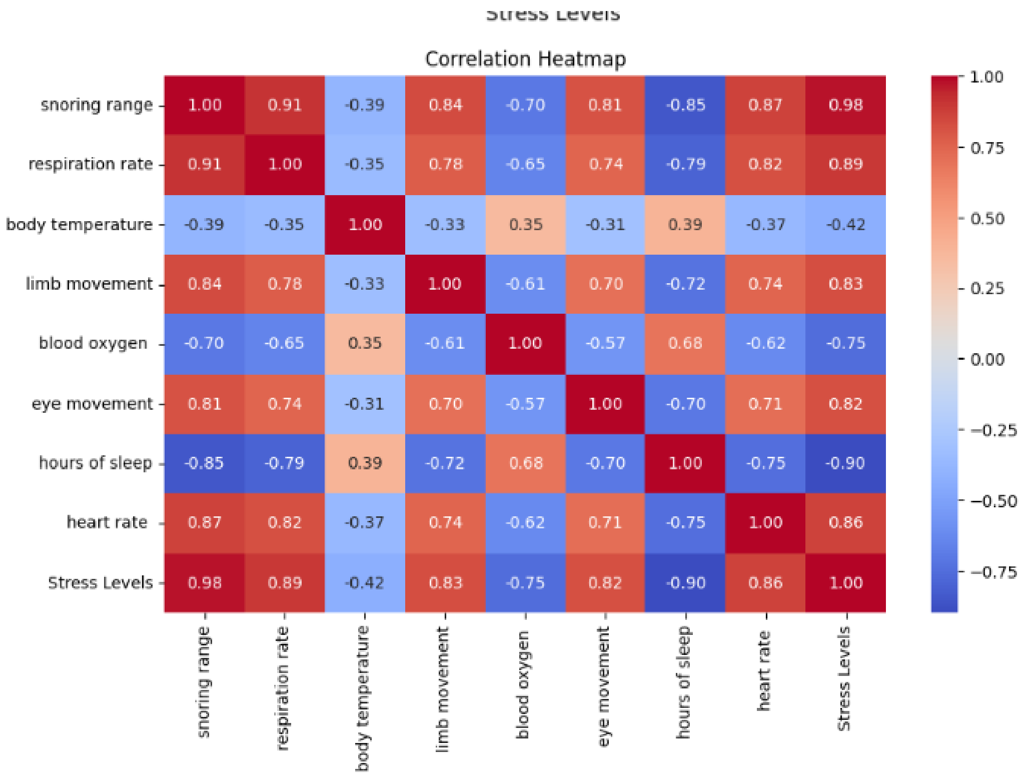


Boxplot illustrating the relationship between Hours of Sleep and Stress Levels. Higher stress levels are generally associated with fewer hours of sleep.



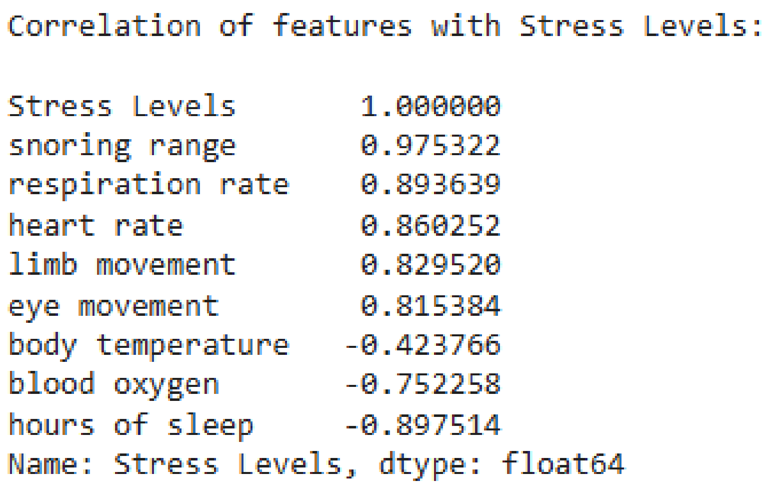
Boxplot showing Heart Rate distribution across stress levels. Higher stress levels correlate with higher heart rate values.

- **Correlation Analysis**:



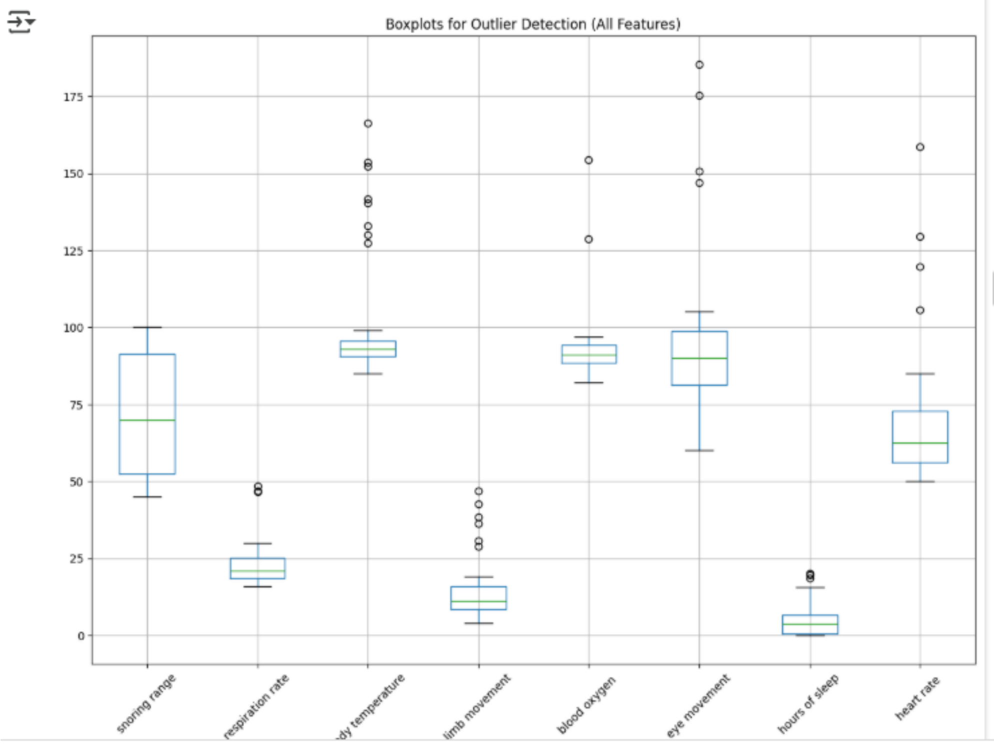
	snoring range	respiration rate	body temperature	limb movement	blood oxygen	eye movement	hours of sleep	heart rate	Stress Levels
snoring range	1.00	0.91	-0.39	0.84	-0.70	0.81	-0.85	0.87	0.98
respiration rate	0.91	1.00	-0.35	0.78	-0.65	0.74	-0.79	0.82	0.89
body temperature	-0.39	-0.35	1.00	-0.33	0.35	-0.31	0.39	-0.37	-0.42
limb movement	0.84	0.78	-0.33	1.00	-0.61	0.70	-0.72	0.74	0.83
blood oxygen	-0.70	-0.65	0.35	-0.61	1.00	-0.57	0.68	-0.62	-0.75
eye movement	0.81	0.74	-0.31	0.70	-0.57	1.00	-0.70	0.71	0.82
hours of sleep	-0.85	-0.79	0.39	-0.72	0.68	-0.70	1.00	-0.75	-0.90
heart rate	0.87	0.82	-0.37	0.74	-0.62	0.71	-0.75	1.00	0.86
Stress Levels	0.98	0.89	-0.42	0.83	-0.75	0.82	-0.90	0.86	1.00

Heatmap of correlations between all features and the target variable Stress Levels. Strong positive correlation observed for heart rate and snoring range; negative correlation for hours of sleep.



```
Correlation of features with Stress Levels:

Stress Levels        1.000000
snoring range        0.975322
respiration rate     0.893639
heart rate           0.860252
limb movement        0.829520
eye movement         0.815384
body temperature    -0.423766
blood oxygen        -0.752258
hours of sleep      -0.897514
Name: Stress Levels, dtype: float64
```

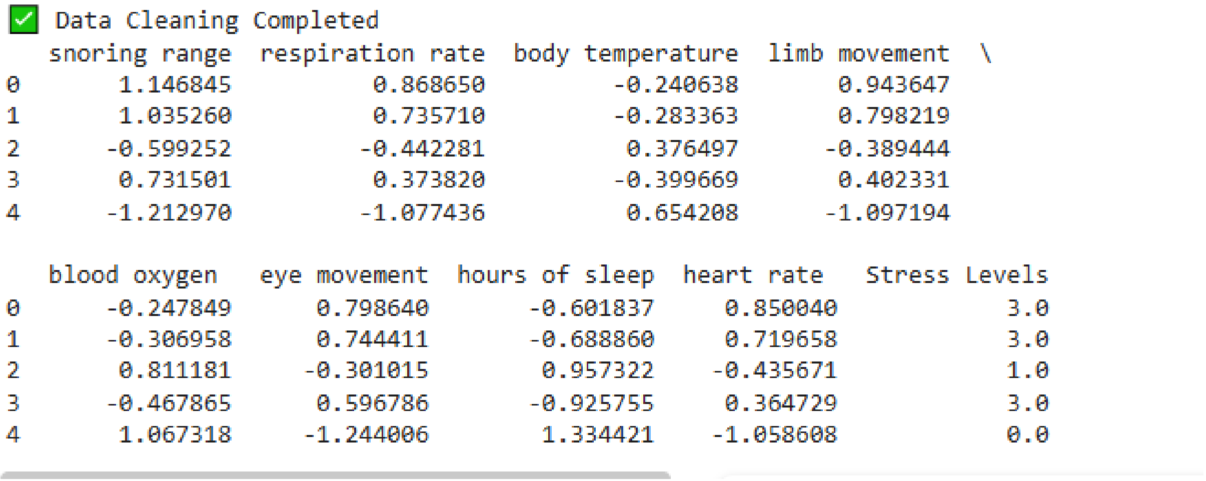
➢ Outlier Detection



Combined boxplot of all numerical features, showing outliers in body temperature and heart rate values.

Data Cleaning:

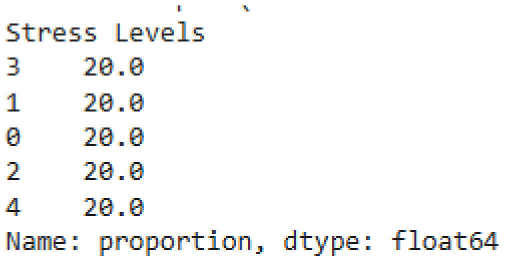
- **Missing value handling** (median imputation applied).
- **Feature scaling** (StandardScaler → mean = 0, std = 1).
- **Target column (Stress Levels)** preserved without scaling.



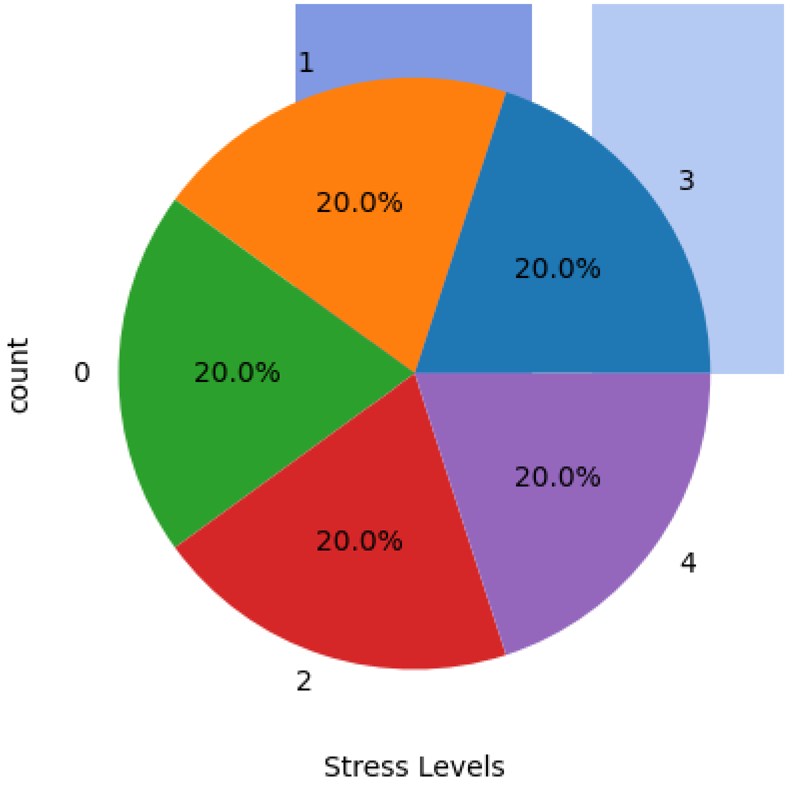
```
✅ Data Cleaning Completed
   snoring range  respiration rate  body temperature  limb movement  \
0       1.146845          0.868650         -0.240638       0.943647
1       1.035260          0.735710         -0.283363       0.798219
2      -0.599252         -0.442281          0.376497      -0.389444
3       0.731501          0.373820         -0.399669       0.402331
4      -1.212970         -1.077436          0.654208      -1.097194

   blood oxygen  eye movement  hours of sleep  heart rate  Stress Levels
0     -0.247849      0.798640       -0.601837    0.850040            3.0
1     -0.306958      0.744411       -0.688860    0.719658            3.0
2      0.811181     -0.301015        0.957322   -0.435671            1.0
3     -0.467865      0.596786       -0.925755    0.364729            3.0
4      1.067318     -1.244006        1.334421   -1.058608            0.0
```

4. Class Balance/Imbalance:

- Checked using bar/pie chart.
- All 5 stress classes are equally distributed (20% each).
- No imbalance handling required.

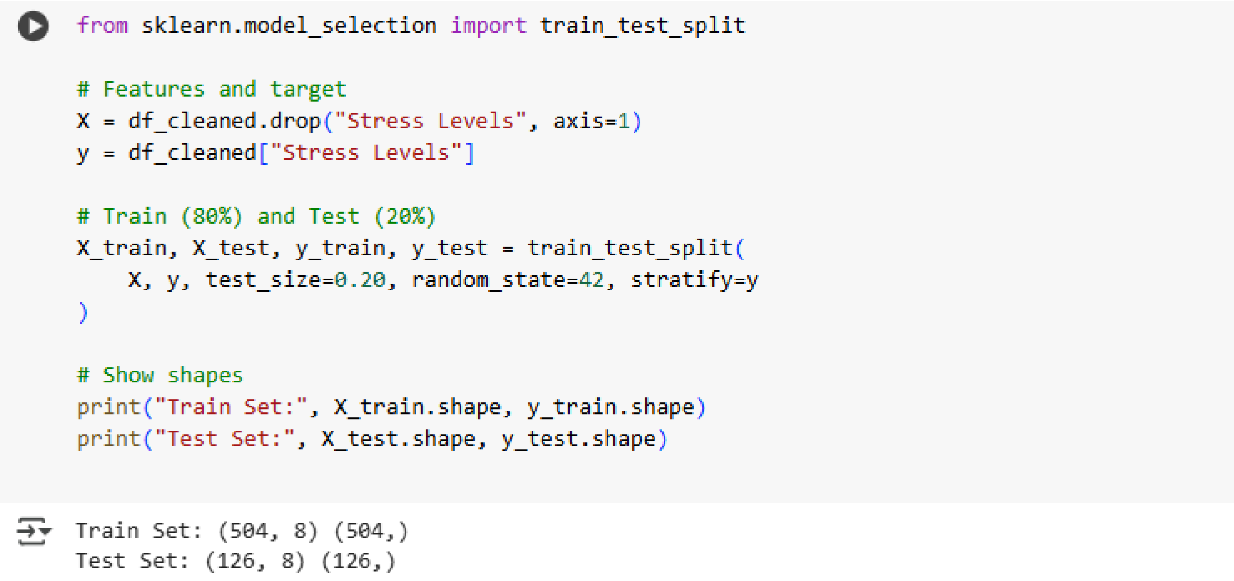


```
Stress Levels
3    20.0
1    20.0
0    20.0
2    20.0
4    20.0
Name: proportion, dtype: float64
```



5. Data Splitting

- Used **80/20 split**.
- **Train Set:** 504 samples (80%)
- **Test Set:** 126 samples (20%)
- Applied **stratified sampling** to preserve class balance.



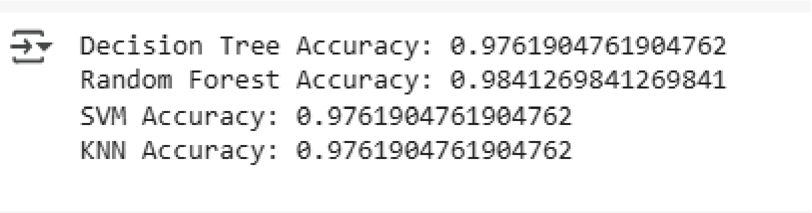
```python
from sklearn.model_selection import train_test_split

# Features and target
X = df_cleaned.drop("Stress Levels", axis=1)
y = df_cleaned["Stress Levels"]

# Train (80%) and Test (20%)
X_train, X_test, y_train, y_test = train_test_split(
    X, y, test_size=0.20, random_state=42, stratify=y
)

# Show shapes
print("Train Set:", X_train.shape, y_train.shape)
print("Test Set:", X_test.shape, y_test.shape)
```

```
Train Set: (504, 8) (504,)
Test Set: (126, 8) (126,)
```

6. Model Building & Training

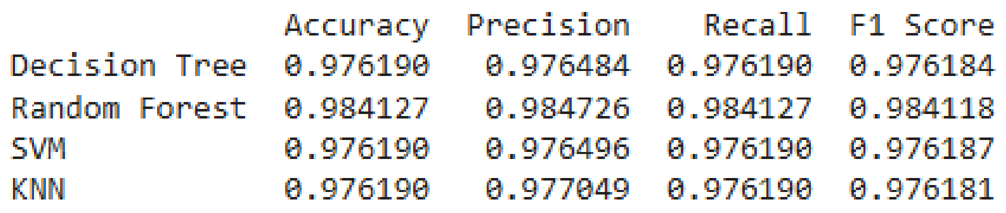
Selected classification algorithms:

- **Decision Tree**
- **Random Forest**
- **Support Vector Machine (SVM)**
- **K-Nearest Neighbors (KNN)**



```
Decision Tree Accuracy: 0.9761904761904762
Random Forest Accuracy: 0.9841269841269841
SVM Accuracy: 0.9761904761904762
KNN Accuracy: 0.9761904761904762
```

➢ Model Evaluation:

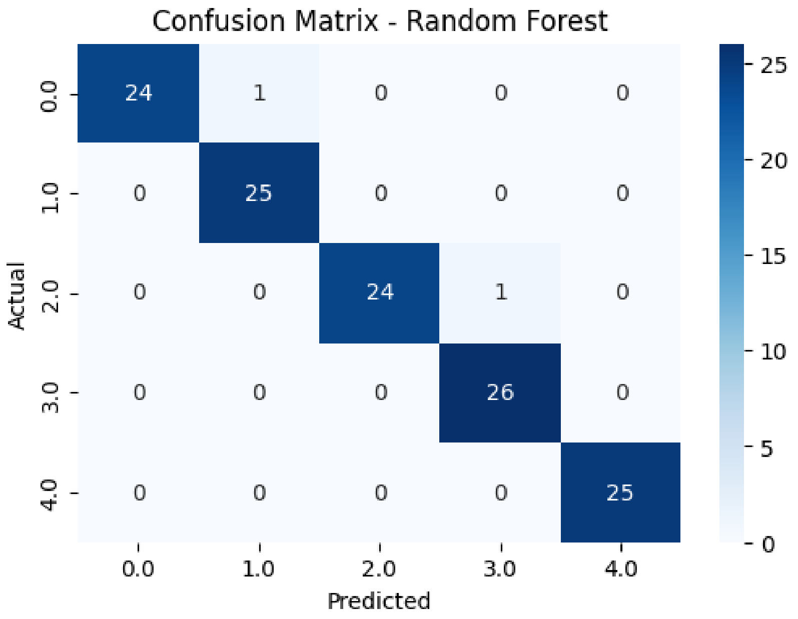


```
               Accuracy  Precision    Recall  F1 Score
Decision Tree  0.976190   0.976484  0.976190  0.976184
Random Forest  0.984127   0.984726  0.984127  0.984118
SVM            0.976190   0.976496  0.976190  0.976187
KNN            0.976190   0.977049  0.976190  0.976181
```

Model	Accuracy	Precision	Recall	F1 score
Decision Tree	97.6%	97.6%	97.6%	97.6%
Random Forest	98.4%	98.5%	98.4%	98.4%
SVM	97.6%	97.6%	97.6%	97.6%
KNN	97.6%	97.7%	97.6%	97.6%

Best Model: Random Forest (highest performance).

Confusion Matrix: Nearly perfect classification with very few errors.



7. Conclusion & Insight

➢ Data Quality & Cleaning

- Missing values were handled with **median imputation**.
- Features were **standardized (mean = 0, std = 1)** for consistency.
- The dataset was confirmed to be **balanced**, with equal representation of all stress levels.

➢ Exploratory Data Analysis (EDA)

- **Hours of Sleep** showed a **negative correlation** with stress , less sleep leads to higher stress.
- **Heart Rate** and **Snoring Range** showed **positive correlation** with stress ,higher values indicate higher stress.
- Outliers in body temperature and heart rate were detected but retained, as they may represent genuine high-stress conditions.

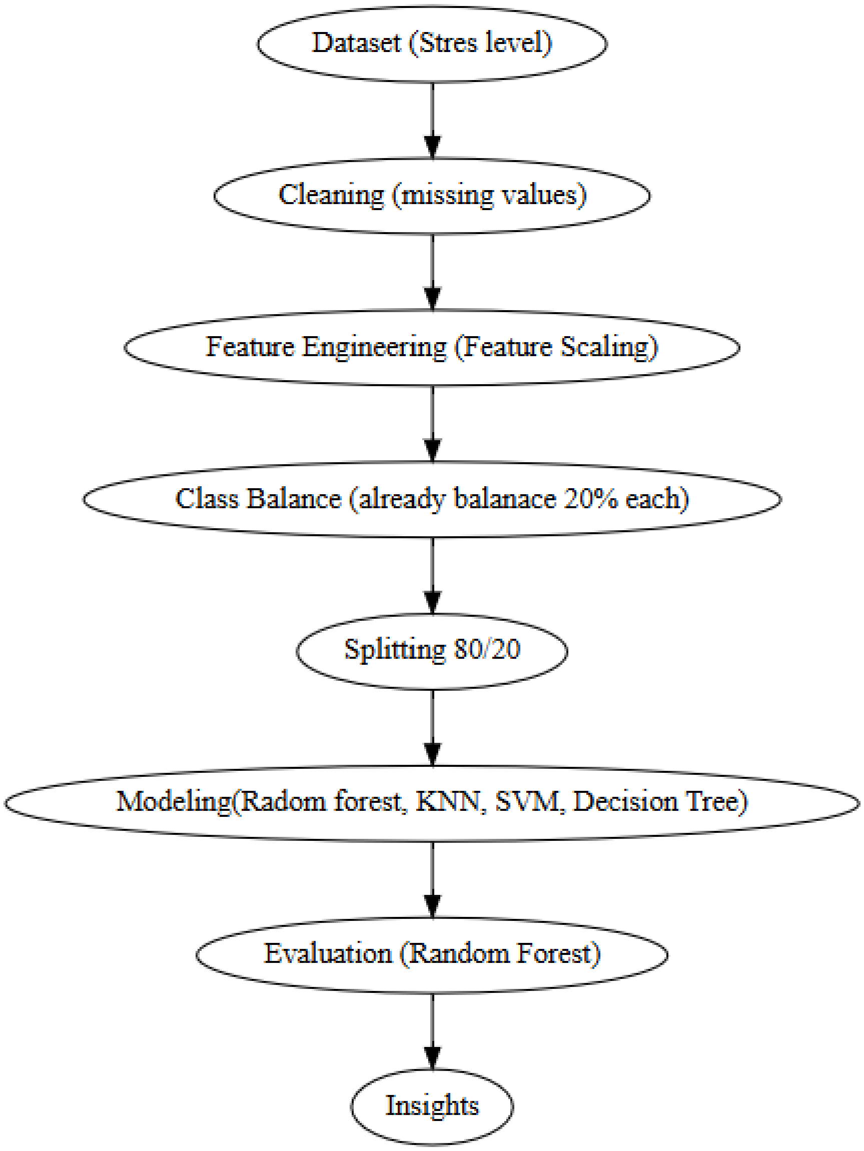
➢ Modeling & Evaluation

- Tested multiple models: **Decision Tree, Random Forest, SVM, KNN**.
- All models achieved strong results (≥97% accuracy).
- **Random Forest** performed the best, with **98.4% accuracy** and highest Precision/Recall/F1 scores.
- Confusion matrix confirmed near-perfect classification across all 5 stress levels.

➢ Business & Practical Relevance

- The model highlights **sleep duration** and **heart rate** as critical indicators of stress.
- Such insights can guide **health monitoring systems**, wearable devices, and mental health programs to provide early stress warnings.

8. ML Pipeline


Dataset (Stres level)
Cleaning (missing values)
Feature Engineering (Feature Scaling)
Class Balance (already balanace 20% each)
Splitting 80/20
Modeling(Radom forest, KNN, SVM, Decision Tree)
Evaluation (Random Forest)
Insights