Big Data Analytics (MET CS 777)

Final Term Project

Human Stress Detection (classification)

Vahid Monfared

October 2024

vahidm@bu.edu

Why is this problem important to solve?

- Human Stress Detection: Human stress level detection using physiological data
- Humidity – Temperature – Step count – Stress levels" represents the titles for Stress-Lysis.csv file
- Importance of Studying This Topic:
- Early detection of stress is essential for preventing health issues like cardiovascular disease, immune dysfunction, and mental health disorders.
- we can develop non-invasive methods to monitor stress in real time.

What We Want to Learn from the Data

1. Determine the relationship between these physiological metrics and stress levels.
2. Develop a model that can classify stress levels (low, normal, and high) based on the input features.
3. Evaluate the accuracy of our classification model and understand which physiological factors contribute most to predicting stress levels.
4. Finding important features
5. How to use GCP for this kind of dataset to build an application on phone or web base systems

Machine Learning Models

- Logistic Regression (Multinomial / Softmax Regression)
- Random Forest Classifier
- Decision Tree Classifier
- Gradient-Boosted Trees (GBT)

Why These Models?

- Suitability in handling numerical tabular data
- These models are well-supported in PySpark (Scalability)
- Performance Comparison (f1 score, accuracy, precision, recall)

Expected outcomes

- Accurate Stress Classification
- Insights for Real-Time Monitoring
- Feature Importance for Physicians

Evaluation plan

- Accuracy
- Precision, Recall, and F1-Score
- Confusion Matrix
- Cross-Validation: To ensure the model's robustness and avoid overfitting, I will use k-fold cross-validation to evaluate the model on multiple splits of the data and assess its generalization capability.

Some samples of dataset

	A	B	C	D
1	humidity	temperature	step_count	stress_level
2	21.33	90.33	123	1
3	21.41	90.41	93	1
4	27.12	96.12	196	2
5	27.64	96.64	177	2
6	10.87	79.87	87	0

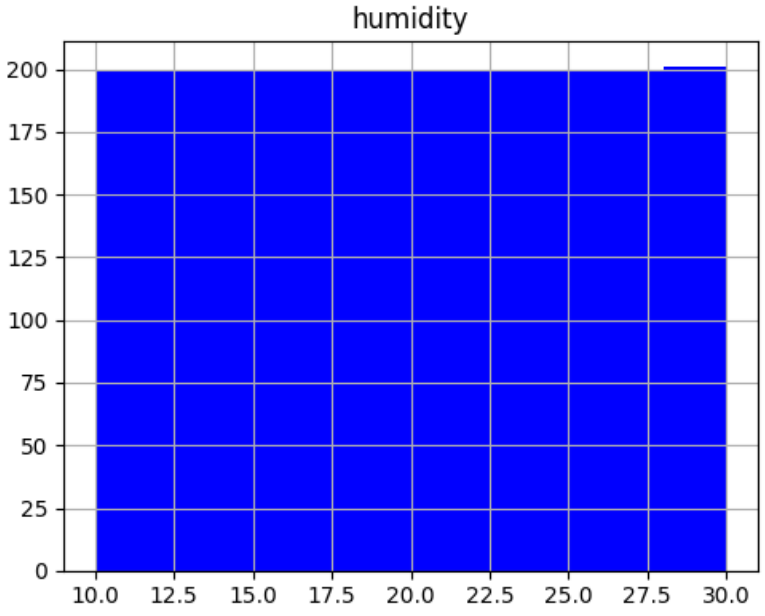
Summary statistics of numerical columns

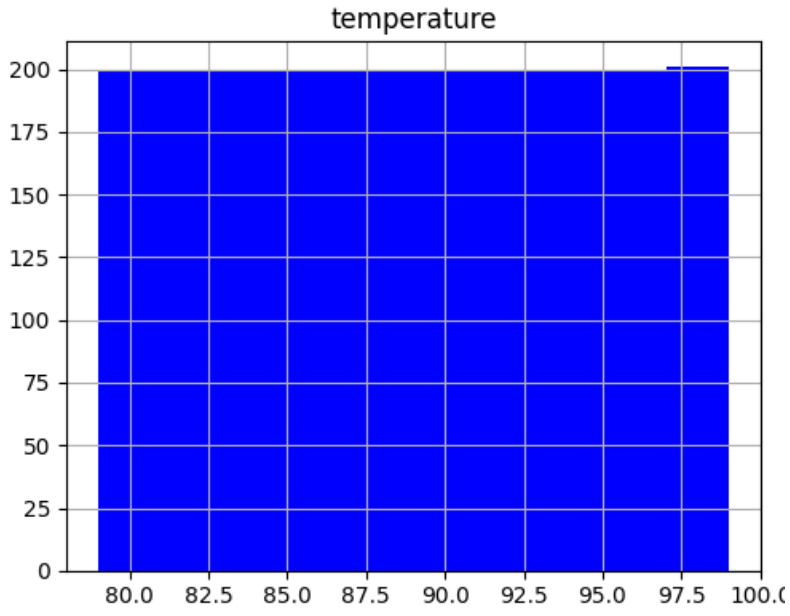
summary	humidity	temperature	step_count	stress_level
count	2001	2001	2001	2001
mean	20.000000000000032	89.000000000001	100.14142928535732	1.104447776111944
stddev	5.777832638628434	5.777832638628431	58.18294842783513	0.771093514041327
min	10.0	79.0	0	0
max	30.0	99.0	200	2

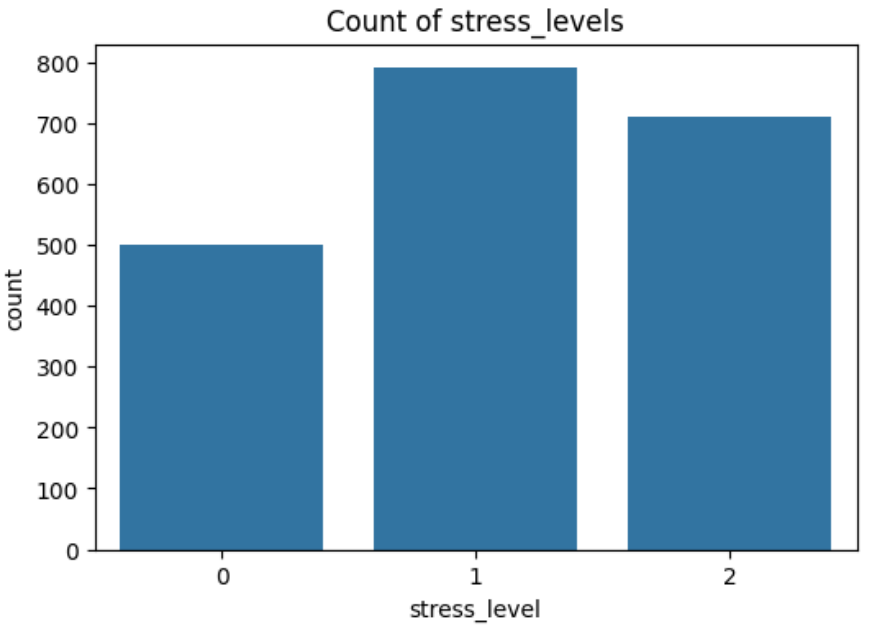
Preprocessing and Cleaning

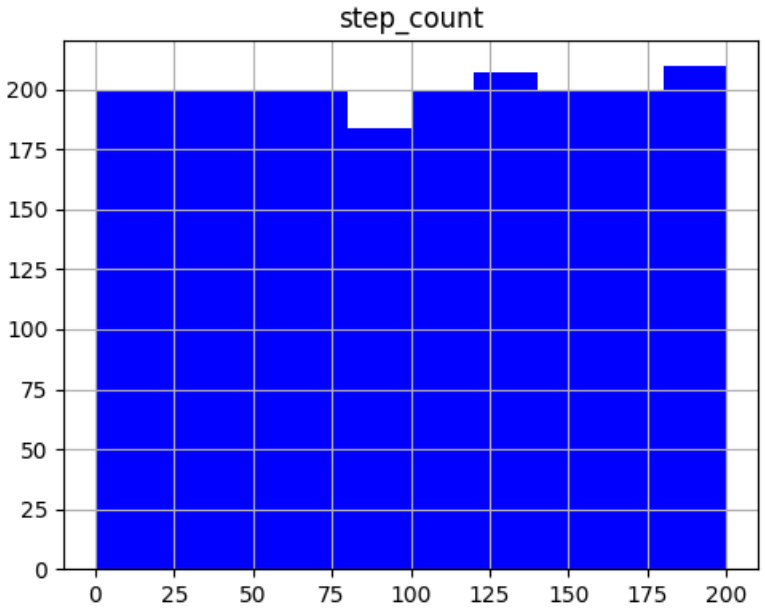
- **Data Visualization**: Utilized to understand data structure and identify anomalies.
- **Handling Missing Values**: Median, KNN imputation exclusively for features
- **Balancing the Dataset on Labels**
- **Removing Duplicates**: Eliminated duplicate entries to maintain data integrity.
- **Addressing Skewed Data**: Used log transformation to correct skewness and approximate normal distribution.
- **Outlier Treatment**: Reduced the impact of outliers through log transformation, Winsorizing, or imputations.
- **Scaling (Normalizing)**: Normalized data to ensure consistency across features.
- **Converting Data Types**: Transformed categorical data into numerical formats.
- **Normal Distribution and Range Checking**: Verified data conforms to expected distributions and ranges.
- **Dropping Unnecessary Data**: Removed irrelevant data to streamline the dataset.

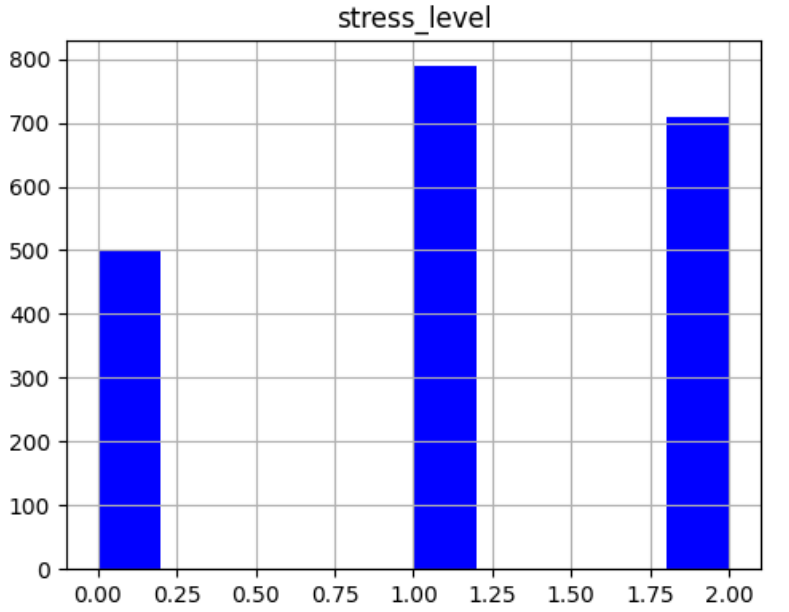
Graphical Presentation of Dataset





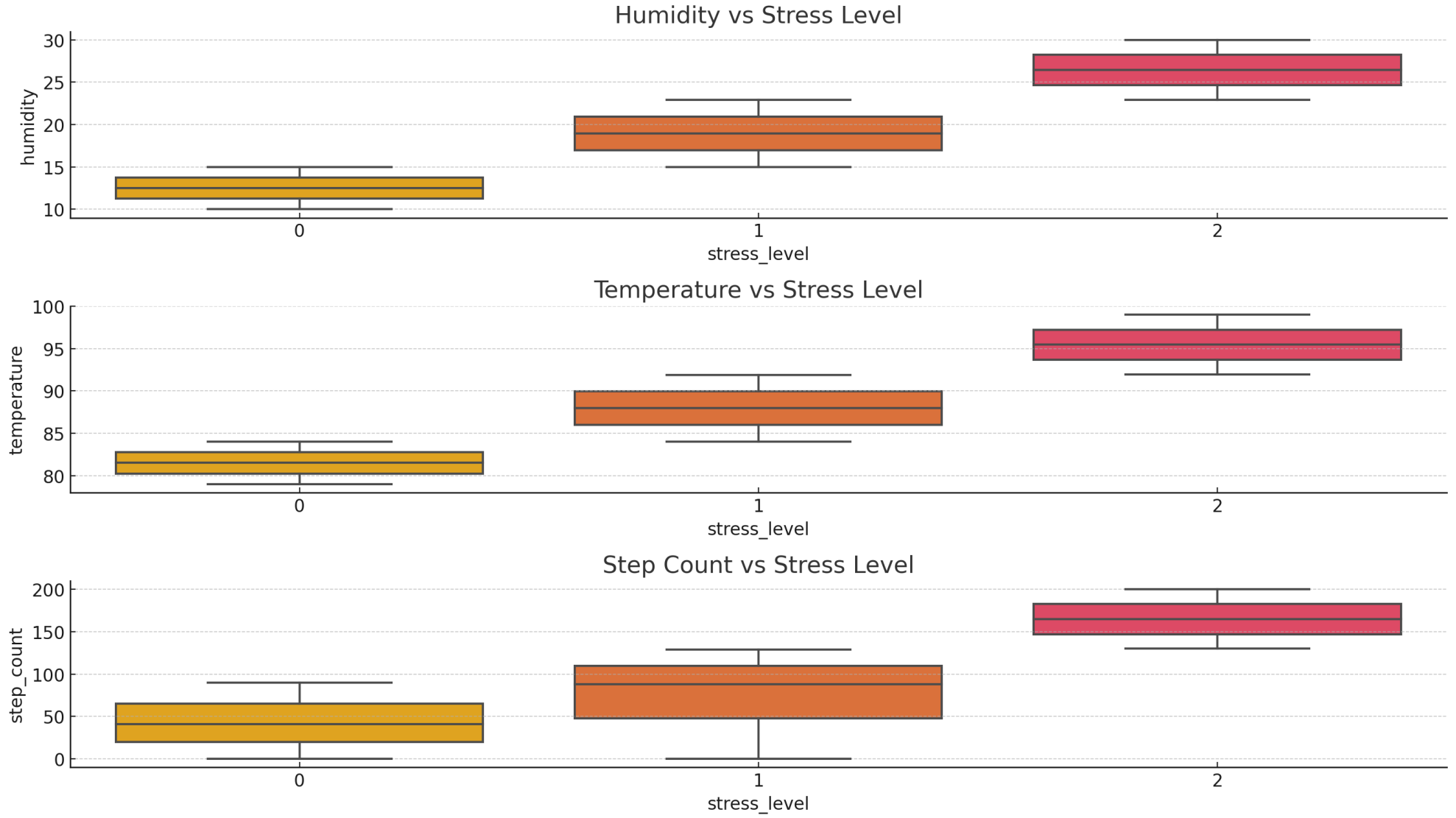






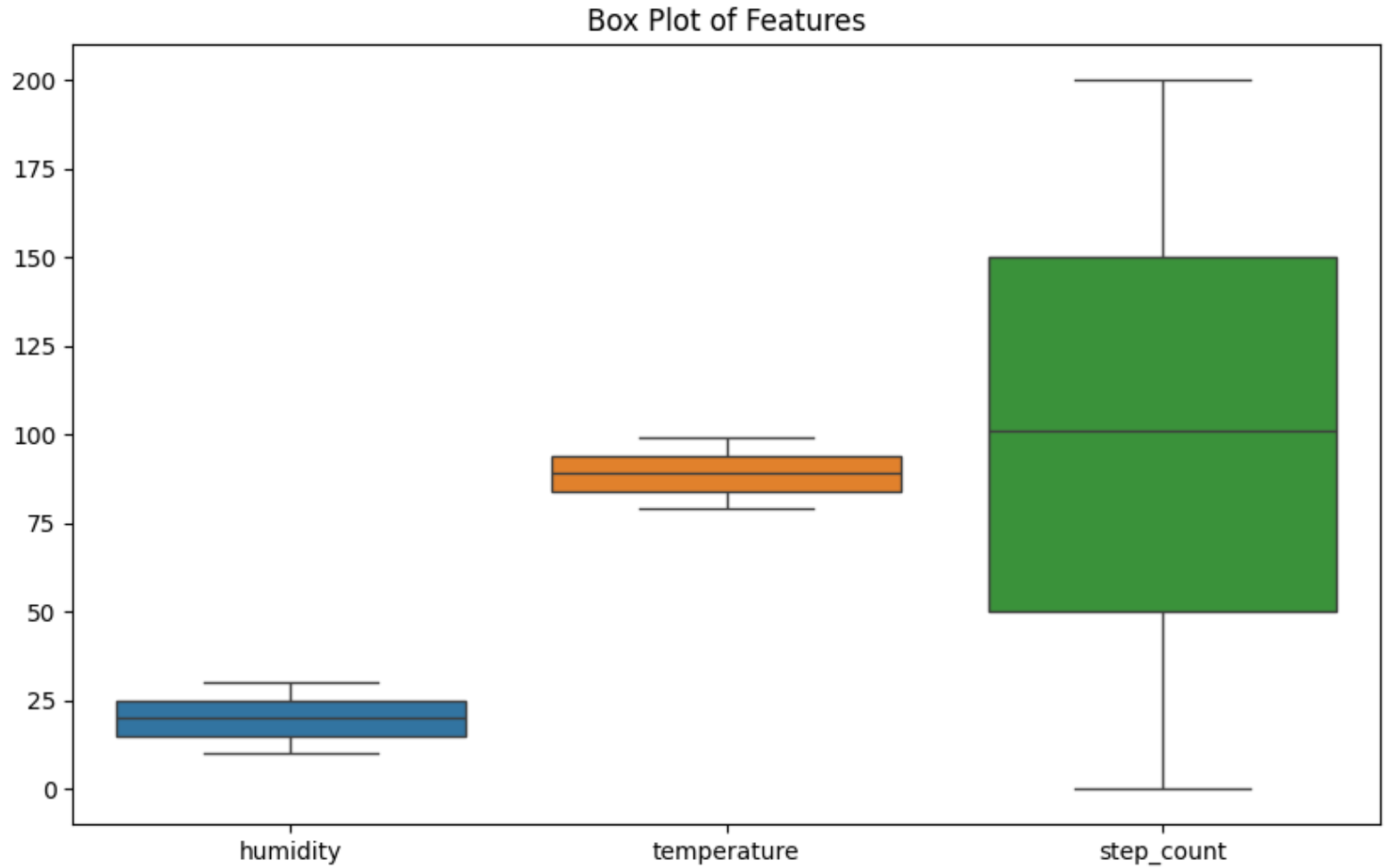
- The bar plot shows moderate label imbalance, with level 1 having the highest count and level 0 the lowest, suggesting potential label balancing.

Features vs. Stress Levels



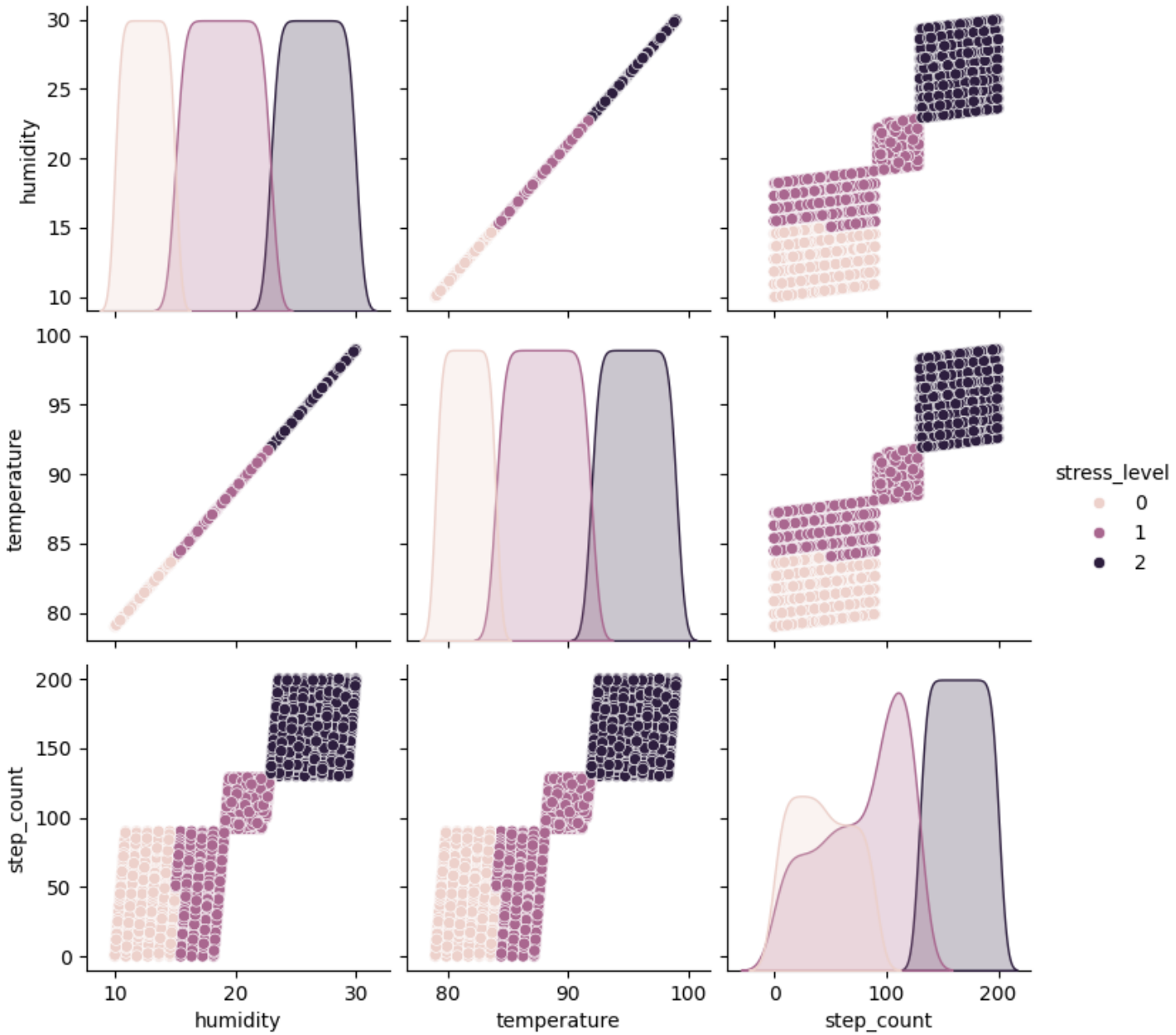
- The box plots show that higher stress levels are linked to increased humidity, temperature, and step count, with distinct differences across stress levels.

Showing Box Plot



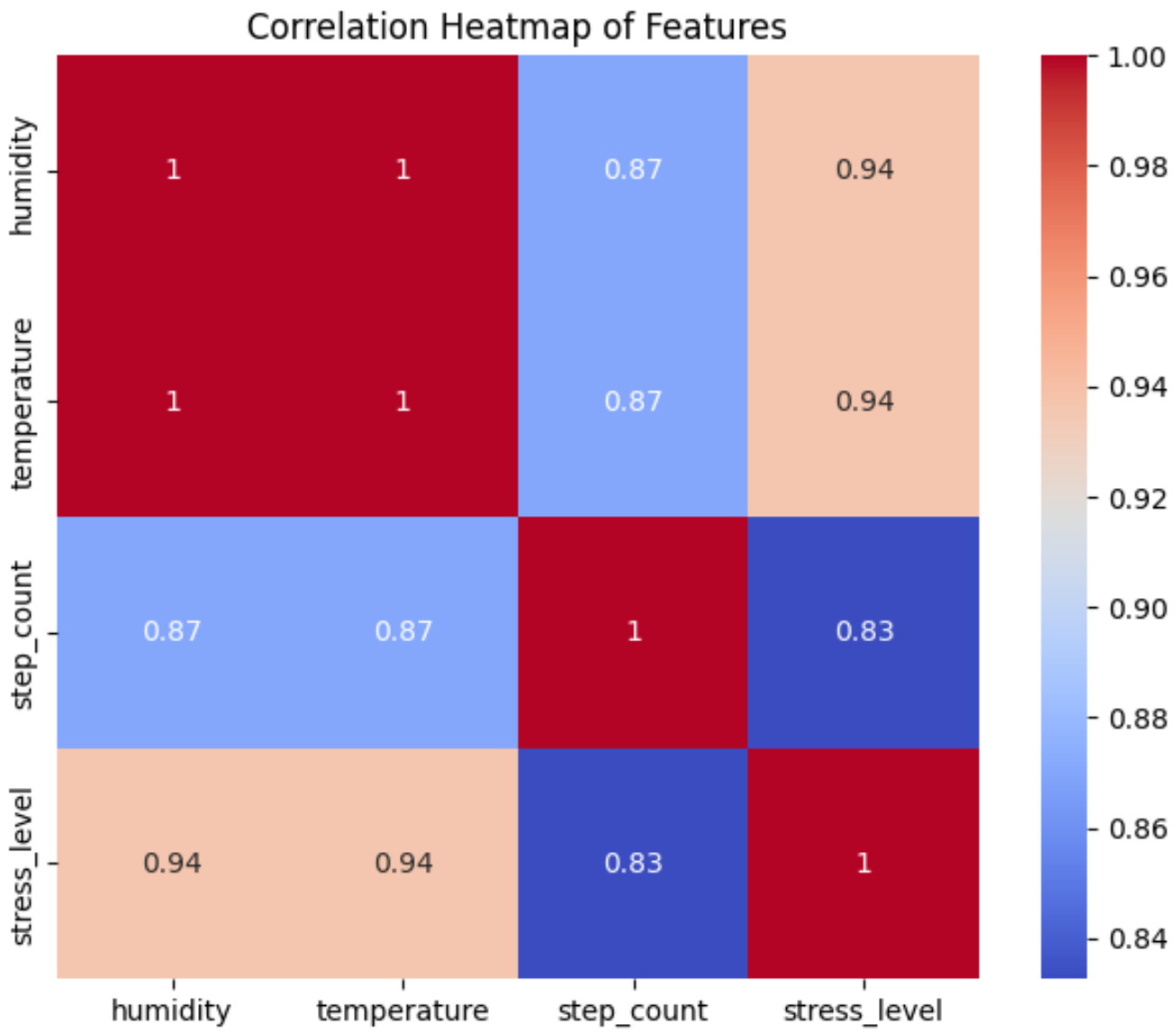
- The box plot displays the distribution of three features: **humidity**, **temperature**, and **step count**.
- **Humidity** shows the lowest range, while **step count** has the largest range and variability with potential outliers; **temperature** is relatively consistent.

Pair plot for Dataset



The pair plot shows clear separation between **stress levels (0, 1, 2)** across **humidity**, **temperature**, and **step count**, indicating strong distinctions in the distribution of features for each stress level.

Pearson Correlation Coefficient Heatmap



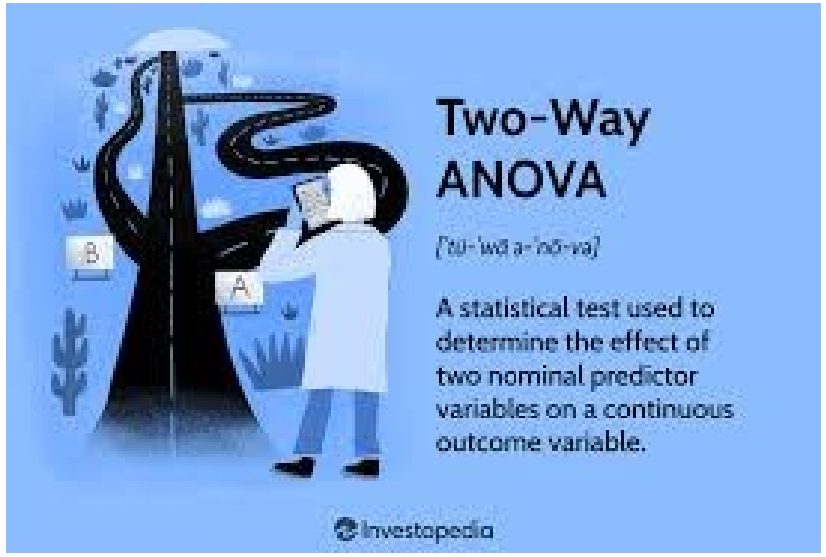
- The heatmap shows strong positive correlations between **humidity**, **temperature**, and the ordinal categorical **stress level** (~0.94), while **step count** also correlates well with these features (~0.83 to 0.87), indicating high overall feature interdependence.

ANOVA for Feature Relationships with the Categorical Label

stress_level	avg_humidity
1	18.95500000000002
2	26.454999999999977
0	12.500000000000002

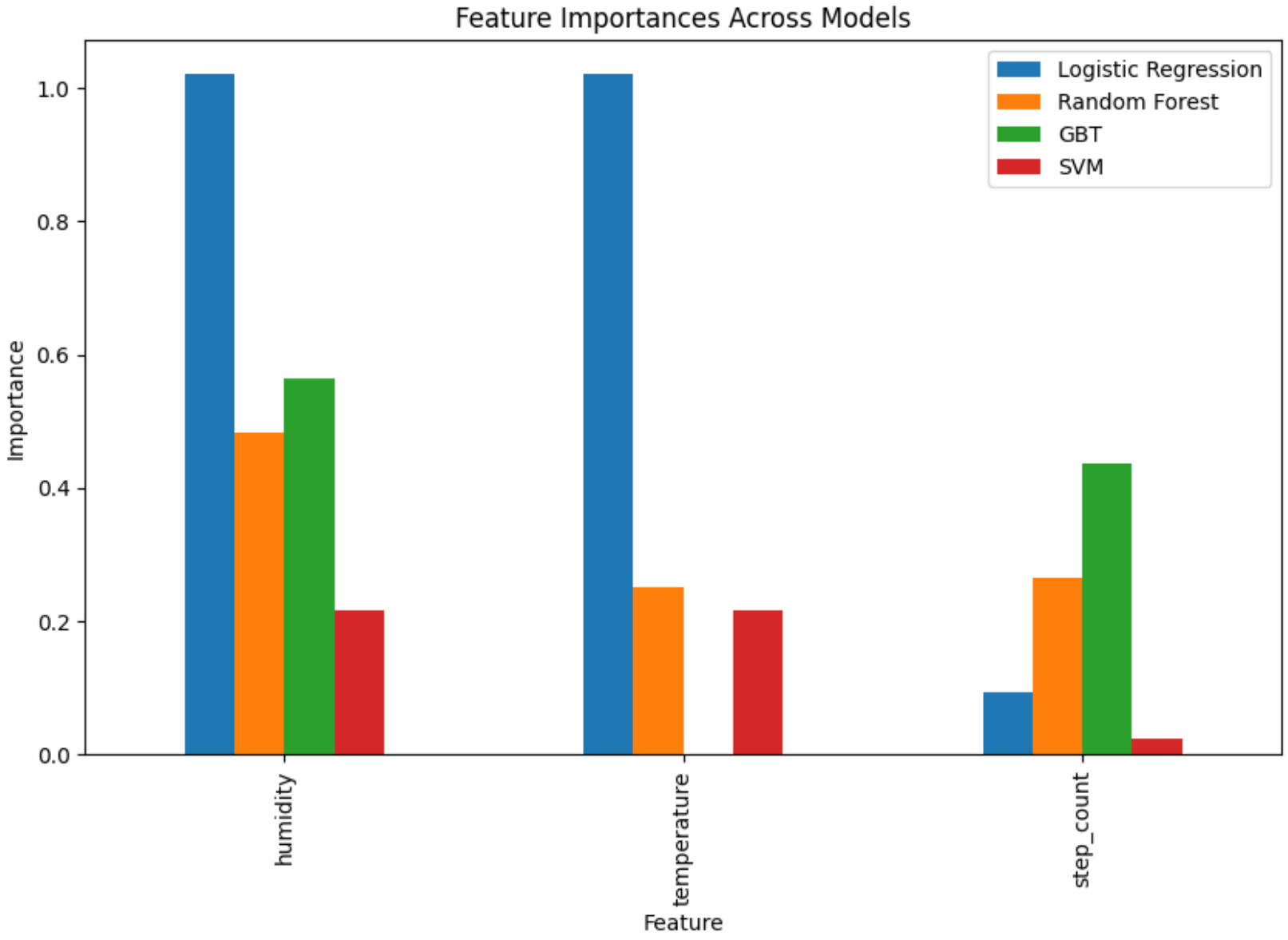
stress_level	avg_temperature
1	87.95499999999994
2	95.45500000000008
0	81.50000000000003

stress_level	avg_step_count
1	78.13037974683544
2	165.0
0	42.93413173652694



- Higher stress levels correlate with increased humidity, temperature, and step count in this dataset. This relationship suggests that environmental factors and physical activity could influence or reflect the stress levels of individuals.

Feature Importance Using Three Methods



- The plot shows **humidity** and **temperature** as the most important features, especially for Logistic Regression, while **step count** is more important in GBT and Random Forest. SVM assigns low importance to all features.
- As an average across the models, the order of importance is:

1.Humidity
2.Temperature
3.Step Count

Humidity and temperature are the most important features, while step count is less important across the models.

Overall accuracy for three models for classification

```
Logistic Regression Model Evaluation Metrics:
Accuracy: 0.9944
F1 Score: 0.9944
Precision: 0.9946
Recall: 0.9944

Random Forest Model Evaluation Metrics:
Accuracy: 0.9972
F1 Score: 0.9972
Precision: 0.9972
Recall: 0.9972

Decision Tree Model Evaluation Metrics:
Accuracy: 0.9972
F1 Score: 0.9972
Precision: 0.9972
Recall: 0.9972
```

- The Random Forest and Decision Tree models both achieved higher performance (Accuracy, F1 Score, Precision, Recall: 0.9972) compared to Logistic Regression (Accuracy, F1 Score: 0.9944). All models performed exceptionally well, with minimal differences in evaluation metrics.

Overall accuracy for three models for classification (tuning, scaling)

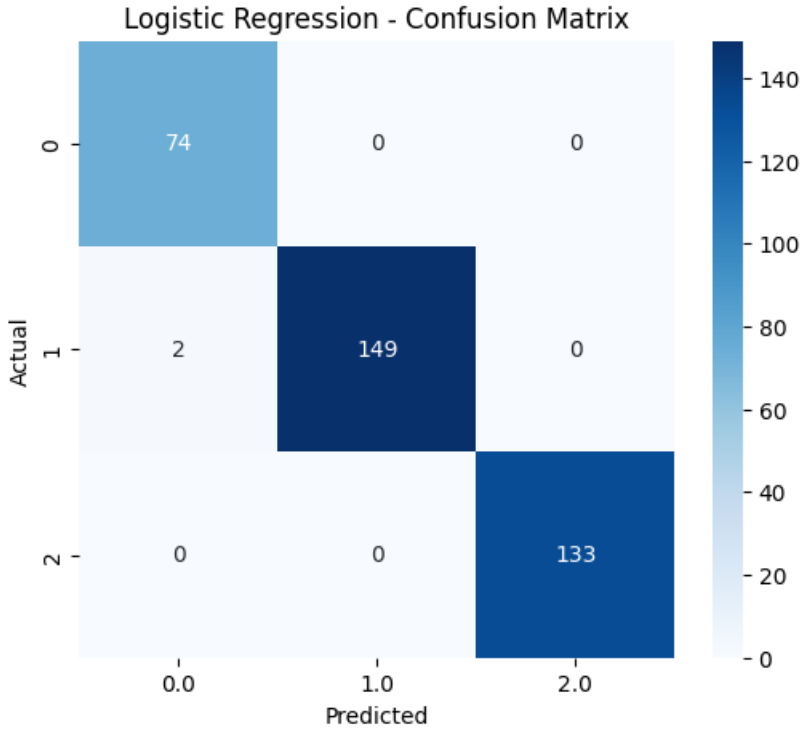
Logistic Regression Model Evaluation Metrics:
Accuracy: 0.9988
F1 Score: 0.9988
Precision: 0.9986
Recall: 0.9988

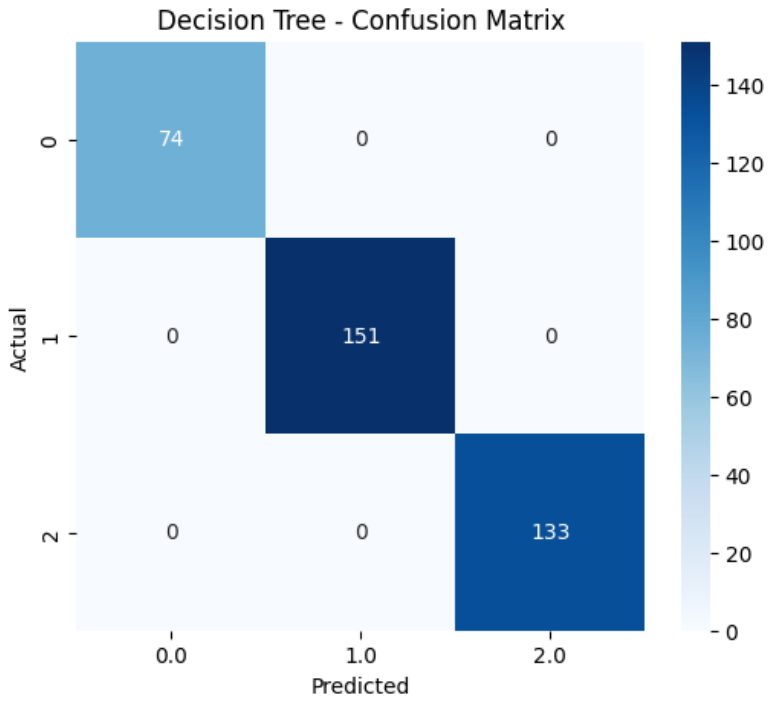
Random Forest Model Evaluation Metrics:
Accuracy: 1.0000
F1 Score: 1.0000
Precision: 1.0000
Recall: 1.0000

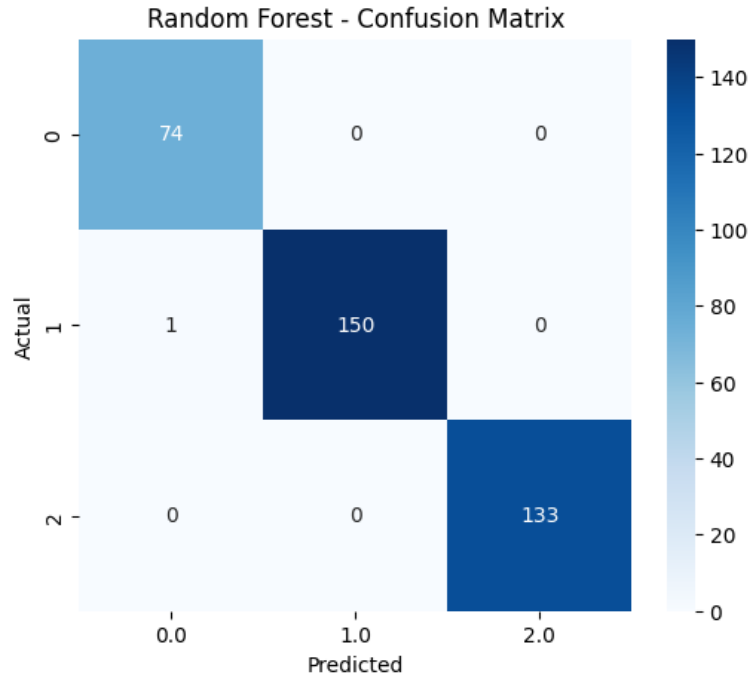
Decision Tree Model Evaluation Metrics:
Accuracy: 1.0000
F1 Score: 1.0000
Precision: 1.0000
Recall: 1.0000

- After fine-tuning and scaling, both **Random Forest** and **Decision Tree** models achieved perfect performance (Accuracy, F1 Score, Precision, Recall: 1.0000), while **Logistic Regression** also performed exceptionally well with near-perfect scores (Accuracy, F1 Score: 0.9988). These results indicate highly accurate and optimized models.

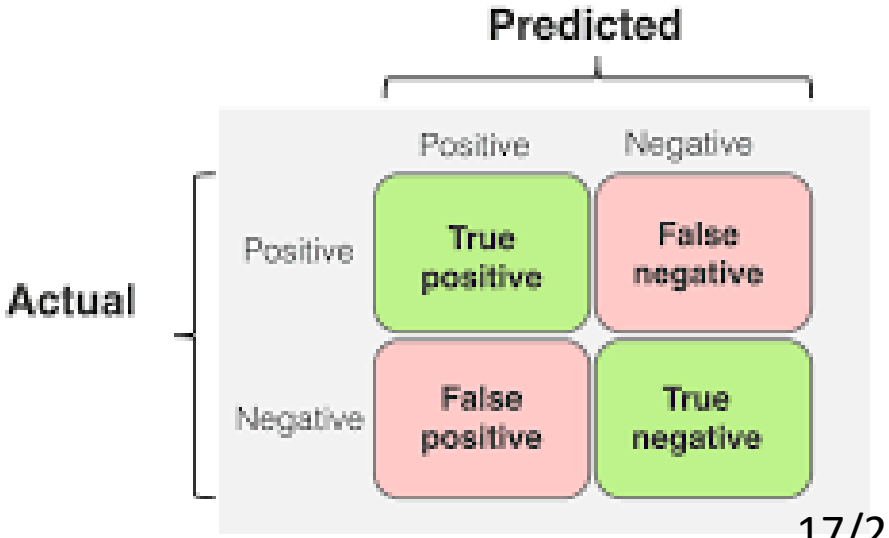
Confusion Matrix for Three Models







- Normalized features (logistic regression)
- Fine tuning for decision tree
- Random forest with no fine tuning



Confusion Matrix

Logistic Regression Confusion Matrix:

	Predicted 0	Predicted 1	Predicted 2
Actual 0	74	0	0
Actual 1	2	149	0
Actual 2	0	0	133

Decision Tree Confusion Matrix:

	Predicted 0	Predicted 1	Predicted 2
Actual 0	74	0	0
Actual 1	1	150	0
Actual 2	0	0	133

Random Forest Confusion Matrix:

	Predicted 0	Predicted 1	Predicted 2
Actual 0	74	0	0
Actual 1	1	150	0
Actual 2	0	0	133

Overall, the most accurate models are as bellow,

- Fine tuned random forest
- Fine tuned decision tree
- Logistic Regression (scaling the features)

How to Use GCP for This Project

- How to convert these models and results for predicting stress levels into a mobile app or website

1. Convert the Model into a Predictive API in Mobile App

2. Create a Website

Humidity:

Temperature:

Step Count:

Predict Stress Level

Conclusion

- Machine learning models can be deployed via a mobile app or website to predict stress levels using real-time data.
- This service provides actionable health insights to users and supports physicians in improving patient outcomes.
- All four models show nearly identical performance after preprocessing and fine-tuning.
- Physicians can use these models as diagnostic tools to track stress levels and prevent stress-related health issues.
- Regular monitoring of environment and activity data helps in refining interventions for better health and well-being.

Thanks for your attention