CO3029 - DATA MINING

Student Stress Analysis - Data Mining Project

GROUP: 22 - L01 - HK251

INSTRUCTORS: DO THANH THAI

VIETNAM NATIONAL UNIVERSITY HO CHI MINH CITY
HO CHI MINH CITY UNIVERSITY OF TECHNOLOGY

TEAM & INSTRUCTOR

Do Thanh Thai
Instructor

Nguyen Trung Hieu
Member - 2113357

Doan Phuong Hung Cuong
Member - 2310381

Nguyen Phu Tue
Member - 2213813

1. INTRODUCTION

Problem Statement

Student stress has become a significant concern in modern education. Understanding the **factors** that contribute to stress can help educational institutions develop **better support systems and interventions**. This project aims to:

- Identify key factors contributing to student stress
- Predict stress levels based on various features
- Discover patterns and groups among stressed students
- Provide actionable insights for stress management

2. DATASET INFORMATION

StressLevelDataset.csv

- **Records:** 1,100 students
- **Features:** 20 input features + 1 target
- **Target:** stress_level (0: Low, 1: Moderate, 2: High)
- **Type:** Numeric features only

First 10 rows [5 rows x 21 columns]

anxiety_level	self_esteem	mental_health_history	depression	headache	blood_pressure	sleep_quality	breathing_problem	noise_level	living_conditions	safety	basic_needs
14	20	0	11	2	1	2	4	2	3	3	2
15	8	1	15	5	3	1	4	3	1	2	2
12	18	1	14	2	1	2	2	2	2	3	2
16	12	1	15	4	3	1	3	4	2	2	2
16	28	0	7	2	3	5	1	3	2	4	3
20	13	1	21	3	3	1	4	3	2	2	1
4	26	0	6	1	2	4	1	1	4	4	4
17	3	1	22	4	3	1	5	3	1	1	1
13	22	1	12	3	1	2	4	3	3	3	3
6	8	0	27	4	3	1	2	0	5	2	2

2. DATASET INFORMATION

Key Features Analyzed

- **Psychological:** anxiety_level, depression, self_esteem
- **Physical:** headache, blood_pressure, sleep_quality
- **Academic:** academic_performance, study_load
- **Social:** social_support, peer_pressure, bullying
- **Environmental:** living_conditions, safety, noise_level

anxiety_level	self_esteem	mental_health_history	depression	headache	blood_pressure	sleep_quality	breathing_problem	noise_level	living_conditions	safety	basic_needs
14	20	0	11	2	1	2	4	2	3	3	2
15	8	1	15	5	3	1	4	3	1	2	2
12	18	1	14	2	1	2	2	2	2	3	2
16	12	1	15	4	3	1	3	4	2	2	2
16	28	0	7	2	3	5	1	3	2	4	3
20	13	1	21	3	3	1	4	3	2	2	1
4	26	0	6	1	2	4	1	1	4	4	4
17	3	1	22	4	3	1	5	3	1	1	1
13	22	1	12	3	1	2	4	3	3	3	3
6	8	0	27	4	3	1	2	0	5	2	2

3. DATA VISUALIZATION AND DATA PREPROCESSING

Data Preprocessing

1. Data loading

2. Data exproration

3. Missing value handling

4. Duplicate removal

5. Categorical encoding

6. Preparation of features (X) and target variable (y)

7. Train-test splitting (80/20)

8. Feature normalization using StandardScaler

3. DATA VISUALIZATION AND DATA PREPROCESSING

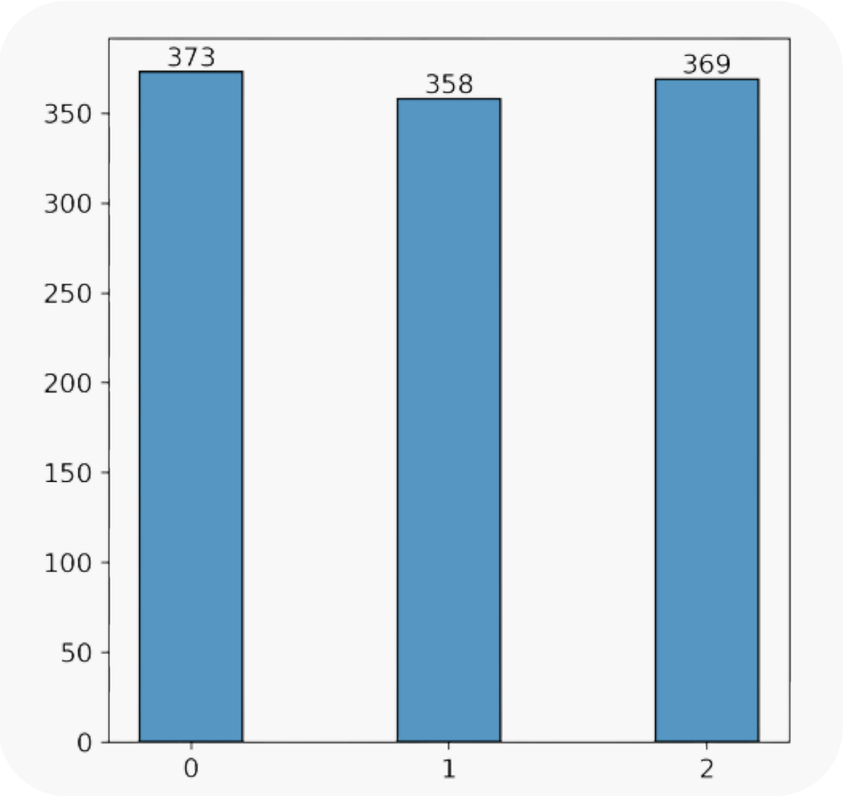
Exploratory Data Analysis

Data Shape

- 1100 rows (instances) and 21 columns (20 features + 1 target variable).
- Sufficient for exploratory analysis and modeling without overwhelming computational resources.

Data Quality

- No missing values or duplicate rows.
- All columns are of type int64, no need for type conversion or imputation.
- No categorical strings or timestamps are present, simplifying preprocessing.



Target Variable Distribution:

- Level 0 (low stress): 373 instances (~33.9%)
- Level 1 (medium stress): 358 instances (~32.5%)
- Level 2 (high stress): 369 instances (~33.6%)

3. DATA VISUALIZATION AND DATA PREPROCESSING

Some Summary Statistics

	anxiety_level	self_esteem	mental_health_history	depression	headache	blood_pressure	sleep_quality	breathing_problem	noise_level
count	1100.000000	1100.000000	1100.000000	1100.000000	1100.000000	1100.000000	1100.000000	1100.000000	1100.000000
mean	11.063636	17.777273	0.492727	12.555455	2.508182	2.181818	2.660000	2.753636	2.649091
std	6.117558	8.944599	0.500175	7.727008	1.409356	0.833575	1.548383	1.400713	1.328127
min	0.000000	0.000000	0.000000	0.000000	0.000000	1.000000	0.000000	0.000000	0.000000
25%	6.000000	11.000000	0.000000	6.000000	1.000000	1.000000	1.000000	2.000000	2.000000
50%	11.000000	19.000000	0.000000	12.000000	3.000000	2.000000	2.500000	3.000000	3.000000
75%	16.000000	26.000000	1.000000	19.000000	3.000000	3.000000	4.000000	4.000000	3.000000
max	21.000000	30.000000	1.000000	27.000000	5.000000	3.000000	5.000000	5.000000	5.000000

Factors

Psychological

Physiological

Academic

Environmental/Social

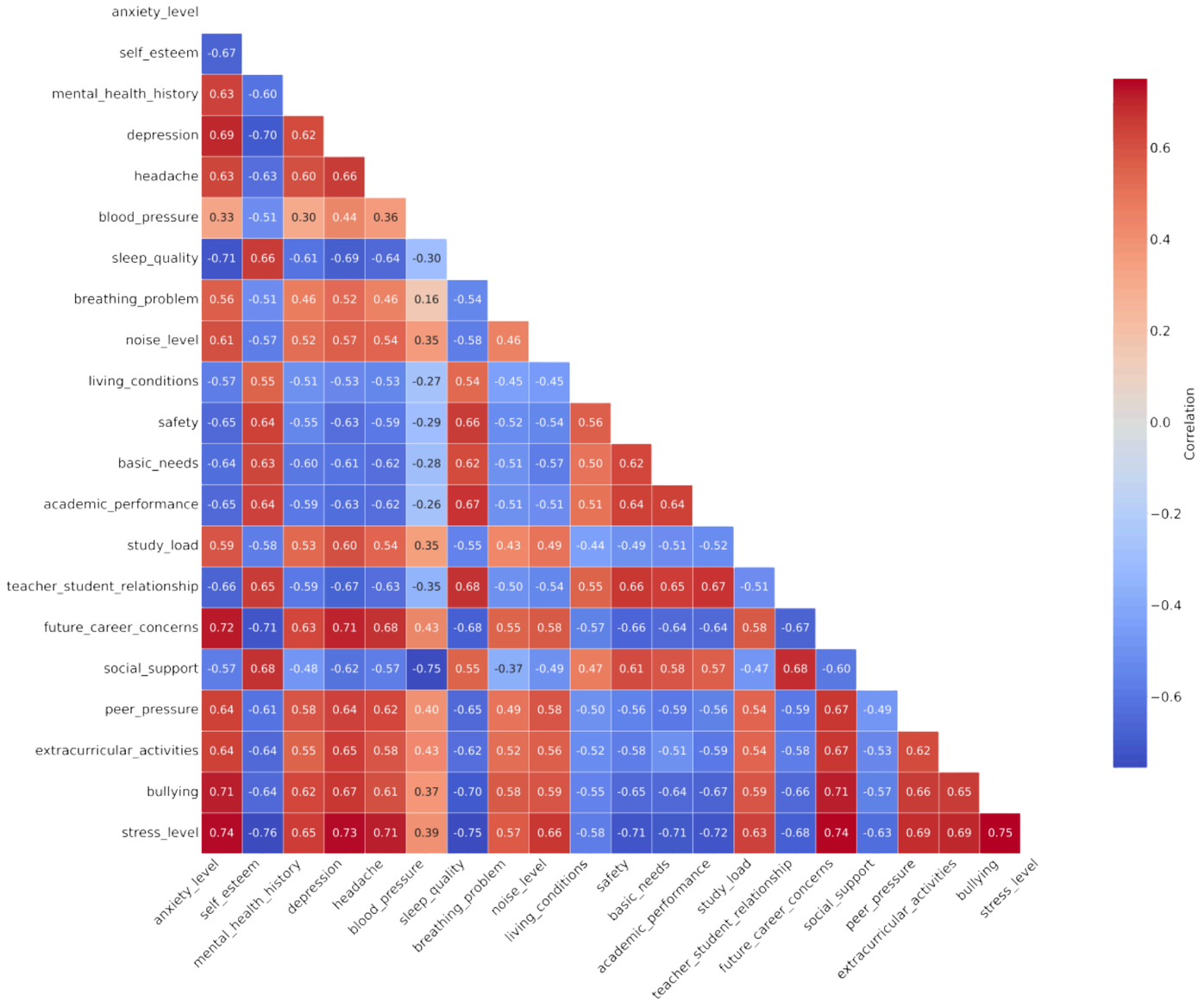
Target Variable (stress_level)

3. DATA VISUALIZATION AND DATA PREPROCESSING

Correlation Heatmap

Strongest Positive Correlations

- Psychological/Social: Bullying (r = 0.75), anxiety_level (r = 0.74), future_career_concerns (r = 0.74), depression (r = 0.73), headache (r = 0.71), and peer_pressure (r = 0.71).
- Environmental/Academic: Secondary contributors include noise_level (r = 0.68), extracurricular_activities (r = 0.67), study_load (r = 0.65), and mental_health_history (r = 0.65).

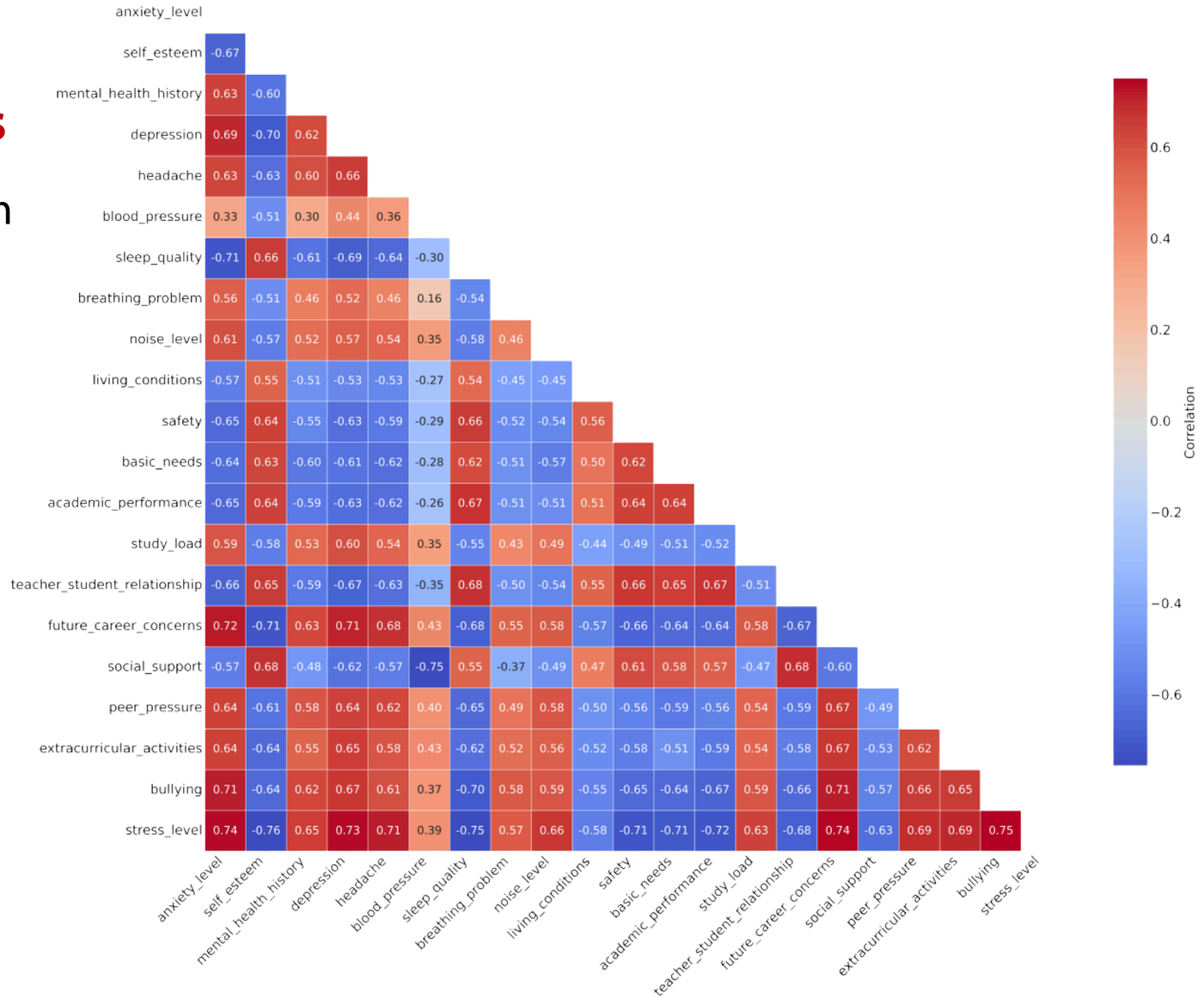


3. DATA VISUALIZATION AND DATA PREPROCESSING

Correlation Heatmap

Strongest Negative Correlations

- Health and Environment: Self_esteem (r = –0.76), sleep_quality (r = –0.75), basic_needs (r = –0.72), living_conditions (r = –0.71), and safety (r = –0.71).
- Academic and Social: Academic_performance (r = –0.71), social_support (r = –0.68), and teacher_student_relationship (r = –0.63).

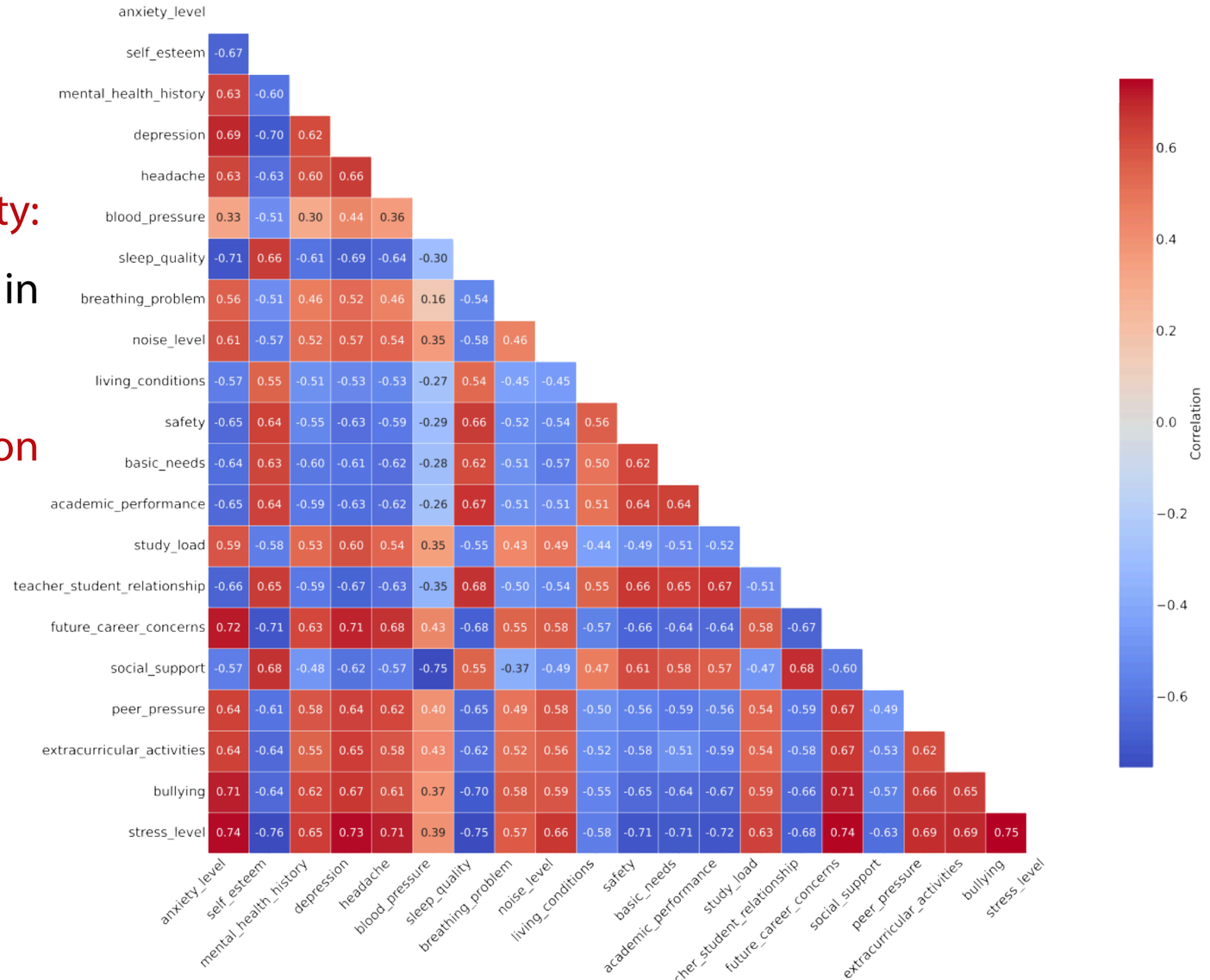


3. DATA VISUALIZATION AND DATA PREPROCESSING

Correlation Heatmap

Modeling Implications

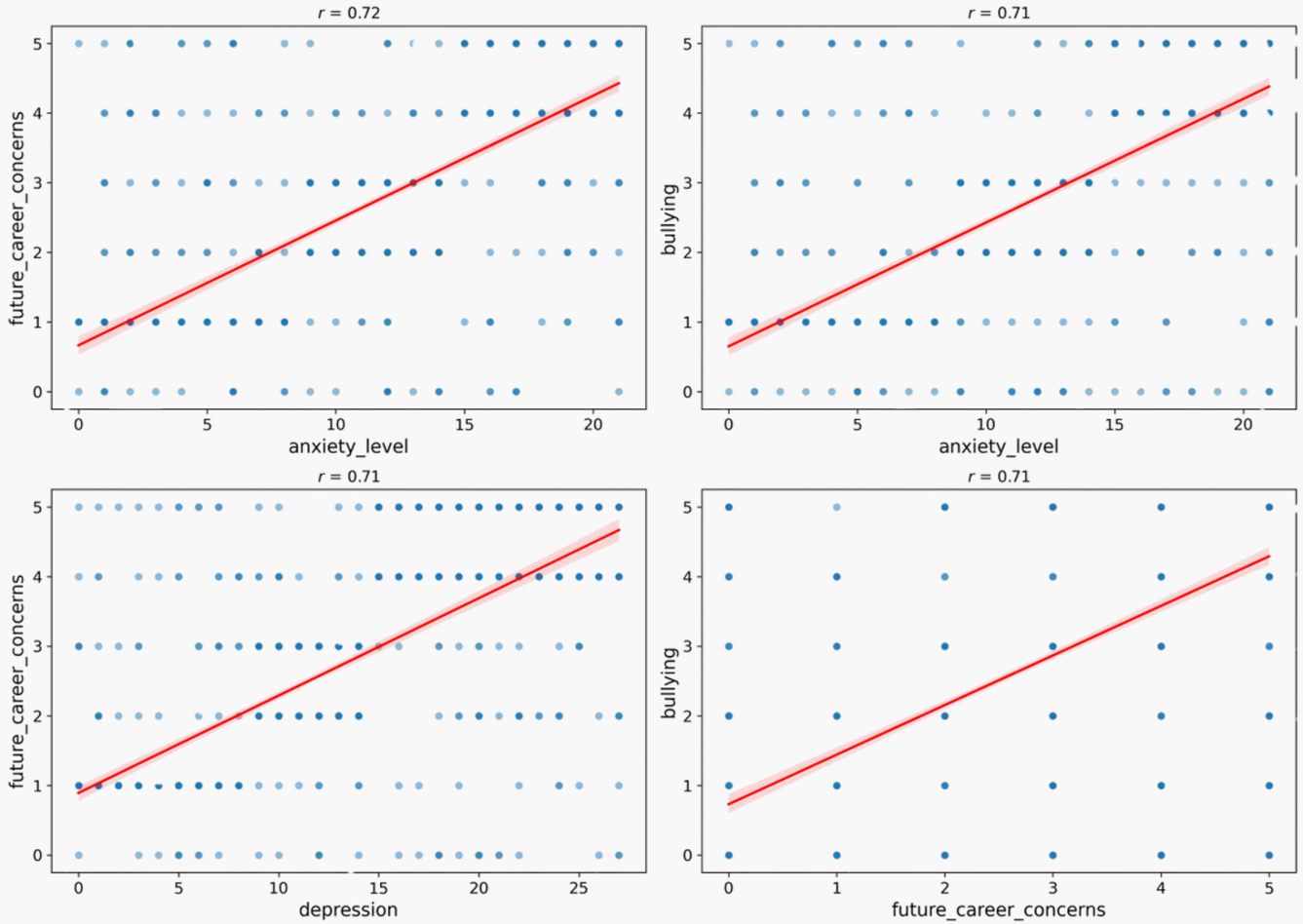
- Inter-Correlation and Multicollinearity: Potential multicollinearity concerns in regression models.
- Feature Selection: Feature selection priority in predictive modeling
- Intervention Targets



3. DATA VISUALIZATION AND DATA PREPROCESSING

Modeling Implications

- Confirmation of Linear Trends: strong positive linear trend, dense clustering and few outliers, interconnectedness of key student stressors.
- Modeling Implications (Multicollinearity): Variance Inflation Factor (VIF) assessment, Principal Component Analysis (PCA), model stability.
- Algorithm Suitability: egression-based predictive models, clustering algorithms.

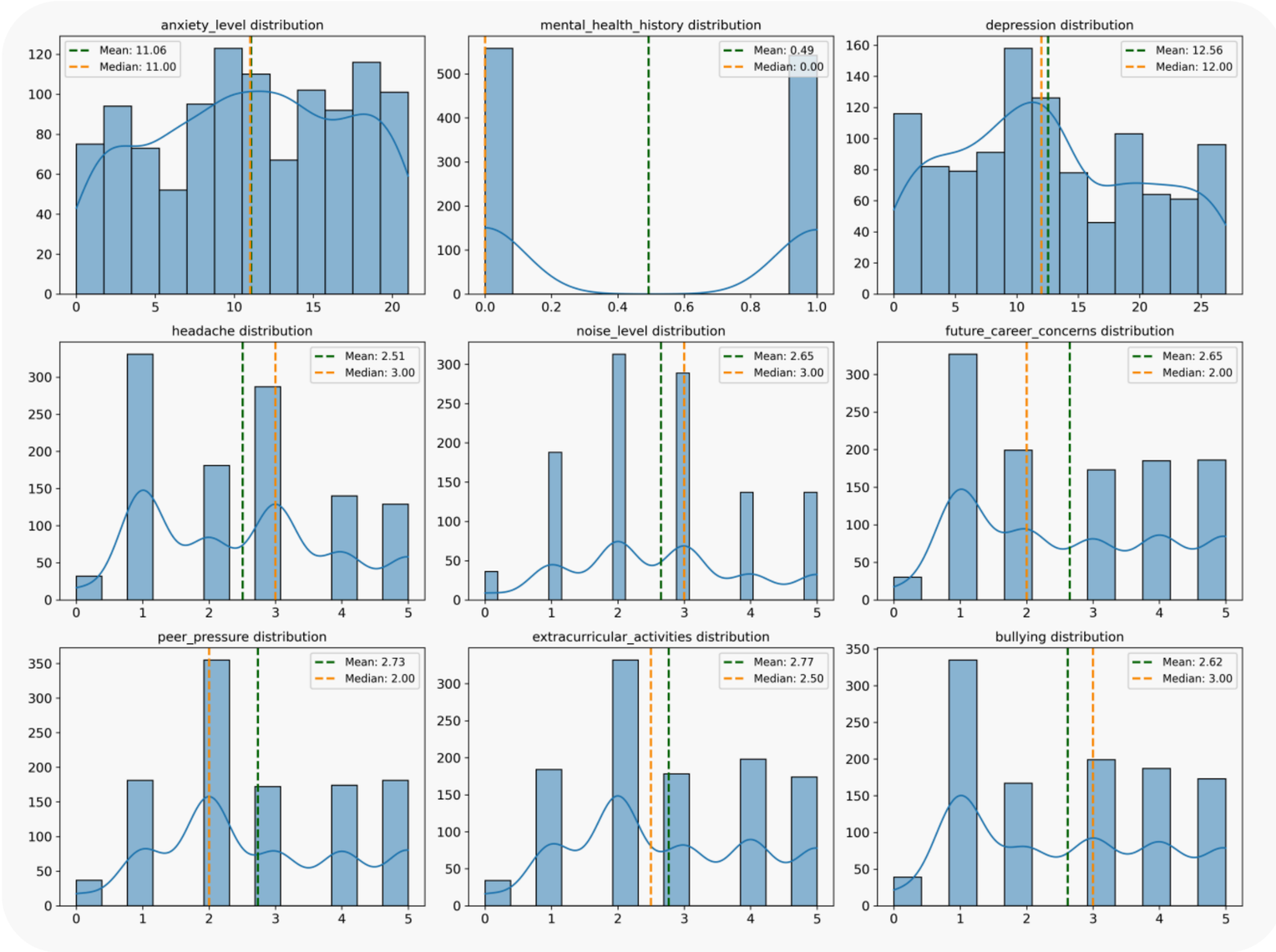


Scatterplots of Highly Correlated Pairs $|r| \geq 0.7$

3. DATA VISUALIZATION AND DATA PREPROCESSING

Key Features Distributions

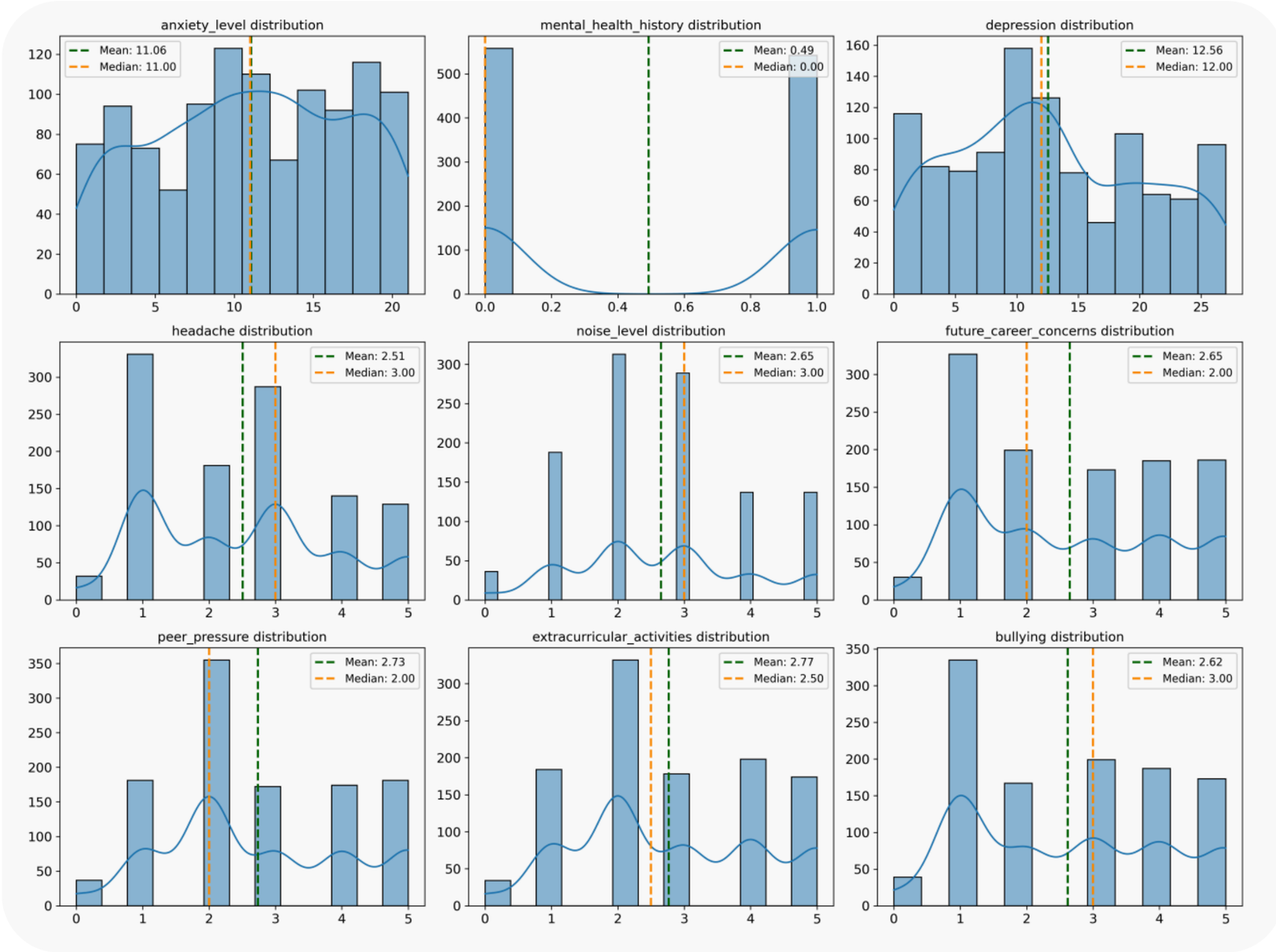
- Anxiety/Depression: Approximately Normal Distributions. Feature Scaling necessary (high variability, std ≈ 6–7).
- Stressors (Bullying, Career Concerns): Positive Skewness (Right-Tailed). Requires Log Transformations or Binning.
- Headache/Activities: Bimodal Patterns. Suggests Clustering Algorithms (K-means) for subgroups.



3. DATA VISUALIZATION AND DATA PREPROCESSING

Key Features Distributions

- Mental Health History: Binary and Slightly Imbalanced (mode at 0). Requires Stratified Sampling or Oversampling.
- Implication: Mixed distributions demand Feature Scaling and handling Skewness.



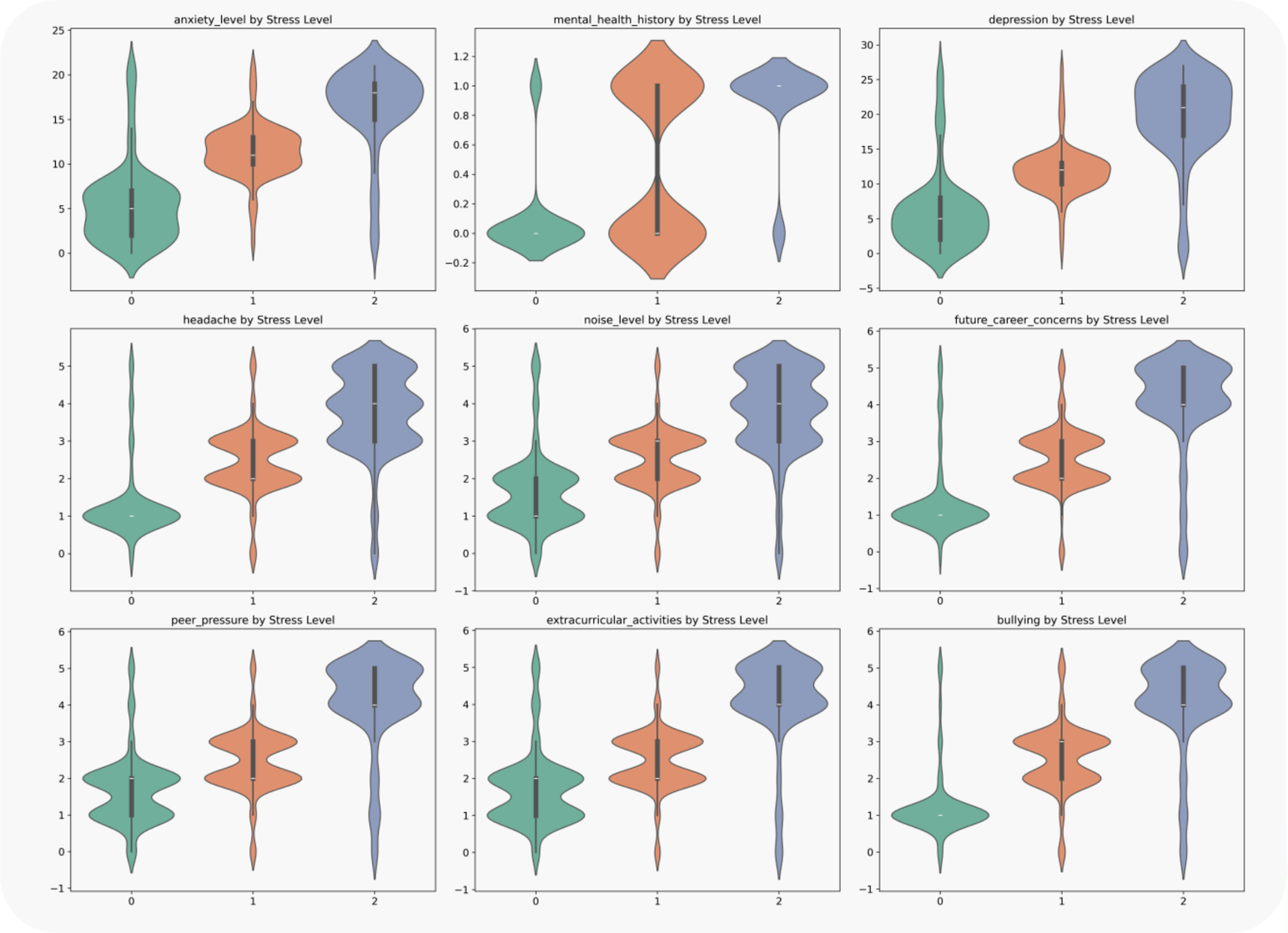
3. DATA VISUALIZATION AND DATA PREPROCESSING

Violin Plots

- Increasing Trends:

  Psychological/Social features (anxiety_level, depression, bullying, peer_pressure, future_career_concerns) show a Clear Positive Shift across stress levels.

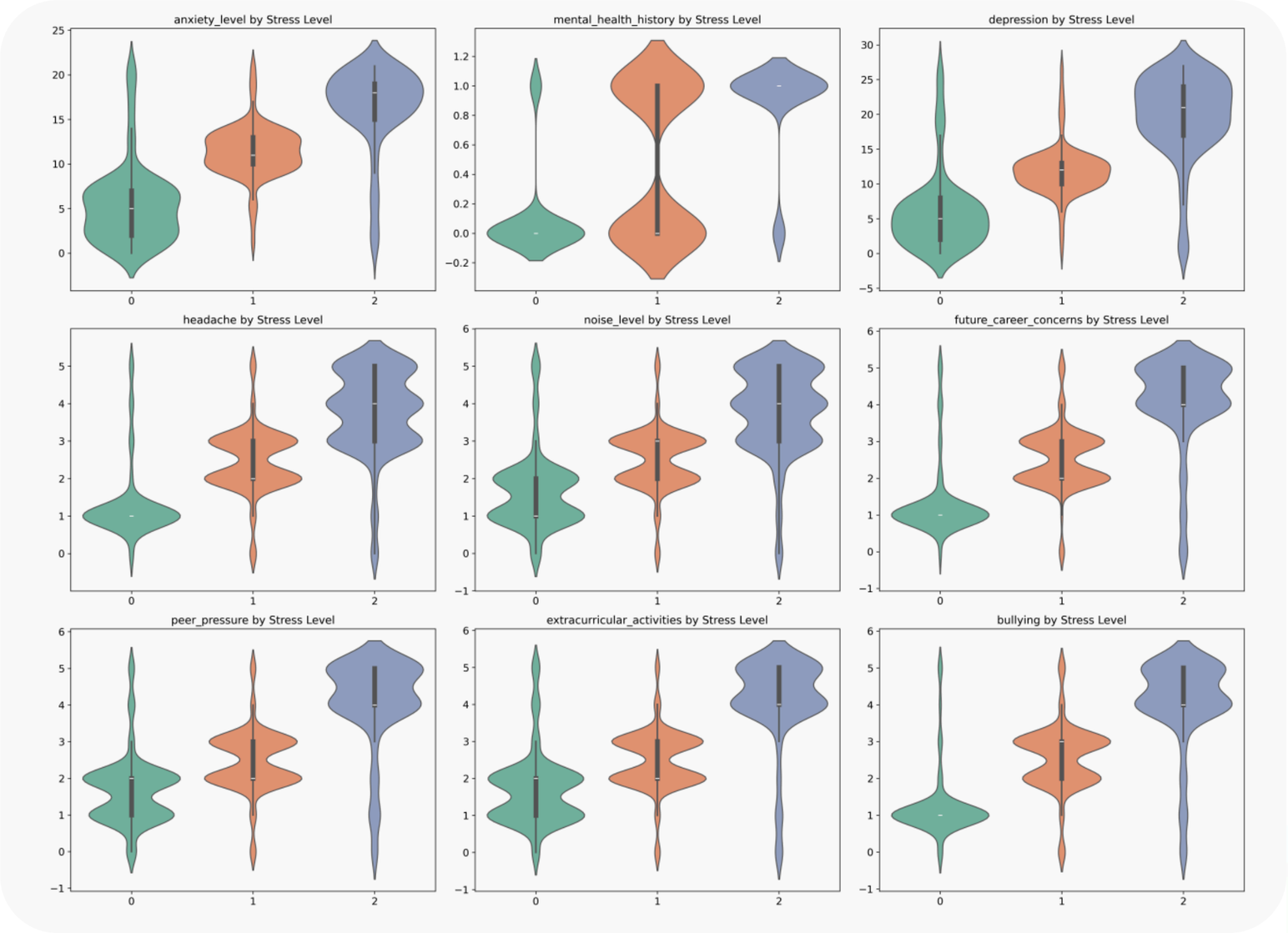
  - Low Stress (Level 0): Narrow, Low-Value Densities.
  - High Stress (Level 2): Broader, Higher-Value Distributions.



3. DATA VISUALIZATION AND DATA PREPROCESSING

Violin Plots

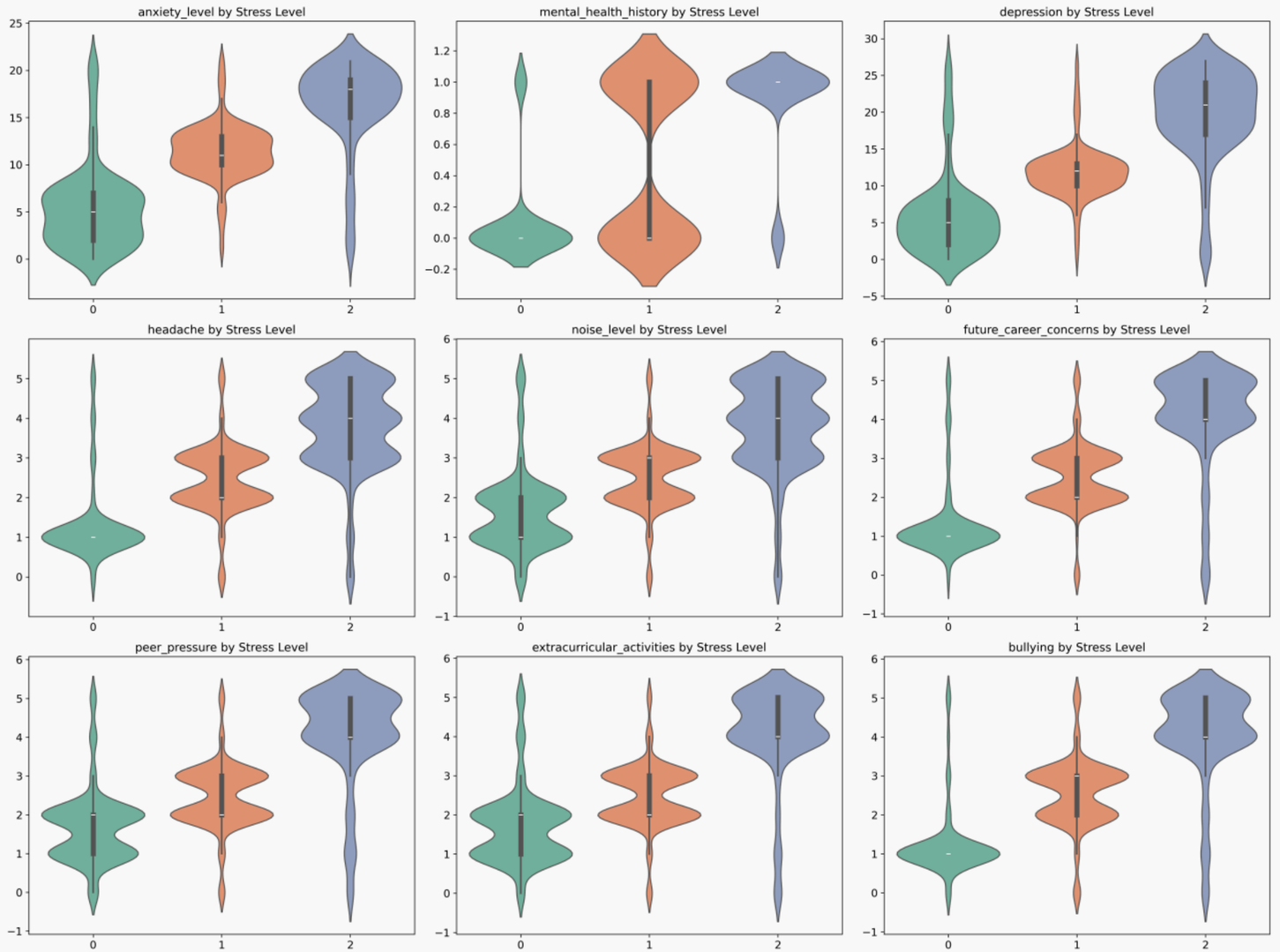
- Other Stressors: Headache and noise_level medians also progress significantly from low to high stress (e.g., ~1–2 → ~3–4).
- Binary Feature: Mental_health_history becomes Bimodal at Stress Level 2. Suggests a Threshold Effect for decision rules.



3. DATA VISUALIZATION AND DATA PREPROCESSING

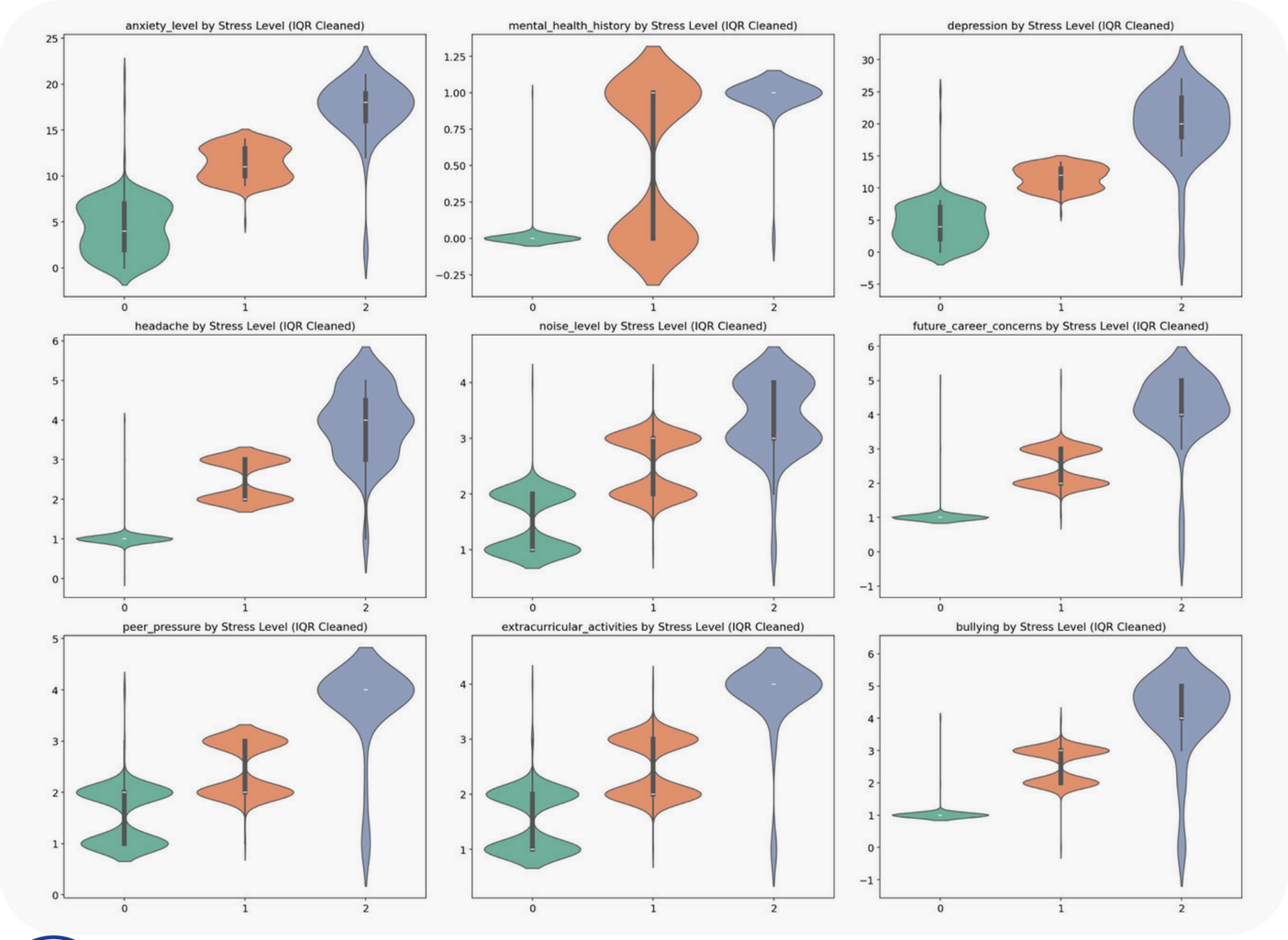
Violin Plots

- Variability/Outliers: Wider Violins (Level 2) indicate Heteroscedasticity (increased variability).
    - Extended Tails highlight potential Outliers in high-stress groups, notably for depression (beyond 20-25) and anxiety.
- Implication: Emphasizes the critical need for Outlier-Aware Preprocessing.



3. DATA VISUALIZATION AND DATA PREPROCESSING

After Applying IQR and SMOTE



- Outliers removed → distributions significantly more compact and realistic.
- Extreme tails eliminated
- Implausible negative values gone.
- Differences between stress levels now clearer and more pronounced without distortion from anomalies.

```
================================================================================
PREPROCESSING PIPELINE COMPLETED
================================================================================

Removing outliers using IQR...
Rows after outlier removal: Train 712, Test 130

Class distribution:
stress_level
0    0.372191
1    0.358146
2    0.269663
Name: proportion, dtype: float64
Apply SMOTE? (y/n): n

Saving processed data...
✓ Data exported to 'data/processed/train_data.csv'
✓ Data exported to 'data/processed/test_data.csv'
✓ Target distribution plot saved to 'results/target_distribution.png'
```

3. DATA VISUALIZATION AND DATA PREPROCESSING

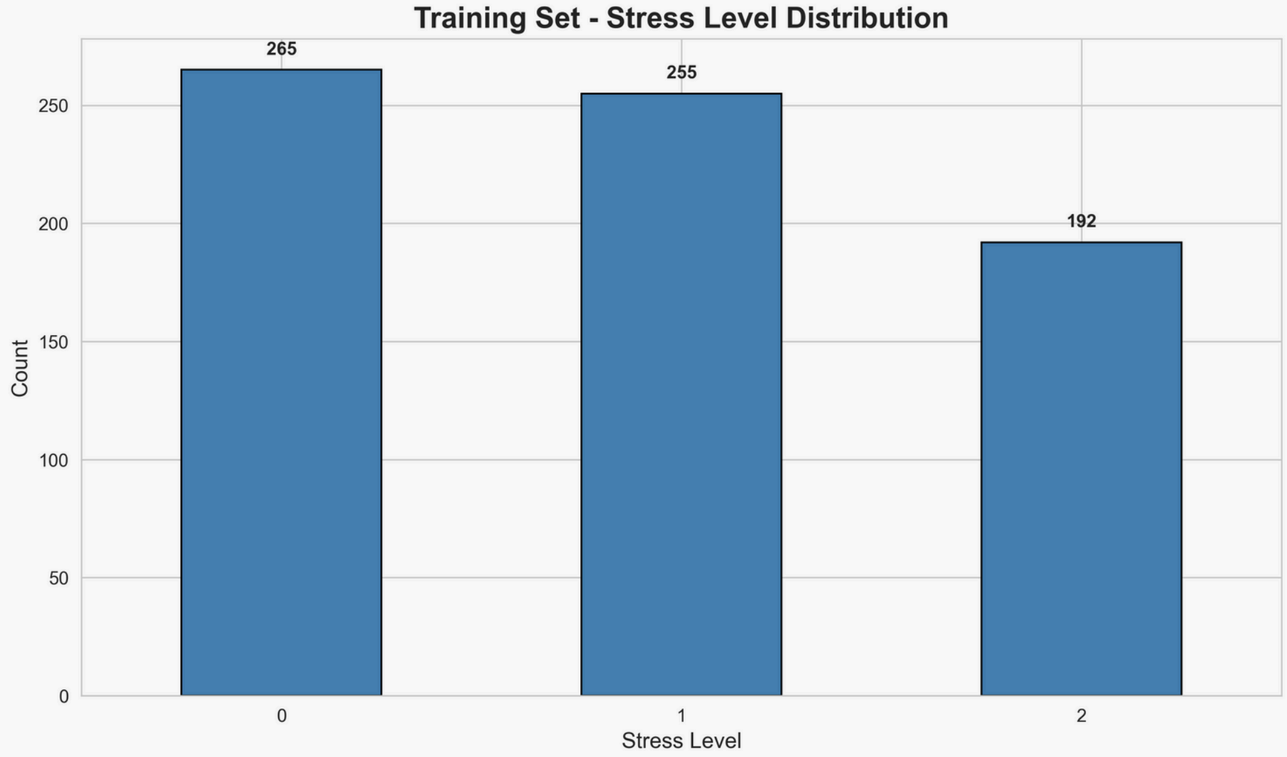
Target distribution plots

Train-test splitting (80/20)

✓ Data split completed

- Training set: 712 samples (80%)
- Testing set: 130 samples (20%)

✓ Normalization completed

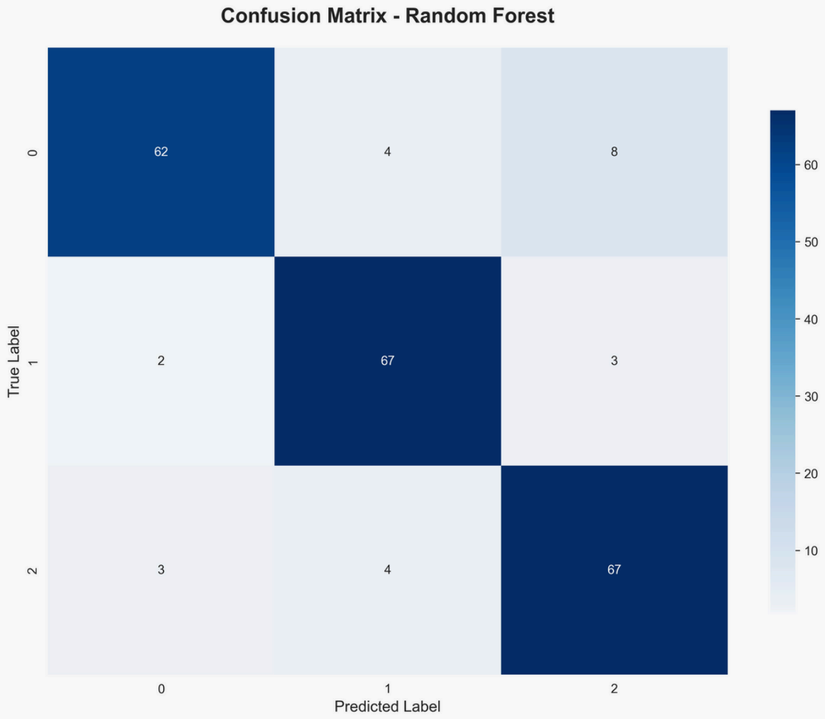


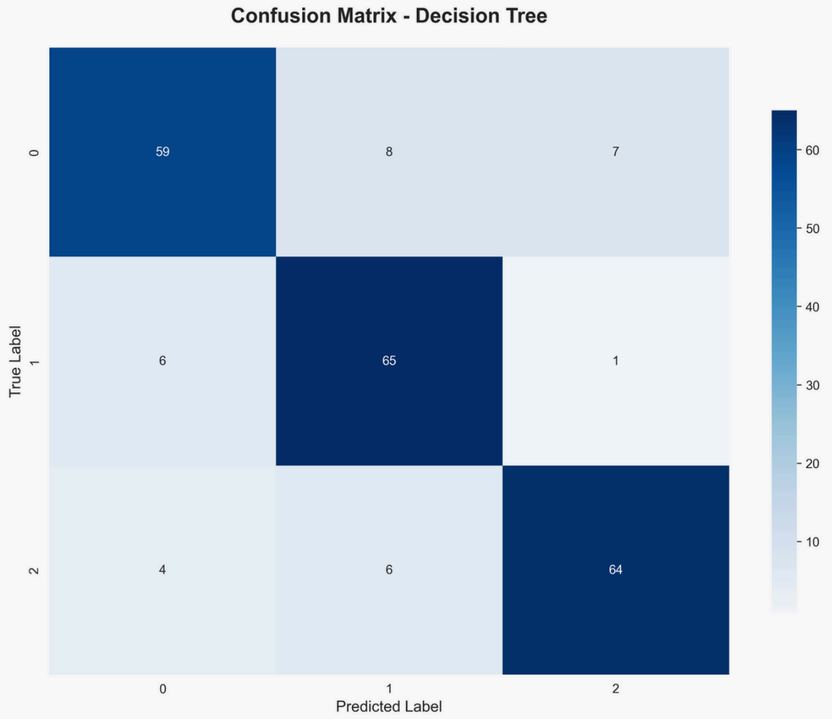
4. CLASSIFICATION MODELS

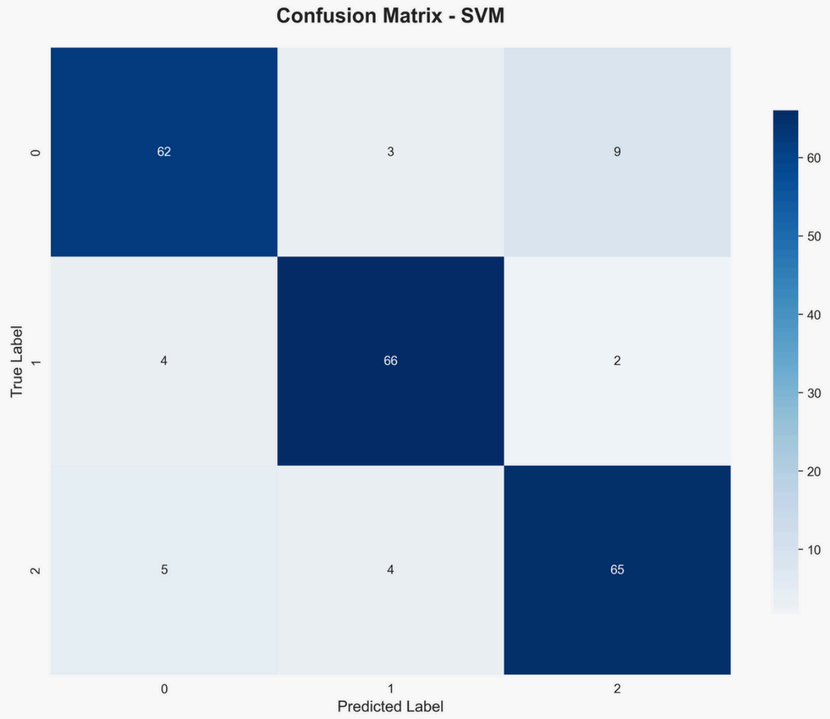
Classifier Models

1. Random Forest
2. Decision Tree
3. SVM
4. KNN
5. Naive Bayes
6. Logistic Regression
7. Gradient Boosting

4. CLASSIFICATION MODELS







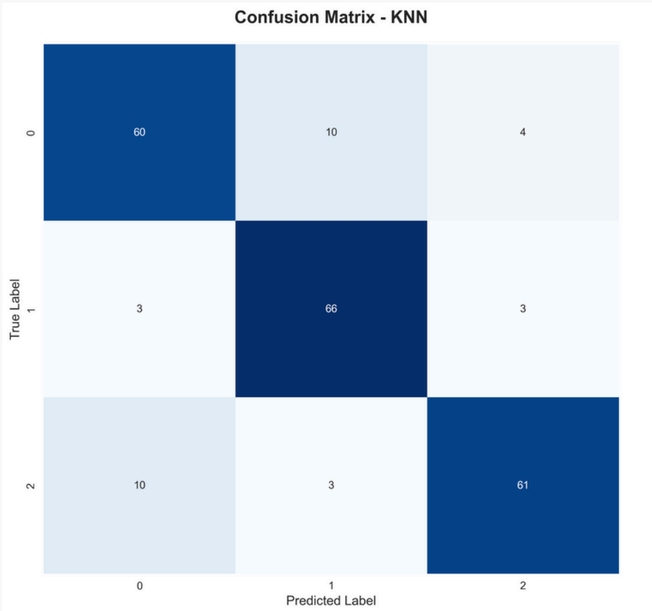

```
Confusion Matrix:
[[62  4  8]
 [ 2 67  3]
 [ 3  4 67]]
```

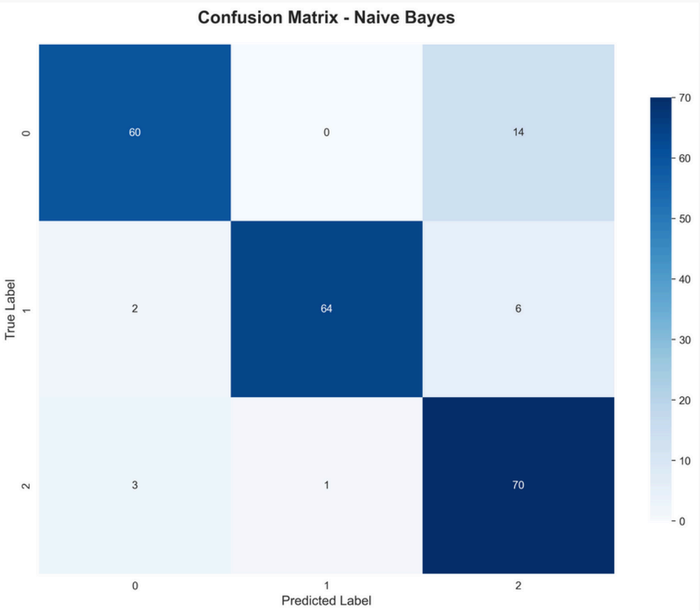
```
Confusion Matrix:
[[59  8  7]
 [ 6 65  1]
 [ 4  6 64]]
```

```
Confusion Matrix:
[[62  3  9]
 [ 4 66  2]
 [ 5  4 65]]
```

4. CLASSIFICATION MODELS

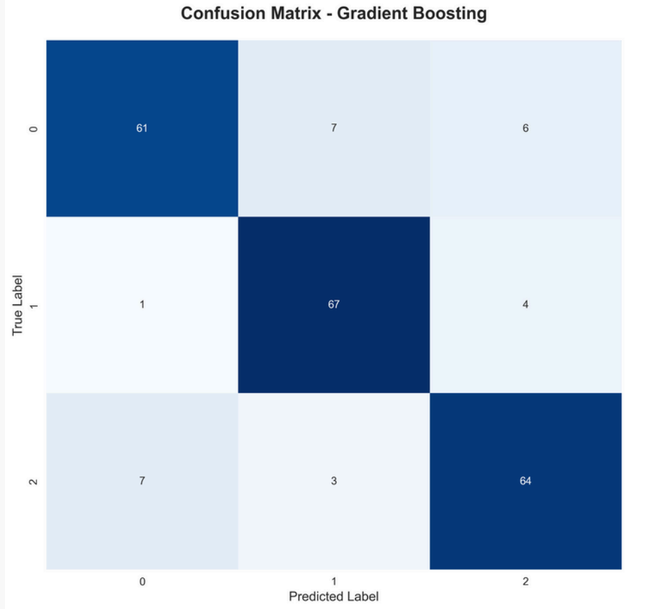
```
Confusion Matrix:
[[60 10  4]
 [ 3 66  3]
 [10  3 61]]
```

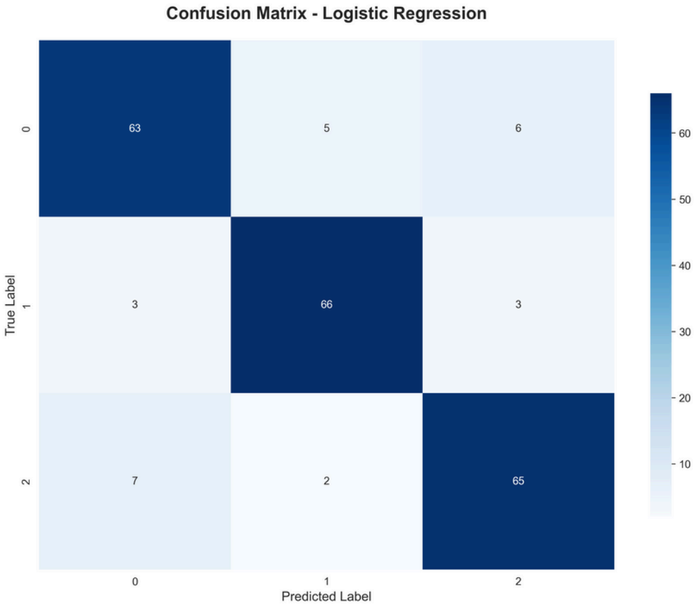





```
Confusion Matrix:
[[60  0 14]
 [ 2 64  6]
 [ 3  1 70]]
```

```
Confusion Matrix:
[[61  7  6]
 [ 1 67  4]
 [ 7  3 64]]
```



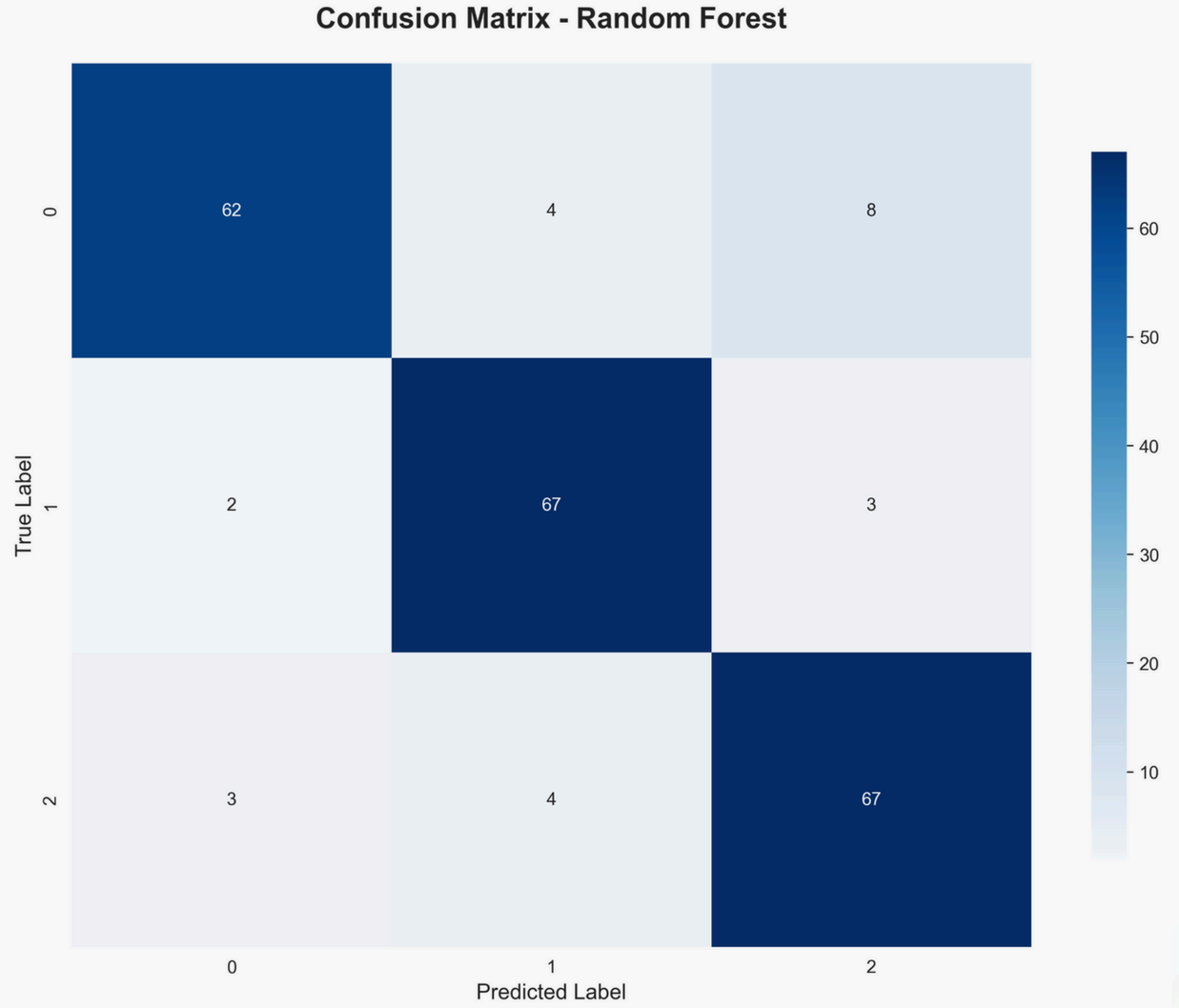



```
Confusion Matrix:
[[63  5  6]
 [ 3 66  3]
 [ 7  2 65]]
```

4. CLASSIFICATION MODELS

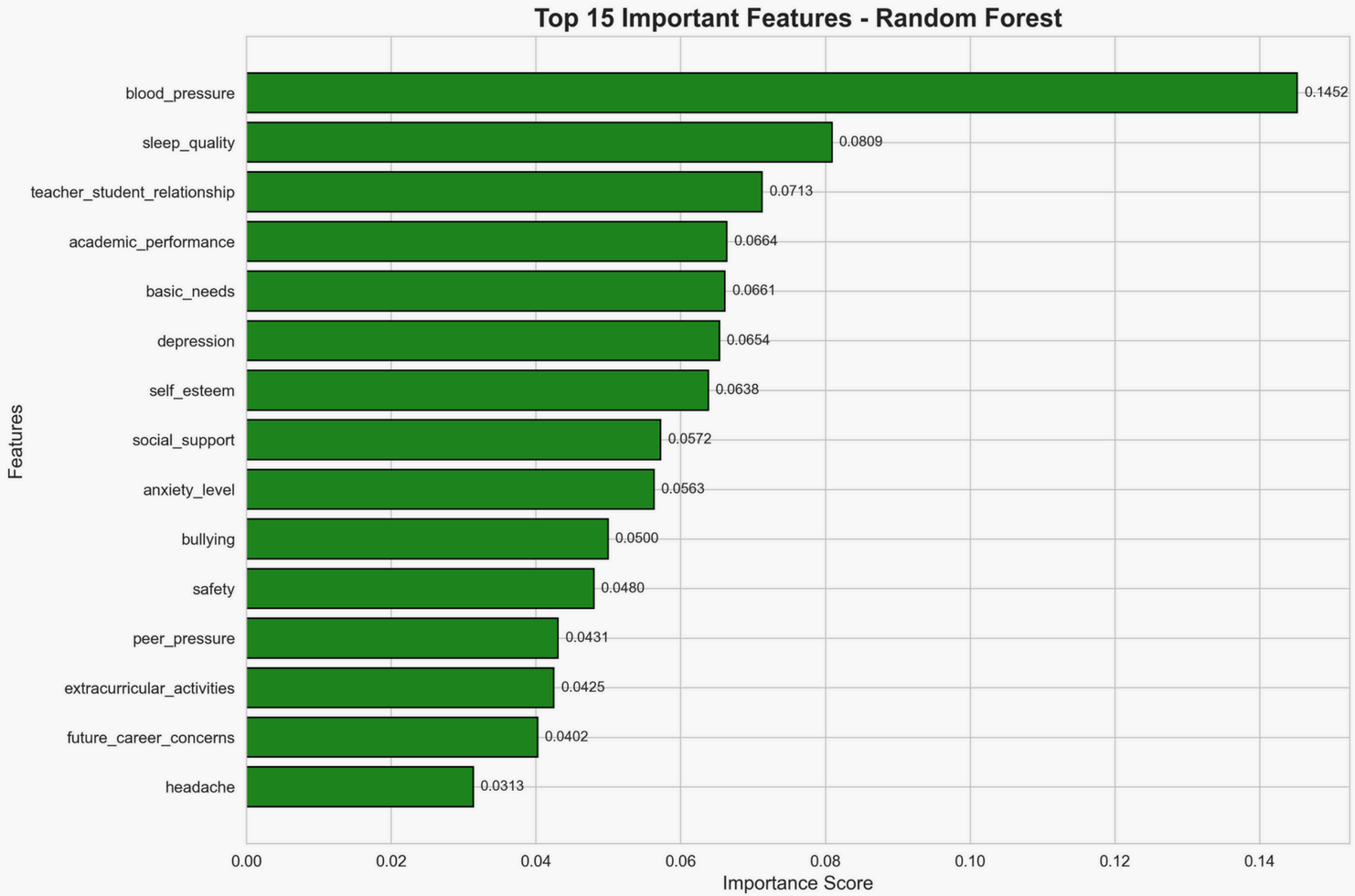
Random Forest

```
Training Random Forest...
✓ Random Forest trained successfully

======================================================================
EVALUATION RESULTS: Random Forest
======================================================================
Accuracy:  0.8909 (89.09%)
Precision: 0.8926
Recall:    0.8909
F1-Score:  0.8907

Confusion Matrix:
[[62  4  8]
 [ 2 67  3]
 [ 3  4 67]]

Classification Report:
              precision    recall  f1-score   support

           0       0.93      0.84      0.88        74
           1       0.89      0.93      0.91        72
           2       0.86      0.91      0.88        74

    accuracy                           0.89       220
   macro avg       0.89      0.89      0.89       220
weighted avg       0.89      0.89      0.89       220

✓ Confusion matrix saved to 'results/confusion_matrix.png'
```



4. CLASSIFICATION MODELS

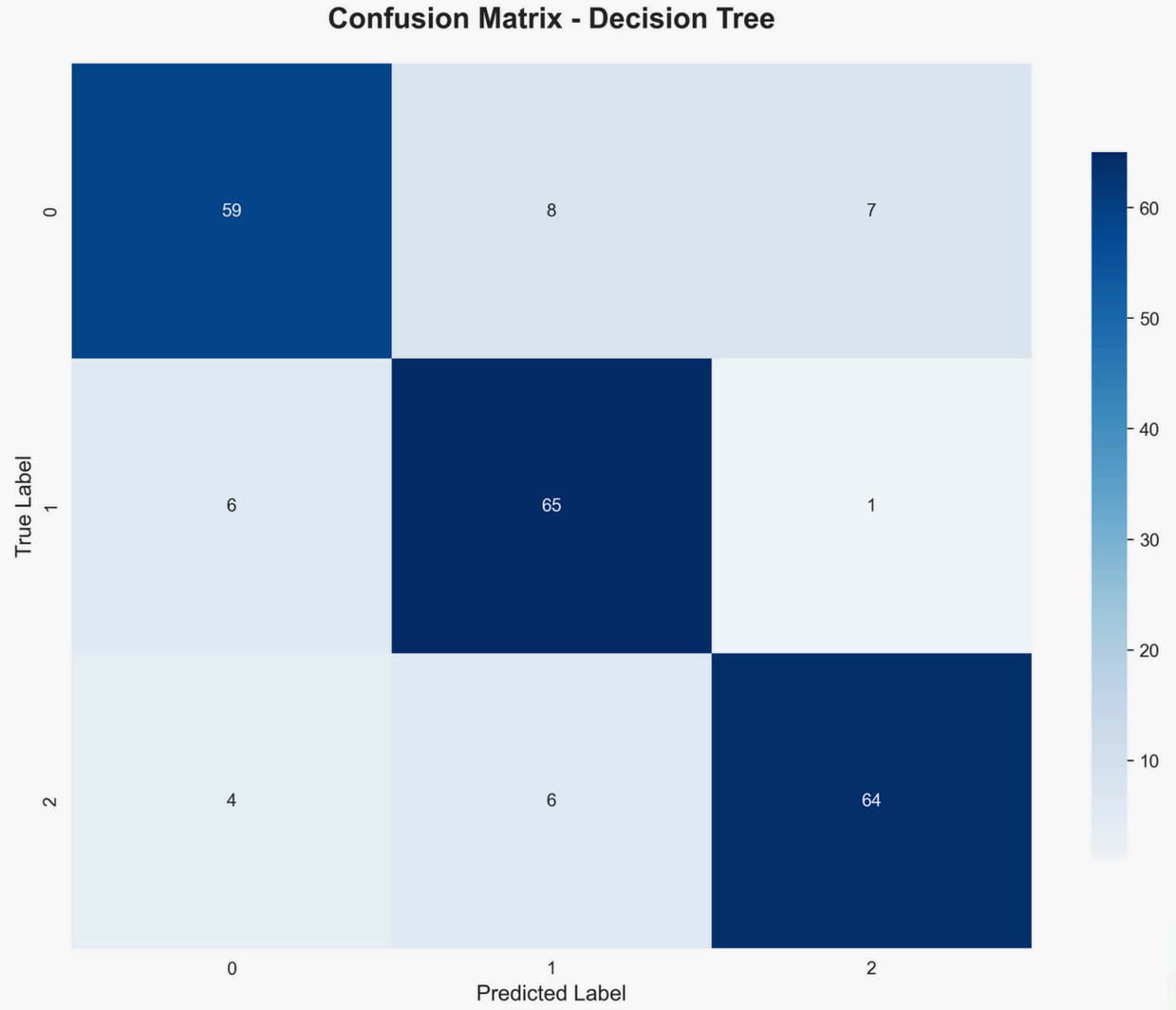
Random Forest



4. CLASSIFICATION MODELS

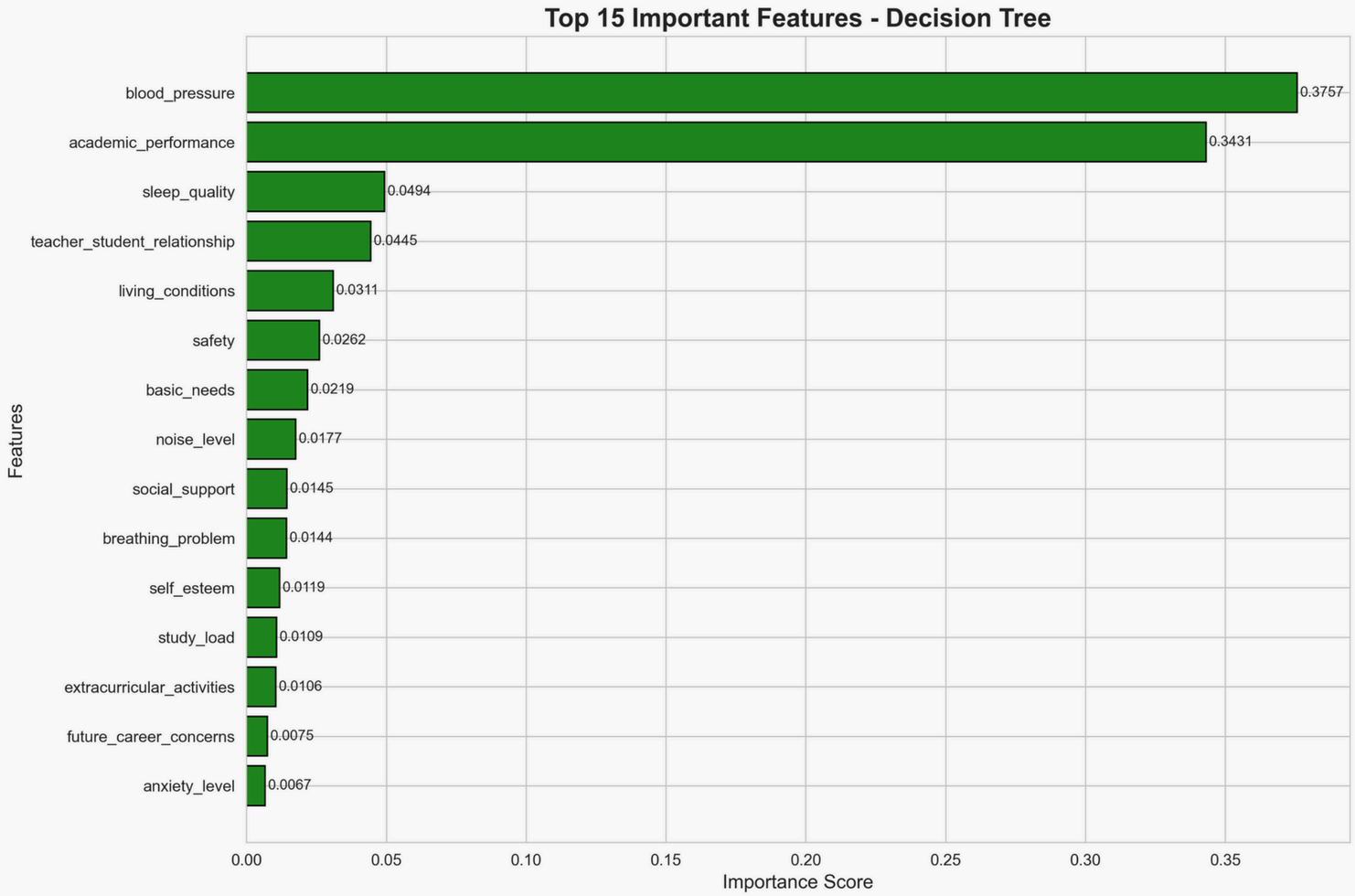
Decision Tree

```
Training Decision Tree...
✓ Decision Tree trained successfully

======================================================================
EVALUATION RESULTS: Decision Tree
======================================================================
Accuracy:  0.8545 (85.45%)
Precision: 0.8559
Recall:    0.8545
F1-Score:  0.8542

Confusion Matrix:
[[59  8  7]
 [ 6 65  1]
 [ 4  6 64]]

Classification Report:
              precision    recall  f1-score   support

           0       0.86      0.80      0.83        74
           1       0.82      0.90      0.86        72
           2       0.89      0.86      0.88        74

    accuracy                           0.85       220
   macro avg       0.86      0.85      0.85       220
weighted avg       0.86      0.85      0.85       220

✓ Confusion matrix saved to 'results/confusion_matrix.png'
```



4. CLASSIFICATION MODELS

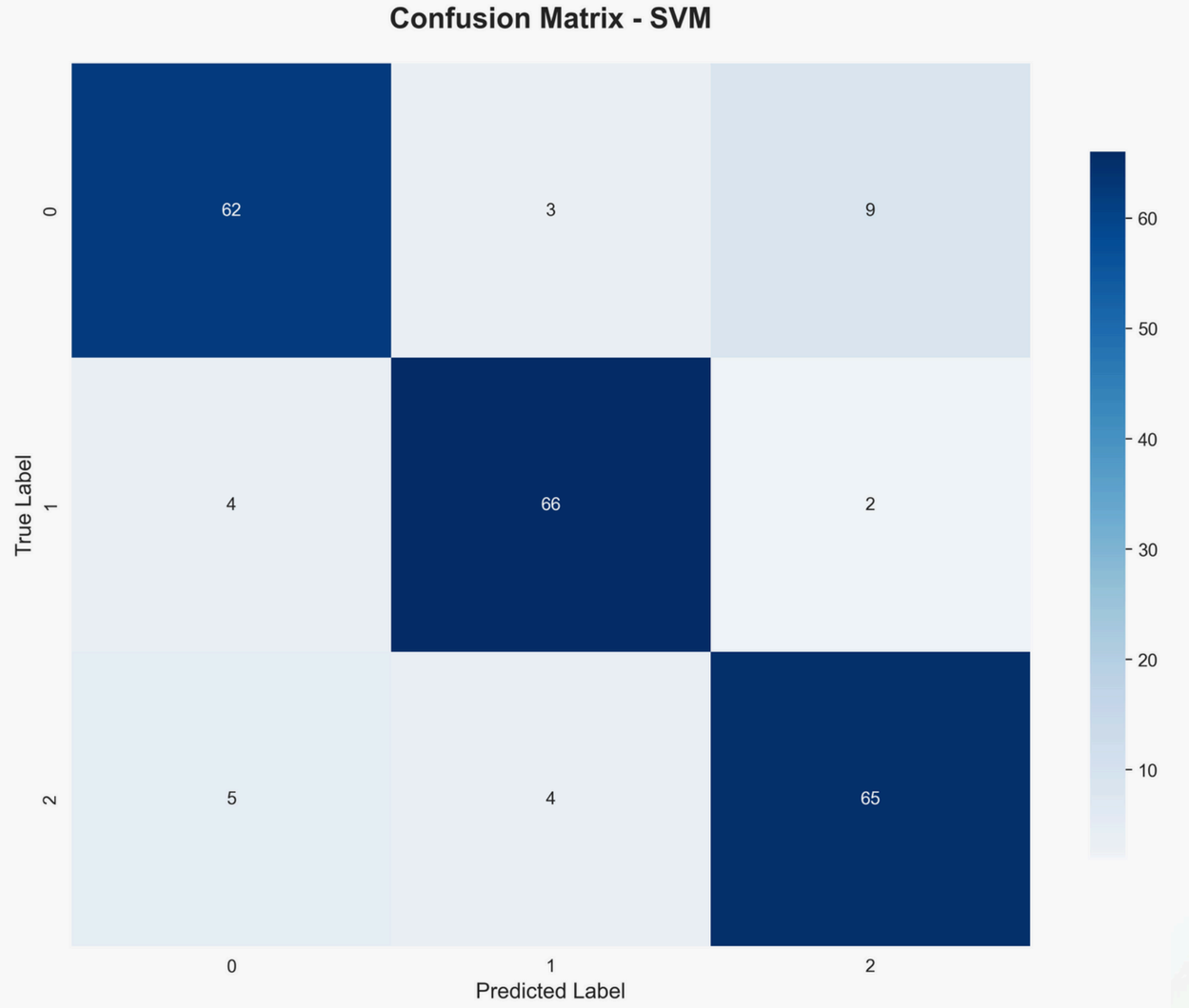
Decision Tree



4. CLASSIFICATION MODELS

SVM

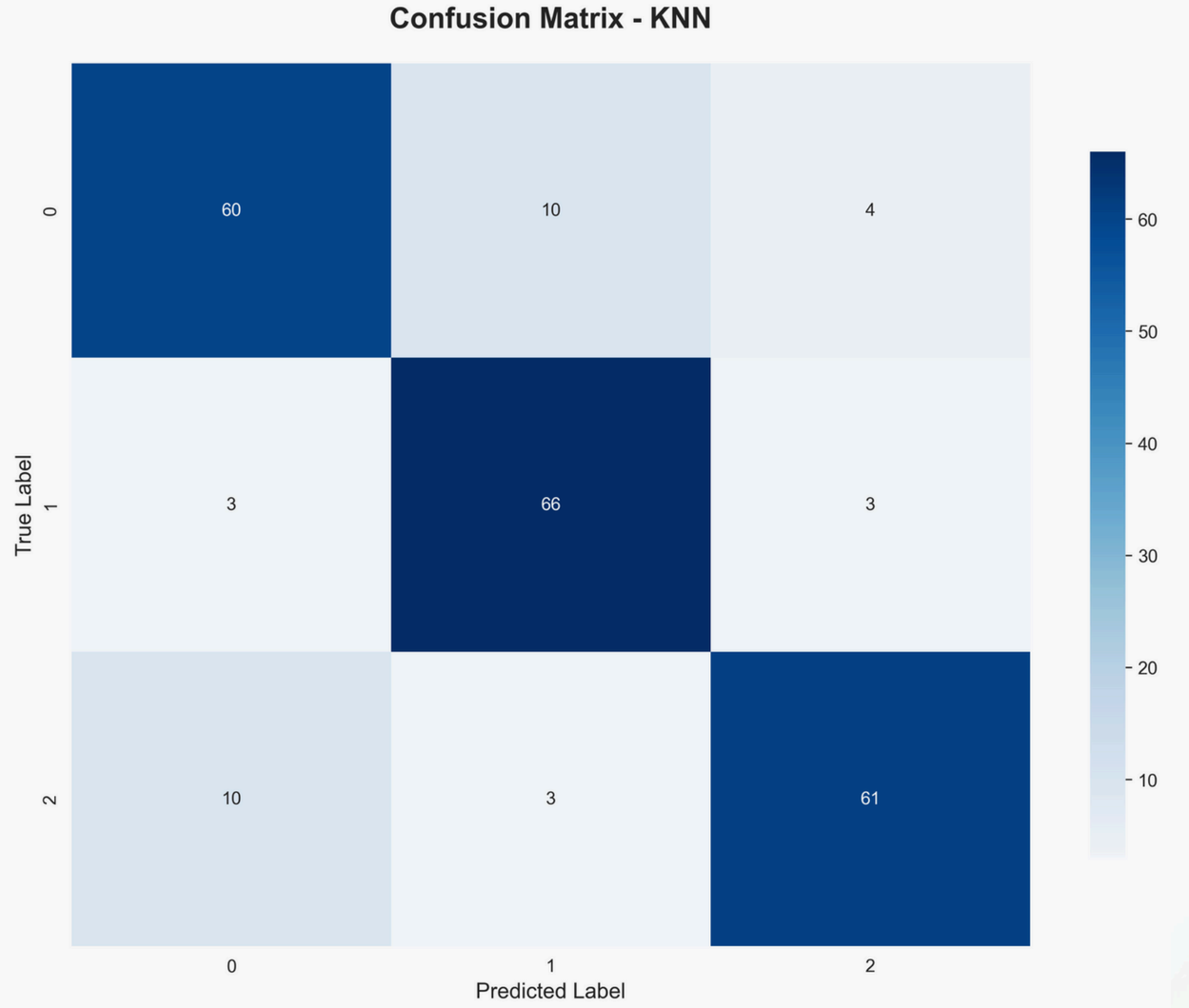
```
Training SVM...
✓ SVM trained successfully

======================================================================
EVALUATION RESULTS: SVM
======================================================================
Accuracy:  0.8773 (87.73%)
Precision: 0.8773
Recall:    0.8773
F1-Score:  0.8771

Confusion Matrix:
[[62  3  9]
 [ 4 66  2]
 [ 5  4 65]]

Classification Report:
              precision    recall  f1-score   support

           0       0.87      0.84      0.86        74
           1       0.90      0.92      0.91        72
           2       0.86      0.88      0.87        74

    accuracy                           0.88       220
   macro avg       0.88      0.88      0.88       220
weighted avg       0.88      0.88      0.88       220

✓ Confusion matrix saved to 'results/confusion_matrix.png'
```



4. CLASSIFICATION MODELS

KNN

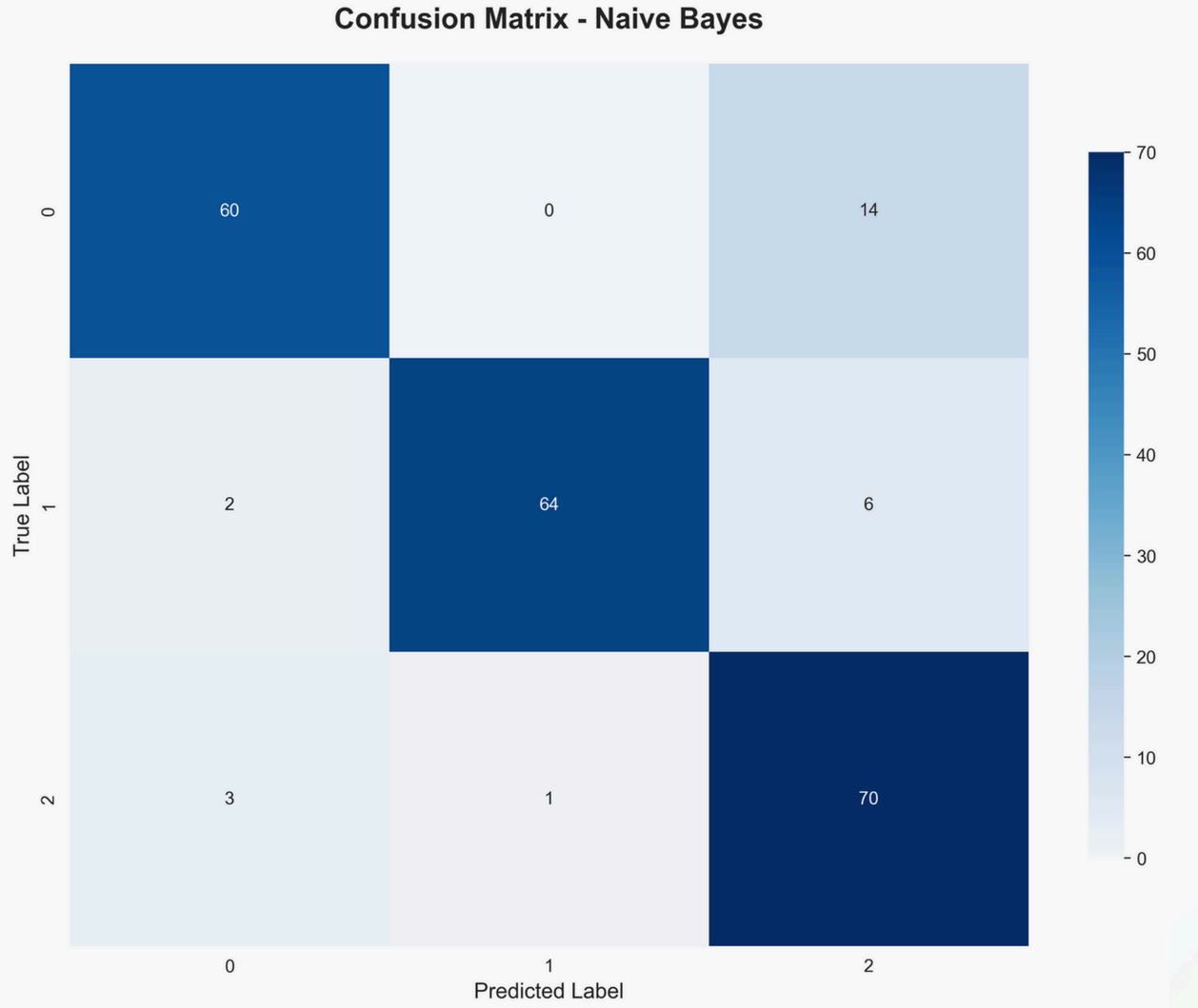
```
Training KNN...
✓ KNN trained successfully

======================================================================
EVALUATION RESULTS: KNN
======================================================================
Accuracy:  0.8500 (85.00%)
Precision: 0.8516
Recall:    0.8500
F1-Score:  0.8497

Confusion Matrix:
[[60 10  4]
 [ 3 66  3]
 [10  3 61]]

Classification Report:
              precision    recall  f1-score   support

           0       0.82      0.81      0.82        74
           1       0.84      0.92      0.87        72
           2       0.90      0.82      0.86        74

    accuracy                           0.85       220
   macro avg       0.85      0.85      0.85       220
weighted avg       0.85      0.85      0.85       220

✓ Confusion matrix saved to 'results/confusion_matrix.png'
```



4. CLASSIFICATION MODELS

Naive Bayes

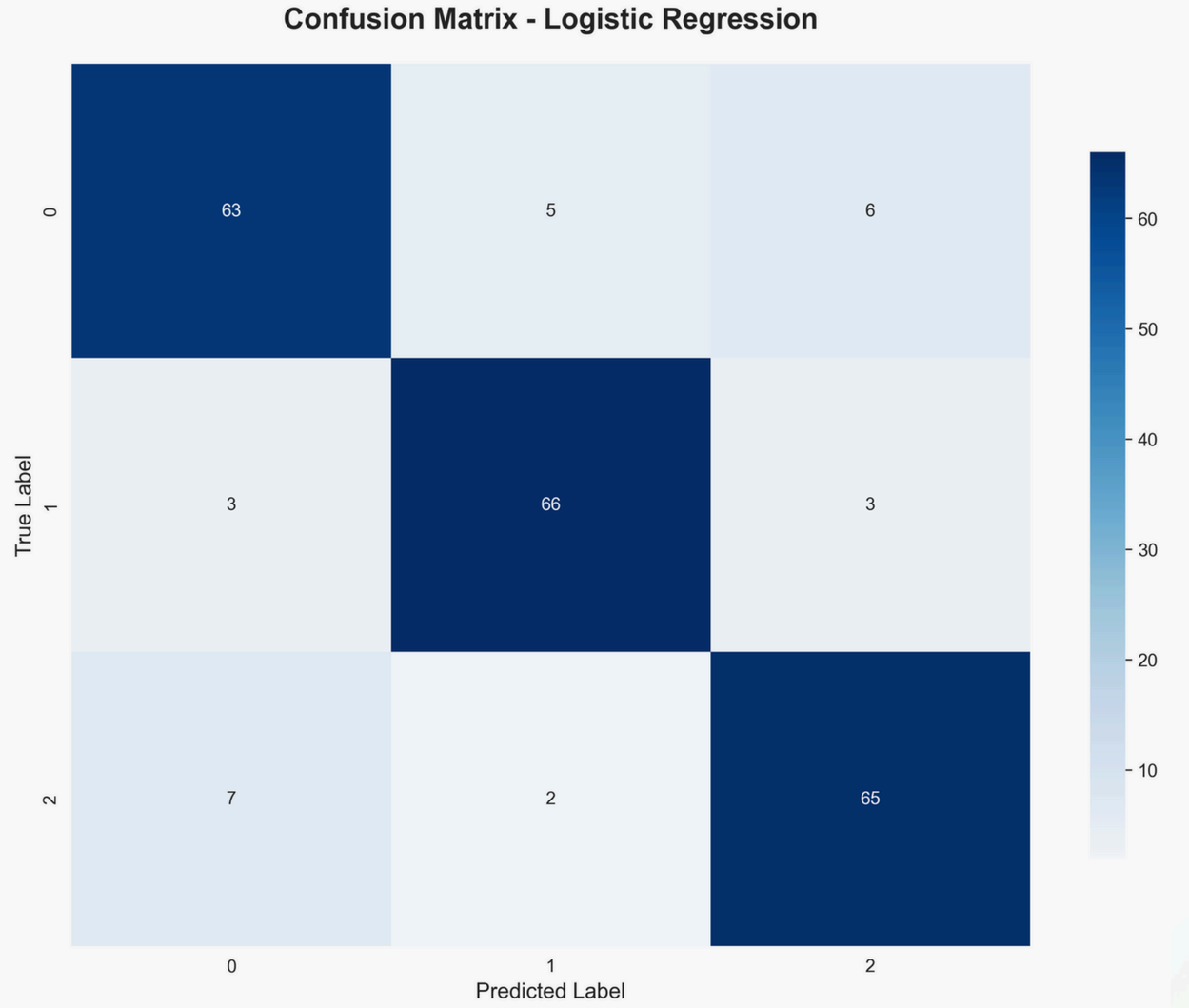
```
Training Naive Bayes...
✓ Naive Bayes trained successfully

======================================================================
EVALUATION RESULTS: Naive Bayes
======================================================================
Accuracy:  0.8818 (88.18%)
Precision: 0.8943
Recall:    0.8818
F1-Score:  0.8833

Confusion Matrix:
[[60  0 14]
 [ 2 64  6]
 [ 3  1 70]]

Classification Report:
              precision    recall  f1-score   support

           0       0.92      0.81      0.86        74
           1       0.98      0.89      0.93        72
           2       0.78      0.95      0.85        74

    accuracy                           0.88       220
   macro avg       0.90      0.88      0.88       220
weighted avg       0.89      0.88      0.88       220

✓ Confusion matrix saved to 'results/confusion_matrix.png'
```



4. CLASSIFICATION MODELS

Logistic Regression

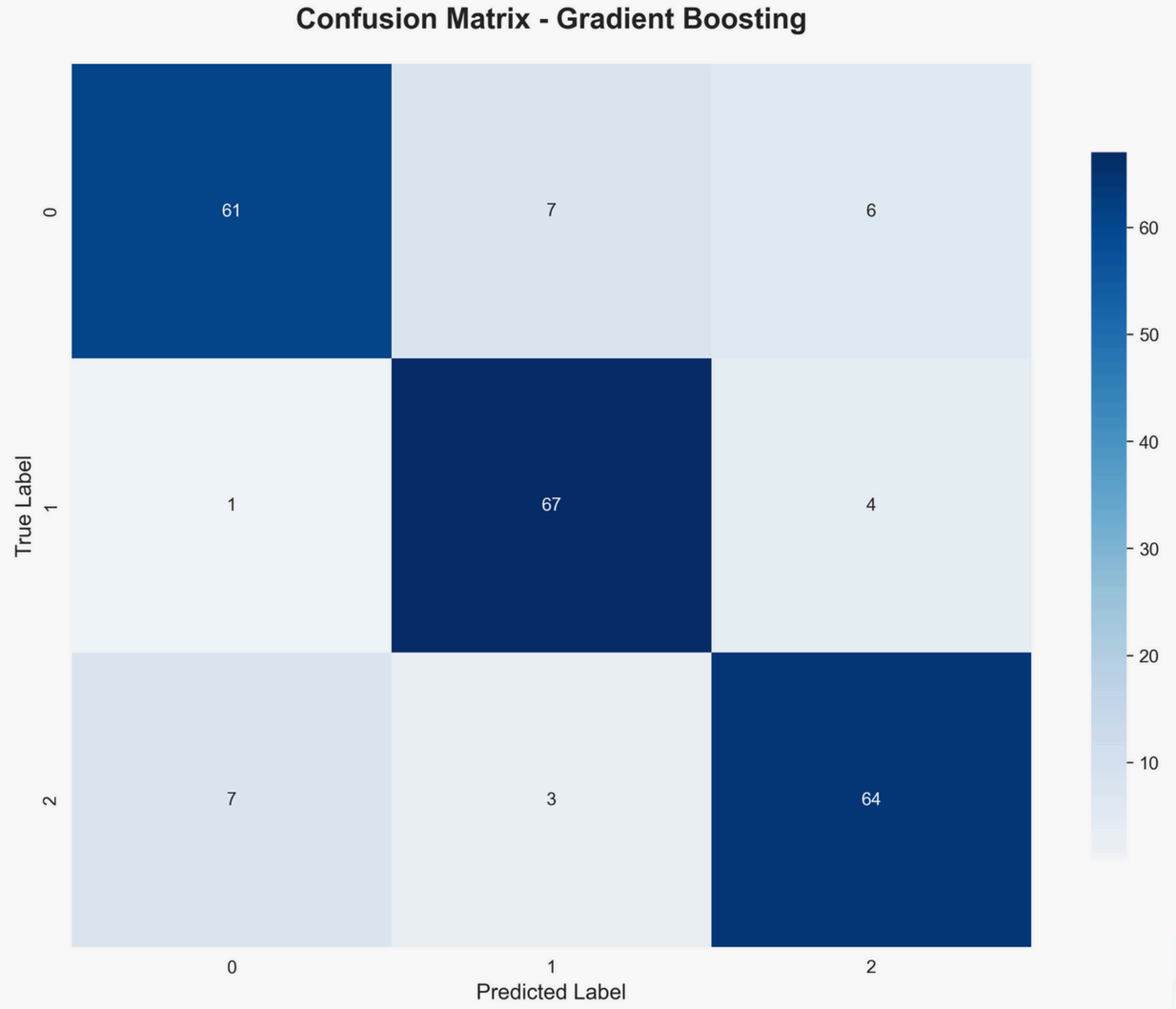
```
Training Logistic Regression...
✓ Logistic Regression trained successfully

======================================================================
EVALUATION RESULTS: Logistic Regression
======================================================================
Accuracy:  0.8818 (88.18%)
Precision: 0.8816
Recall:    0.8818
F1-Score:  0.8817

Confusion Matrix:
[[63  5  6]
 [ 3 66  3]
 [ 7  2 65]]

Classification Report:
              precision    recall  f1-score   support

           0       0.86      0.85      0.86        74
           1       0.90      0.92      0.91        72
           2       0.88      0.88      0.88        74

    accuracy                           0.88       220
   macro avg       0.88      0.88      0.88       220
weighted avg       0.88      0.88      0.88       220

✓ Confusion matrix saved to 'results/confusion_matrix.png'
```



4. CLASSIFICATION MODELS

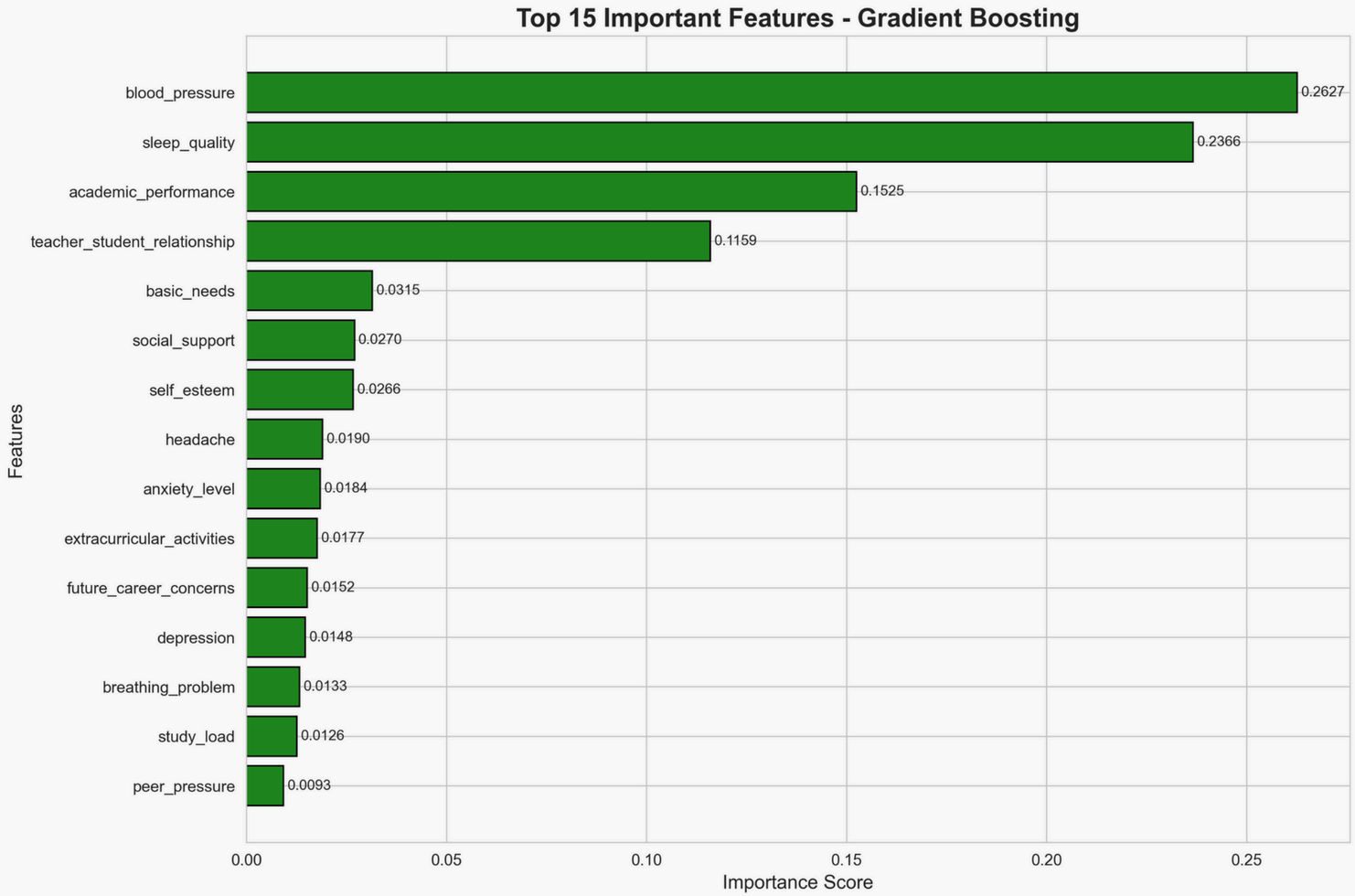
Gradient Boosting

```
Training Gradient Boosting...
✓ Gradient Boosting trained successfully

======================================================================
EVALUATION RESULTS: Gradient Boosting
======================================================================
Accuracy:  0.8727 (87.27%)
Precision: 0.8730
Recall:    0.8727
F1-Score:  0.8722

Confusion Matrix:
[[61  7  6]
 [ 1 67  4]
 [ 7  3 64]]

Classification Report:
              precision    recall  f1-score   support

           0       0.88      0.82      0.85        74
           1       0.87      0.93      0.90        72
           2       0.86      0.86      0.86        74

    accuracy                           0.87       220
   macro avg       0.87      0.87      0.87       220
weighted avg       0.87      0.87      0.87       220

✓ Confusion matrix saved to 'results/confusion_matrix.png'
```



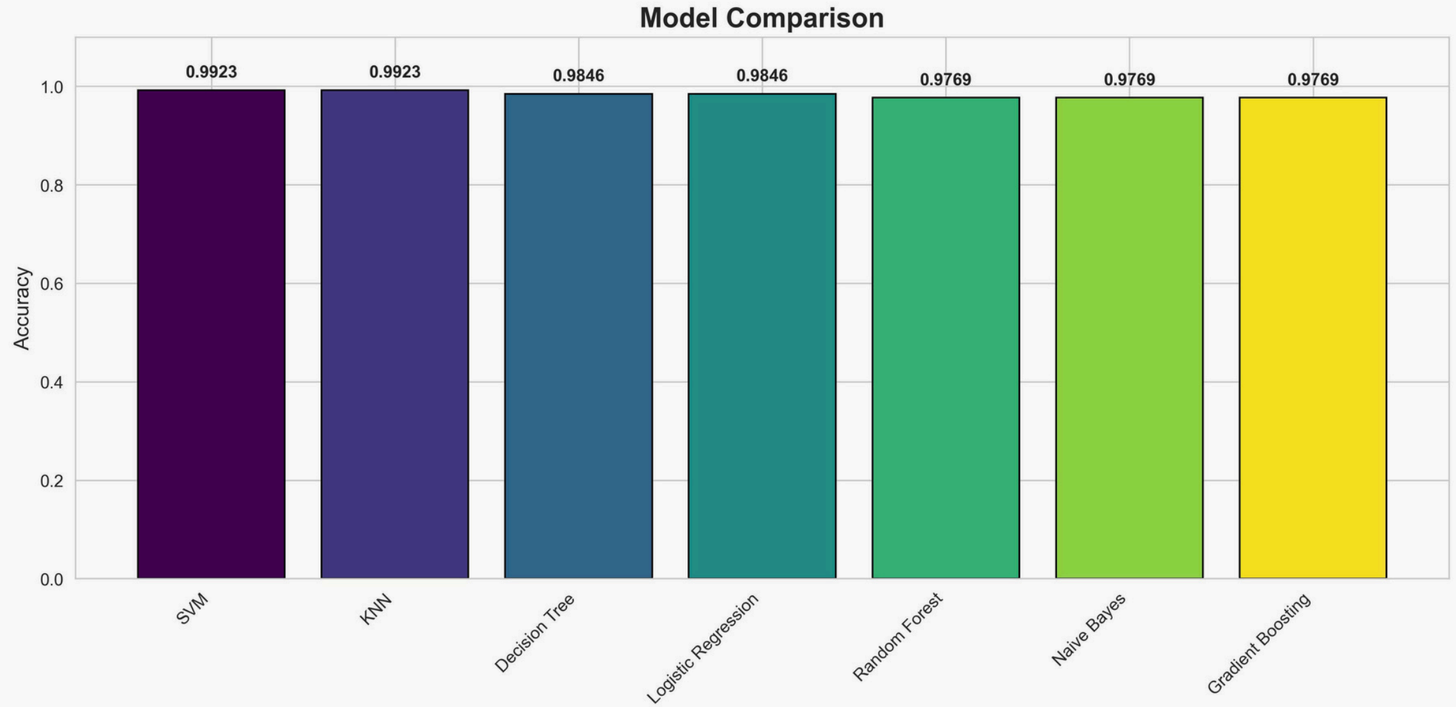
4. CLASSIFICATION MODELS

Random Forest



4. CLASSIFICATION MODELS

```
======================================================================
TRAINING ALL CLASSIFICATION MODELS
======================================================================

Training Random Forest...
✓ Random Forest trained successfully

Training Decision Tree...
✓ Decision Tree trained successfully

Training SVM...
✓ SVM trained successfully

Training KNN...
✓ KNN trained successfully

Training Naive Bayes...
✓ Naive Bayes trained successfully

Training Logistic Regression...
✓ Logistic Regression trained successfully

Training Gradient Boosting...
✓ Gradient Boosting trained successfully

✓ All models trained and evaluated

======================================================================
MODEL COMPARISON
======================================================================
                Model Accuracy Precision Recall F1-Score
                  SVM   0.9923    0.9933 0.9923   0.9925
        Random Forest   0.9846    0.9847 0.9846   0.9843
                  KNN   0.9846    0.9847 0.9846   0.9846
  Logistic Regression   0.9846    0.9851 0.9846   0.9835
    Gradient Boosting   0.9846    0.9851 0.9846   0.9844
        Decision Tree   0.9769    0.9762 0.9769   0.9761
          Naive Bayes   0.9769    0.9761 0.9769   0.9761

🏆 Best Model: SVM
   Accuracy: 0.9923 (99.23%)
✓ Confusion matrix saved to 'results/confusion_matrix.png'
✓ Model comparison plot saved to 'results/model_comparison.png'
✓ Results saved to 'results/classification_results.json'
```



5. CLUSTERING ALGORITHMS

Clustering Algorithms

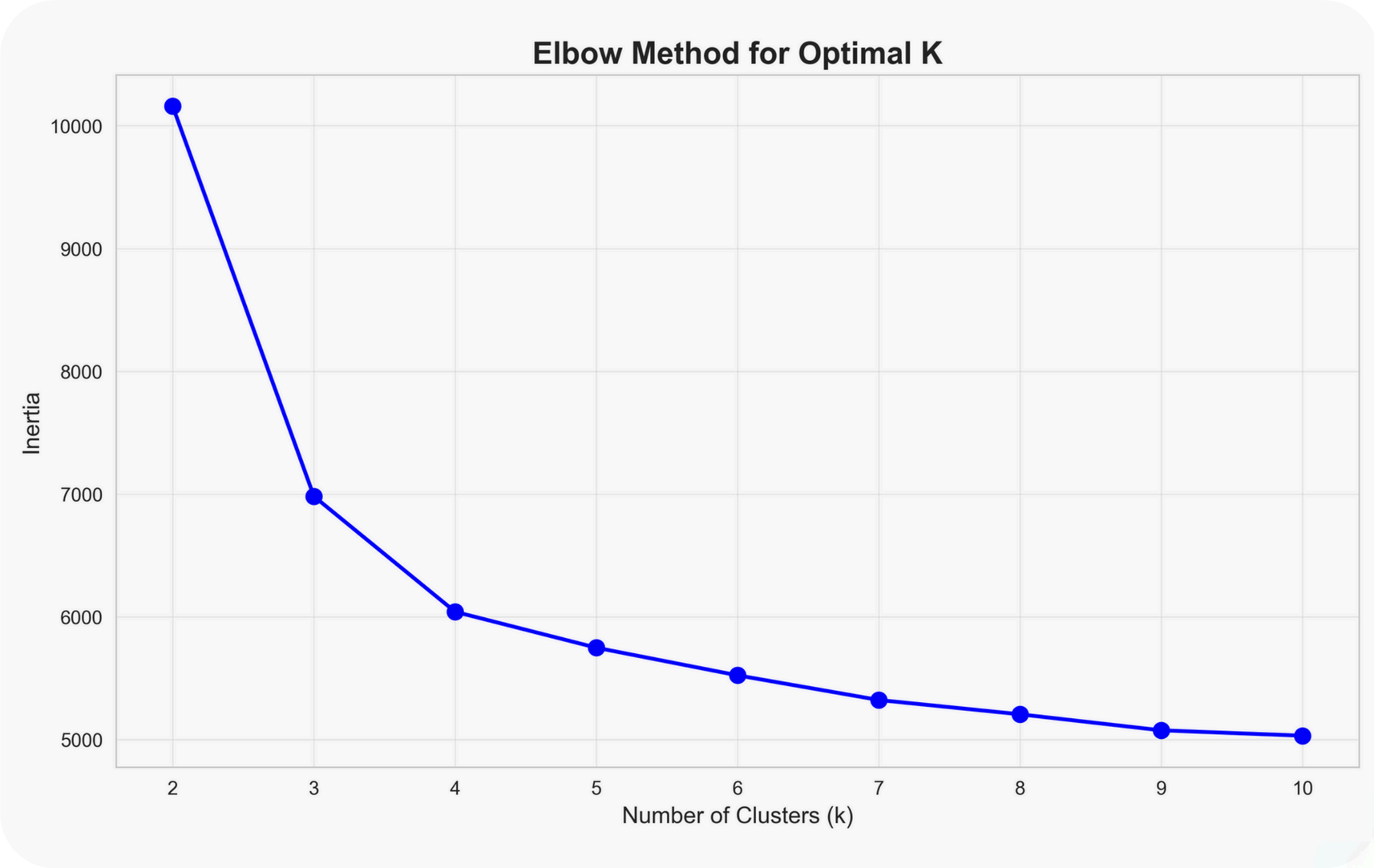
1. K-means
Centroid-based clustering

2. Hierarchical
Agglomerative clustering

2. DBSCAN
Density-based clustering

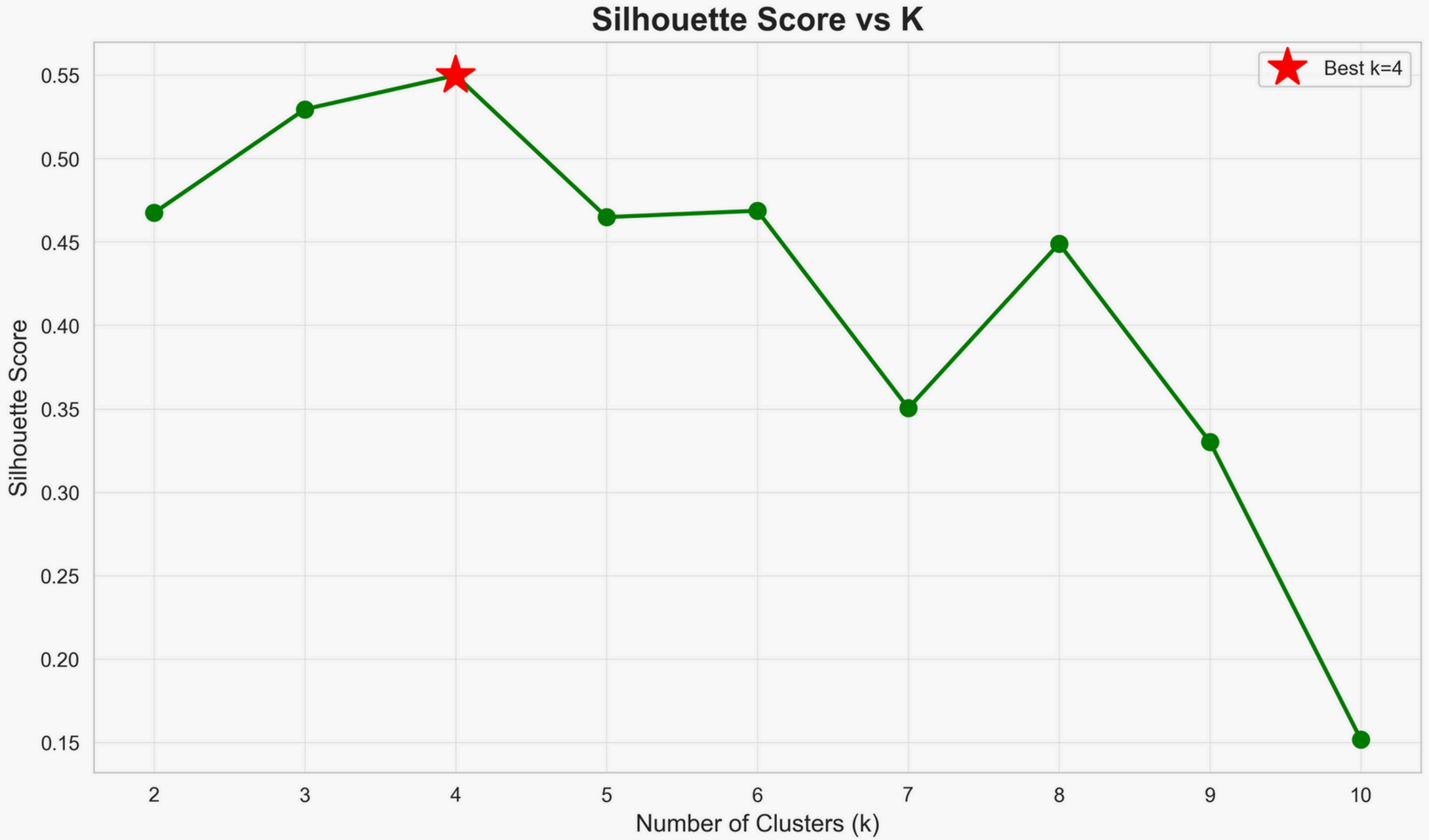
5. CLUSTERING ALGORITHMS

Elbow curves



5. CLUSTERING ALGORITHMS

Silhouette score plots

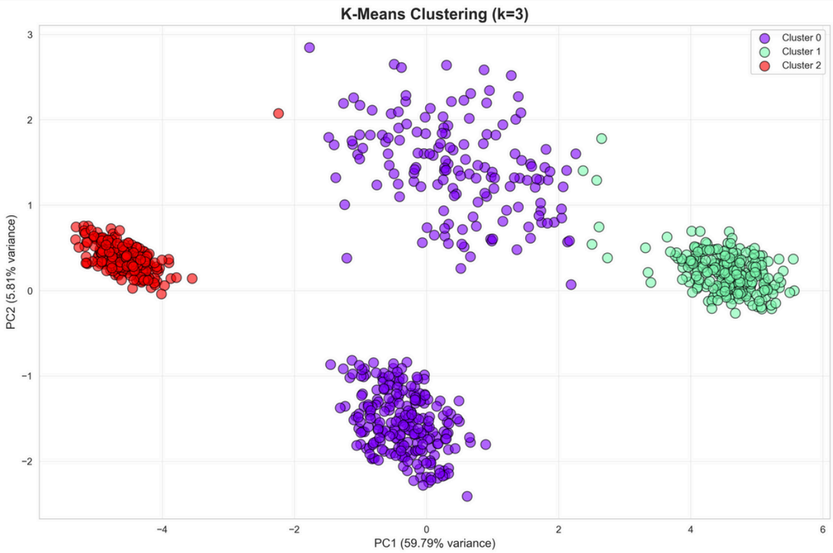


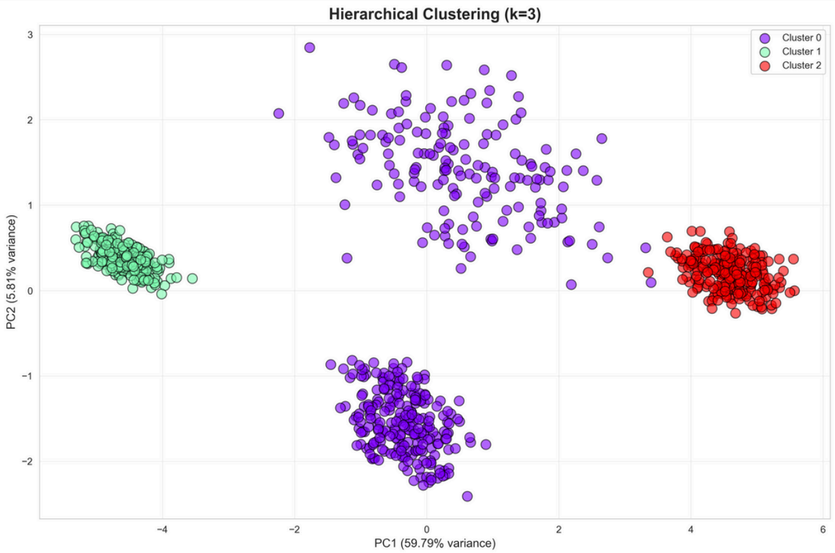
5. CLUSTERING ALGORITHMS

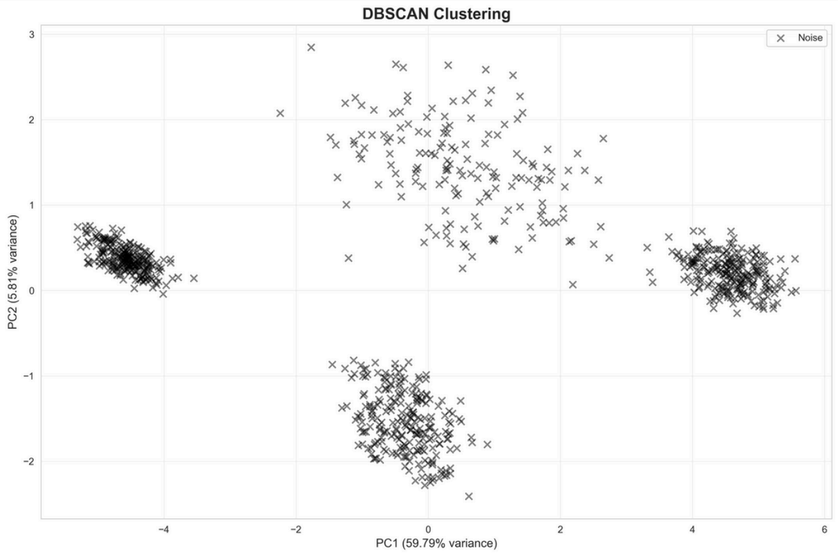
K-Means
Centroid-based clustering

Hierarchical
Agglomerative clustering

DBSCAN
Density-based clustering








```
================================================================================
CLUSTERING METHODS COMPARISON
================================================================================
        Method  N_Clusters Silhouette Davies-Bouldin Calinski-Harabasz
        KMeans           3     0.5295         0.8077           1028.86
Hierarchical             3     0.5267         0.8001           1008.77
        DBSCAN           0        N/A            N/A               N/A
```

5. CLUSTERING ALGORITHMS

K-Means

Centroid-based clustering

```
Enter number of clusters (default=3): 3

Performing K-Means Clustering (k=3)...
✓ K-Means clustering completed
  Silhouette Score: 0.5295
  Davies-Bouldin Score: 0.8077
  Calinski-Harabasz Score: 1028.8556

======================================================================
CLUSTER ANALYSIS
======================================================================

Cluster Distribution:
Cluster
0    180
1    243
2    289
Name: count, dtype: int64

Cluster Statistics (Mean values):
                0         1         2         3         4         5   ...         8         9         10        11        12        13
Cluster                                                               ...
0        4.911312  0.709410  0.212461  0.208488 -0.090482 -0.033618  ... -0.009427  0.011504  0.027175 -0.013071 -0.007725 -0.018097
1       -4.051101  0.573548  0.155410  0.184741 -0.035066 -0.029819  ... -0.019764 -0.007742  0.010885 -0.007225  0.007081 -0.008172
2        0.347340 -0.924103 -0.263002 -0.285190  0.085840  0.046012  ...  0.022489 -0.000655 -0.026078  0.014216 -0.001143  0.018143

[3 rows x 14 columns]
```

5. CLUSTERING ALGORITHMS

Hierarchical

Agglomerative clustering

```
Enter number of clusters (default=3): 3

Performing Hierarchical Clustering (k=3, linkage=ward)...
✓ Hierarchical clustering completed
  Silhouette Score: 0.5267
  Davies-Bouldin Score: 0.8001
  Calinski-Harabasz Score: 1008.7729

======================================================================
CLUSTER ANALYSIS
======================================================================

Cluster Distribution:
Cluster
0    299
1    243
2    170
Name: count, dtype: int64

Cluster Statistics (Mean values):
                0         1         2         3         4         5         6         7         8         9         10        11        12        13
Cluster
0        0.431171 -0.856671 -0.262587 -0.286765  0.099770  0.051563  0.025271 -0.044987  0.031209  0.000346 -0.011123  0.014163 -0.019916  0.010614
1       -4.051101  0.573548  0.155410  0.184741 -0.035066 -0.029819 -0.014715  0.020171 -0.019764 -0.007742  0.010885 -0.007225  0.007081 -0.008172
2        5.032337  0.686898  0.239700  0.240298 -0.125355 -0.048067 -0.023413  0.050292 -0.026641  0.010459  0.004004 -0.014583  0.024906 -0.006987
```

5. CLUSTERING ALGORITHMS

DBSCAN

Density-based clustering

```
Enter eps value (default=0.5): 0.5
Enter min_samples (default=5): 5

Performing DBSCAN Clustering (eps=0.5, min_samples=5)...
✓ DBSCAN clustering completed
  Number of clusters: 0
  Number of noise points: 712

======================================================================
CLUSTER ANALYSIS
======================================================================

Cluster Distribution:
Cluster
-1    712
Name: count, dtype: int64

Cluster Statistics (Mean values):
                  0             1             2             3             4             5   ...             8             9            10            11            12            13
Cluster                                                                                     ...
-1       1.284865e-16 -5.808400e-17 -3.243348e-17  3.485040e-17 -2.814540e-17 -3.134197e-17  ... -2.978267e-17  2.190819e-17 -4.420621e-17 -4.623330e-17 -2.237598e-17  2.572848e-18

[1 rows x 14 columns]
```

Thank you for your listening!