F21DL- DATA MINING AND MACHINE LEARNING

SPORTS EDITION - GROUP 18

REQUIREMENT 2

Data Analysis and Exploration

Data Preprocessing:

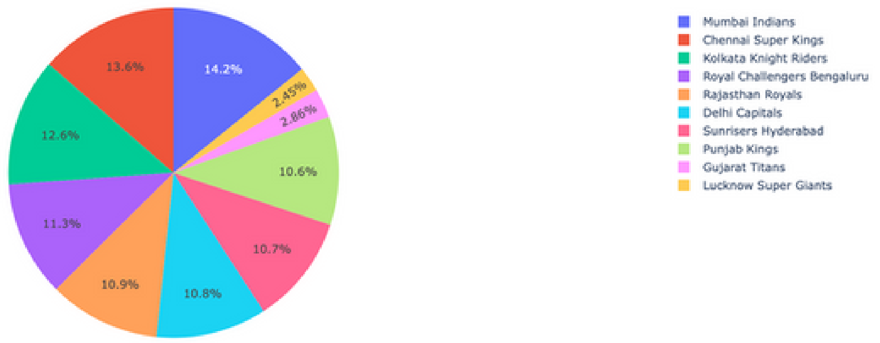
1. Comprehensive cleaning to ensure data quality:
    - Removed redundant teams no longer part of IPL.
    - Standardized team and venue names for consistency.
2. Addressed missing values and ensured feature completeness.

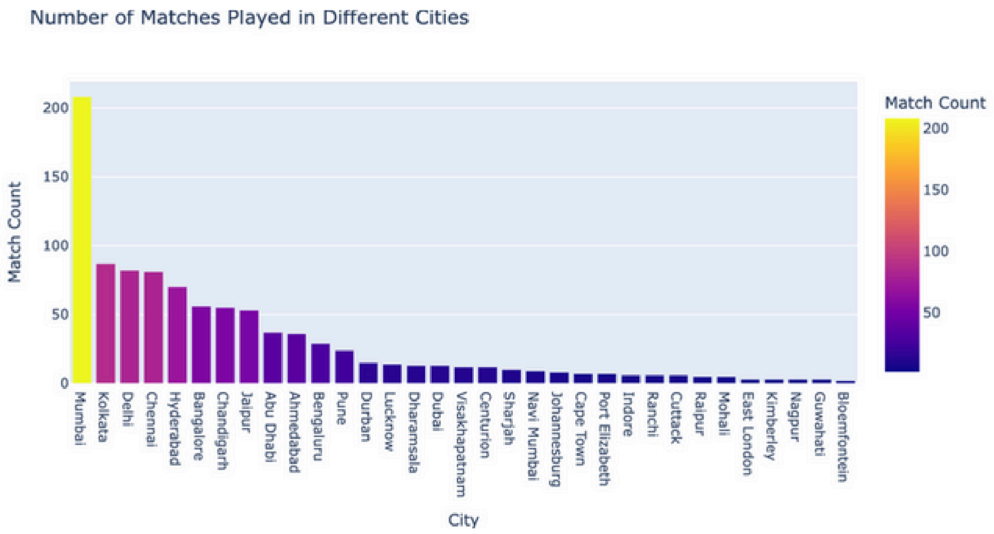
Exploratory Data Analysis (EDA):

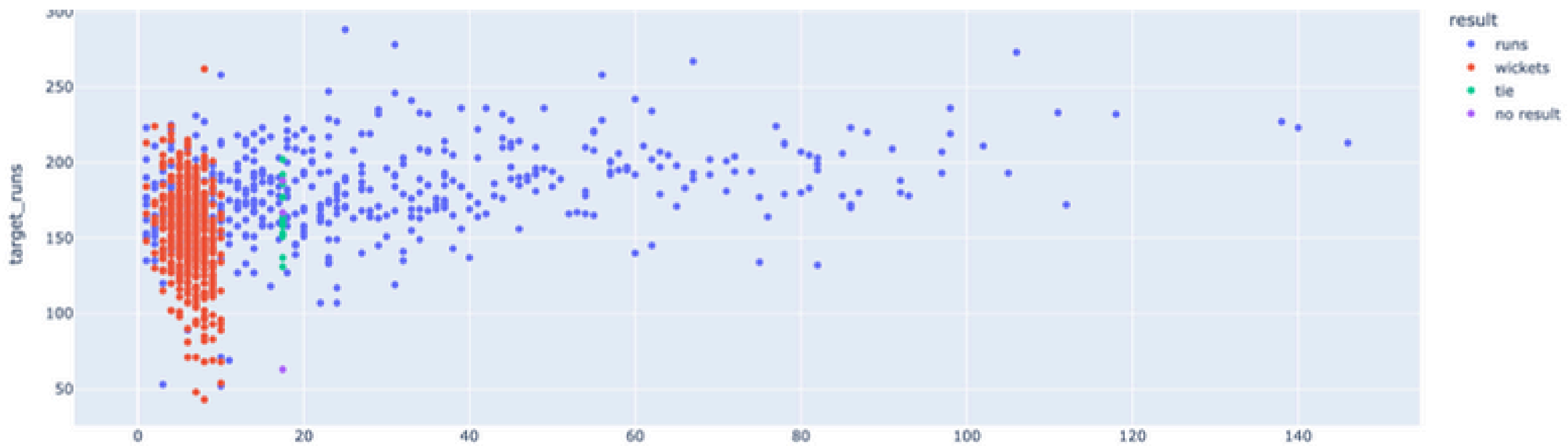
1. Winners Over Time (Pie Chart):
    - The chart highlights the distribution of match wins among IPL teams over time. Teams like Mumbai Indians and Chennai Super Kings show dominance, while newer teams like Gujarat Titans have smaller shares.
2. Result Margin vs. Target Runs (Scatter Plot):
    - This visualization examines the relationship between target runs and result margins. It highlights patterns for matches won by runs or wickets, showcasing trends in high-scoring matches.
3. Number of Matches Played in Different Cities (Bar Chart):
    - The bar chart shows the frequency of matches played across cities, with Mumbai and Kolkata hosting the highest numbers, reflecting their importance in IPL history.

Feature Selection and Evaluation:

1. Selected key features and evaluated their importance for clustering and prediction tasks:
    - Toss-Winner Decisions: Encoded combinations of toss winners and their decisions to analyze strategic outcomes.
    - Venue-Team Interactions: Developed features combining venues with teams to assess performance influences.
    - Team-to-Team Interactions: Encoded team match-ups to explore rivalry dynamics and win probabilities.







Data Preprocessing, Augmentation, EDA

Data Preprocessing

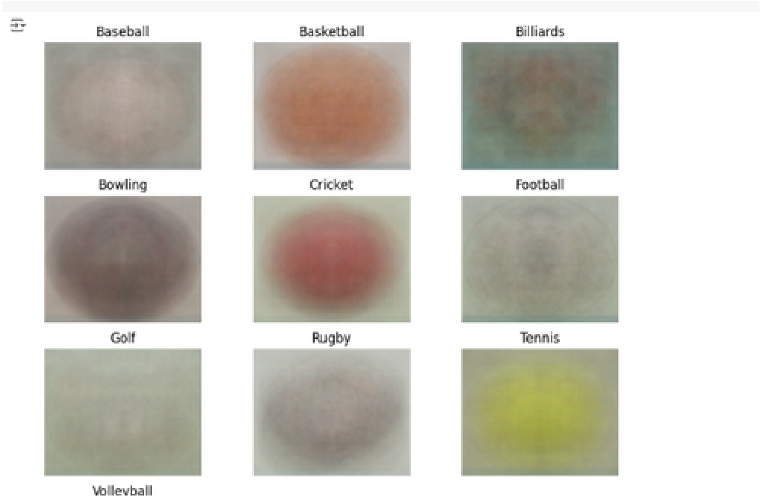
- Metadata: Extracted image paths, class labels, dimensions, and aspect ratios, saving the processed data in a CSV format.
- Image Resizing: Standardized images to 123x100 pixels with a 20px banner for consistency.
- Bounding Boxes: Calculated height, width, and aspect ratios for feature extraction.
- Class Balancing: Capped each class at 400 images to ensure equal representation.

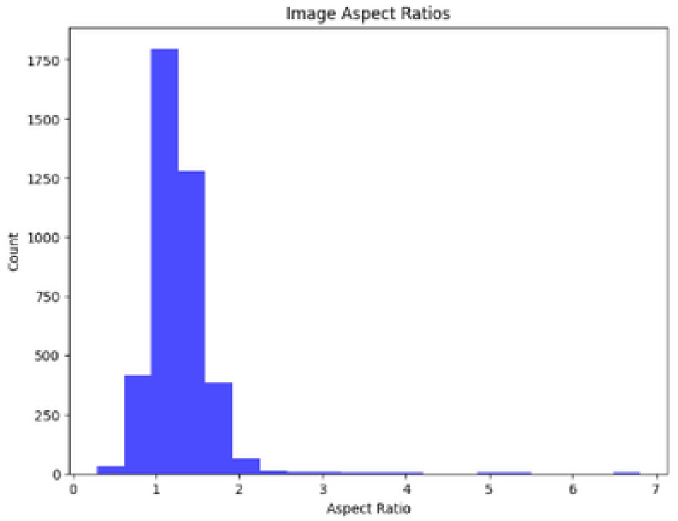
Data Augmentation

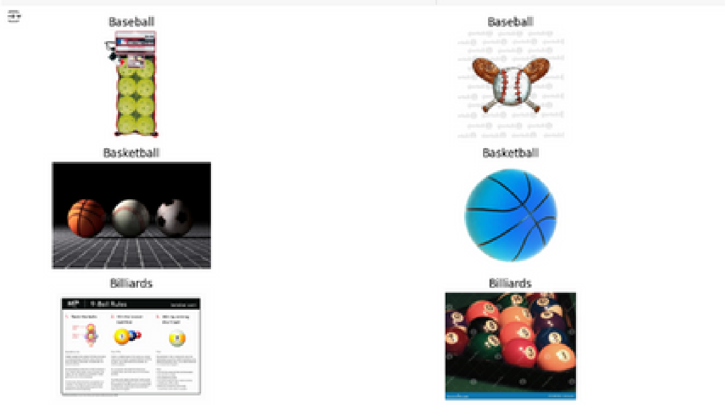
- Underrepresented Classes: Applied horizontal flipping to balance classes with fewer images.
- Outcome: All 10 classes balanced with 400 images each, ensuring uniform representation.

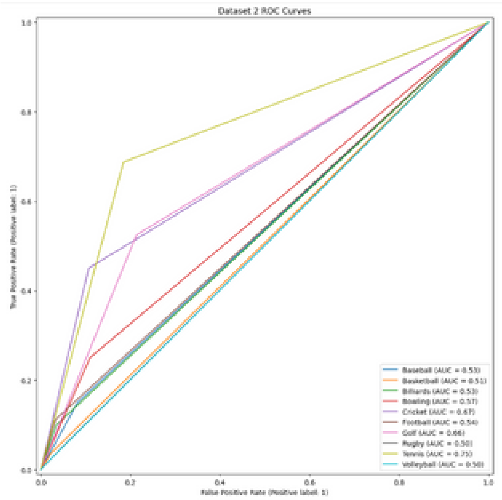
Exploratory Data Analysis (EDA)

- Image Analysis: Scatter plots showed height vs. width and aspect ratio distribution across images.
- Class Visualization:
    - Mean Images: Highlighted dominant features like shape and color for each class.
    - Pairwise Differences: Showed inter-class distinctions (e.g., Rugby vs. Football overlaps).
    - Standard Deviation: Captured intra-class variations.
- Keypoint Detection: Used SIFT to extract top 10 keypoints per class, identifying significant spatial and texture features.
- Feature Correlation: Pearson maps highlighted strong correlations for distinct classes like Tennis and overlaps in similar classes like Football and Rugby.









REQUIREMENT 3

CLUSTERING WAS USED TO UNCOVER HIDDEN PATTERNS AND SEGMENT IPL DATA INTO MEANINGFUL GROUPS BASED ON:

Clustering Algorithm: K-Means

- ○ Optimized number of clusters using the Silhouette Score.
- ○ Results showed six clusters provided the best balance between intra-cluster similarity and inter-cluster separation.

Venue-Based Clustering:

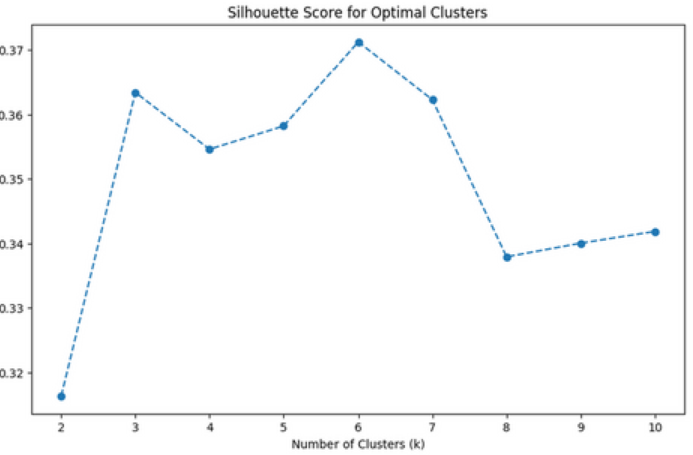
- Clusters formed based on utilization (frequent vs. rare), scoring trends, and result margins.Examples:
  - ○ High-scoring venues like Wankhede Stadium.
  - ○ Low-margin venues with balanced outcomes like M. Chinnaswamy Stadium.

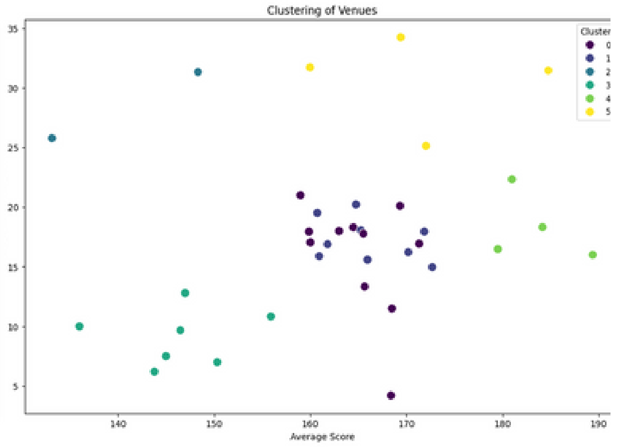
Team Performance Clustering:

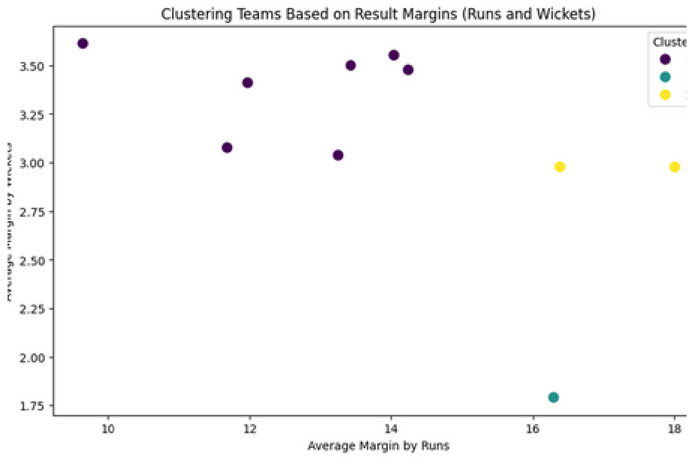
- Teams were grouped into:
  - ○ Dominant Teams: High victory margins, e.g., Mumbai Indians.
  - ○ Consistent Teams: Moderate margins but strategic wins, e.g., Chennai Super Kings.
  - ○ Balanced Teams: Moderate performance across metrics.

Toss-Based Clustering:

- Clusters highlighted:
  - ○ Teams with high toss success rates but fewer matches played (e.g., new franchises).
  - ○ Teams with consistent conversion of toss wins into match wins.
  - ○ Teams with low toss dependency, showing strategic flexibility.


Silhouette Score for Optimal Clusters


Clustering of Venues

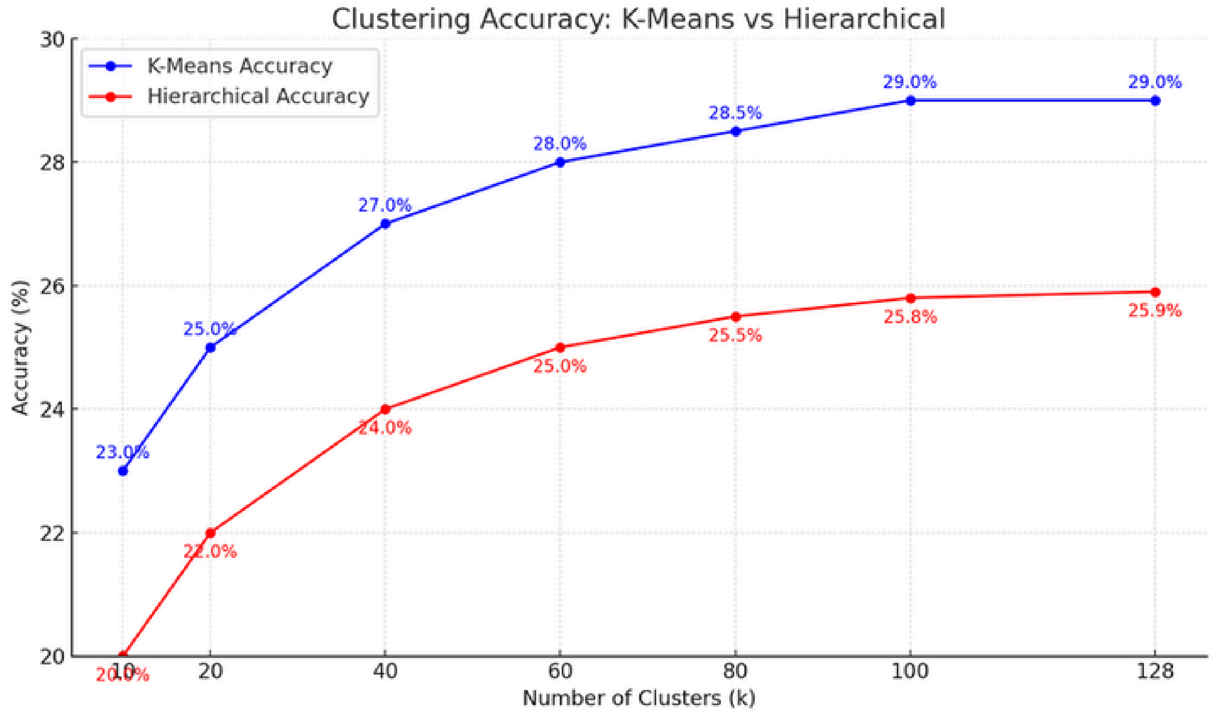

Clustering Teams Based on Result Margins (Runs and Wickets)

Clustering (K-Means & Hierarchical)

- **Purpose:** Group data points based on similarities and evaluate separability.
- **Performance:**
  - ○ PCA-Reduced Data:
    - ▪ Clustering Accuracy: 25.9% (indicating moderate separability).
    - ▪ Inertia decreased with higher clusters, showing compactness within clusters.
- **Insights:**
  - ○ Clusters showed overlap due to similar RGB features, making separation difficult.
  - ○ Visualization revealed intra-cluster similarity and inter-cluster overlap, emphasizing the challenges of unsupervised methods for complex RGB datasets.

Clustering Metrics

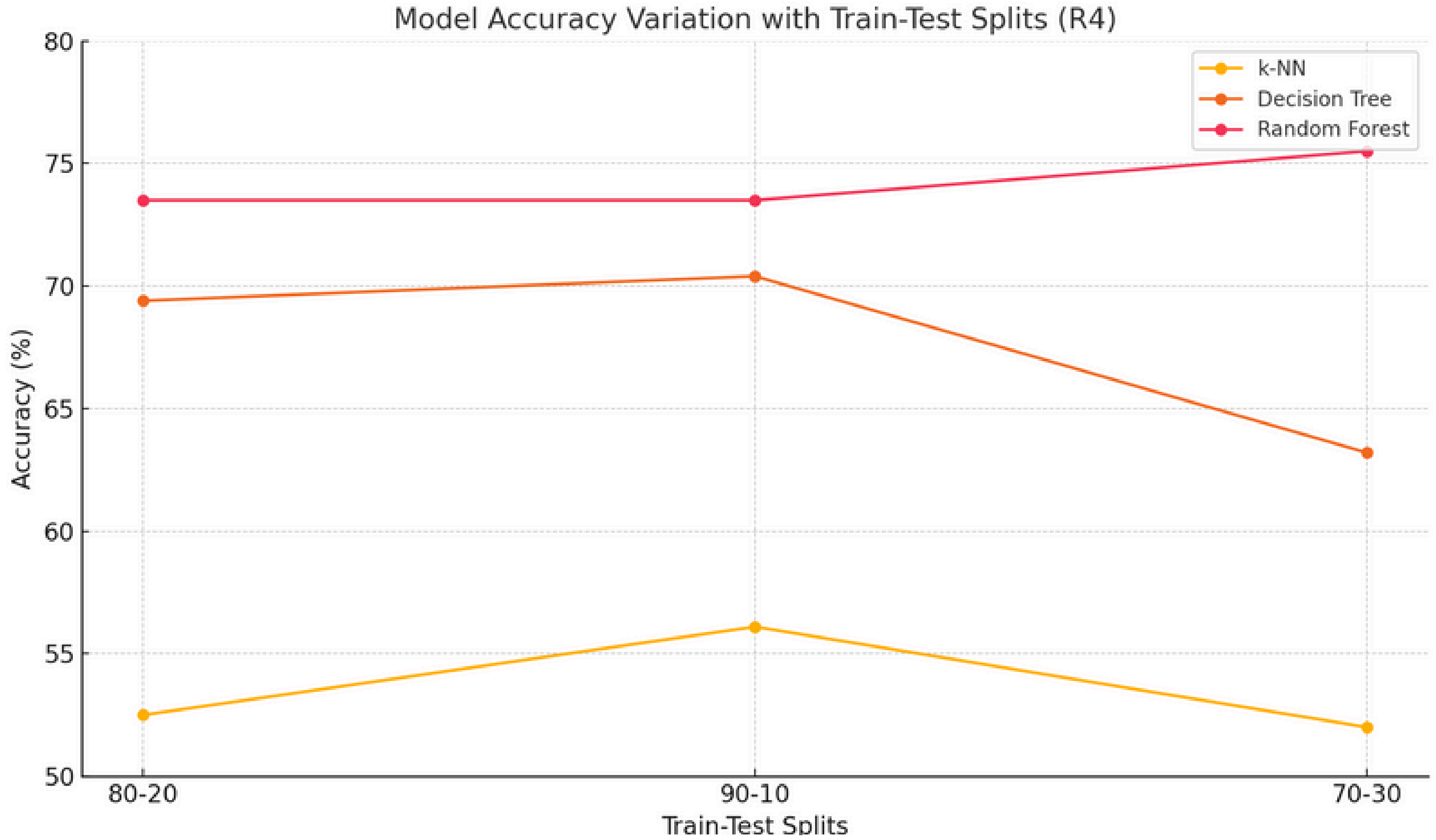
- **K-Means:**
  - ○ Accuracy improved with the number of clusters, peaking at 29% for k=128.
  - ○ Inertia reduction demonstrated compact clusters but diminishing returns beyond a certain k value.
- **Hierarchical Clustering:**
  - ○ Achieved 25.9% accuracy with 40 clusters, reflecting limited separability of RGB data without supervision.


Clustering Accuracy: K-Means vs Hierarchical

REQUIREMENT 4

EVALUATE THE PERFORMANCE OF CLASSIFICATION MODELS FOR PREDICTING IPL MATCH OUTCOMES.

- Models Tested:
  - k-Nearest Neighbors (k-NN)
  - Decision Tree
  - Random Forest
- Train-Test Splits:
  - Evaluated using 80-20, 90-10, and 70-30 splits.
  - Random Forest achieved the best performance across splits, with an accuracy of up to 75.5% (70-30 split).
- Model Performance:
  - k-NN: Accuracy of 52.5% (80-20); limited due to sensitivity to scaling and noise.
  - Decision Tree: Accuracy ranged from 63.2% (70-30) to 70.4% (90-10); prone to overfitting on smaller test sets.
  - Random Forest: Best model with accuracy between 73.5% (80-20) and 75.5% (70-30); robust due to ensemble learning.
- Insights:
  - Larger training sets improve performance (e.g., 90-10 split boosts Decision Tree to 70.4%).
  - Smaller test sets (90-10) may inflate accuracy but reduce generalizability.



Logistic Regression

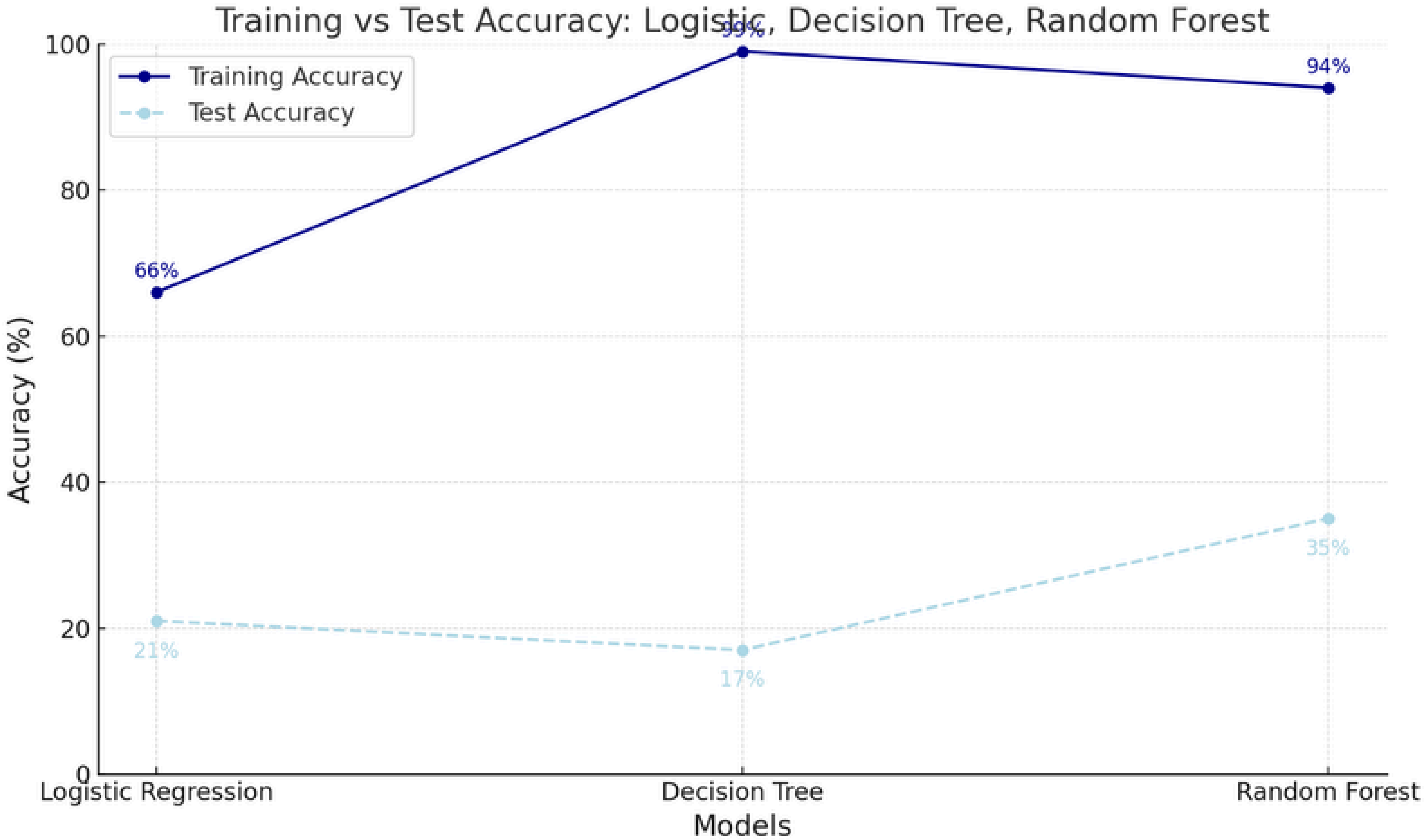
- Purpose: Baseline model for comparison.
- Performance: Training: 66%, Test: 21%, Cross-Validation: 20.3%.
- Insights: Poor generalization and spatial feature handling; highlight the need for MLPs and CNNs.

Decision Trees

- Purpose: Analyze depth and constraints for tuning.
- Performance:
  - Base Model: Training: 99%, Test: 17%, Cross-Validation: 22%.
  - Reduced Training Sets: The test improved to 62% (30% set) and 75% (60% set).
  - Reduced Depth: Training: 85%, Test: 23%.
  - Increased Constraints: Training: 44%, Test: 29%.
- Insights: Reduced overfitting but limited generalization for RGB datasets.

Random Forest

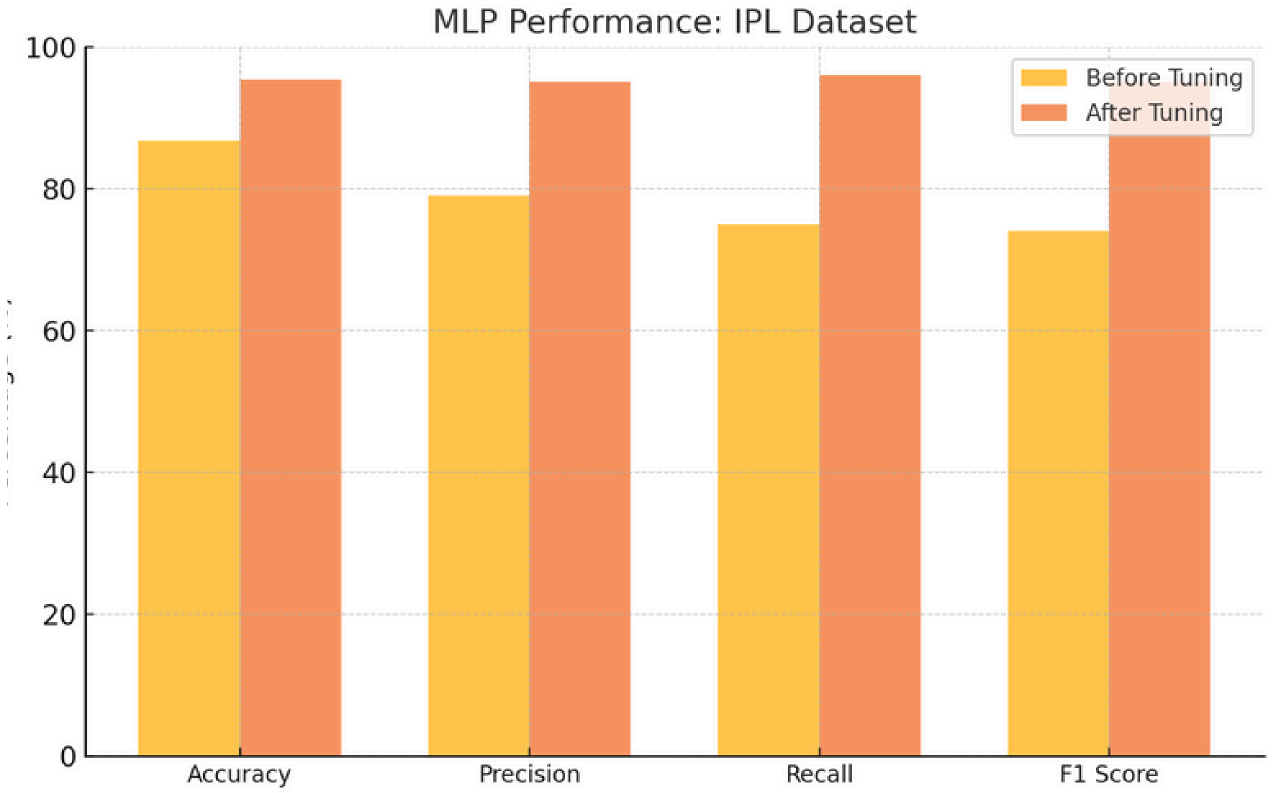
- Purpose: Ensemble methods to improve Decision Trees.
- Performance: Training: 94%, Test: 35%, ROC AUC: 0.70–0.82.
- Insights: Learned feature patterns well but struggled with spatial dependencies; better than Logistic Regression and Decision Trees but inferior to CNNs.



REQUIREMENT 5

USE NEURAL NETWORKS TO IMPROVE PREDICTIVE ACCURACY FOR IPL MATCH OUTCOMES.

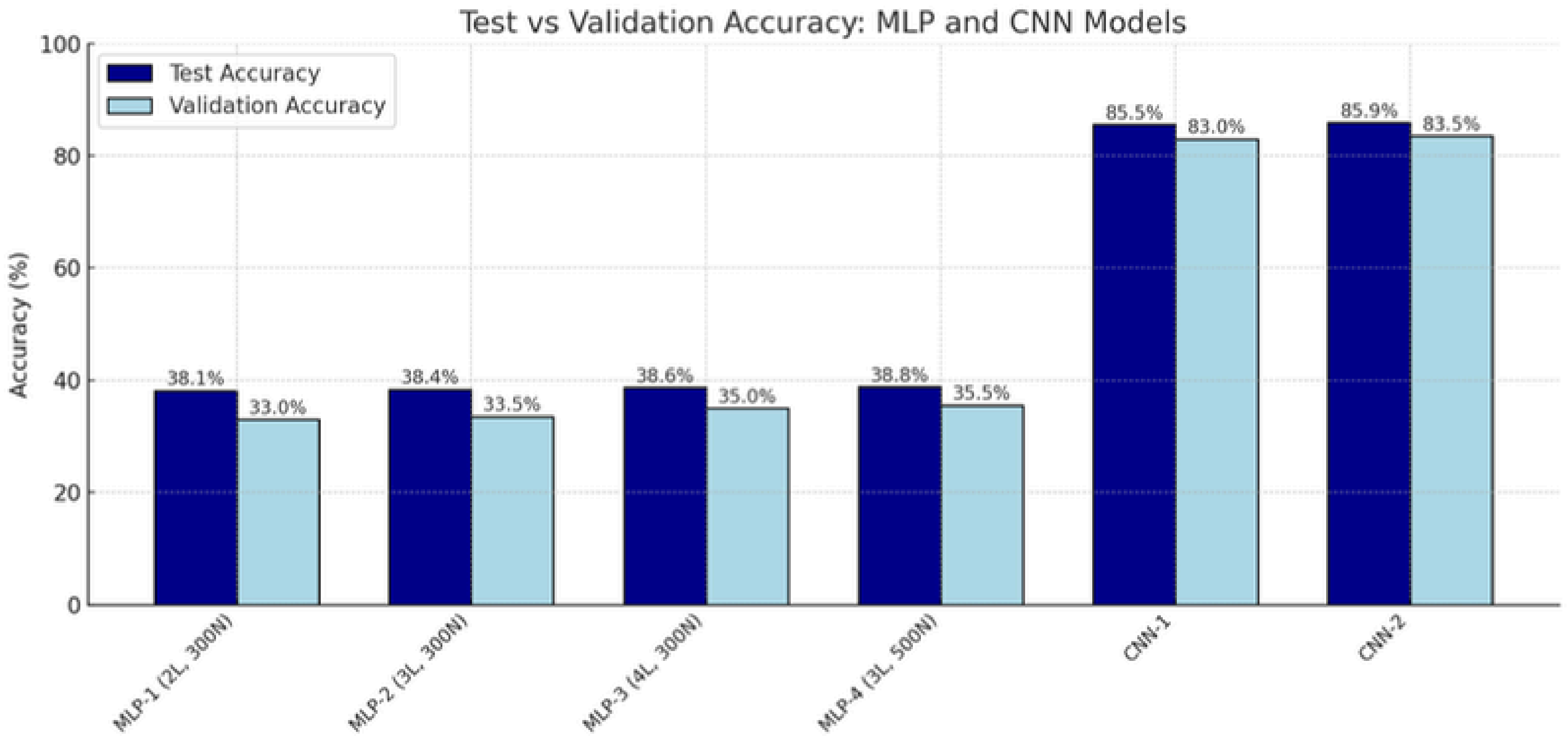
- **Model Tested: Multi-Layer Perceptron (MLP)**
  - Configured with two hidden layers (50 and 50 neurons) and Logistic function.
  - Optimized using the Adam optimizer.
- **Performance Before Tuning:**
  - Achieved an accuracy of 86.7% on an 80-20 train-test split.
  - Precision: 79%, Recall: 75%, F1 Score: 74%.
  - Demonstrated strong ability to capture non-linear patterns but showed slight misclassifications.
- **Hyperparameter Tuning:**
  - Adjusted to: Hidden layers: (50, 30, 30), Activation function: Logistic, Learning rate: 0.005, Alpha (regularization rate): 0.0001, Maximum iterations: 1000
  - Achieved a significant boost in accuracy to 95.4%, with an F1 score of 95%.
  - However, this high performance raised concerns about potential overfitting.
- **Strengths:**
  - MLP effectively modeled complex relationships in the IPL dataset.
  - Clear improvement in frequently occurring classes, as reflected in the confusion matrix.
- **Insights:**
  - Hyperparameter tuning plays a critical role in maximizing neural network performance.
  - MLP offers a robust alternative to traditional machine learning models for structured data like IPL.



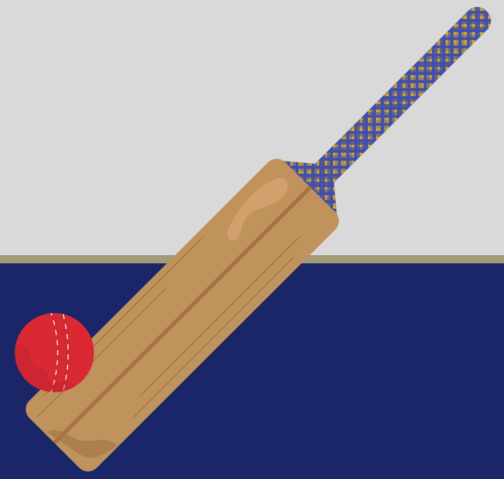
MLP and CNN Architectures and Results

- **Multi-Layer Perceptron (MLP)**
  - **Training Performance:** Model accuracy improved with increased neurons and layers. Best training accuracy: 77.3% (Model 4, 3 layers with 500 neurons).
  - **Validation and Test Performance:** Validation accuracy stabilized between 33–35.5%, indicating limited generalization capacity due to RGB feature dependencies. Test accuracy remained around 38% for all MLP models, suggesting that adding neurons and layers had diminishing returns.
  - **Key Observations**: Increasing layers and neurons improved training accuracy but had minimal impact on validation and test performance. The MLP models struggled with RGB-specific spatial dependencies, which CNNs better handle.

- **Convolutional Neural Networks (CNNs)**
  - CNN-1: Two convolutional layers with ReLU activation and max-pooling, followed by fully connected layers with softmax output.
  - CNN-2: An optimized CNN with similar architecture but adjusted parameters for better feature extraction.
  - **Training Performance:** Learning Features: CNNs effectively learn spatial and color-based patterns.
  - **Validation/Test Performance:** Validation accuracy stabilized at 83%, and test accuracy reached 85.5%, demonstrating strong generalization and ability to handle complex features.
  - **Key Observations:** CNNs outperformed MLPs due to superior spatial pattern recognition, effectively managing challenges in distinguishing similar classes in a multi-class dataset.



CONCLUSION



Summary of Findings:

  - Random Forest was the best-performing traditional model for the IPL dataset, achieving 75.5% accuracy on the 70-30 split.
  - MLP significantly outperformed traditional models, reaching 95.4% accuracy after hyperparameter tuning.
  - Clustering provided valuable insights into venue characteristics, team performance, and toss dynamics, enriching feature engineering.
- Strengths:
  - Robust preprocessing and feature engineering improved model performance.
  - Neural networks demonstrated their ability to model non-linear relationships and outperform traditional methods.
- Challenges:
  - k-NN struggled due to sensitivity to feature scaling and noise.
  - Overfitting in the MLP model post-tuning requires careful mitigation strategies.

This project showcased how data analysis and machine learning can uncover important IPL insights and improve prediction accuracy. By cleaning the data, analyzing toss decisions, and studying venue dynamics, we identified key factors affecting match outcomes. Clustering revealed strategic patterns, while models like Random Forest and optimized neural networks performed well, with the latter achieving 95.4% accuracy. These results highlight how data-driven methods can enhance IPL strategies and support real-time decision-making in sports analytics.

Summary of Findings:

  - Clustering: Provided key insights into class separability and overlaps for RGB patterns, achieving ~29% accuracy.
  - Random Forest: Achieved 35% test accuracy and ROC AUC scores of 0.70–0.82, excelling in feature-based learning but struggling with spatial dependencies.
  - MLP: Improved training accuracy with additional layers and neurons, reaching a test accuracy of 38%. The performance highlighted its limitations in modeling spatial and hierarchical dependencies.
  - CNN: Outperformed all other models with 85% test accuracy and ~83% validation accuracy, leveraging its ability to capture spatial and hierarchical features effectively.
- Strengths:
  - Advanced models like CNNs and MLPs effectively captured complex patterns and relationships in the data, achieving superior performance.
- Challenges:
  - Traditional models like Logistic Regression and Decision Trees struggled with high-dimensional data and spatial dependencies.
  - Deep learning models required significant computational resources and were prone to overfitting without careful regularization.

the project showcased how combining robust preprocessing, exploratory techniques, and advanced modeling can address challenges in high-dimensional datasets, paving the way for effective data-driven strategies and informed decision-making.