# Project Final Report

**Project Title:** Organism Prediction based on Codon Usage

## 1. Abstract

The Codon Usage Dataset provides insights into the frequency of codons across various organisms. This project aims to predict the organism's kingdom based on codon usage patterns using classification techniques. Models such as K-Nearest Neighbors (KNN), Logistic Regression, Random Forest, and Support Vector Machines (SVM) were implemented and compared. Extensive preprocessing, exploratory data analysis, feature selection, and hyperparameter tuning were performed to enhance model performance. KNN emerged as the top-performing model with a 97% accuracy rate.

## 2. Dataset Information:

| Dataset Characteristics | Multivariate |
|---|---|
| Subject Area | Biology |
| Associated Tasks | Classification, Clustering |
| Feature Type | Real |
| Number of Instances | 13028 |
| Number of Features | 69 |
| No of missing values | 69 |

**Dataset:** https://archive.ics.uci.edu/dataset/577/codon+usage

# 3. Dataset Overview

The Codon Usage Dataset contains data on the relative usage of codons across multiple organisms categorized by their DNA type and kingdom. The dataset includes 69 features capturing codon frequency distributions and additional metadata such as species names, DNA type, and species ID.

**Rationale**

Codon usage plays a crucial role in understanding molecular biology, species evolution, and synthetic biology. Predicting the organism's kingdom based on codon usage can assist in biological data classification and genomics research.

**Target Variable**

The target variable for this project is the **Kingdom** of the organism. This is a **classification problem** where each observation is categorized into one of the predefined kingdom classes.

The dataset consists of 13,028 samples and 69 features. Key features include:

- **Kingdom:** Target variable with classes such as "primate," "virus," "mammal," etc.
- **Codon Frequencies (UUU, UUC, UUA, ...):** Frequency of codons within a genome.
- **Species Metadata (Species ID, DNA Type):** Identification and classification of species.

**Summary Statistics:**

- Numerical features: 62 columns
- Categorical features: 4 columns
- No missing values or duplicate rows.

# 4. Data Cleaning and Preprocessing

**Preprocessing Steps**

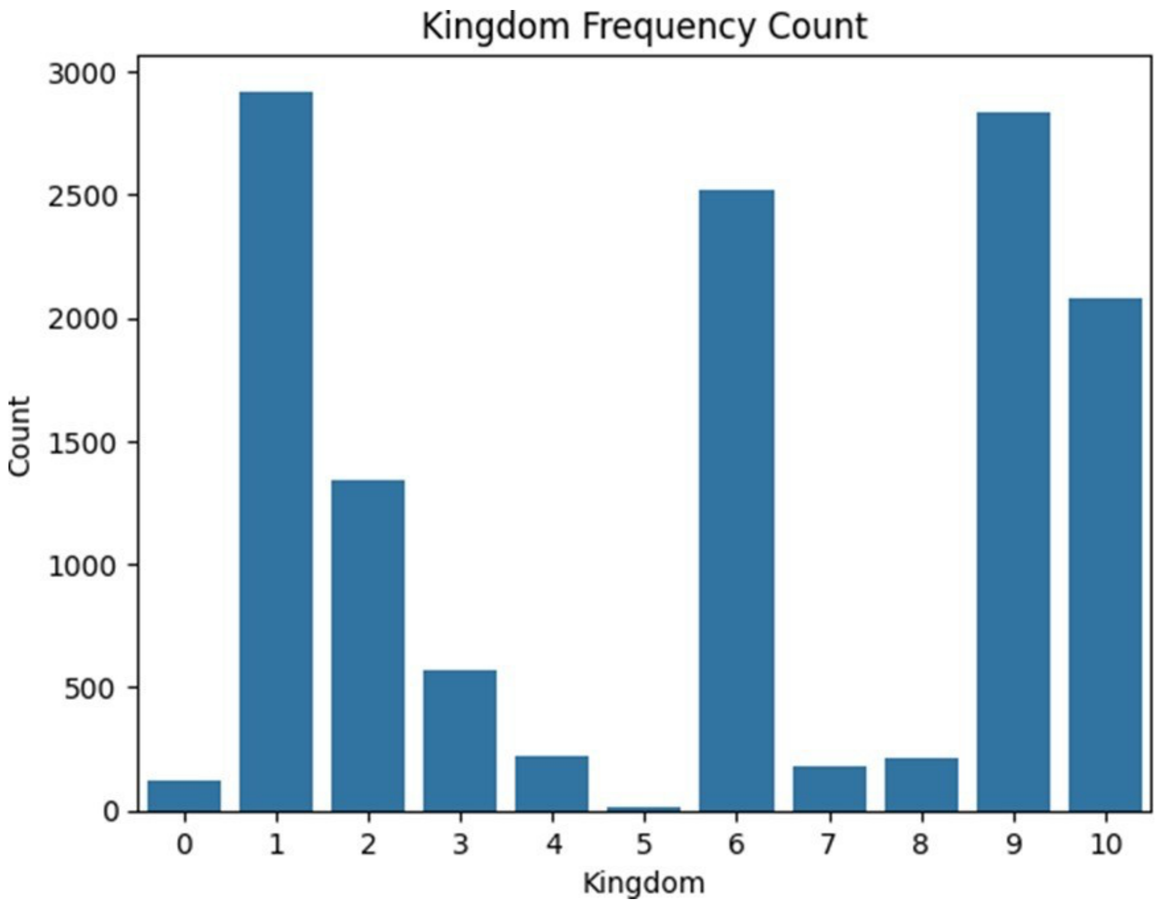
1. **Outlier Detection and Removal:**
   - Used the IQR method, reducing the dataset to 3,744 rows while preserving diversity.

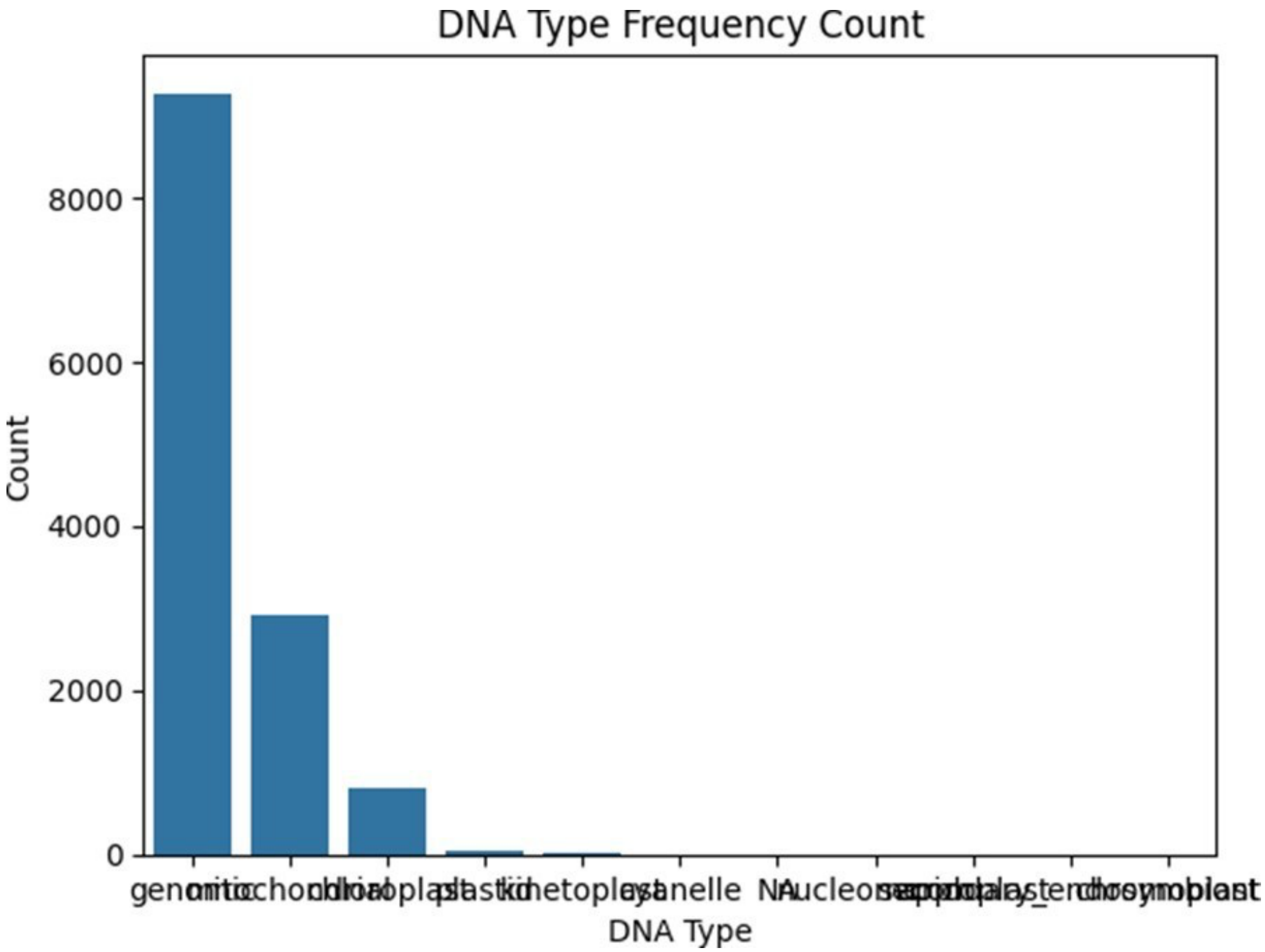
2. **Encoding:**
   - Converted categorical variables (e.g., Kingdom) to numeric codes for modeling.

3. **Normalization:**
   - Scaled numerical features for uniformity across codon frequency data.

**Exploratory Data Analysis (EDA)**

- **Frequency Analysis:**
  - Dominance observed in bacteria, vertebrate, primate, and virus kingdoms.
  - Archaea showed the highest variability in codon counts.

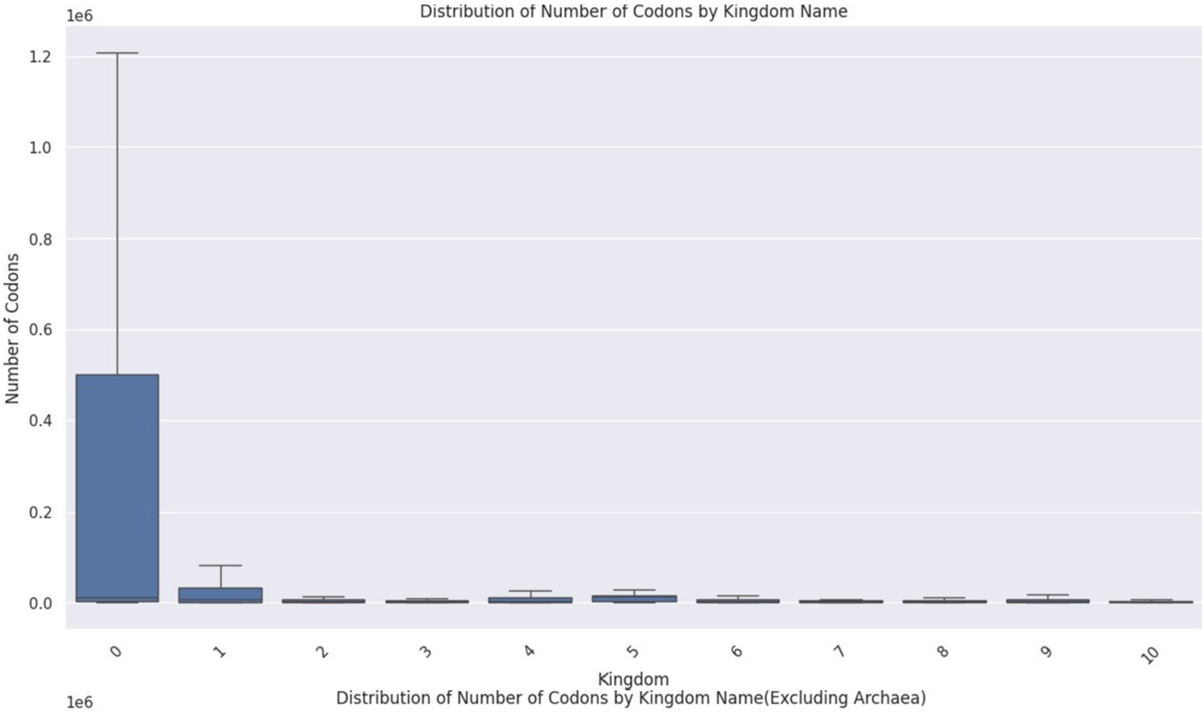




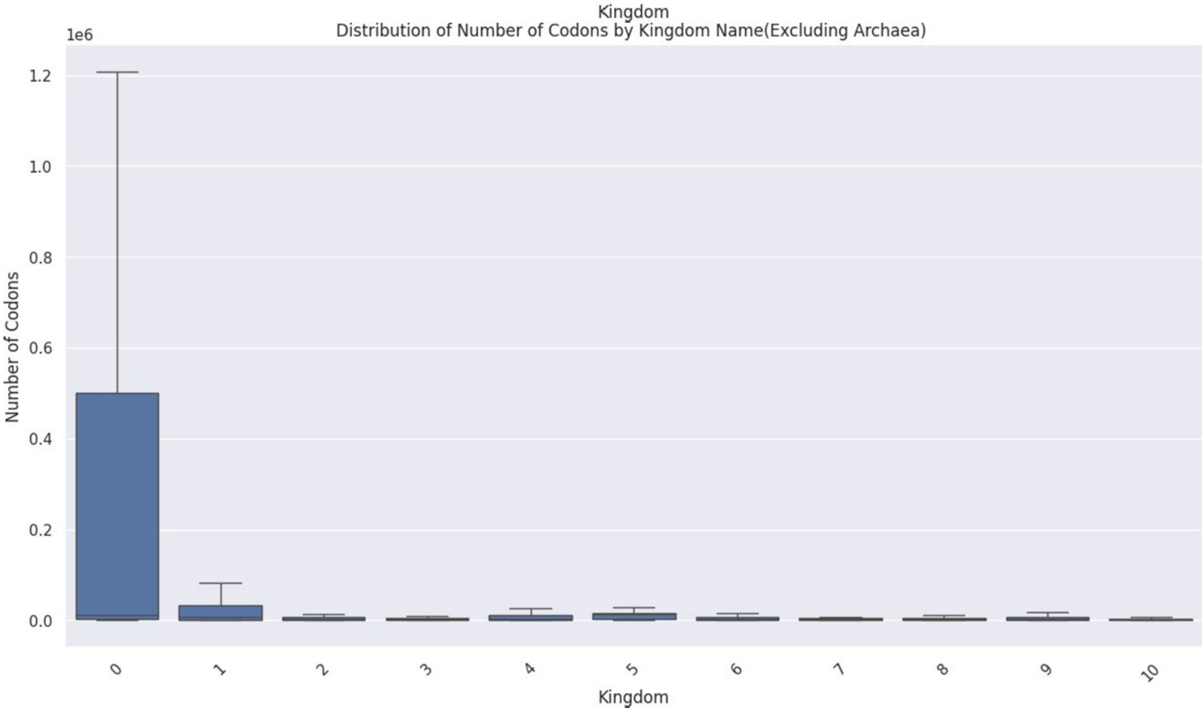
0: archaea
1: bacteria
2: bacteriophage
3: plasmid
4: plant
5: invertebrate
6: vertebrate
7: mammal
8: rodent
9: primate
10: virus

- **Correlation Heatmap:**
  - Codon frequencies exhibited interdependencies, informing feature selection.

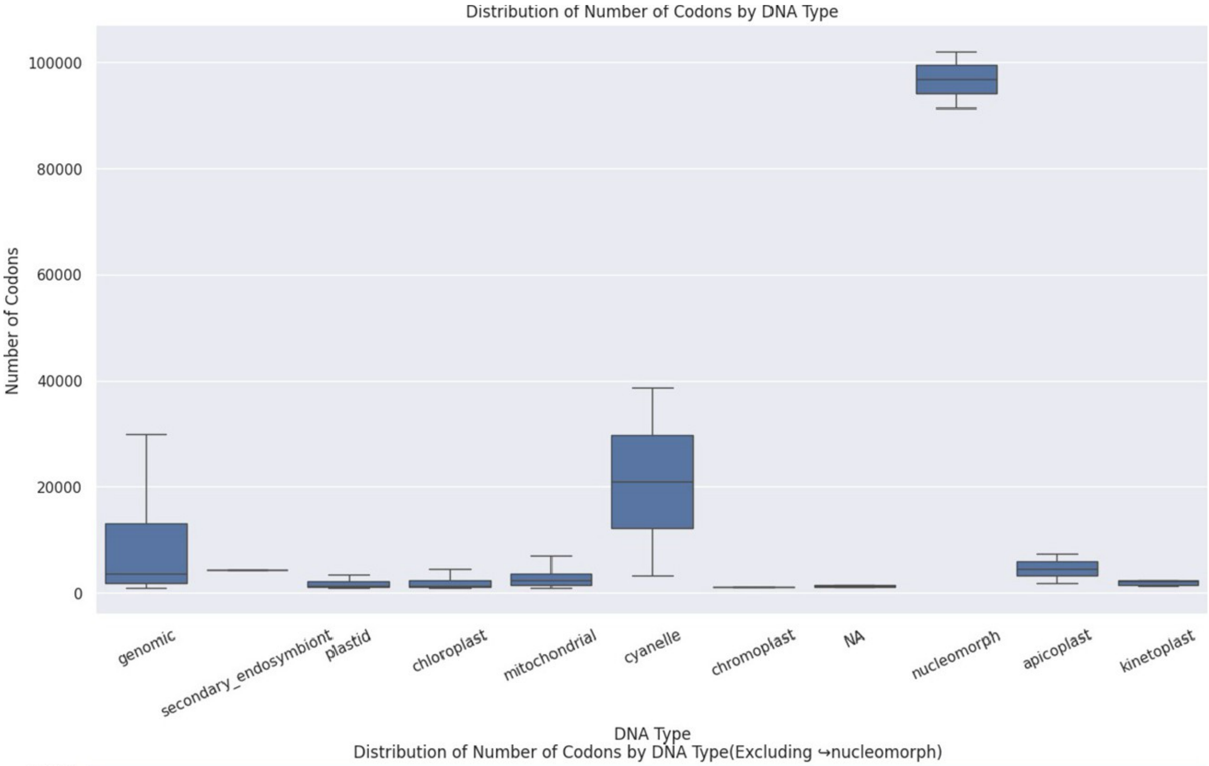
- **Boxplots:**
  - Highlighted the distribution of codons across kingdoms, emphasizing variability in specific taxa.

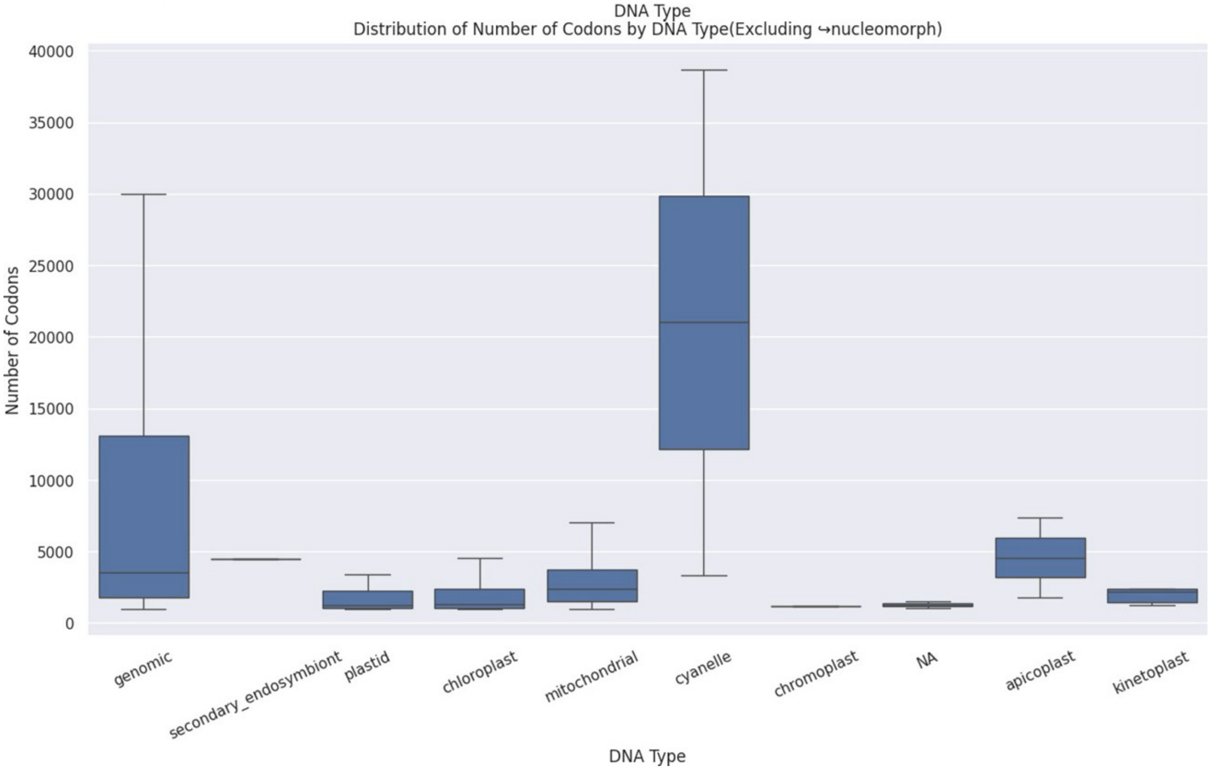


Distribution of Number of Codons by Kingdom Name



Distribution of Number of Codons by Kingdom Name(Excluding Archaea)


Distribution of Number of Codons by DNA Type
Number of Codons
100000
80000
60000
40000
20000
0
genomic
secondary_endosymbiont
plastid
chloroplast
mitochondrial
cyanelle
chromoplast
NA
nucleomorph
apicoplast
kinetoplast
DNA Type


Distribution of Number of Codons by DNA Type(Excluding ↪nucleomorph)
Number of Codons
40000
35000
30000
25000
20000
15000
10000
5000
0
genomic
secondary_endosymbiont
plastid
chloroplast
mitochondrial
cyanelle
chromoplast
NA
apicoplast
kinetoplast
DNA Type

# 5. Model Building

### Selected Models

1. **K-Nearest Neighbors (KNN):**
   - Tuned with n_neighbors=6.
   - Achieved 97% accuracy.
2. **Logistic Regression (One-vs-Rest):**
   - Multi-class approach.
   - Accuracy: 76%.
3. **Random Forest:**
   - Decision-tree-based ensemble method.
   - Achieved 97% accuracy.
4. **Logistic Regression (One-vs-Rest):**
   - Linear Kernel: 62% accuracy
   - Nonlinear Kernel: 91% accuracy

### Feature Selection and Hyperparameter Tuning

### Feature Selection

- **Technique Used:** Correlation Analysis.
- **Impact:** Reduced multicollinearity and improved model interpretability.

### Hyperparameter Tuning

- Applied Grid Search to optimize parameters for each model.
- Visualizations compared model performances before and after tuning.

### Example (KNN):

- Pre-tuning accuracy: 90%
- Post-tuning accuracy: 97%

### Bi-Directional Elimination

Performed wrapper-based feature selection using bi-directional elimination on the top-performing model (KNN). This reduced the feature set further, marginally improving accuracy and reducing training time.

### Additional Models

1. **XGBoost**
   - Gradient boosting implementation for robust classification.
2. **Extreme Learning Machine (ELM)**
   - Single-layer feedforward neural network for faster learning.
3. **Ensemble Model**
   - Combines the top 3 models (KNN, Random Forest, and Logistic Regression) for improved accuracy.

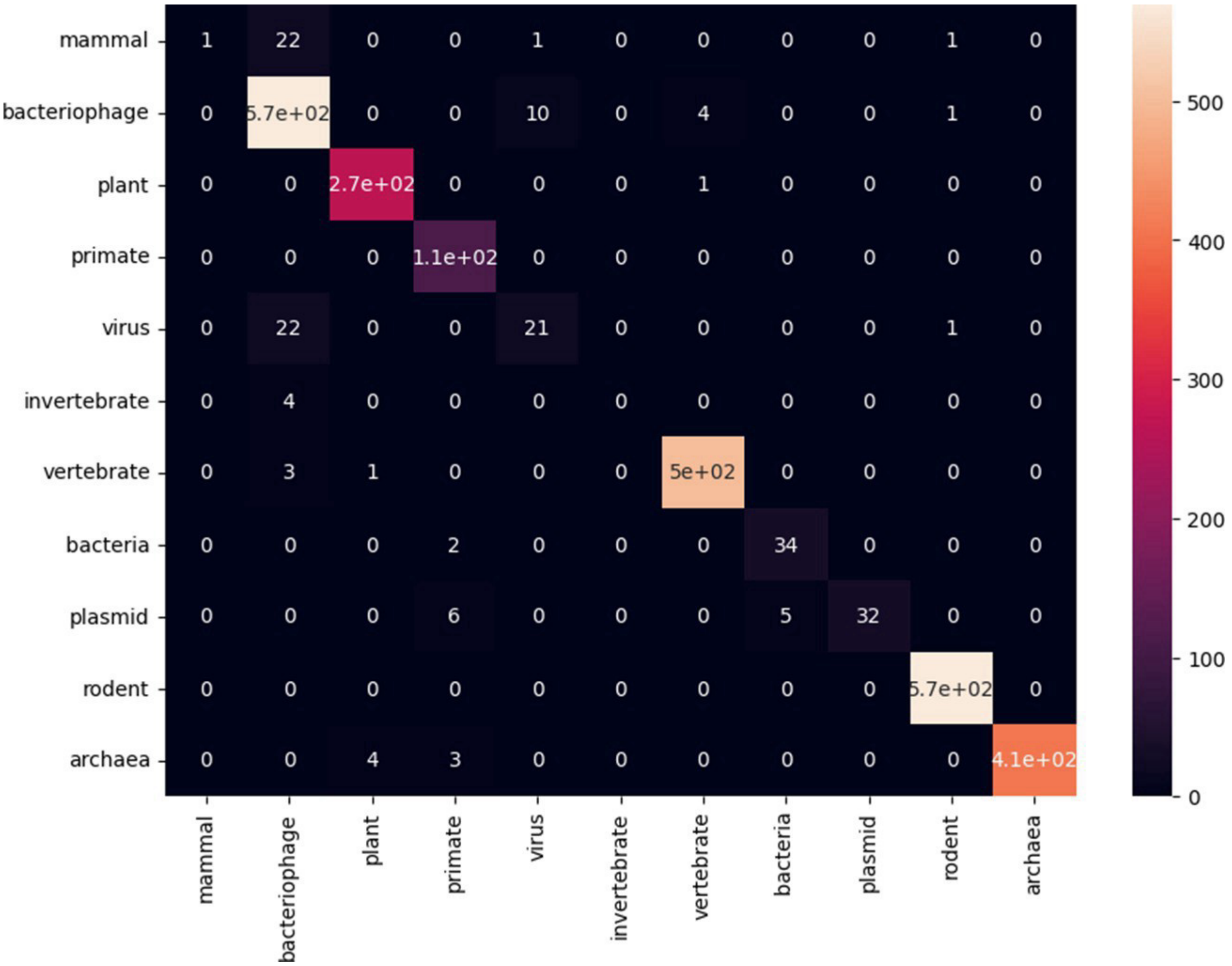
# 6. Results and Evaluation

## Performance Metrics

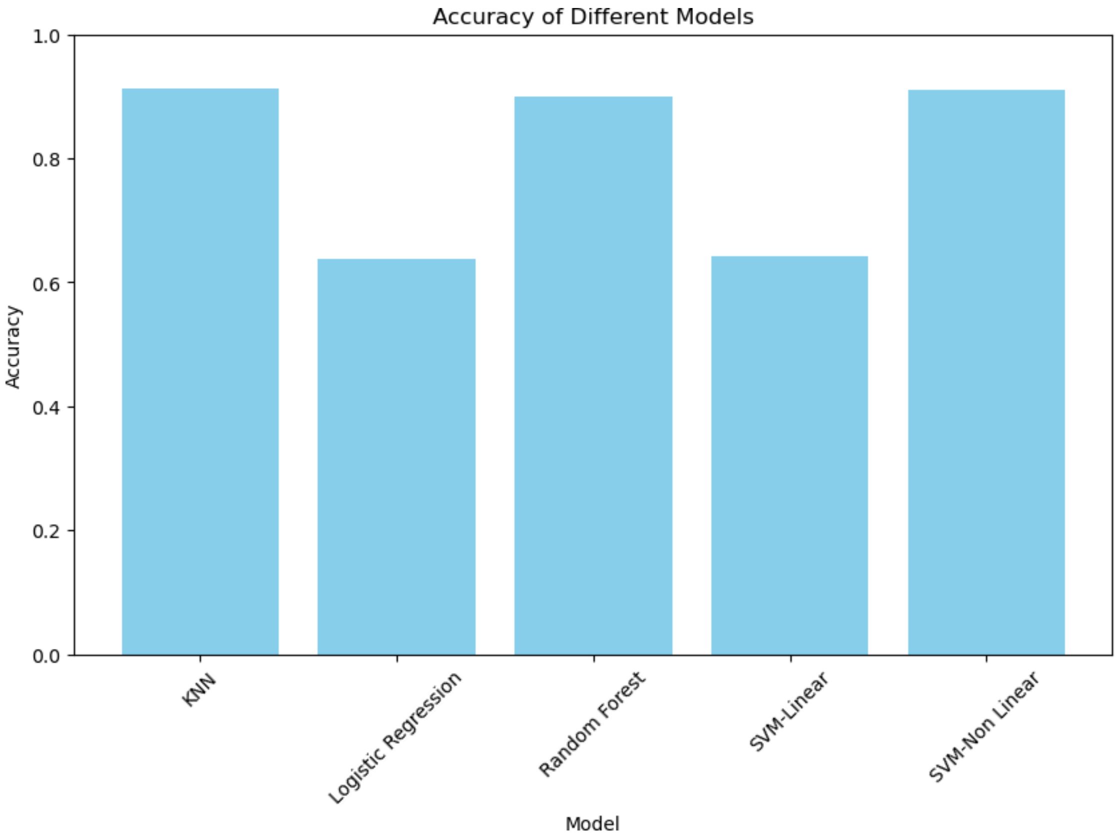
| Model | Accuracy | Precision | Recall | F1-Score |
|---|---|---|---|---|
| KNN | 97% | 96% | 97% | 97% |
| Logistic Regression | 76% | 75% | 76% | 75% |
| Random Forest | 90% | 90% | 90% | 89% |
| SVM (Linear) | 62% | 64% | 58% | 64% |
| SVM (Non-Linear) | 91% | 91% | 91% | 91% |

## Key Visualizations

- **KNN Confusion Matrix:** Demonstrated strong predictive performance across all classes.
- **Classification Reports:** Provided granular insights into precision, recall, and F1-scores.



| | mammal | bacteriophage | plant | primate | virus | invertebrate | vertebrate | bacteria | plasmid | rodent | archaea |
|---|---|---|---|---|---|---|---|---|---|---|---|
| mammal | 1 | 22 | 0 | 0 | 1 | 0 | 0 | 0 | 0 | 1 | 0 |
| bacteriophage | 0 | 5.7e+02 | 0 | 0 | 10 | 0 | 4 | 0 | 0 | 1 | 0 |
| plant | 0 | 0 | 2.7e+02 | 0 | 0 | 0 | 1 | 0 | 0 | 0 | 0 |
| primate | 0 | 0 | 0 | 1.1e+02 | 0 | 0 | 0 | 0 | 0 | 0 | 0 |
| virus | 0 | 22 | 0 | 0 | 21 | 0 | 0 | 0 | 0 | 1 | 0 |
| invertebrate | 0 | 4 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 |
| vertebrate | 0 | 3 | 1 | 0 | 0 | 0 | 5e+02 | 0 | 0 | 0 | 0 |
| bacteria | 0 | 0 | 0 | 2 | 0 | 0 | 0 | 34 | 0 | 0 | 0 |
| plasmid | 0 | 0 | 0 | 6 | 0 | 0 | 0 | 5 | 32 | 0 | 0 |
| rodent | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 5.7e+02 | 0 |
| archaea | 0 | 0 | 4 | 3 | 0 | 0 | 0 | 0 | 0 | 0 | 4.1e+02 |



## 7. Additional Insights

**Impact of Preprocessing**

- Outlier removal significantly improved model accuracy and reduced computational complexity.
- Encoding and normalization facilitated effective learning in models requiring numerical inputs.

**Future Work**

- Extend feature engineering with domain expert insights to improve classification performance.
- Explore ensemble approaches combining top models.

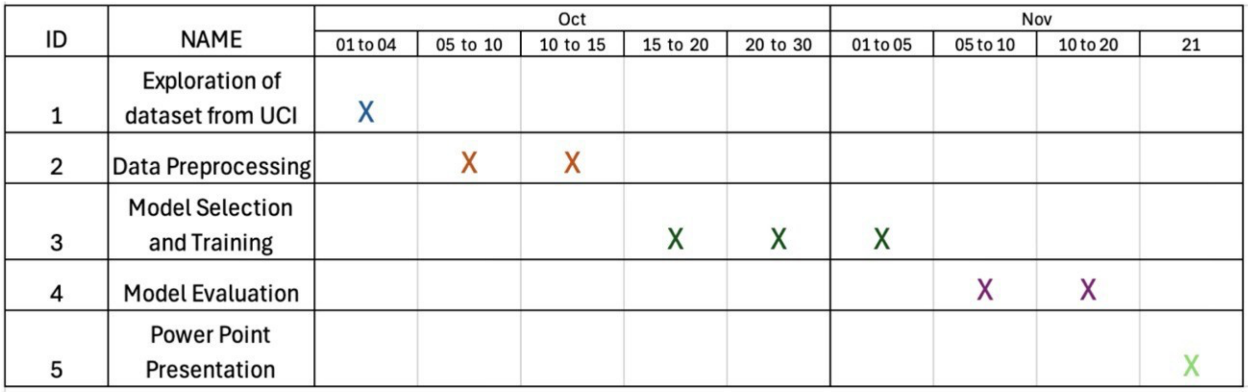
# 8. Conclusion

This study demonstrated the effective use of the Codon Usage Dataset for multi-class biological classification. The application of preprocessing techniques, feature selection, and hyperparameter tuning contributed significantly to enhancing model performance. Among the models tested, the K-Nearest Neighbors (KNN) algorithm stood out, achieving an impressive 97% accuracy. The results underline the potential of KNN in handling complex, high-dimensional biological datasets. Future efforts could focus on integrating ensemble methods, extending feature engineering through domain expertise, and exploring advanced algorithms like XGBoost and Extreme Learning Machines to further refine prediction accuracy.

# 9. References

1. UCI Machine Learning Repository: Codon Usage Dataset. Available at: [https://archive.ics.uci.edu/dataset/577/codon+usage]
2. Scikit-learn documentation. Comprehensive guide for implementing KNN, Logistic Regression, and hyperparameter tuning.
3. Relevant domain literature on codon usage patterns and their significance in biological classification and genomics research.

**PROGRESS GANTT CHART**



| ID | NAME | Oct | | | | | Nov | | | |
|---|---|---|---|---|---|---|---|---|---|---|
| | | 01 to 04 | 05 to 10 | 10 to 15 | 15 to 20 | 20 to 30 | 01 to 05 | 05 to 10 | 10 to 20 | 21 |
| 1 | Exploration of dataset from UCI | X | | | | | | | | |
| 2 | Data Preprocessing | | X | X | | | | | | |
| 3 | Model Selection and Training | | | | X | X | X | | | |
| 4 | Model Evaluation | | | | | | | X | X | |
| 5 | Power Point Presentation | | | | | | | | | X |