

# Monkeypox Classification Analysis

This document provides a detailed explanation and analysis of the Monkeypox Classification project. The project involved applying machine learning techniques to predict Monkeypox outcomes using real-world clinical and demographic data. Various classification algorithms were used to analyze the dataset and evaluate the predictive capabilities of each model.

## Dataset Overview

The dataset included features such as Age, HIV Infection status, Rectal Pain, Sore Throat, Oral Lesions, and others. The target variable was MPOX\_Result, indicating whether a Monkeypox PCR result was positive (1) or negative (0). Missing values were handled using appropriate imputation techniques.

## Data Cleaning

The data underwent extensive cleaning. Age values were corrected for inconsistencies, missing categorical values were filled with their mode, and numerical columns were imputed with their median values. Irrelevant columns were dropped to ensure the dataset was suitable for analysis.

## Visualizations

Exploratory Data Analysis revealed insights into the dataset:

1. Distribution of MPOX PCR Results showed a higher proportion of positive results.
2. A correlation heatmap highlighted relationships between features and the target variable.
3. Feature importance plots (e.g., SHAP values) demonstrated the significance of variables like Age and HIV Infection.

## Model Evaluation

Several machine learning models were tested, including Logistic Regression, Decision Tree, Naïve

Bayes, SVM, Random Forest, Gradient Boosting, AdaBoost, and XGBoost. The evaluation metrics are as follows:

- Logistic Regression: Accuracy = 68.14%
- Gradient Boosting: Accuracy = 70.06% (Best Performing Model)
- AdaBoost: Accuracy = 69.56%
- XGBoost: Accuracy = 67.86%
- Decision Tree: Accuracy = 58.46%

The Gradient Boosting model showed the highest accuracy and was further fine-tuned using Grid Search.

## **Advanced Techniques**

The following advanced techniques were employed to enhance model performance:

1. Hyperparameter Tuning: Grid Search and Bayesian Optimization were used to find optimal parameters for Gradient Boosting and XGBoost.
2. Cross-Validation: Achieved a mean accuracy of 69.10% with Gradient Boosting across 10 folds.
3. Bootstrap Confidence Intervals: Calculated for XGBoost model to validate accuracy.
4. Ensemble Learning: Stacking and Voting classifiers combined models to improve predictions.

## **Public Health Relevance**

This project has significant implications for public health, such as:

- Early Detection: Identifying high-risk patients based on clinical features.
- Resource Allocation: Assisting healthcare providers in prioritizing patients.
- Epidemiological Insights: Understanding symptoms and risk factors linked to Monkeypox.

## **Future Work**

The following areas are recommended for future exploration:

- Developing a web-based application for real-time Monkeypox risk assessment.

- Applying deep learning models to improve classification accuracy.
- Integrating time-series analysis to monitor Monkeypox outbreaks.

## **Conclusion**

The Monkeypox Classification project successfully demonstrates the application of machine learning techniques to predict outcomes and identify key risk factors. The insights generated are valuable for public health initiatives and clinical decision-making.