



# Predicting AIDS Progression with Data Insights

Rahul Gade

# Table of Content

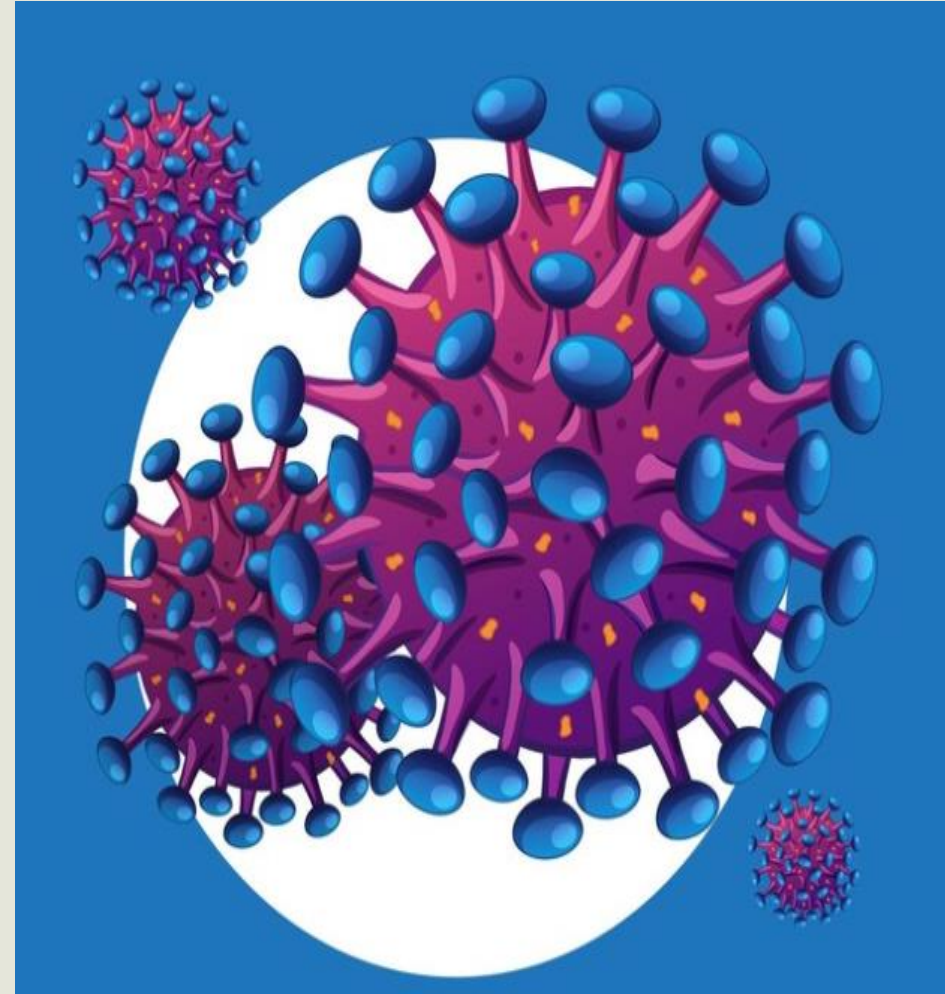
- Project Overview
- Data Cleaning
- Visualizations
- Data Examination
- Machine Learning
- Recommendations



# Project Overview

## Objective

A machine learning model capable of accurately predicting HIV progression to AIDS.



# Dataset

- **Source:** Kaggle – AIDS Clinical Trials Group
- **Task:** to predict whether or not each patient is infected with AIDS at end of the trial
- **Target:** infected – yes/no – a binary classification problem
- **Features:** current treatments, previous treatments, CD4/CD8 cell count at baseline and after 20 weeks (immunity measure), symptomatic, sex, age, weight, race, history of drug use, etc.

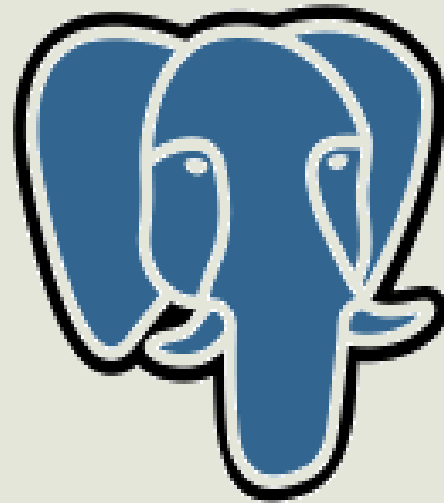
# Data Cleaning

Extract, transform, and load (ETL)



Data cleaning

PostgreSQL



Connection to SQL Database

SQLAlchemy

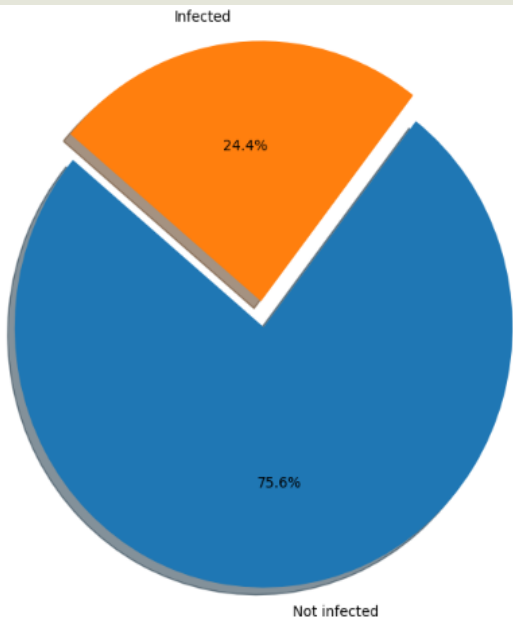




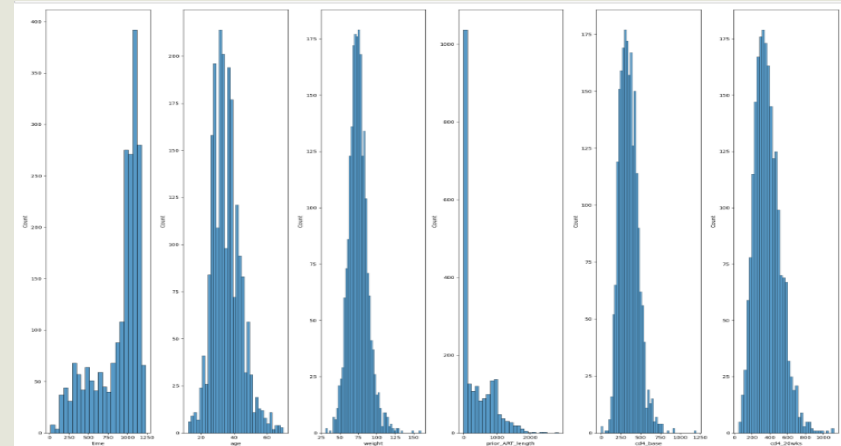
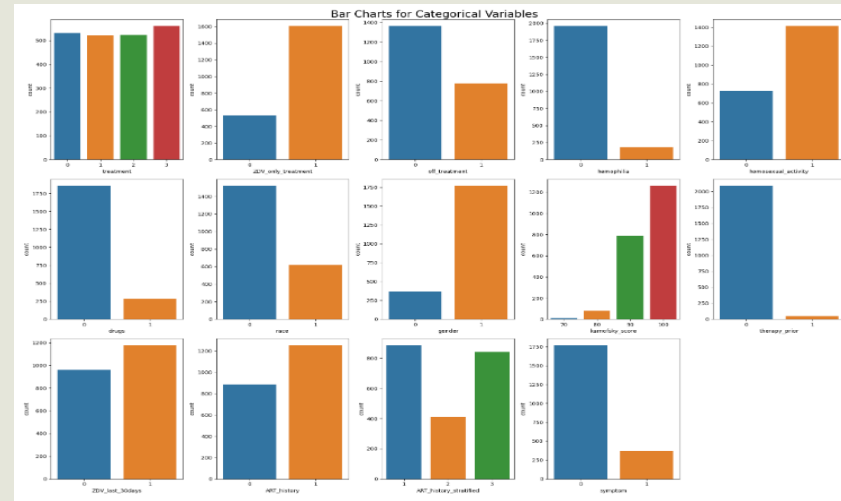
# Visualisations

## Data Insights

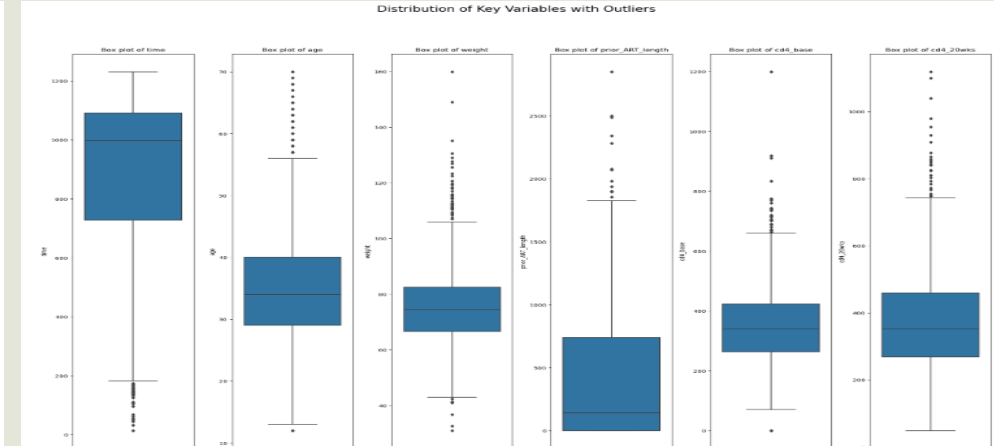
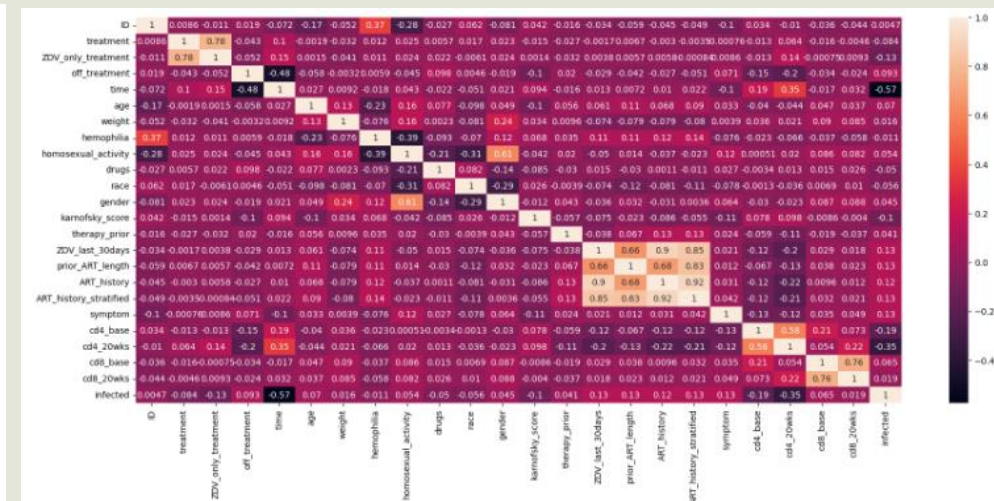
Infected vs Not infected Distribution



Categorical and Numerical Variables



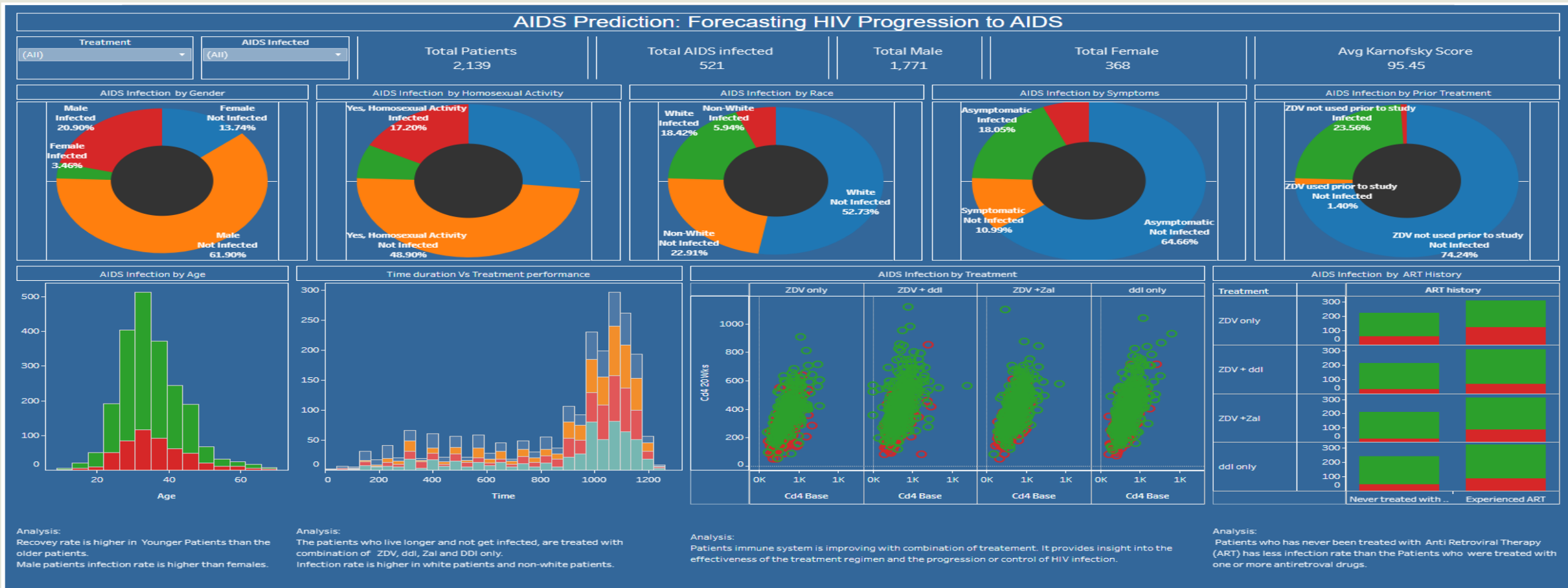
Correlation matrix and Variable Outliers



# Visualisations

## Data Insights

### Variables affecting AIDS infection Rate



# Handling class Imbalance

1. **SMOTE (Synthetic Minority Over-sampling Technique)**
2. **Algorithmic techniques: class weight parameter to “balance”**  
e.g. Logistic Regression, SVM(Support Vector Machine)
3. **Machine Learning Model selection like Random Forest**
4. **Evaluation Metrics. i.e. precision, and recall**



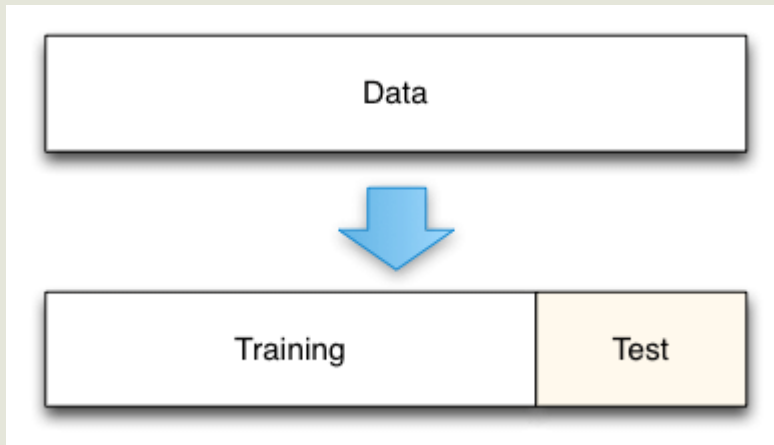
# Machine Learning Models

## KNN, SVM and Logistic Regression

### Methodology

#### Train-Test Split:

Split the dataset into 80% training and 20% testing using 'train\_test\_split' with stratification to maintain the class distribution.



Metric	Logistic Regression (Train)	Logistic Regression (Test)	K-Nearest Neighbors (Train)	K-Nearest Neighbors (Test)	Support Vector Machine (Train)	Support Vector Machine (Test)
Accuracy	0.8597	0.8388	0.9447	0.8248	0.8636	0.8364
Precision	0.8606	0.7807	0.9451	0.7624	0.8645	0.7774
Recall	0.8597	0.8184	0.9447	0.7863	0.8636	0.8104
F1-score	0.8597	0.7954	0.9447	0.7726	0.8635	0.7908
Support	None	None	None	None	None	None

# Machine Learning Models

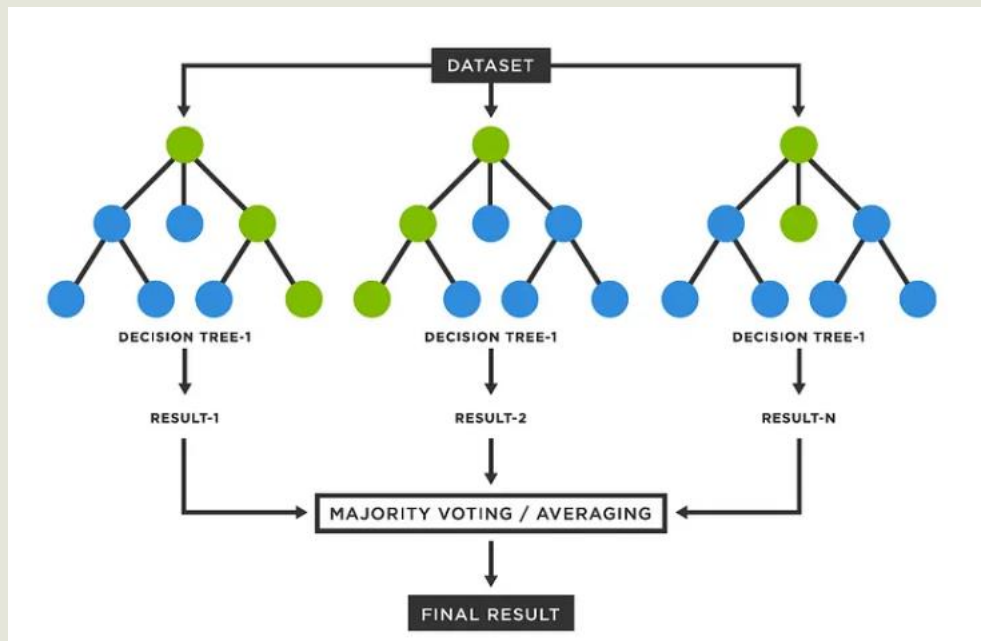
## KNN, SVM and Logistic Regression

Consistent accuracy across models - approx. 82% to 84%

- **Logistic Regression** and **SVM** performed similarly, with Logistic Regression having a slightly better balance between precision and recall for the test set.
- **KNN** had the highest training accuracy, but slightly lower test accuracy compared to Logistic Regression and SVM, indicating potential overfitting.
- **Logistic Regression** showed the highest F1-score for the infected class on the test set, which may be crucial depending on the importance of minimizing false negatives or positives.
- Overall, **Logistic Regression** and **SVM** appear to be more reliable, with Logistic Regression having a slight edge in handling imbalanced data better.

# Machine Learning Models

## Random Forest



	precision	recall	f1-score	support
0	0.92	0.93	0.92	327
1	0.76	0.73	0.75	101
accuracy			0.88	428
macro avg	0.84	0.83	0.84	428
weighted avg	0.88	0.88	0.88	428

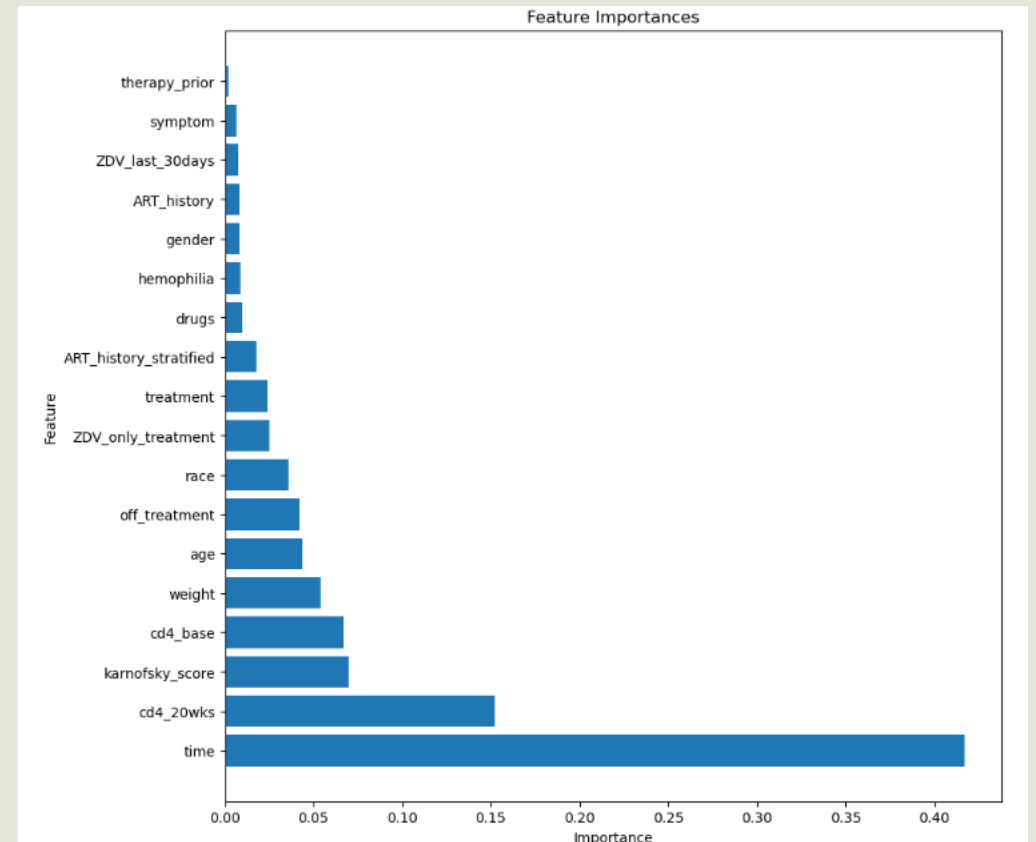
Image source: <https://medium.com/@denizgunay/random-forest-af5bde5d7e1e>



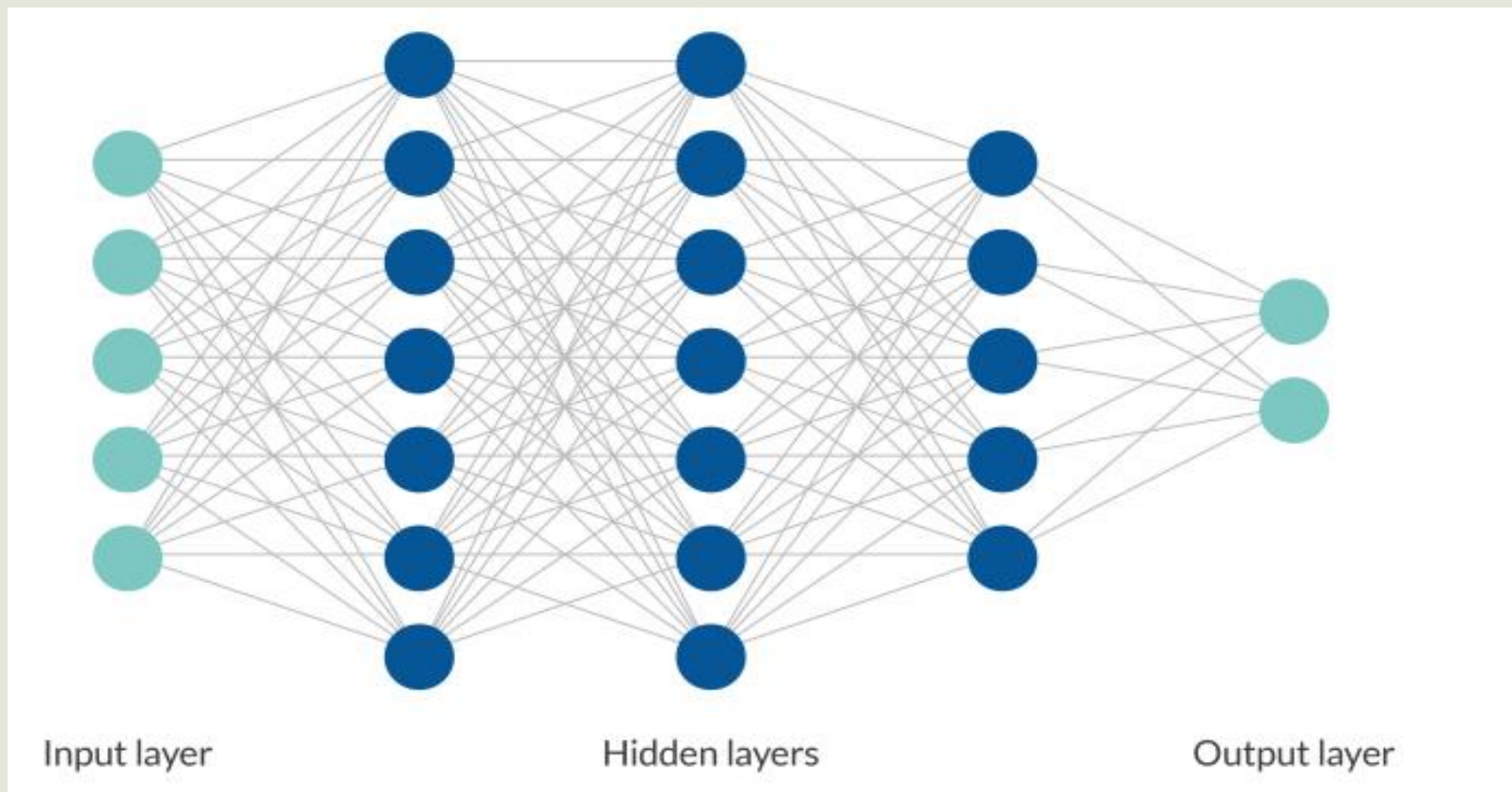
# Machine Learning Models

## Random Forest

- Extracts feature importances from the trained Random Forest model.



# Deep Neural Network (DNN)



# Deep Neural Network (DNN)

## DNN Model (**Attempt 1**)

optimizer='adam'  
epochs=50

Measured metrics:  
acc train: 100%  
acc test: 87%  
precision: 74%  
recall: 73%

- Model is over fitted

80 neuron  
Relu

30 neuron  
Relu

1 neuron  
Sigmoid



# Deep Neural Network (DNN)

## DNN Model (Attempt 2)

optimizer='adam'  
epochs=100

Measured metrics:  
acc train: 96%  
acc test: 86%  
precision: 71%  
recall: 76%

- Model is less overfitted
- 3% increase in recall

14 neuron  
Relu

7 neuron  
Relu

1 neuron  
Sigmoid

# Deep Neural Network (DNN)

## DNN Model (Attempt 3)

The structure optimised using the **Keras Tuner**

optimizer='adam'  
epochs=25

Measured metrics:

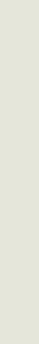
acc train: 96%

acc test: 88%

precision: 73%

recall: 81%

- Model is less overfitted
- 5% increase in recall



30 neuron  
Relu



12 neuron  
Relu



24 neuron  
Relu



1 neuron  
Sigmoid

# Deep Neural Network (DNN)

## DNN Model (Attempt 4)

The structure optimised using the **Keras Tuner**  
We just kept **7 first important features** offered by Random Forest model.

optimizer='adam'

epochs=**25**

Measured metrics:

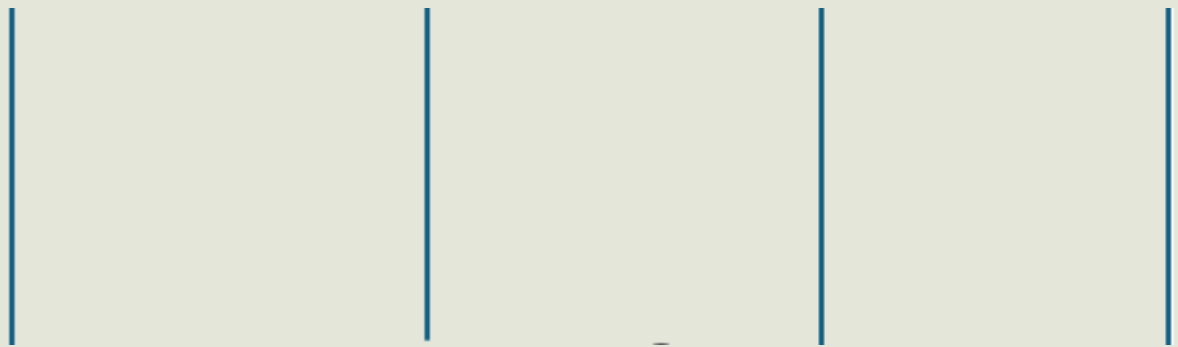
acc train: 95%

acc test: 89%

precision: 79%

recall: 81%

- Model was not over fitted
- This was the best model with highest accuracy, precision, and recall



30 neuron  
Relu

12 neuron  
Relu

24 neuron  
Relu

1 neuron  
Sigmoid



# Evaluation of Models

Logistic Regression

K-Nearest Neighbours

SVM

Random Forest

Deep Neural Network

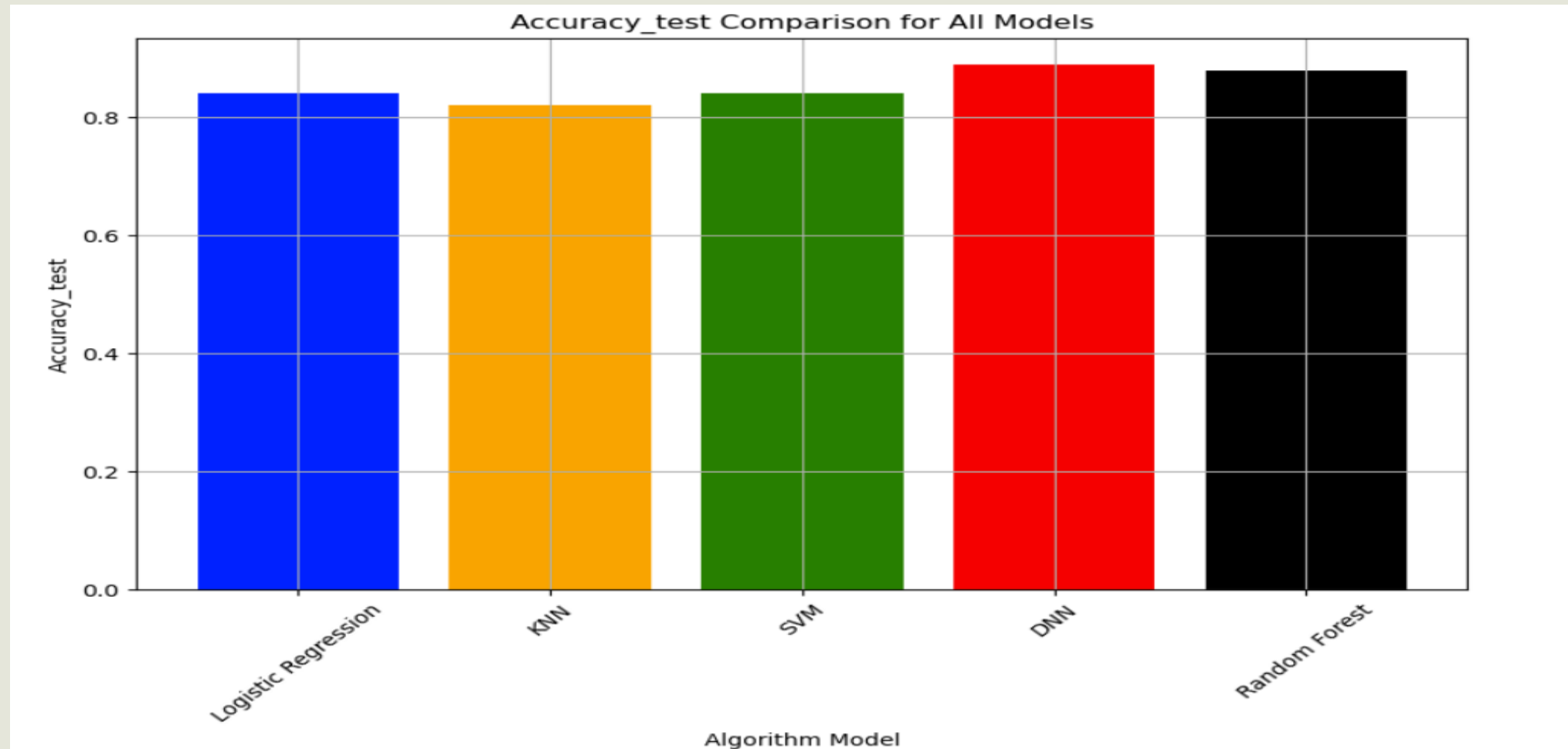


Test Accuracy

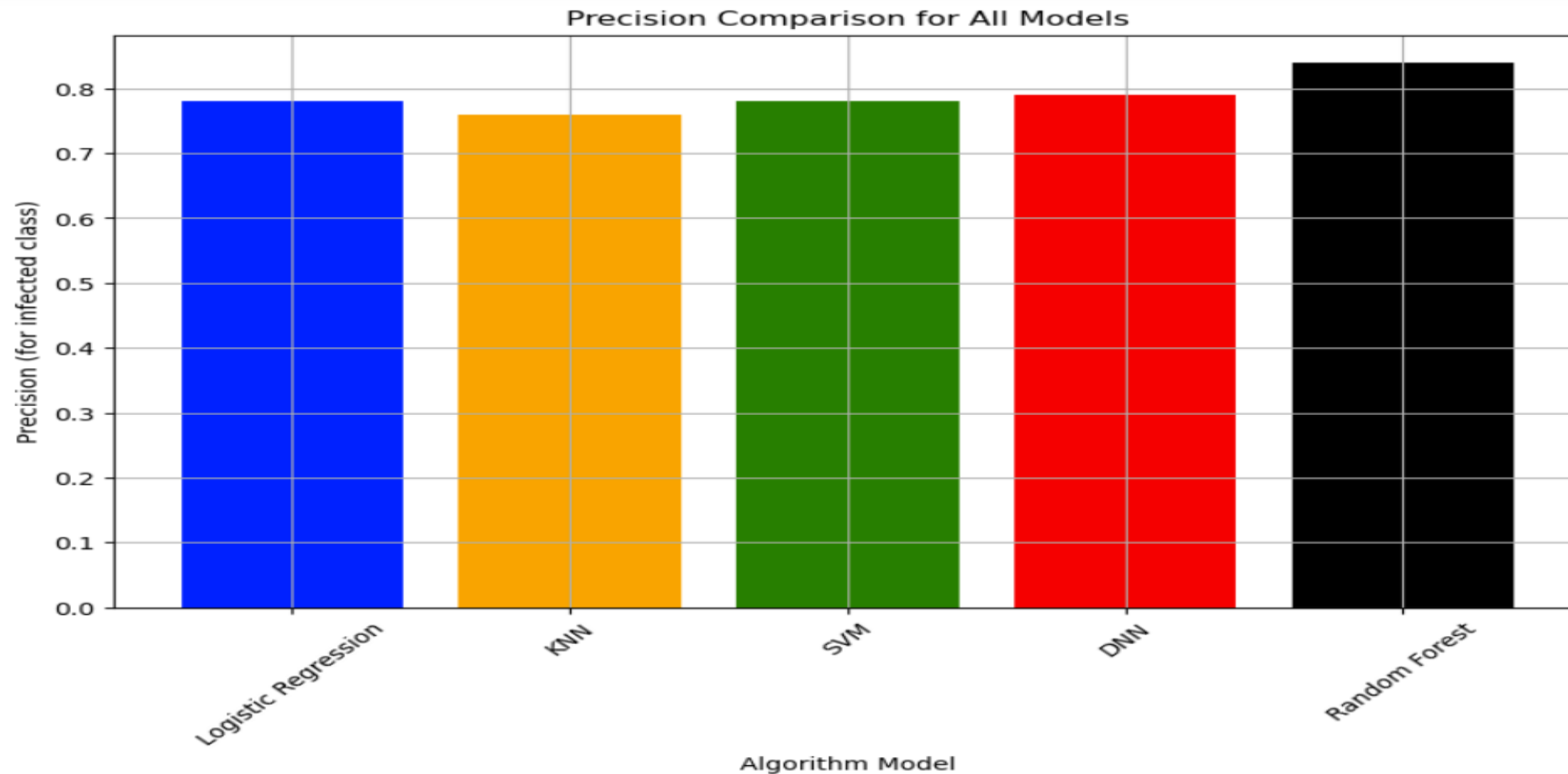
Precision

Recall

# Statistical Metrics Comparison - Test Accuracy

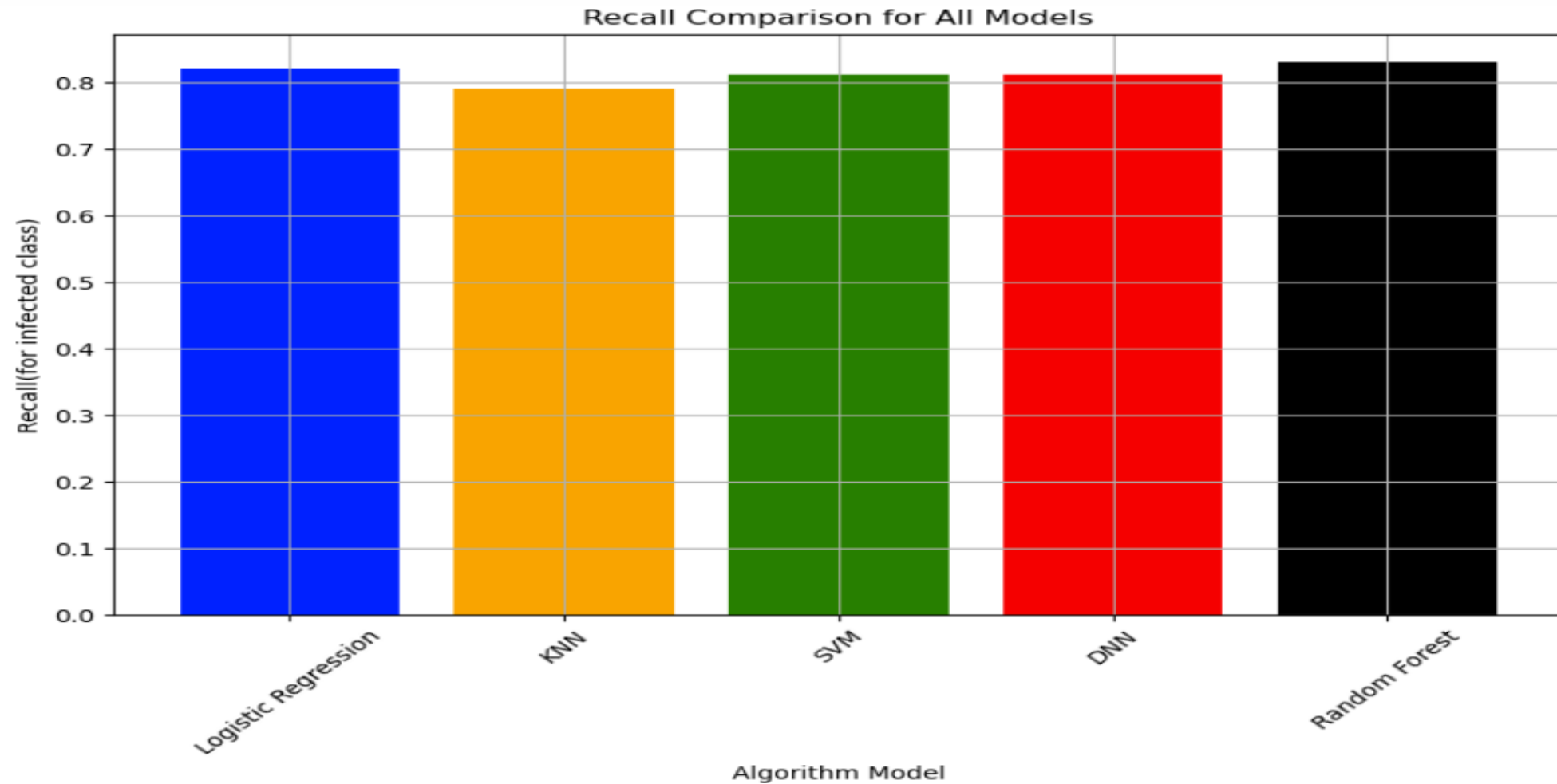


# Statistical Metrics Comparison - Precision





# Statistical Metrics Comparison - Recall



# Predictive Machine Learning Models Discussion

## Overall Conclusion:

Based on the evaluation metrics:

- Random Forest:** This model has strong performance across all metrics, particularly in test accuracy (0.88), precision (macro avg: 0.84), recall (macro avg: 0.83), and F1-score (macro avg: 0.84).
- Neural Network:** This model also performs very well, with the highest accuracy (0.89) and strong precision (0.79) and recall (0.81).

## Recommendation:


The **Random Forest** model is the best choice overall because it offers a good balance of accuracy, precision, recall, and F1-score, making it highly reliable for predicting both infected and not infected cases. The Neural Network is a close second and might be preferable in scenarios where slightly higher accuracy is crucial.

Any Questions?

# Thank you

Rahul Gade

 rahulmg2002@yahoo.com

 +614032732966

 <https://github.com/RahulG2381>



**RAHUL GADE**

Reporting Analyst | Data and Analytics | PowerBI



[linkedin.com/in/rahul-gade-b9514119](https://www.linkedin.com/in/rahul-gade-b9514119)