Predicting AIDS Progression with Data Insights

Shiva Bajelan
Jeremy Hooper
Rahul Gade
Linh Doan

Table of contents

01

Project Overview

02

Data Cleaning

03

Visualizations

04

Data Examination

05

Machine Learning

06

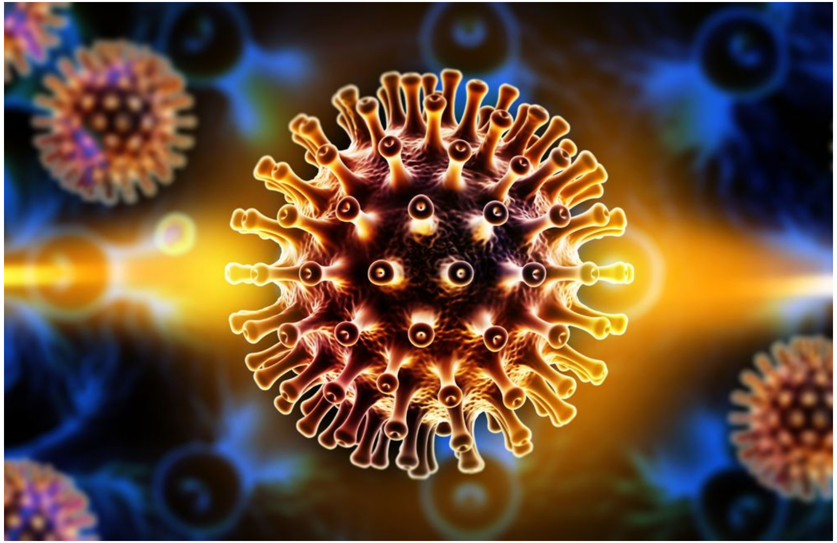
Recommendations

01

Project Overview

Objectives

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.

02

Data Cleaning

Extract, transform, and load (ETL)

PostgreSQL
SQLAlchemy
.py

Data cleaning

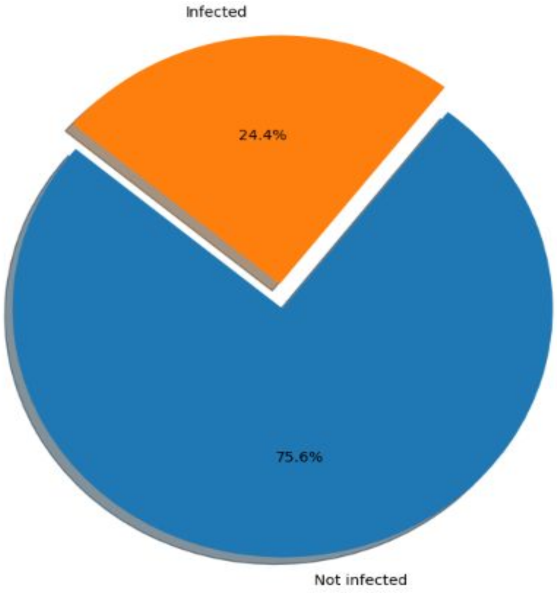
Connection to SQL Database

03

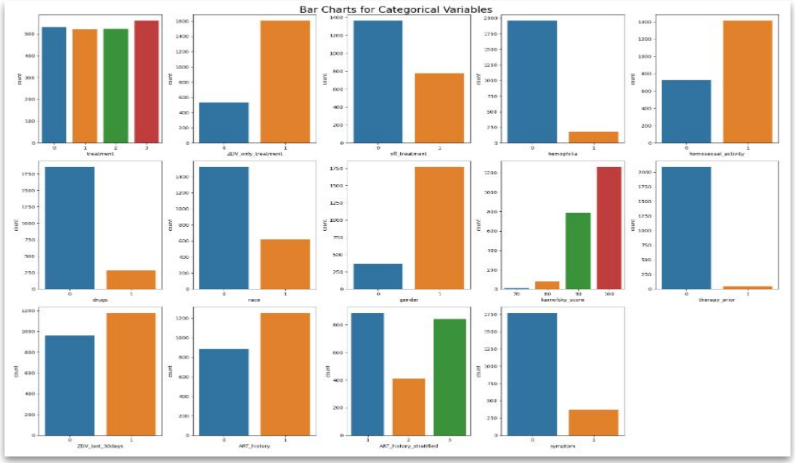
Visualisations

Data Insights

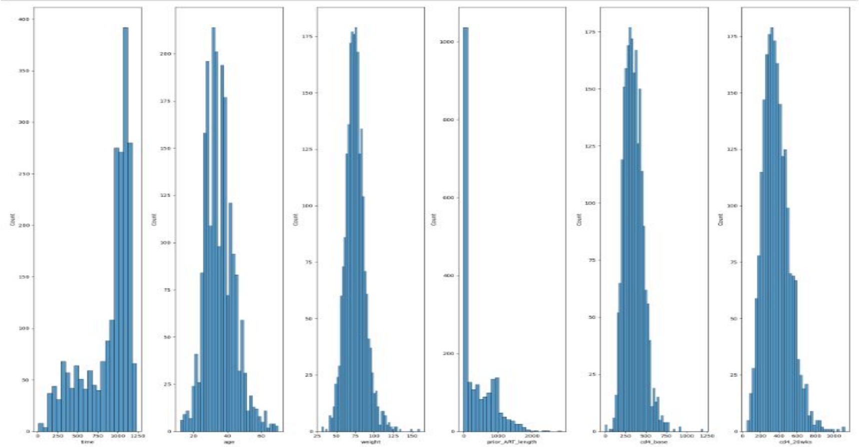
Infected vs Not infected Distribution



Categorical Variables

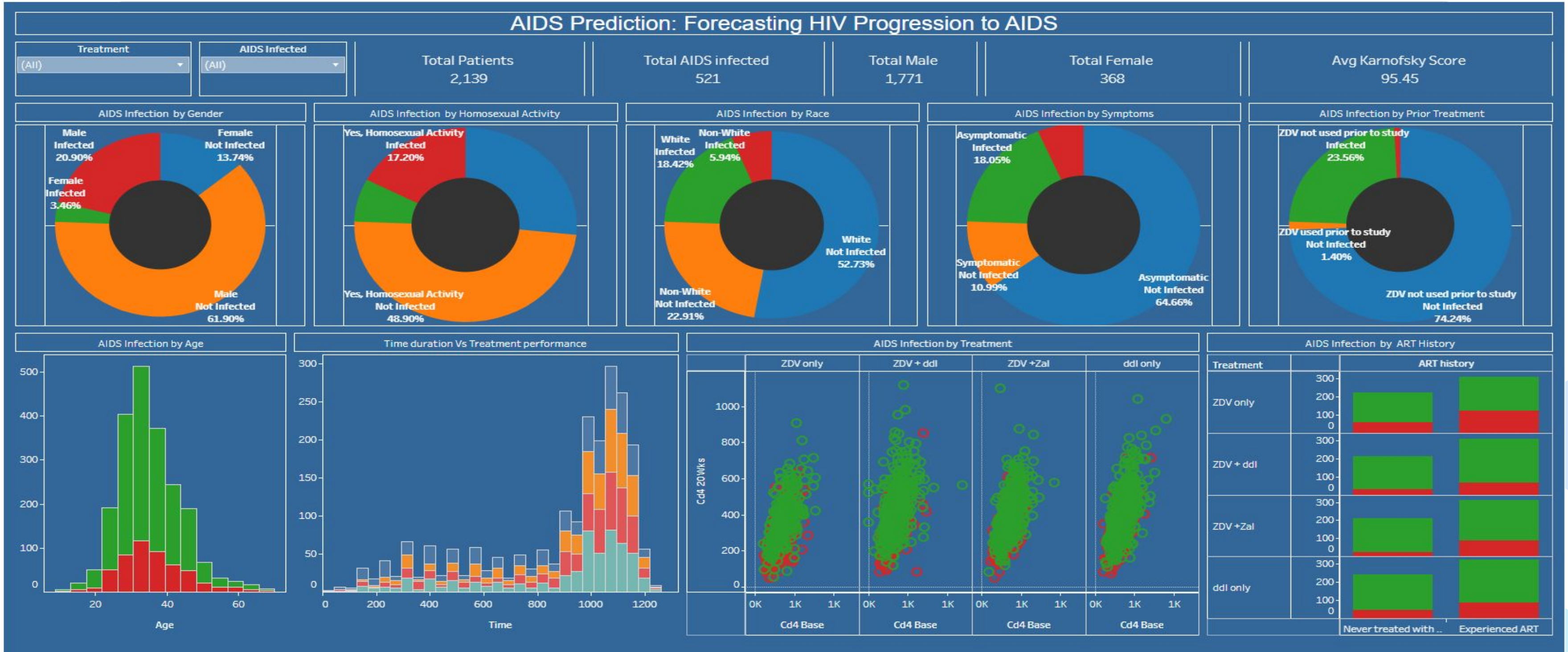


Numerical Variables



Data Insights

Variables affecting AIDS infection Rate



AIDS Prediction: Forecasting HIV Progression to AIDS

Treatment: (All)

AIDS Infected: (All)

Total Patients 2,139

Total AIDS infected 521

Total Male 1,771

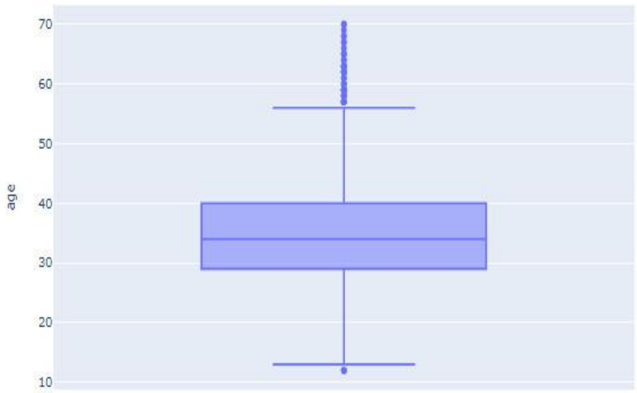
Total Female 368

Avg Karnofsky Score 95.45

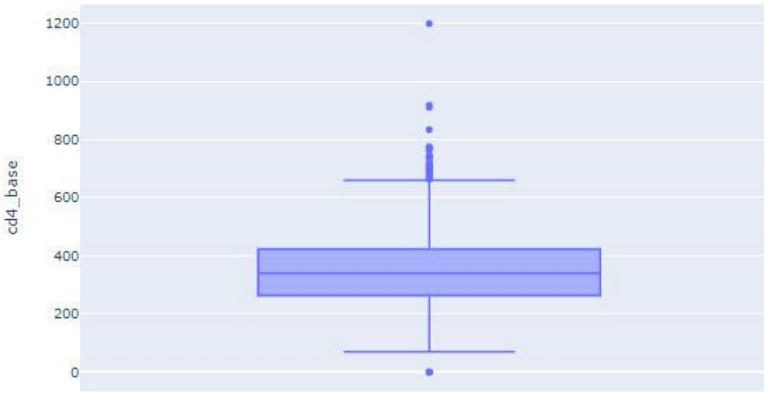
AIDS Infection by Gender
Male Infected 20.90%
Female Not Infected 13.74%
Female Infected 3.46%
Male Not Infected 61.90%
AIDS Infection by Homosexual Activity
Yes, Homosexual Activity Infected 17.20%
Yes, Homosexual Activity Not Infected 48.90%
AIDS Infection by Race
White Infected 18.42%
Non-White Infected 5.94%
White Not Infected 52.73%
Non-White Not Infected 22.91%
AIDS Infection by Symptoms
Asymptomatic Infected 18.05%
Symptomatic Not Infected 10.99%
Asymptomatic Not Infected 64.66%
AIDS Infection by Prior Treatment
ZDV not used prior to study Infected 23.56%
ZDV used prior to study Not Infected 1.40%
ZDV not used prior to study Not Infected 74.24%
AIDS Infection by Age
500
400
300
200
100
0
20
40
60
Age
Time duration Vs Treatment performance
300
250
200
150
100
50
0
0
200
400
600
800
1000
1200
Time
AIDS Infection by Treatment
ZDV only
ZDV + ddI
ZDV +Zal
ddI only
Cd4 20Wks
1000
800
600
400
200
0
0K
1K
1K
Cd4 Base
AIDS Infection by ART History
Treatment
ART history
ZDV only
ZDV + ddI
ZDV +Zal
ddI only
300
200
100
0
Never treated with ..
Experienced ART

04
Data Examination

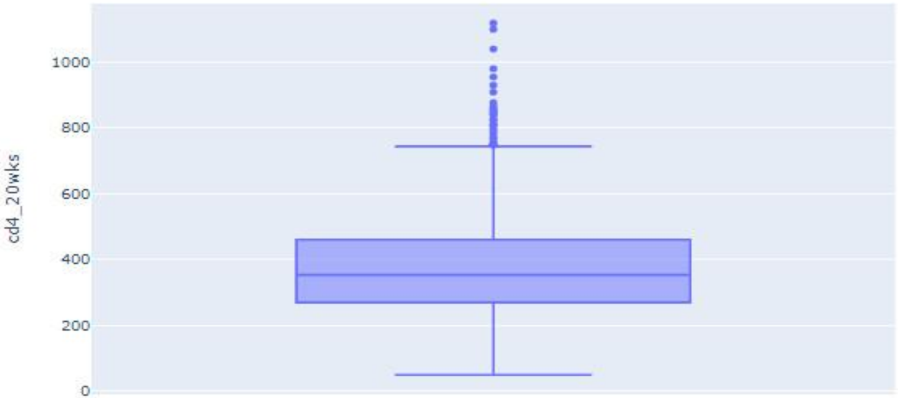
Box Plot of age (Participant's age)



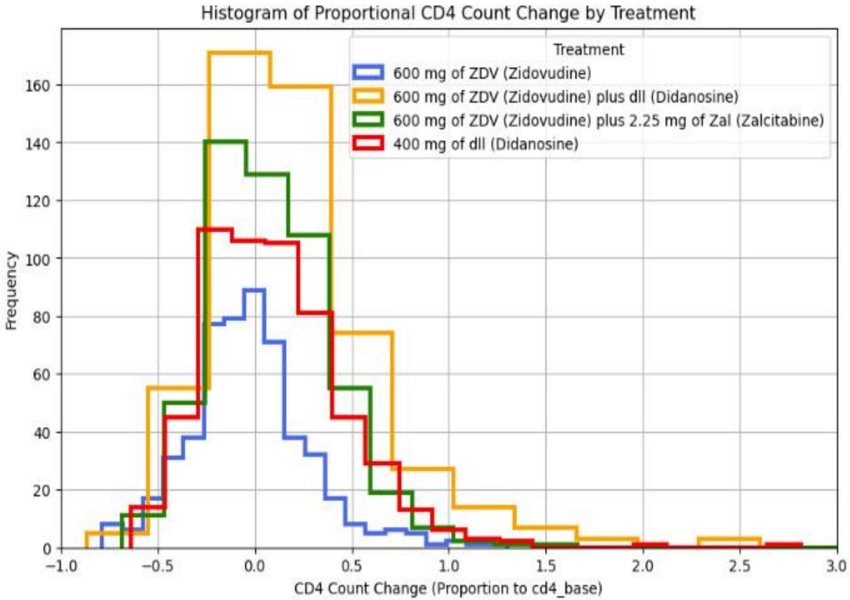
Box Plot cd4_base (Baseline count of cd4 cells in the blood)



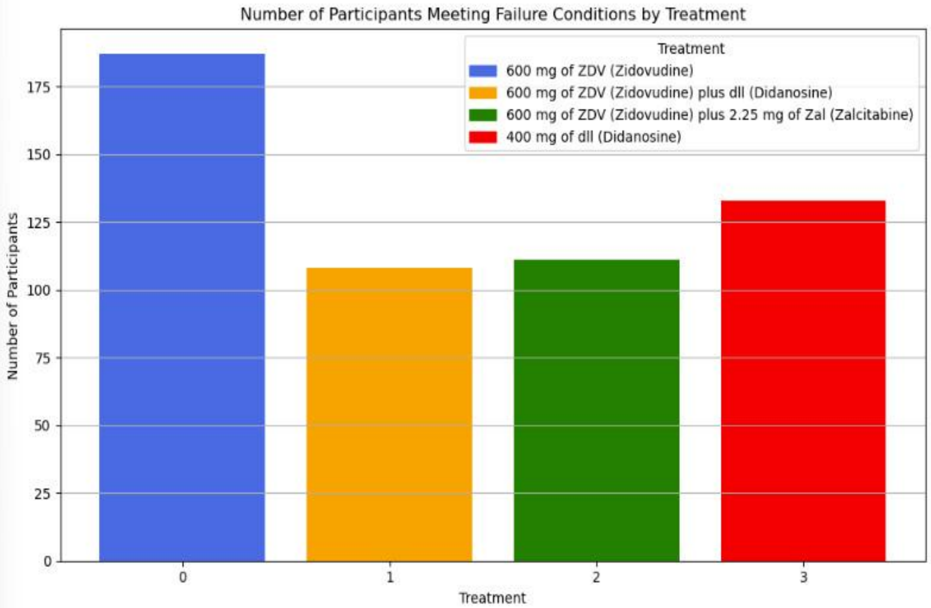
Box Plot cd4_20wks (20 weeks count of cd4 cells in the blood)



Data Examination

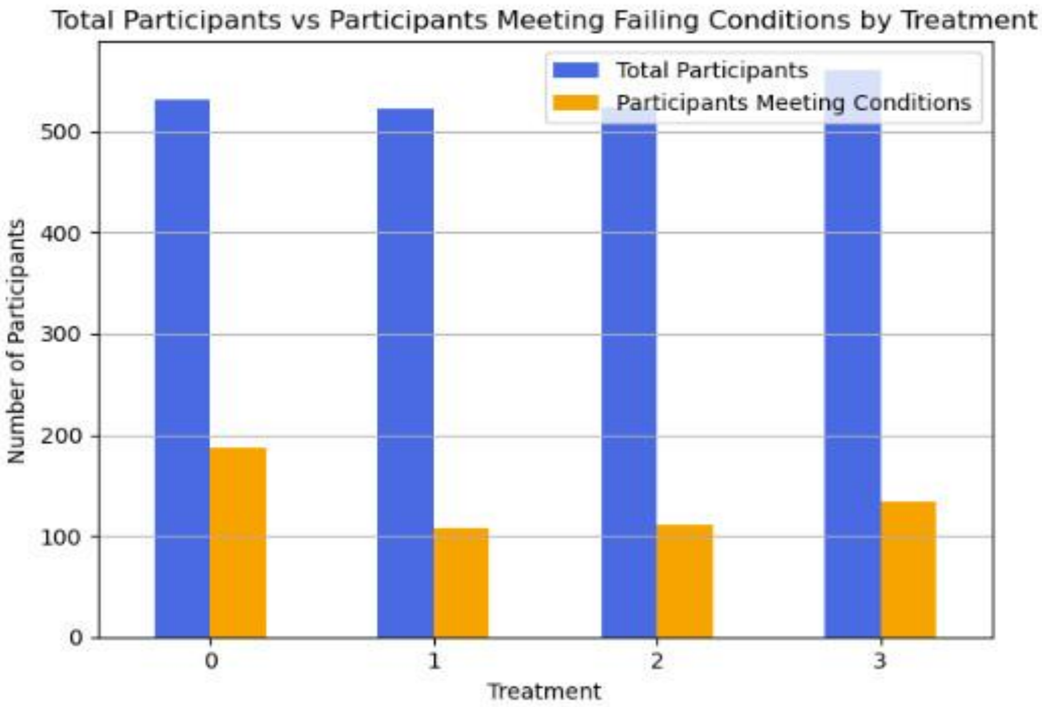


Proportional CD4 Count Change (cd4_20wks - cd4_base)



Number of Participants Meeting Failure Conditions by Treatment (cd4 < 50%, Infected)

Data Examination



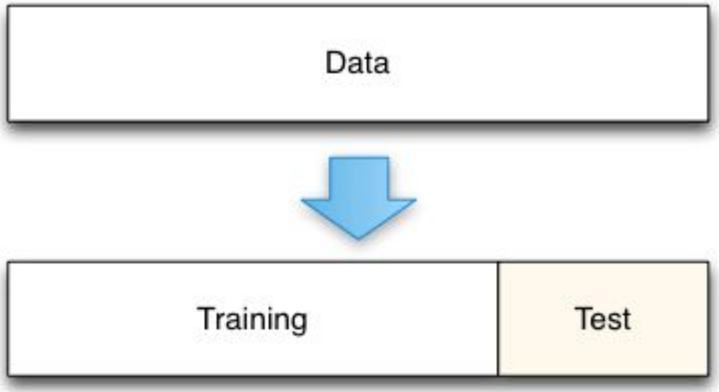
Total Participants vs Participant's
Meeting Failure Conditions by Treatment

05

Machine Learning

KNN, SVM and Logistic Regression

Methodology



Dataset split into training and testing sets using train-test split ratios: 70:30, 50:50, and 30:70.

KNN, SVM and Logistic Regression

Consistent accuracy across models - approx. 81% to 88%

Logistic Regression: Balanced precision and recall for the "not infected" class; not so great for the "infected" class

➔ challenges in correctly classifying infected individuals

KNN: Demonstrated relatively lower precision and recall for both classes

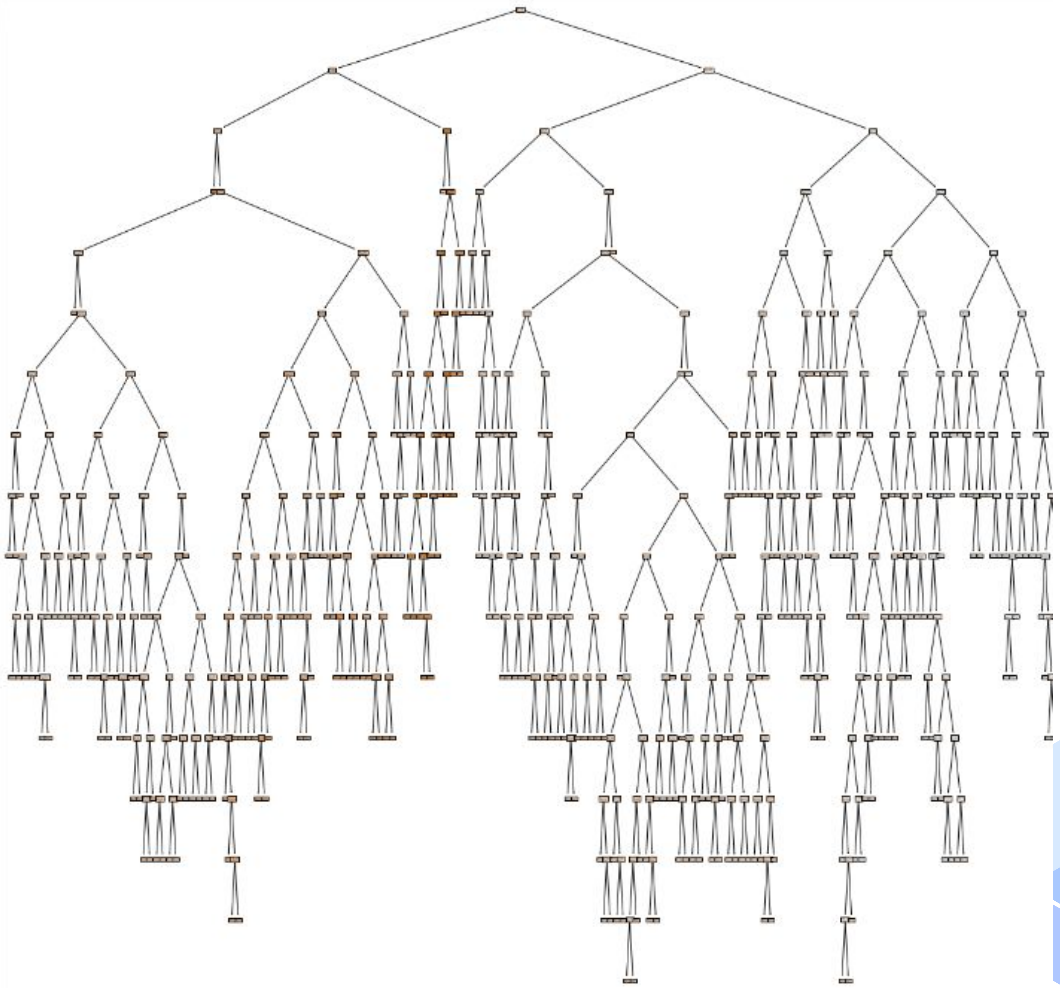
➔ suboptimal performance in classification

SVM: Showed balanced precision and recall for both classes

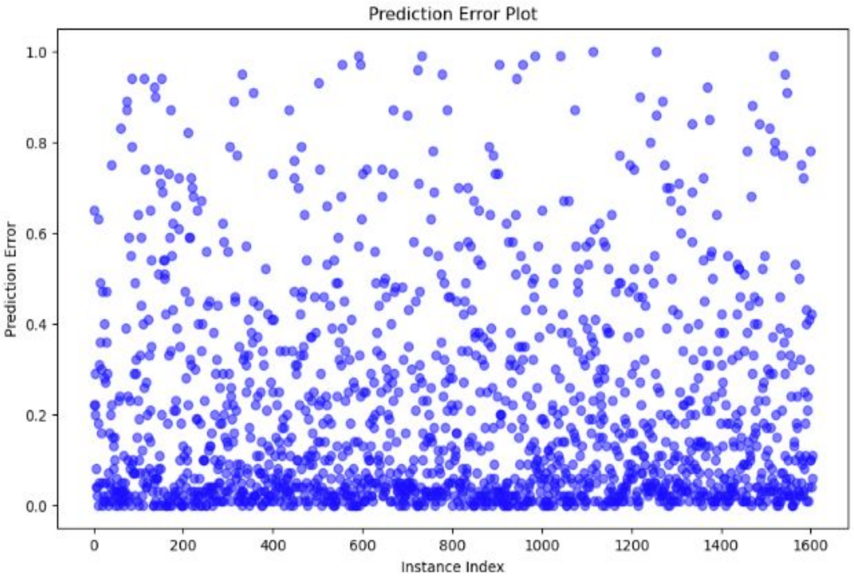
➔ effective classification performance.

Random Forest

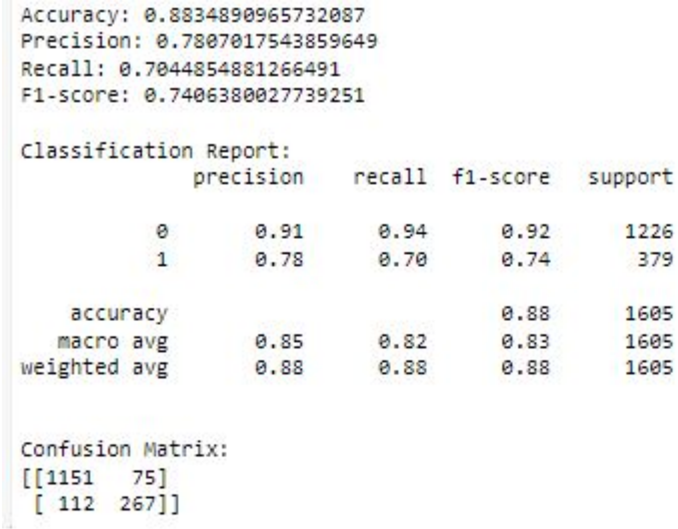
Decision Tree



Random Forest - Prediction Error



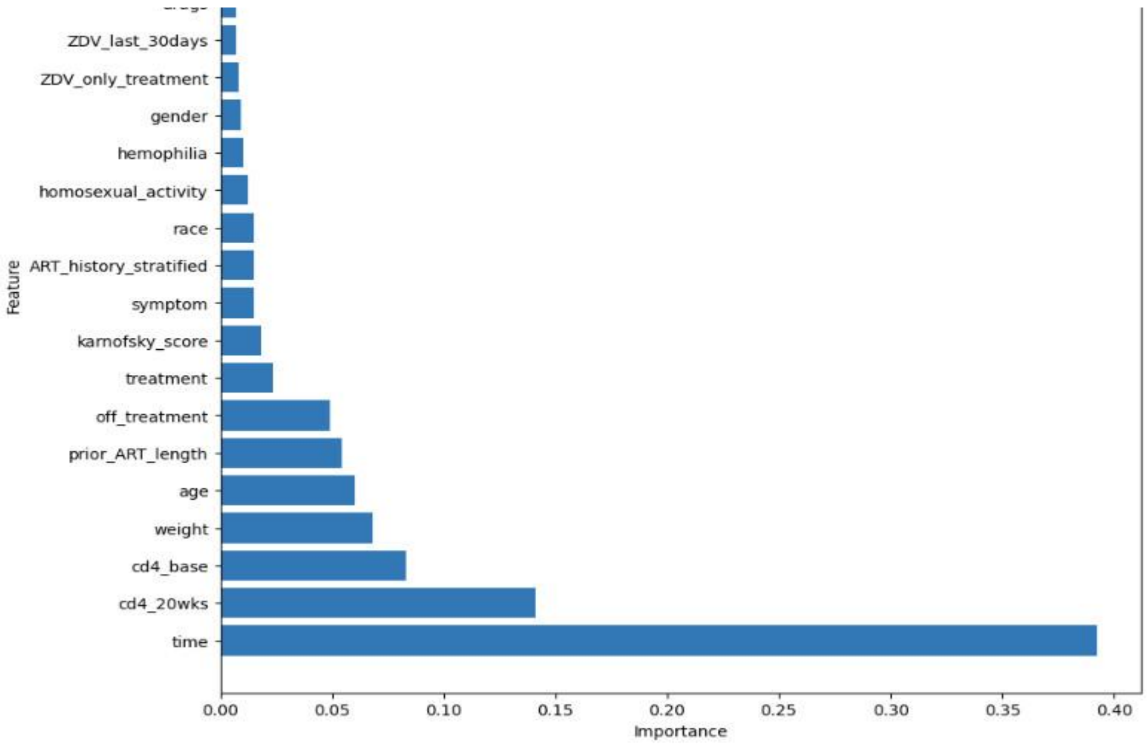
RandomForest – Prediction Error – 75%



```
Accuracy: 0.8834890965732087
Precision: 0.7807017543859649
Recall: 0.7044854881266491
F1-score: 0.7406380027739251

Classification Report:
              precision    recall  f1-score   support

           0       0.91      0.94      0.92      1226
           1       0.78      0.70      0.74       379

    accuracy                           0.88      1605
   macro avg       0.85      0.82      0.83      1605
weighted avg       0.88      0.88      0.88      1605


Confusion Matrix:
[[1151   75]
 [ 112  267]]
```

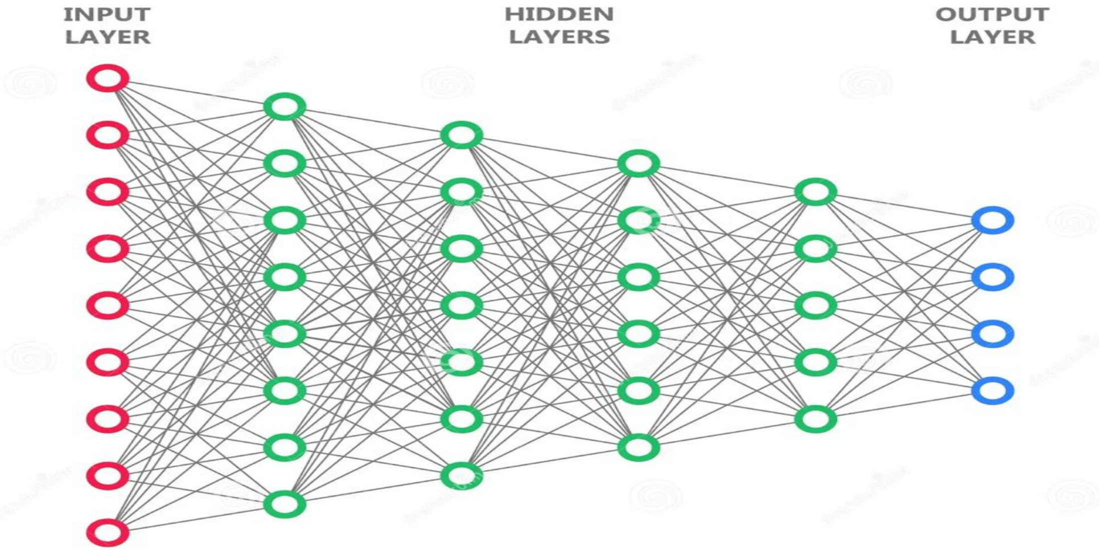
RandomForest – Results Statistics

Random Forest



Random Forest - Features Importance

DNN Model

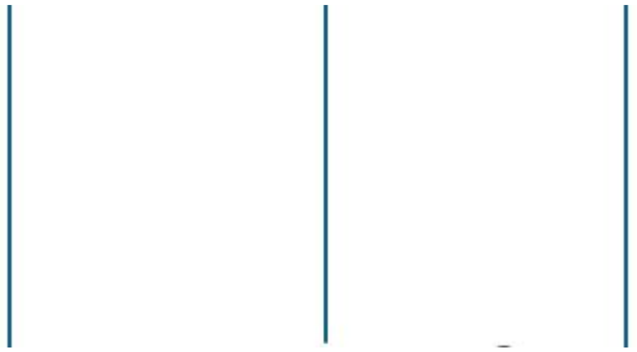


DNN Model (Attempt 1)

optimizer='adam'
epochs=50

Measured metrics:
acc train: %99
acc test: %89
precision: %82
recall: %75

- Model was over fitted



80 neuron
Relu

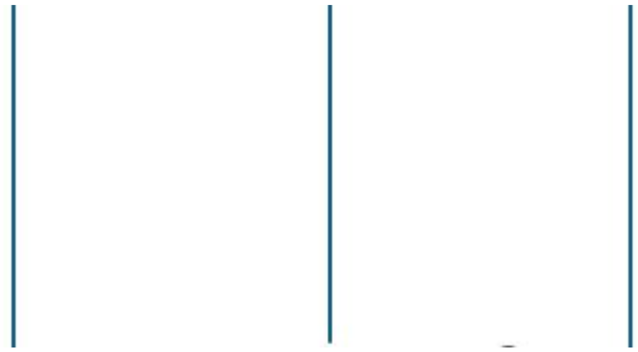
30 neuron
Relu

1 neuron
Sigmoid

DNN Model (Attempt 2)

optimizer='adam'
epochs=100

Measured metrics:
acc train: %95
acc test: %89
precision: %81
recall: %76



14 neuron
Relu

7 neuron
Relu

1 neuron
Sigmoid

- Model was less over fitted
- %1 increase in recall

DNN Model (Attempt 3)

The structure optimised using the **Keras Tuner**

optimizer='adam'
epochs= 25

Measured metrics:
acc train: %94
acc test: %89
Precision: %84
recall: %74

30 neuron
Relu

12 neuron
Relu

24 neuron
Relu

1 neuron
Sigmoid

- Model was less over fitted
- %3 increase in precision

DNN Model (Attempt 4)

We still use the structure offered by **Keras Tuner** In attempt 3

We just kept 7 first important features offered by Random Forest model.

Measured metrics:
acc train: %90
acc test: %90
precision: %86
recall: %76

30 neuron
Relu

12 neuron
Relu

24 neuron
Relu

1 neuron
Sigmoid

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

Evaluation of Models

Logistic Regression

K-Nearest Neighbours

SVM

Random Forest

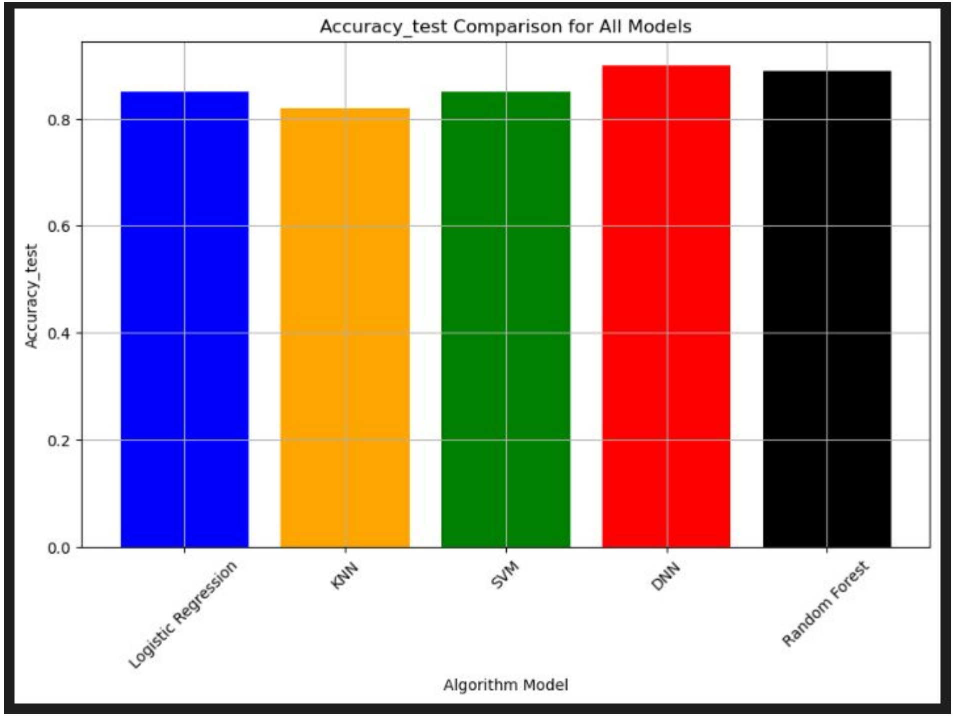
Deep Neural Network

Test Accuracy

Precision

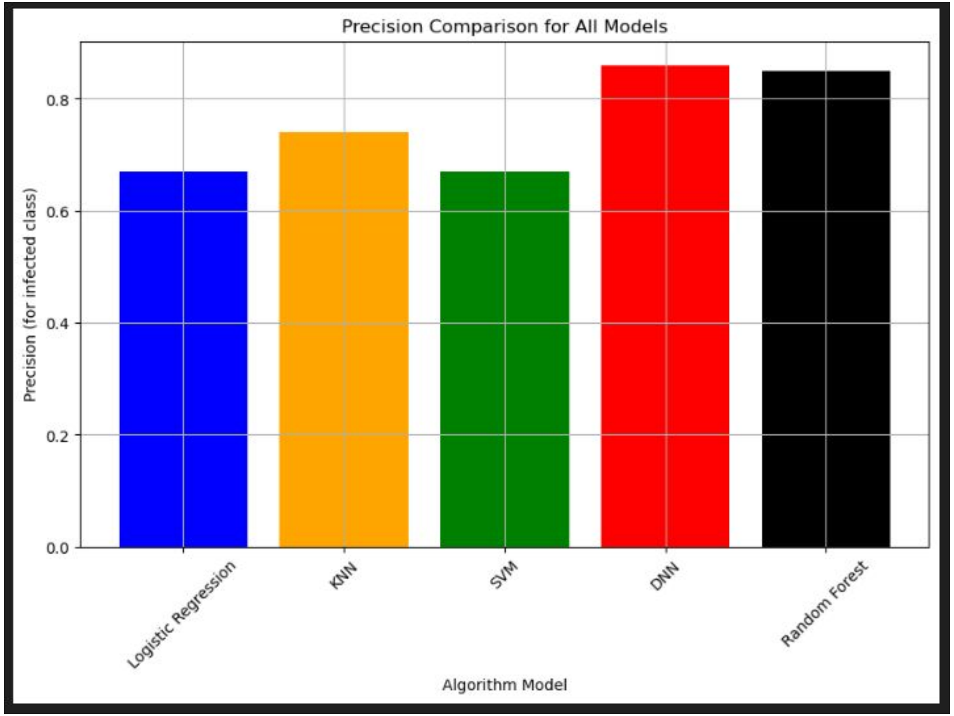
Recall

Statistical Metrics Comparison - Test Accuracy



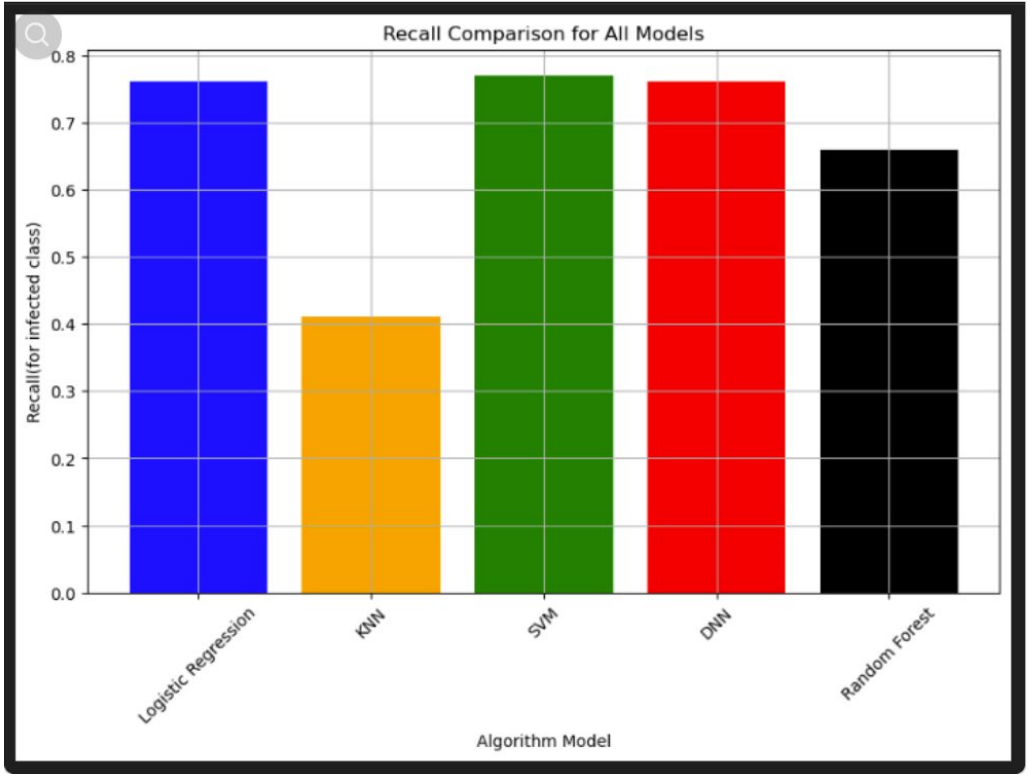
Accuracy Test Comparison - All Models

Statistical Metrics Comparison - Precision



Precision Comparison - All Models

Statistical Metrics Comparison - Recall



Recall Comparison - All Models

06

Predictive Machine Learning Models Discussion

Predictive Machine Learning Models Discussion

Handling Class Imbalance

The code used in our predictive learning models indicates that our classes are imbalanced. To handle this issue, we have used the following strategies:

- **Using algorithmic techniques**, adjusting the model's class weights to penalise misclassifications of the minority class more than the majority class.
- **Using ensemble methods,** such as Random Forest, which inherently handles class imbalance by combining multiple weak learners.
- **Evaluation Metrics,** using appropriate evaluation metrics less sensitive to class imbalance, such as precision and recall rather than accuracy.

Conclusions

- **Neural networks** displays strong performance, with high accuracy and precision. Despite this, the fourth attempt of the Neural Network model shows promising performance, boasting high accuracy_train, accuracy_test, precision, and a recall of for the infected class (class 1).

- **Optimal Prioritisation:** If maximising accuracy and precision is paramount, Neural Network is the top contender due to its impressive performance in these metrics.

- **Recall Priority:** The Support Vector Machine (SVM) with recall stands out, making it the preferred choice for scenarios where maximising recall is crucial.

Conclusions

Random Forest Analysis: Random Forest yields high accuracy and precision but exhibits a lower recall. This suggests potential for precise predictions but with limitations capturing all relevant instances.

Alternative Options: Logistic Regression and KNN offer accuracy and recall. Both provide viable alternatives with balanced performance metrics for consideration.

Given the concept of our medical dataset and the importance of predicting individuals infected with AIDS in order to get timely treatment, Neural Network is the best model, with an accuracy of 89% and a recall of 78%.

Thank you!