Heart attack prediction machine learning

by Lorena, Eliska, Owen, Filip and Camil

Data selection and preparation

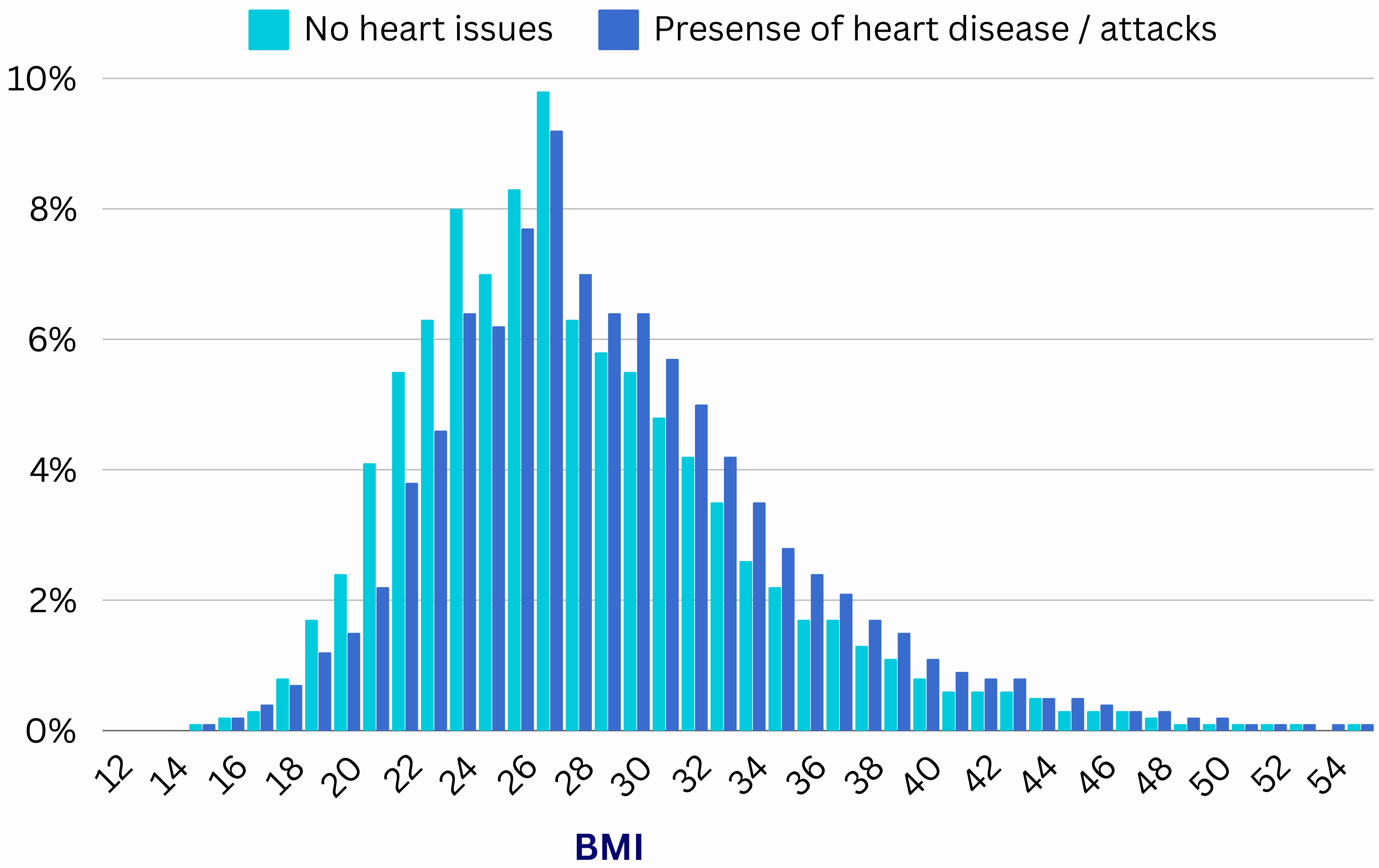
- Heart disease is a leading cause of death in the developed world, and early diagnosis can prevent severe outcomes.
- This project uses the cleaned **2015 BRFSS dataset** (250,000+ survey responses) to explore and predict heart disease risk, assessing whether survey data can aid in preventive health screening.
- Dataset includes **key health metrics** (e.g., blood pressure, cholesterol, BMI), **chronic conditions** (e.g., diabetes, stroke), **healthcare access issues**, and **demographic data** (e.g., age, sex, income) .

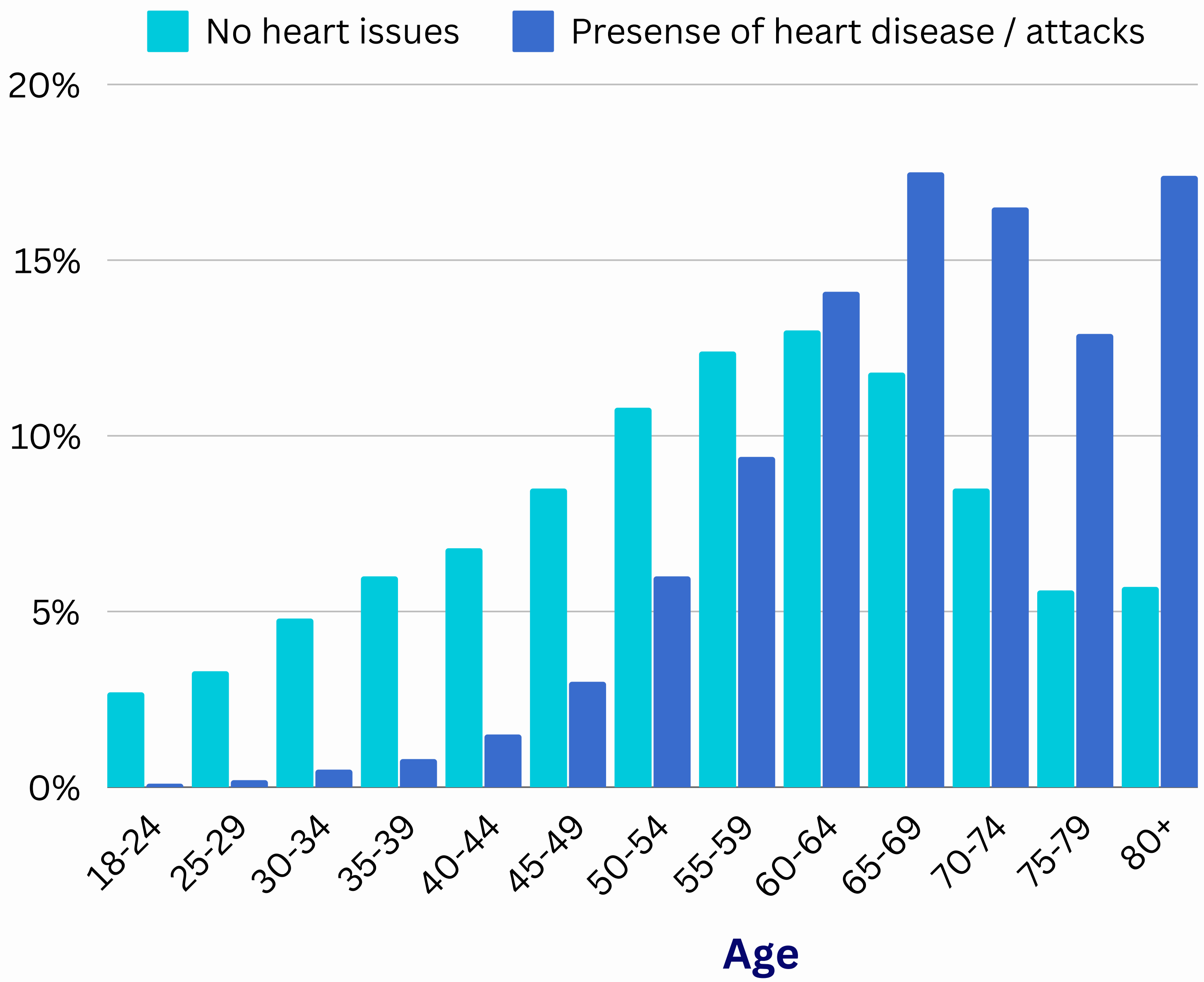
Data Cleaning

- **highly correlated columns** “physical health” and “difficulty walking or climbing stairs” (with general health) and “education” (with income) **were dropped** in order to **improve model performance**
- dataset has been pre-processed, and **categorical variables were already encoded**

Initial data evaluation

- Splitting the data into those who have experienced heart disease or heart attacks and those who haven’t immediately made it clear that there were several key differences in the distribution of metrics between these groups.
- We have highlighted a couple of key factors, but there were many more.





Feature Engineering and Selection

Feature and Target separation

dataset was split into **features** (independent variables) and the **target** (HeartDiseaseorAttack)

Train-Test Split

data divided into 80% training and 20% testing sets, ensuring reproducibility with a fixed random state

Feature Normalization

numerical features were scaled using **Min-Max Scaling** to normalize values between 0 and 1, ensuring that all variables contribute equally to the model's performance

Initial evaluation with KNN

baseline KNN model was used to determine key features

optimized models with normalized features **improved accuracy on both training and test set**

Class imbalance & resampling

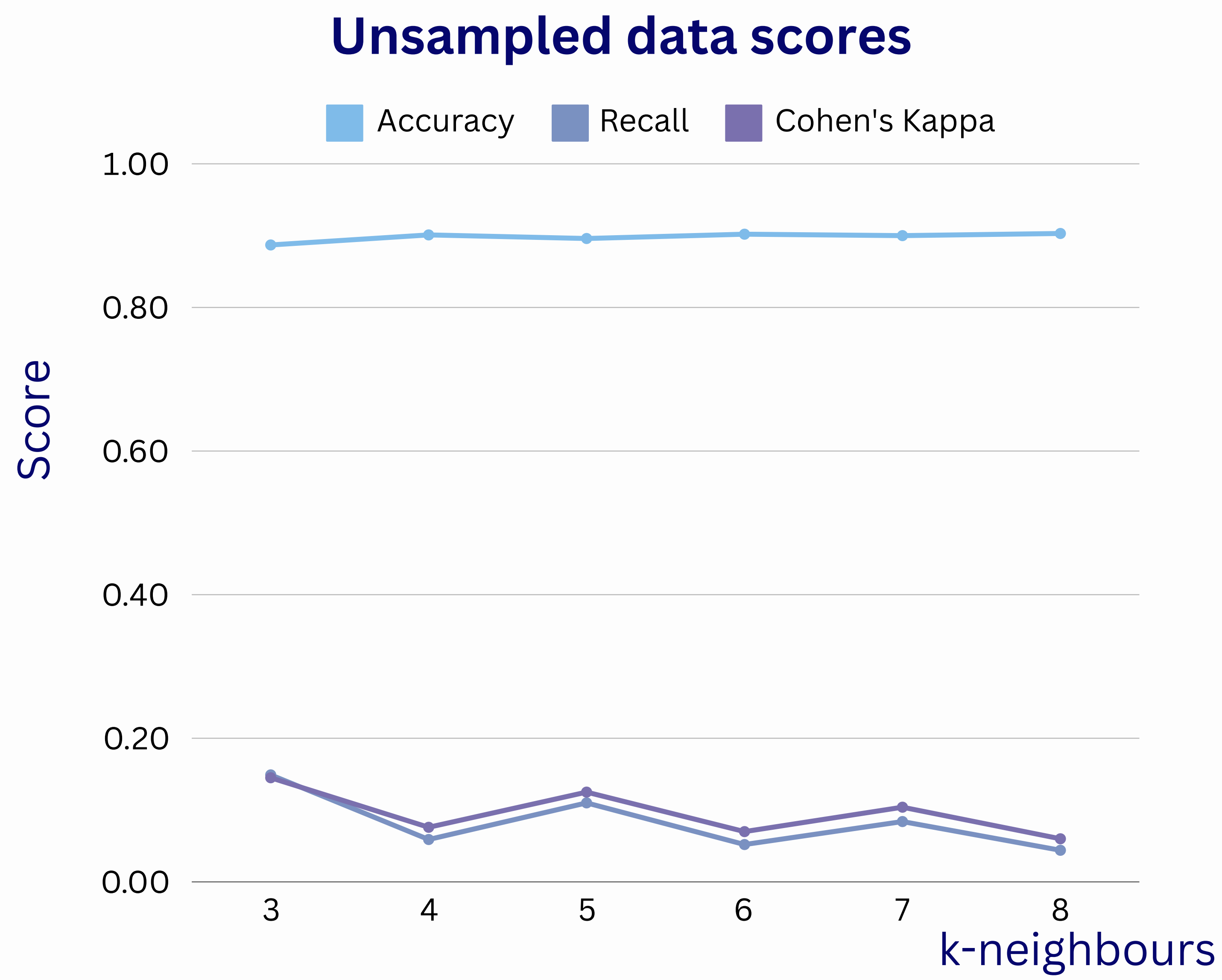
01 Initial KNN evaluation showed high accuracy (~90%), however we discovered this was **due to the class imbalance,** and the Recall scores were very poor

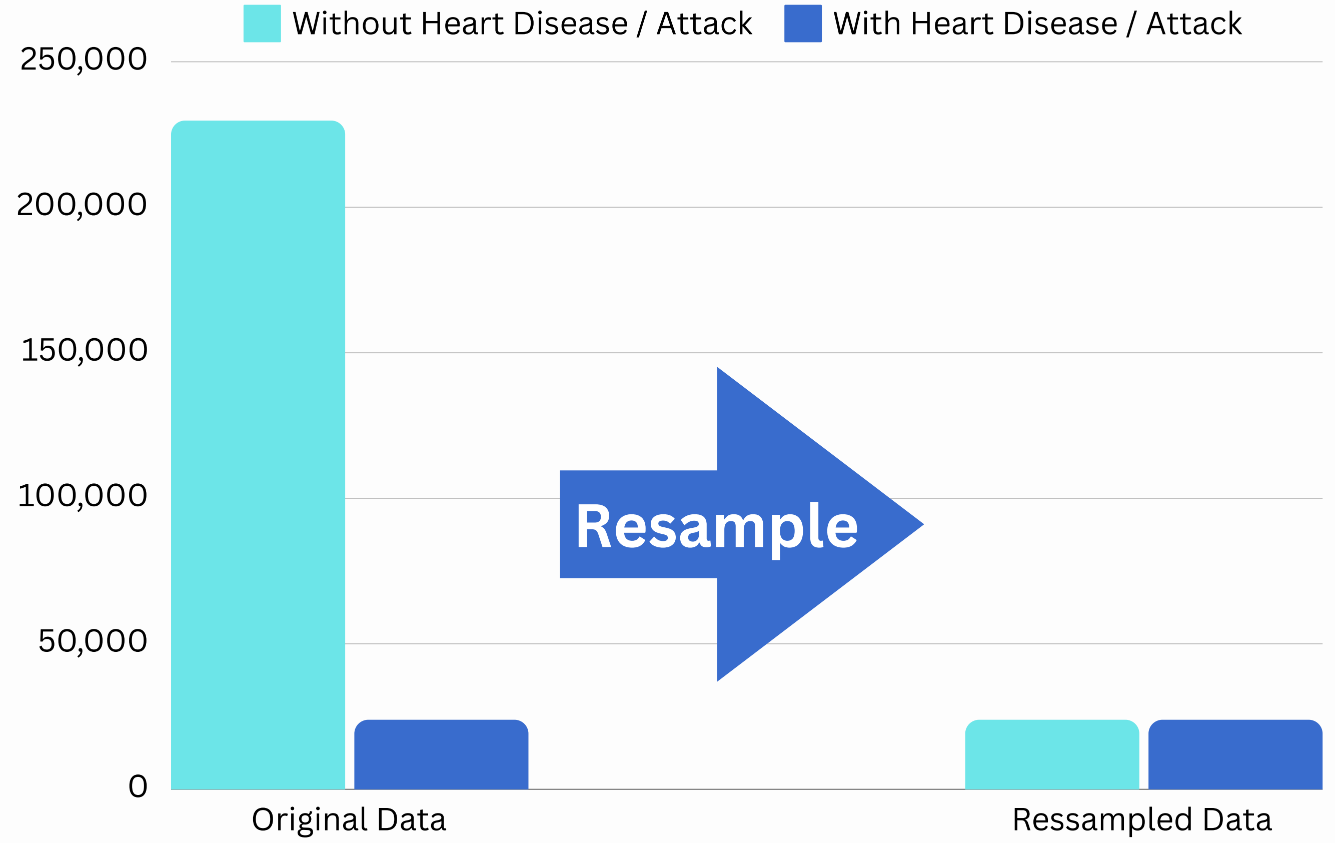
02 **Class imbalance identified** using a count plot of our target “**HeartDiseaseorAttack**” column, revealing significant imbalance, **with ~90% of instances labeled as "No" (0)**

03 The majority class (0) was **undersampled to match the minority class** (1)

- Undersampling was deemed appropriate due to the size of the dataset

04 **Resampled dataset is used for further machine learning applications** to ensure fair and unbiased model performance

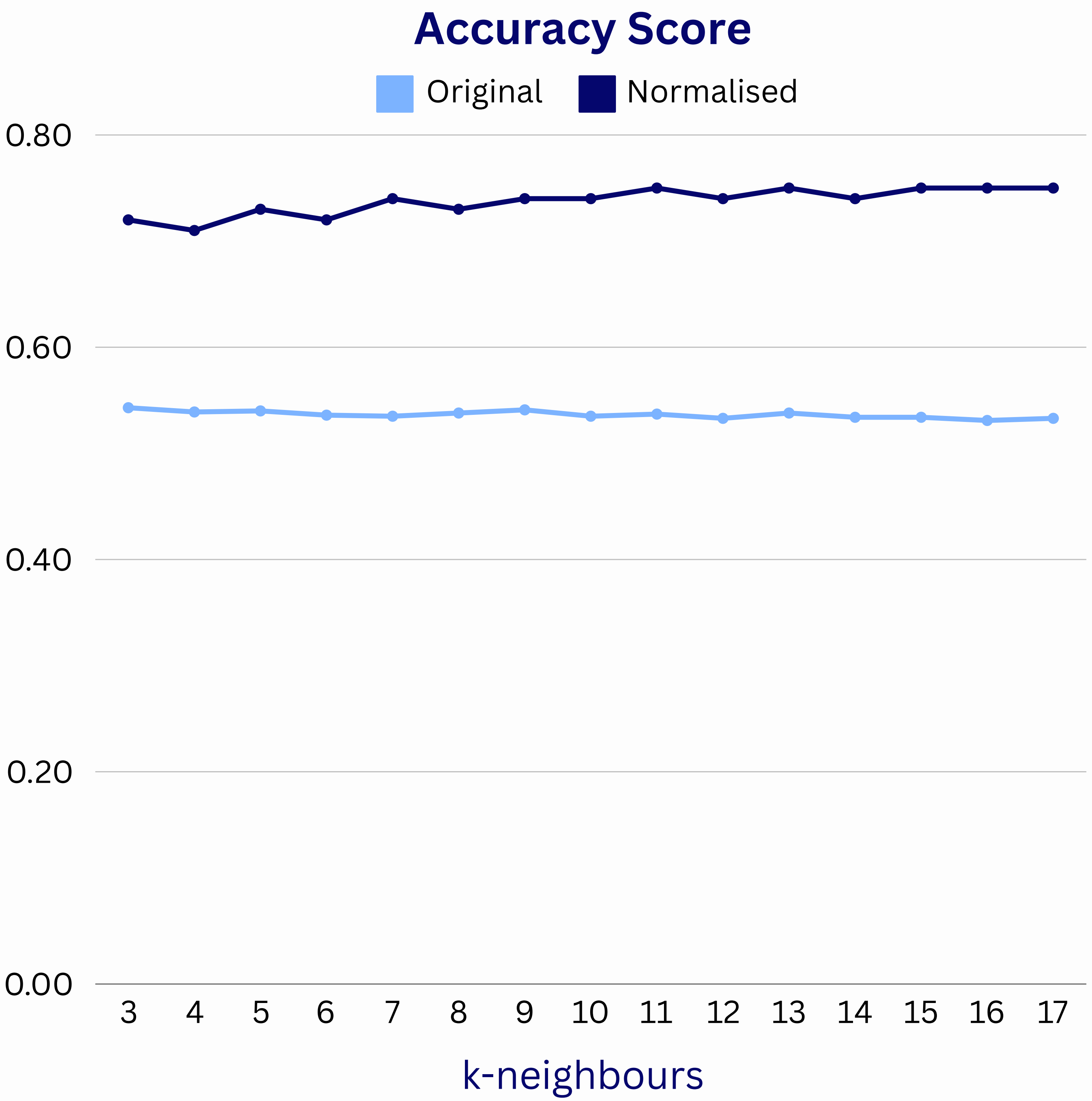


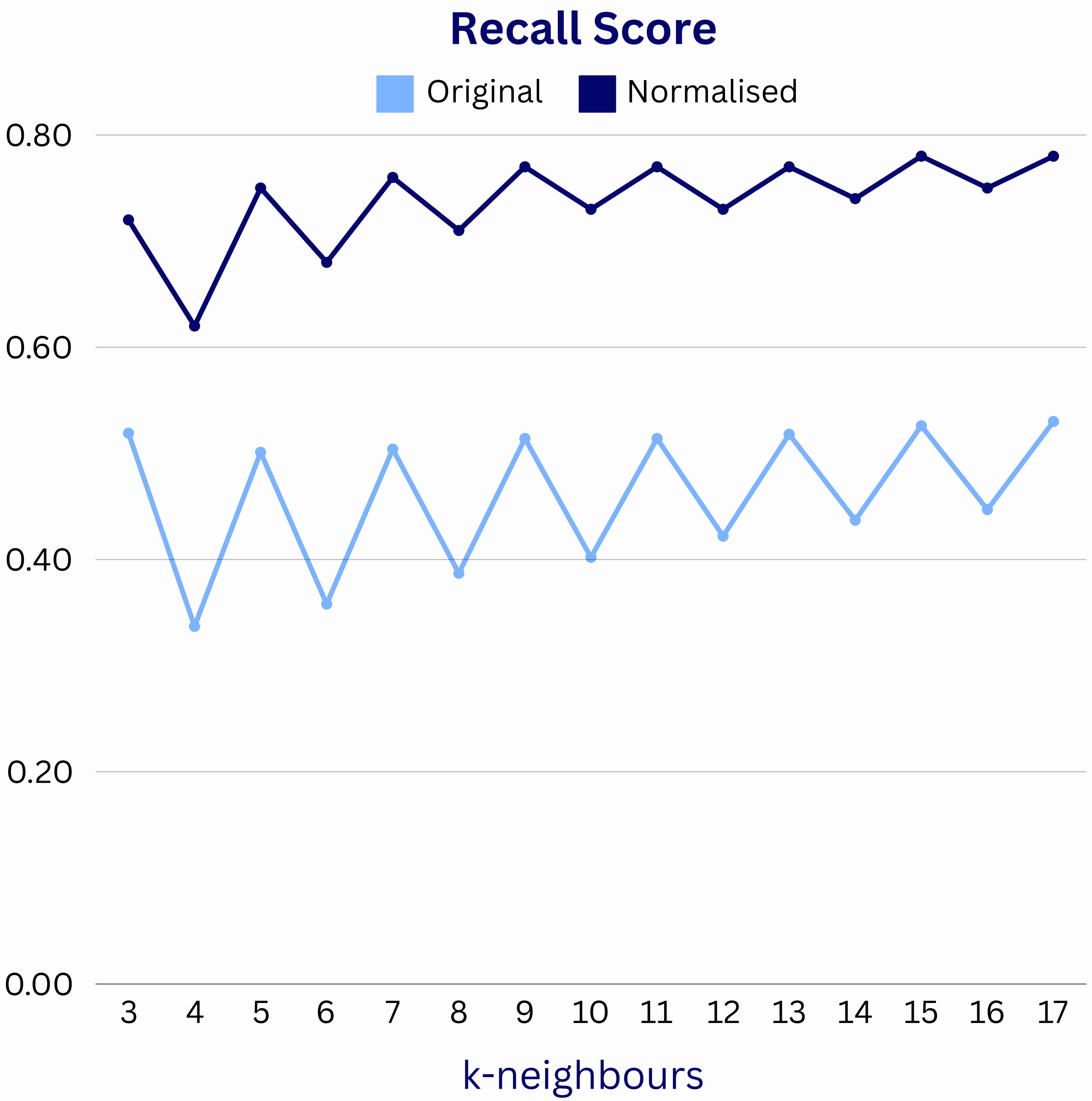


Data Normalisation

01

We employed min-max scaling on the resampled data to improve the machine learning models

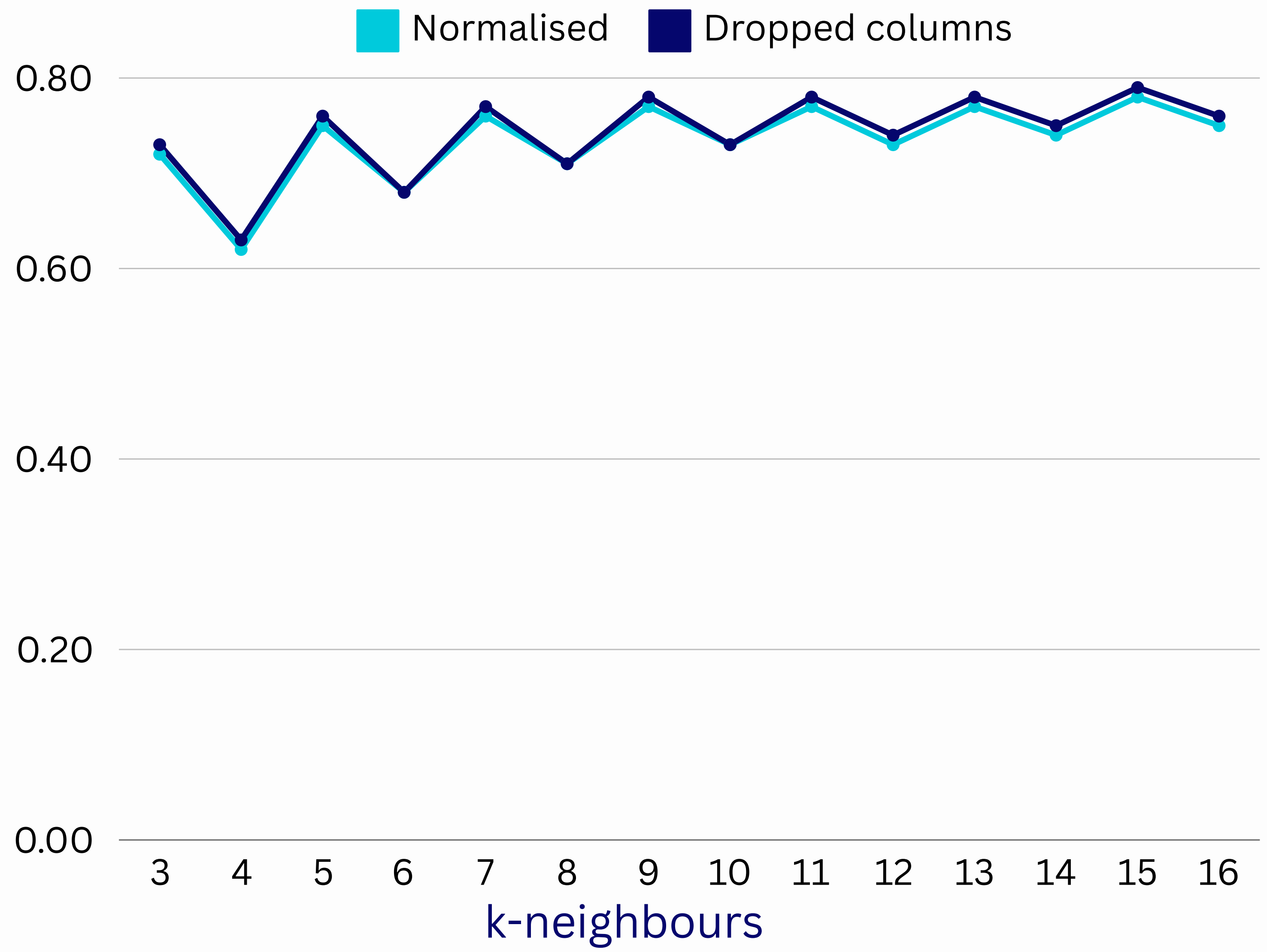


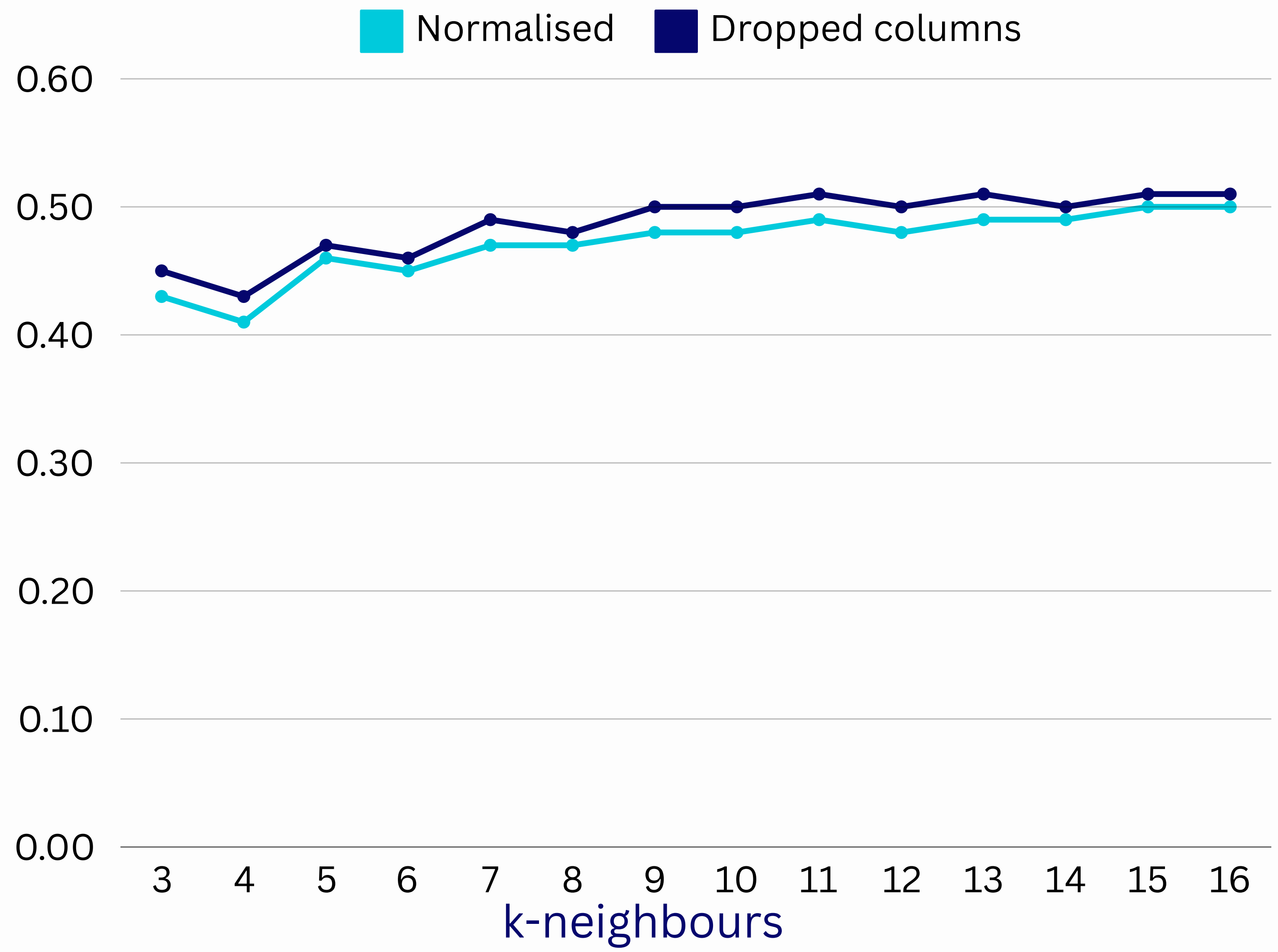


Feature Selection

01 After checking the correlation matrix for our model features we discovered some were multi-correlated and others were not contributing to the model performance

02 Removing these columns made for more efficient analysis, and improved performance metrics


Recall Score


Cohen’s Kappa Score

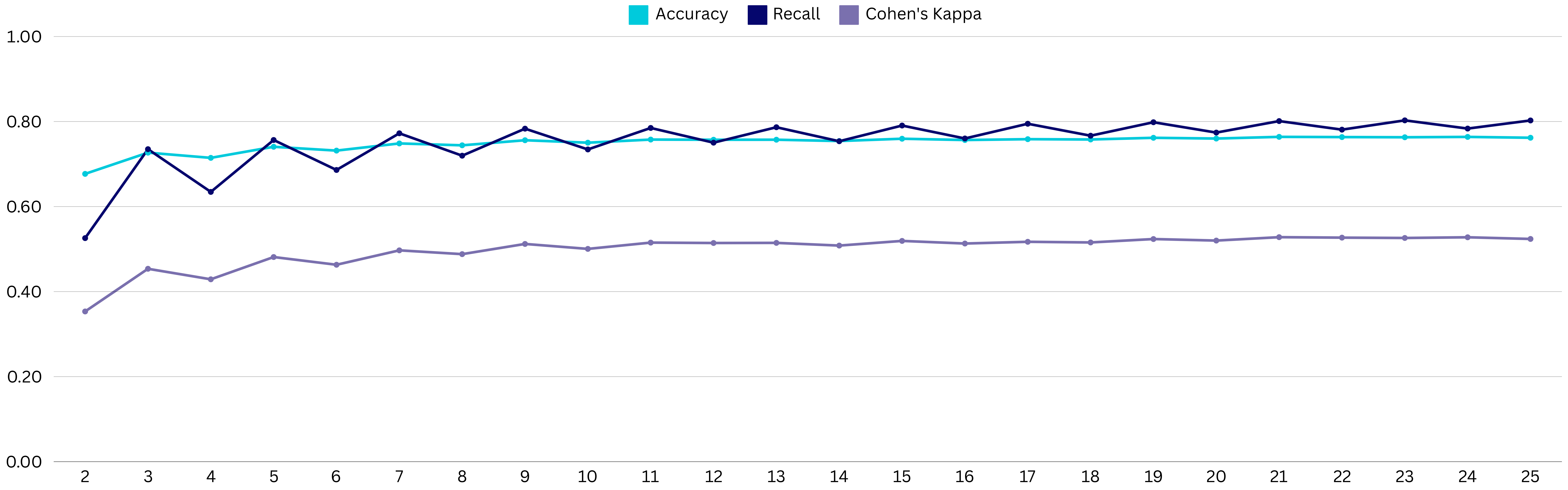
Model Building and Evaluation

K-Nearest Neighbours

We optimised for **Recall score** using the parameters in the **KNN model** for the resampled, normalised data for **n_neighbours** 2-25 and **weighting** either uniformly or by distance to find the best superparameters.

Best Configuration:

Number of neighbours **= 23**
Weight: **uniform**
Accuracy: **80.3%**
Recall**: 74.3%**
Cohen's Kappa: **0.526**



Evaluation of other models:

Model	Accuracy	Cross validation	ROC-AUC
Decision tree	0.76	0.75	0.82
Random Forests	0.76	0.75	0.82
Gradient Boosting(XGB)	0.77	0.77	0.85
Adaptive Boosting	0.77	0.77	0.85
Logistic	0.77	0.77	0.85
Voting (hard)	0.78	0.77	-
Voting (soft)	0.76	0.74	-

Comparison report

Model (0/1)	Precision	Recall	F1 score
Decision tree	0.77/0.75	0.74/0.78	0.75/0.77
Random Forests	0.77/0.74	0.73/0.79	0.75/0.76
Gradient Boosting(XGB)	0.80/0.75	0.73/0.81	0.76/0.78
Adaptive Boosting	0.79/0.76	0.75/0.79	0.77/0.78
Logistic	0.79/0.76	0.75/0.80	0.77/0.78
Voting (hard)	0.80/0.76	0.74/0.81	0.77/0.78
Voting (soft)	0.76/0.75	0.74/0.77	0.75/0.76

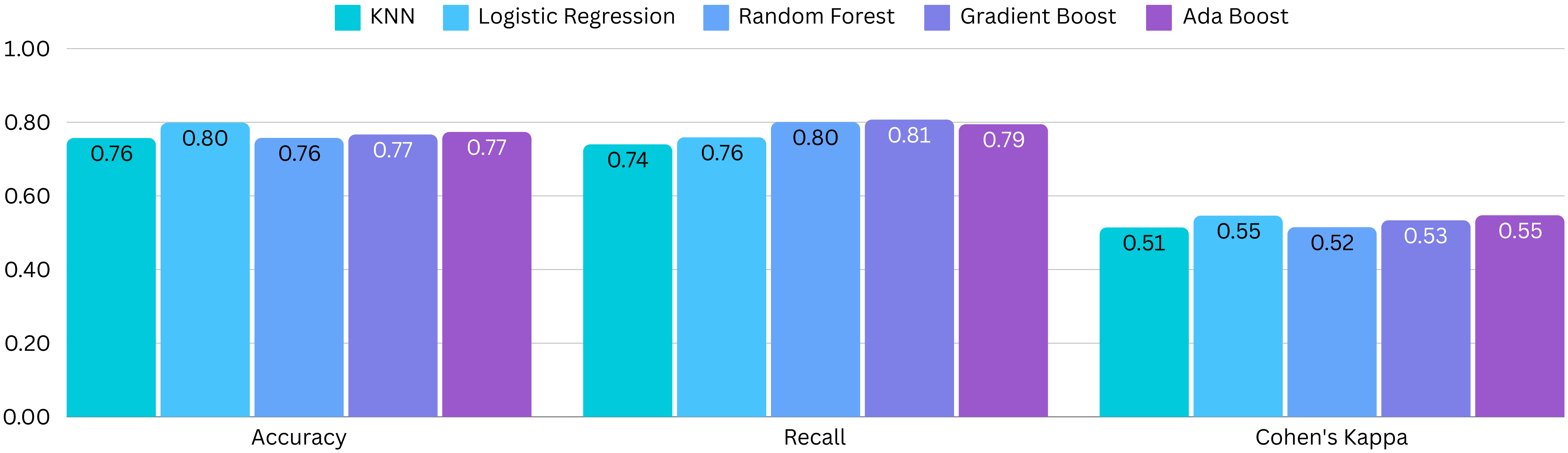
Model Building and Evaluation

Testing other models

We tested several other models to see the difference in performance

The ensemble methods were more effective than the KNN and Logisitc Regression model, so we took these and optimised further.

- **KNN**: As before, with k=23, and uniform weights
- **Log. Regression**: Default parameters
- **Random Forest**: 100 estimators
- **Gradient Boost**: 50 estimators, max. depth of 10
- **Adaptive Boost**: 100 estimators



Model Optimisation

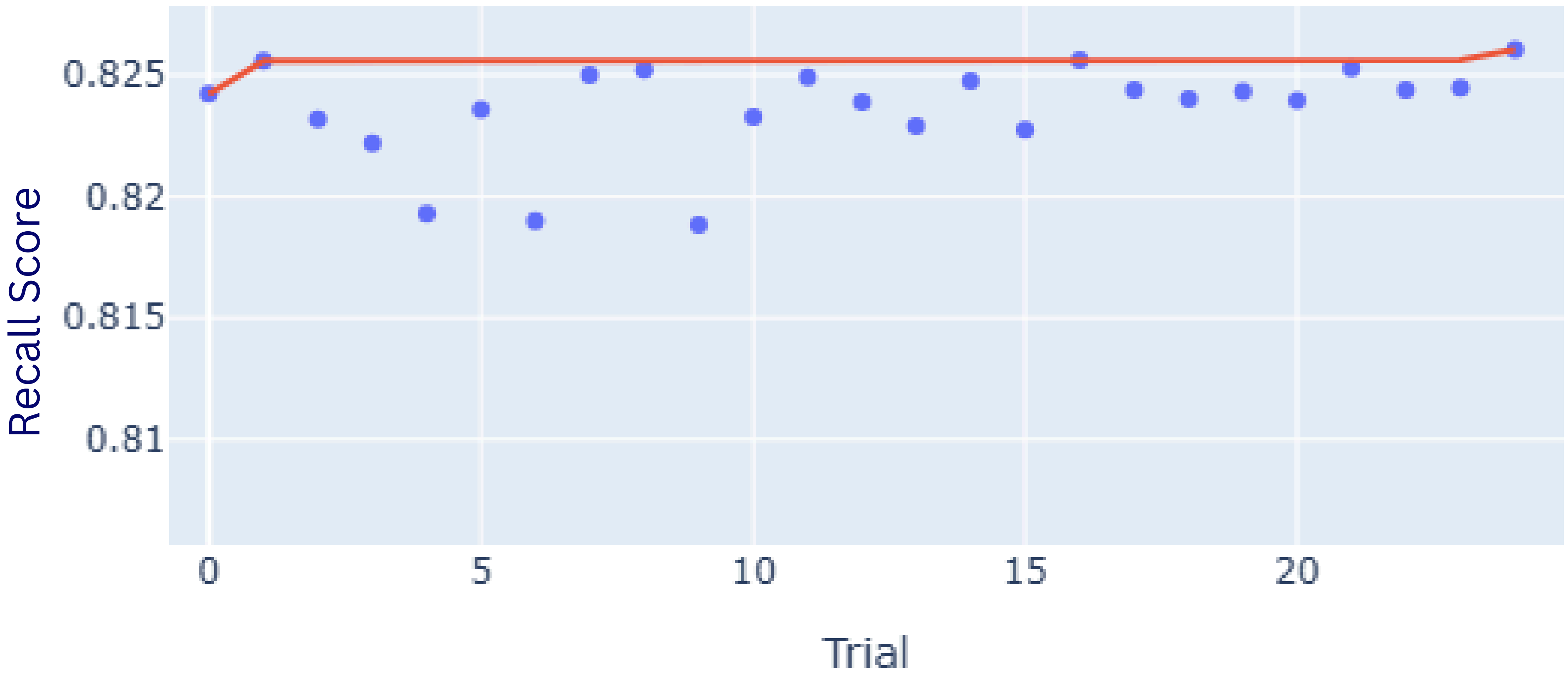
Random Forest

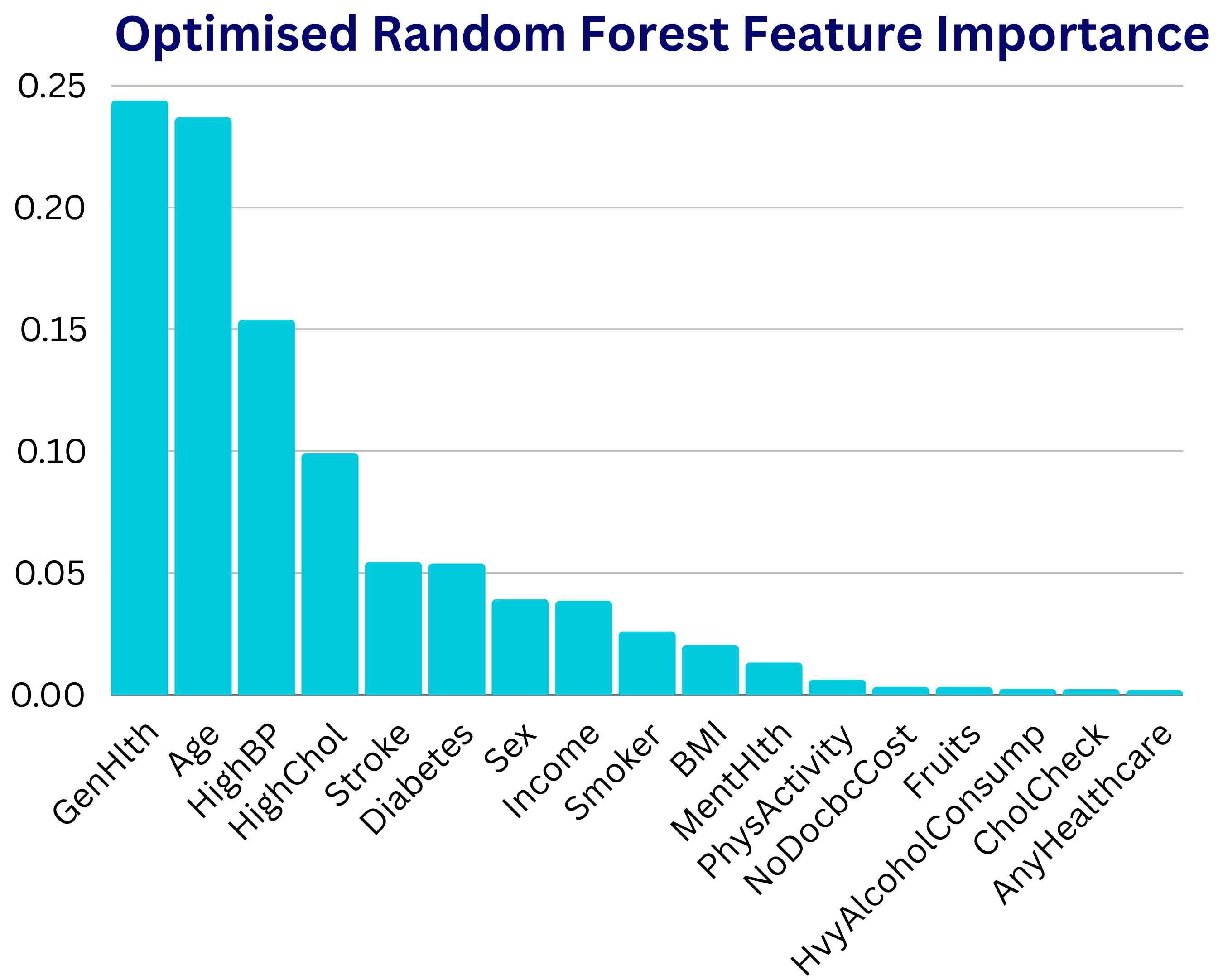
We optimised the Random Forest model to the best Recall score using a Bayesian optimisation on several hyperparameters, with 24 trials.

Whilst the optimisation significantly improved recall performance for the training data, it decreased the test result. Suggesting over-fitting: Accuracy: **82.6%**, Recall: **75.0%**, Cohen’s Kappa: **0.553**

Best Configuration:

n_estimators: **141**
Max. depth: **9**
Min. Samples split: **6**
Min. Samples Leaf: **3**
Max. features: **sqrt**
Train recall score: **82.6%**





Model Optimisation

Adaptive Boosting

We also optimised an Adaptive Boosting model, with a Decision Tree Base, to the best Recall score using a Bayesian optimisation with 29 trials.

Whilst the training data recall score was stronger than the Random Forest model, it was lower in the test data (vs. 75.0%), suggesting over-fitting: Accuracy: **83.0%**, Recall: **73.4%**, Cohen's Kappa: **0.531**

The optimisation also resulted in many parameters not contributing to the model, even though we know many of these were correlated with the target.

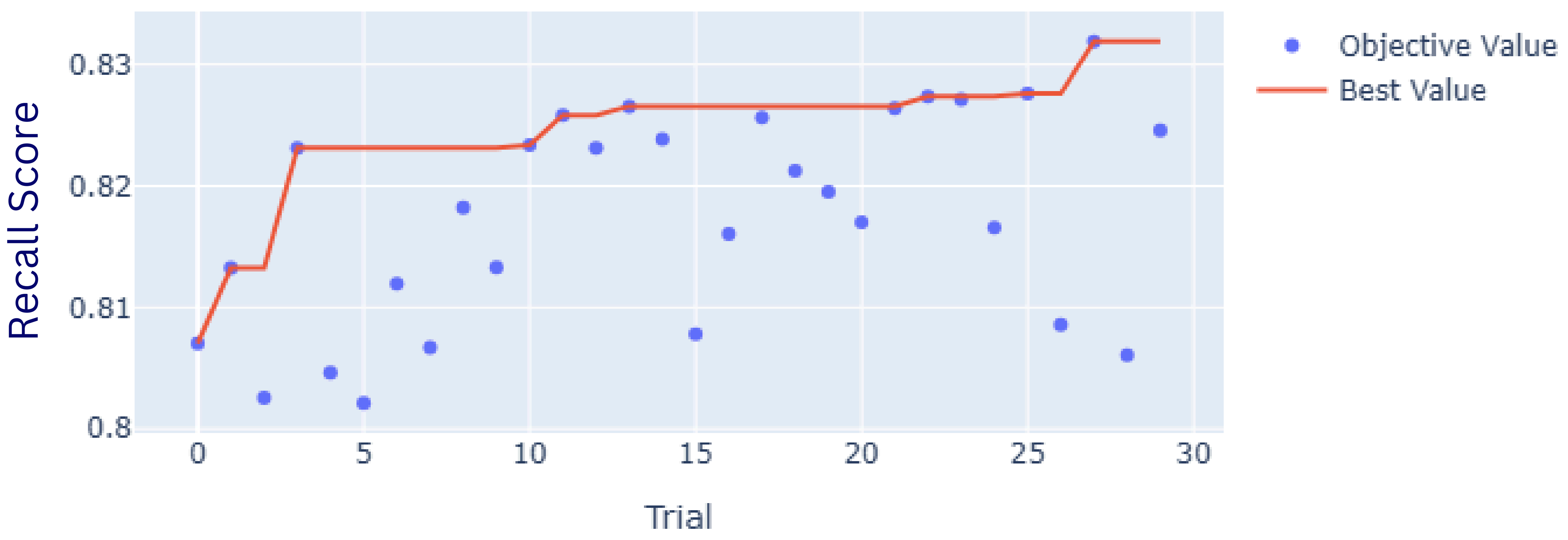
Best Configuration:

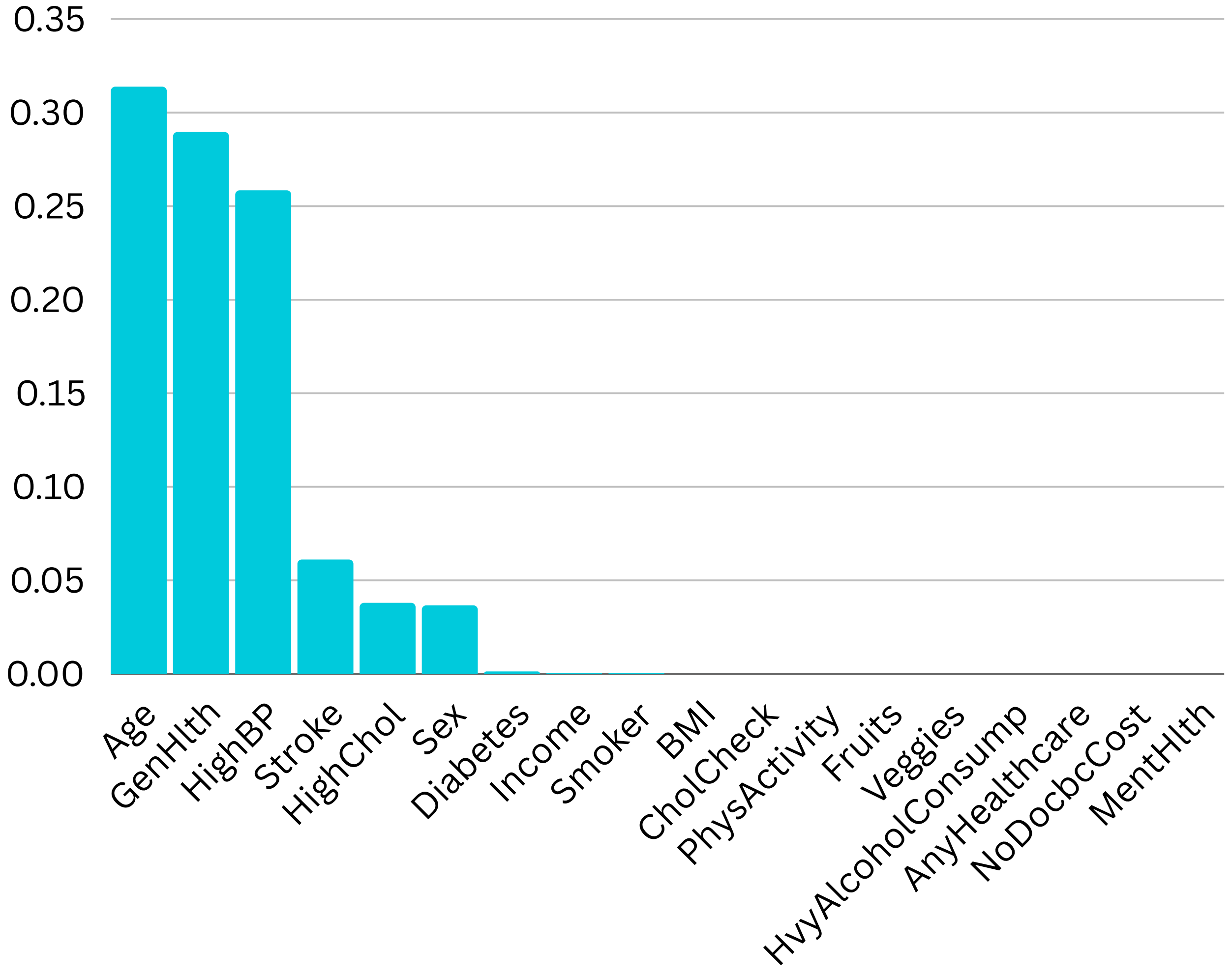
Number of estimators: **70**

Learning rate: **0.0147**

Max depth: **4**

Train Recall score: **83.2%**





Real-World Application & Suggestions

- By employing these models, you could input your personal data and medical test results to determine if you are at high risk of heart disease.

- We can see that key factors that make people more prone to heart conditions are: other health issues, old age, high blood pressure, high cholesterol, diabetes (especially type 2), having had a stroke and smoking.

Key Insights

- **RECALL** is crucial in medical diagnosis as it measures the model's ability to correctly identify true positives (e.g., patients with a condition), minimizing the risk of false negatives, which could result in undiagnosed cases and delayed or missed treatment

- All ensemble methods were effective for high recall models, at around 80%.
    - **Adaptive Boost** had the greatest **Cohen's Kappa** value of 0.55
    - However, **Random Fores**t had stronger **Recall** (0.80 vs 0.79)

- The Bayesian-optimized AdaBoost and Random Forest models achieved strong recall for the training data, this was at the expense of test results. This suggests the models were over-fitted to the training data

Thank You

GitHub