Introduction:

In this project report, we employed supervised and unsupervised machine learning methodologies on a heart disease dataset sourced from Kaggle. The dataset encompasses crucial attributes including Resting Blood Pressure, Cholesterol levels, Age, Sex, and Chest Pain Type, among others. The report delves into comprehensive analysis and insightful visualizations. Through the development of precise predictive models, medical practitioners may identify patients more susceptible to heart disease, facilitating prompt interventions and better patient outcomes.

Analysis:

Supervised Learning:

In our pursuit to predict the likelihood of heart disease, we focused on a key variable indicating whether someone developed the condition or not. We started by using supervised learning techniques to improve our model's performance. Initially, we tested several algorithms like KNN, SVM, Naive Bayes, Logistic Regression, Decision Trees, and Random Forests with the raw data. Later, we cleaned the data, selected important features, and ran the algorithms again to get better accuracy.

When applying the K Nearest Neighbors (KNN) algorithm with 30 nearest neighbors, we achieved an accuracy of 70.11%. Upon setting the value of k to 3, the accuracy remained consistent, indicating that k=3 is well-suited for this dataset. KNN was chosen for its simplicity and effectiveness in classifying data points based on their proximity to neighboring points in the feature space.

```
> # Print accuracy scores for each k
> print(accuracy_scores)
 [1] 0.7249576 0.6604414 0.7011885 0.6672326 0.6604414 0.6417657 0.6298812 0.6332767 0.6451613
[10] 0.6536503 0.6332767 0.6519525 0.6451613 0.6570458 0.6519525 0.6519525 0.6587436 0.6689304
[19] 0.6672326 0.6485569 0.6519525 0.6570458 0.6621392 0.6859083 0.6740238 0.6706282 0.6587436
[28] 0.6536503 0.6570458 0.6655348
```

Fig. 1 Accuracies for each value of k

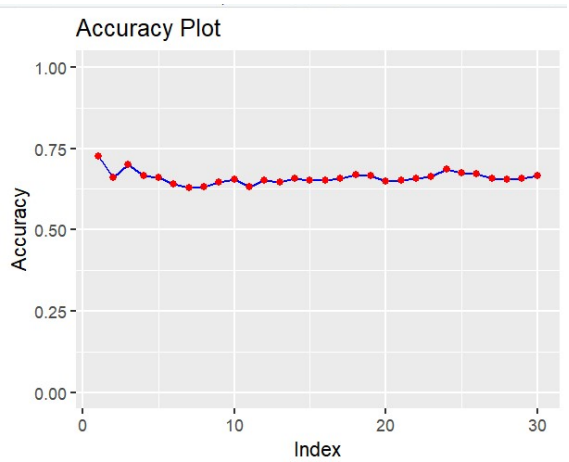


Fig. 2 Accuracy Plot for knn

To determine the likelihood that a given instance falls into a specific class—in this case, the possibility of heart disease—logistic regression estimates that probability. The logistic regression model displayed a testing accuracy of 75.55% after being trained on the training set of data. However, SVM resulted in an accuracy of 65.70%.

```
> logit_model <- glm(HeartDisease ~ ., data=train_data, family=binomial)
>
> # Evaluating the Model
> predicted <- predict(logit_model, newdata=test_data, type="response")
> predicted_class <- ifelse(predicted > 0.5, 1, 0)
>
> # evaluation metrics
> confusion_matrix <- table(test_data$HeartDisease, predicted_class)
> accuracy <- sum(diag(confusion_matrix)) / sum(confusion_matrix)
> cat("Accuracy:", accuracy, "\n")
Accuracy: 0.7555178
```

Fig. 3 Accuracy for Logistic Regression

```
> confusion_matrix
      y_pred
y_test   0   1
     0 169  93
     1 109 218
> accuracy <- sum(diag(confusion_matrix)) / sum(confusion_matrix)
> print(accuracy)
[1] 0.6570458
```

Fig. 4 Accuracy for SVM

Next, we applied Naive Bayes as it leverages conditional probability for event prediction which returned an accuracy of 73.01% on the test dataset. After that, we used the decision tree algorithm which culminates in 70.63% accuracy. A decision tree plot is shown below:

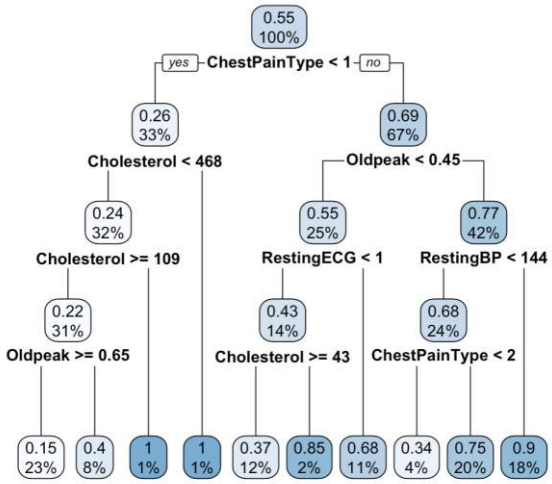
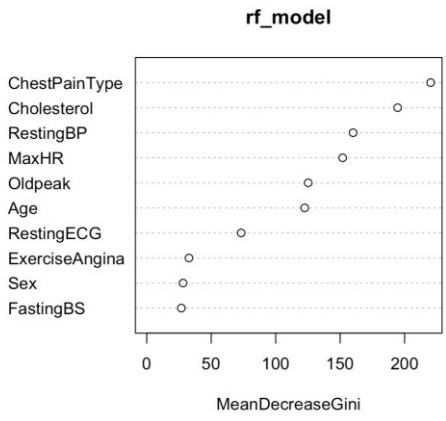


Fig. 5 Decision Tree Result

In order to improve the analysis further, we used the 500-tree Random Forest approach to determine the best model for predicting heart disease. The Random Forest model achieved 91% accuracy and the visualisation shows that the most significant factor in predicting heart disease is the type of chest pain, which is followed by cholesterol, resting blood pressure, and maximum heart rate.




```
> accuracy <- confusion_matrix$overall['Accuracy']
> print(paste("Accuracy:", round(accuracy * 100, 2), "%"))
[1] "Accuracy: 91 %"
> varImpPlot(rf_model)
```

Fig. 6 Random Forest Results

With all the results of the above implemented algorithms, we further cleaned the data by identifying outliers and removing them and selecting the best features based on the correlation plot for enhancing the model performance. The results of each model after applying these steps are tabulated below:

	Initial Acc ıracy	**Accuracy after Cleaning**	**Accuracy after Feature Selection**
KNN	0.70118 35	0.6672794	0.7657046
SVM	0.65704 58	0.7261029	0.7317487
Logistic Regression	0.75551 78	0.71875	0.7555178
Naïve Classifier	0.70628 l8	0.6691176	0.7300509
Decision Tree	0.70628 l8	0.6691176	0.7300509
Random Forest	0.8695	0.8713	0.9010

Table. 1 Results of supervised learning algorithms

Unsupervised Learning:

We attempted to identify clusters of patients based on similar symptoms using unsupervised learning. We employed k-means clustering to identify these clusters and analyzed the characteristics of each cluster. Our goal was to discern patterns in the data.

We decided to use the k-means clustering method, a form of vector quantization. This approach partitions observations into k clusters, where each observation belongs to the cluster with the nearest mean cluster center or centroid. Using this algorithm, we were able to identify clusters of patients with similar symptoms and study them further.

We began by cleaning the data and then executed the algorithm with varying numbers of centers, from 1 to 20. For each iteration, we recorded the total within-cluster sum of squares. This sum represents the squared distances of each data point to its respective cluster centroid. Essentially, it measures the compactness of the clusters, reflecting how well the data points are grouped within their assigned clusters. The results of our algorithm are depicted in the following image.

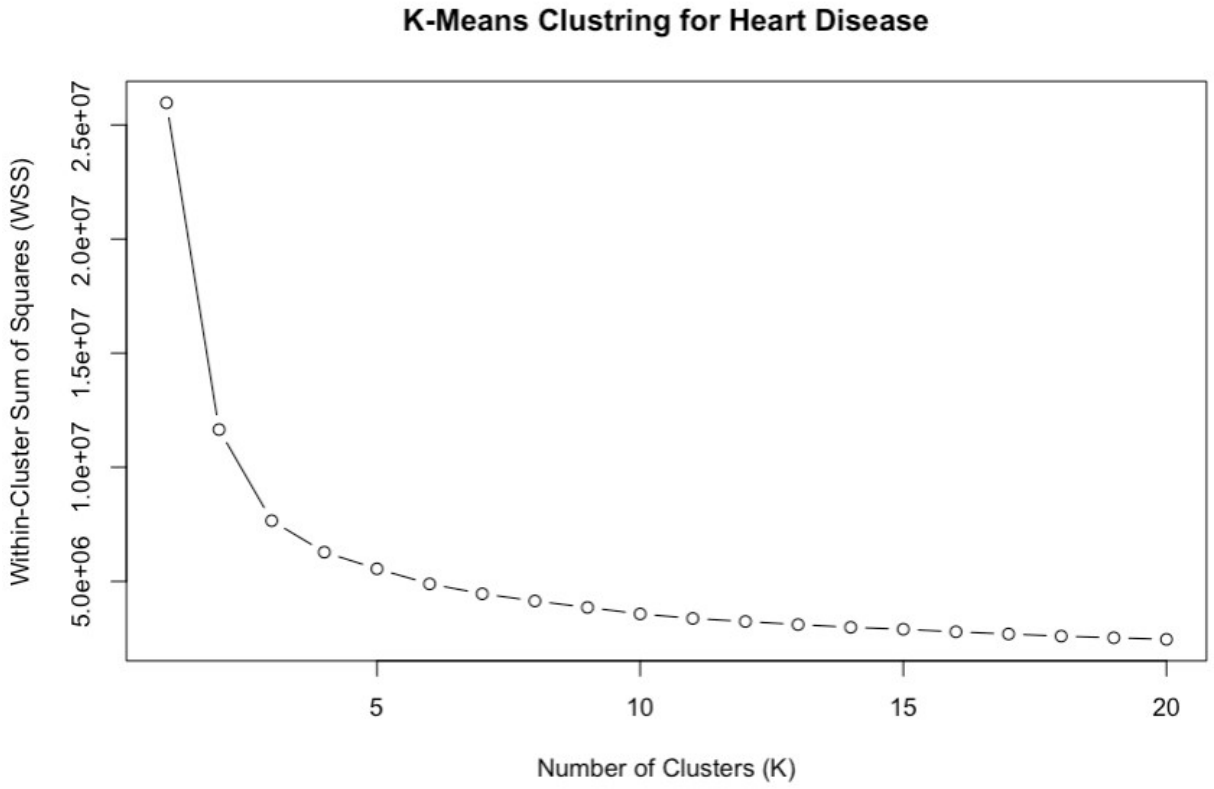


Fig.7 KMeans clustering Results

Our results reveal that the within-cluster sum of squares (WSS) significantly improved as we increased the number of centers from 1 to 4. However, beyond 4 centers, the improvement slowed down. Consequently, we opted to proceed with 3 clusters of patients for further analysis. Using the selected model, we assigned each data point to its respective cluster using k-means clustering and added a new data column indicating the cluster for each observation. Subsequently, we created a summary for each cluster and conducted a detailed study of their characteristics.

	Group.1	Age	Sex	ChestPainType	RestingBP	Cholesterol	FastingBS	RestingECG	MaxHR	ExerciseAngina	Oldpeak	HeartDisease
1	1	51.53333	0.5909091	1.342424	159.1091	462.4182	0.4242424	0.8878788	149.1545	0.4545455	2.487273	0.8090909
2	2	52.02683	0.7697466	1.341282	133.2876	208.7437	0.1535022	0.5901639	146.6073	0.3517139	1.280999	0.4910581
3	3	52.68356	0.7390057	1.202677	139.4120	299.5822	0.2179732	0.5994264	142.2629	0.4646272	1.710038	0.4617591

Fig. 8 Summary of Clusters

When we created clusters from the data, we uncovered a significant and insightful pattern. Cluster 1 showed a striking characteristic: nearly 89% of patients in this cluster were at high risk of heart

disease. The other two clusters contained almost an equal number of patients prone to heart disease.

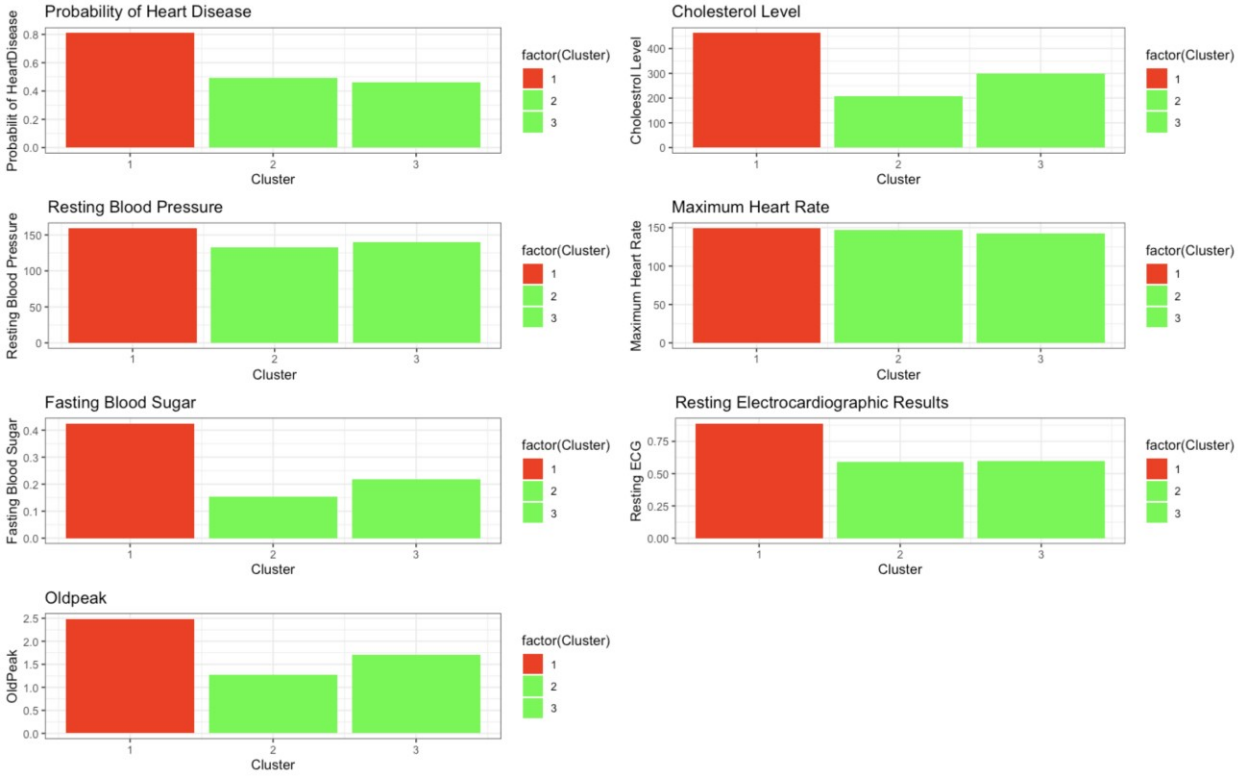


Fig. 9 Visualization of clusters

Cluster 1 provided a clear picture of symptoms in patients that could lead to heart disease. According to our results, cholesterol levels, fasting blood sugar levels, resting ECG, and old peak were crucial indicators of heart disease.

In the other two groups, cholesterol levels averaged around 208. However, for individuals with heart disease, this shot up to an average of 462. Similarly, fasting blood sugar levels reaching 42 were a significant concern for patients.

Doctors can preliminarily identify heart disease using resting ECG and old peak measurements. A normal range for ECG is around 59-60. If it jumps to 85 or higher, it signals an emergency. Likewise, if angina, a measure of pain, doubles from the normal level during a fitness test of the heart, or if the old peak observed during the exercise stress test doubles from normal, it serves as a significant indicator of heart disease.

Recommendations from our study

For Patient

- Regularly assess maximum heart rate and resting blood pressure, with particular concern if blood pressure is around 160 mmHg and maximum heart rate reaches 150 bpm, consult the doctor and get following test

For Doctor

- Monitor cholesterol levels closely, especially if they exceed 208 mg/dL, as elevated levels are associated with a higher risk of heart disease.
- Pay attention to fasting blood sugar levels, particularly if they reach 42 mg/dL, as this could indicate potential heart issues.
- Keep track of resting ECG results, as readings above 85 bpm may signal an emergency situation.

Conclusion

The draft report presents a comprehensive analysis of heart disease prediction using supervised and unsupervised machine learning techniques applied to a dataset sourced from Kaggle. Supervised learning algorithms such as KNN, logistic regression, and random forest yielded promising results in predicting heart disease likelihood, with the latter demonstrating the highest accuracy of 91%. Unsupervised learning through k-means clustering uncovered distinct patient clusters, with one cluster showing a high risk of heart disease. Crucial indicators identified include cholesterol levels, fasting blood sugar levels, resting ECG, and old peak measurements. Recommendations based on the study emphasize regular monitoring of key health metrics for both patients and doctors, aiding in the early identification and management of heart disease risks. Overall, the findings underscore the potential of machine learning in enhancing heart disease prediction and management strategies, thereby improving patient outcomes and healthcare delivery.