جامعة بيرزيت

BIRZEIT UNIVERSITY

FACULTY OF ENGINEERING AND TECHNOLOGY
DEPARTMENT OF COMPUTER ENGINEERING

MACHINE LEARNING AND DATA SCIENCE
ENCS5341

Title: Heart Failure Prediction

Prepared by: Shereen Ibdah – Jehad Hamayel
IDs: 1200373 – 1200348

Instructor: Dr. Yazan Abu Farha

Section: 1,2
January 24, 202

Table of Contents

Table of Figures

Table of Tables

Introduction

Heart failure is a significant global health concern, with a high mortality rate. The percentage of deaths caused by heart and vascular diseases is approximately 33%, which translates to around nineteen million deaths each year. Therefore, for prompt intervention and better patient outcomes, heart failure early detection and prediction are essential.

The application of artificial intelligence (AI) technology in the medical field has revolutionized the field, especially in terms of heart failure prediction and prevention. AI classifiers have transformed the field of cardiac treatment because of their ability to evaluate and understand a wide range of variables. These advanced algorithms look at a number of attributes to assess a person's risk of heart failure. This opens the door for individualized treatment regimens in addition to facilitating early identification and action. It is anticipated that these AI systems will achieve previously unheard-of levels of prediction accuracy and efficiency in treating heart health as they continue to learn from and adapt to enormous datasets, completely changing the field of cardiovascular medicine in the process.

In the analysis of the heart dataset, Random Forests (RF) and Support Vector Machines (SVM) have been chosen as the primary classifiers due to their suitability for complex medical datasets. Their performance is being evaluated using key metrics such as Recall, Precision, F1-score, and Accuracy. The identification of true positive cases is facilitated by Recall, while Precision ensures the accuracy of positive predictions. The F1-score, combining both Recall and Precision, provides a balanced measure, and Accuracy offers an overall view of the model's performance. These metrics are instrumental in assessing the accuracy of the models and their capability to correctly detect heart disease cases, which is crucial for facilitating early intervention and patient care.

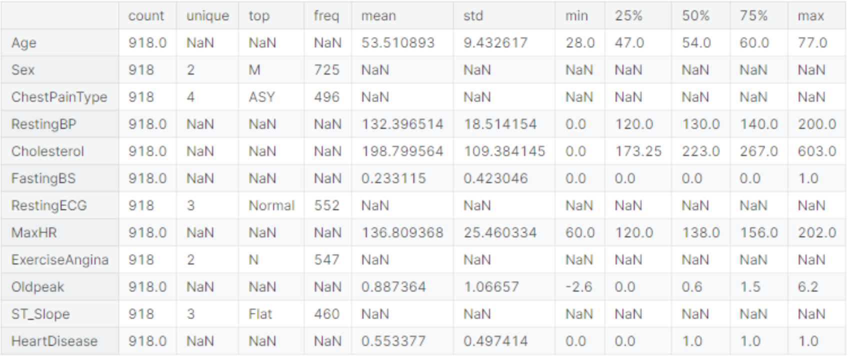
Dataset

This dataset was created by combining different datasets that were already available independently but had not been combined before. In this dataset, five heart datasets are combined over 11 common features, which makes it the largest heart disease dataset available so far for research purposes. The five datasets used for its curation are:

Cleveland: 303 observations, Hungarian: 294 observations, Switzerland: 123 observations, Long Beach VA: 200 observations, Stalog (Heart) Data Set: 270 observations

Total: 1190 observations, Duplicated: 272 observations, Final dataset Size: 918 observations [1]

Our dataset's exploratory data analysis (EDA) includes quantitative measures and visualizations, as shown in the below figure:



	count	unique	top	freq	mean	std	min	25%	50%	75%	max
Age	918.0	NaN	NaN	NaN	53.510893	9.432617	28.0	47.0	54.0	60.0	77.0
Sex	918	2	M	725	NaN	NaN	NaN	NaN	NaN	NaN	NaN
ChestPainType	918	4	ASY	496	NaN	NaN	NaN	NaN	NaN	NaN	NaN
RestingBP	918.0	NaN	NaN	NaN	132.396514	18.514154	0.0	120.0	130.0	140.0	200.0
Cholesterol	918.0	NaN	NaN	NaN	198.799564	109.384145	0.0	173.25	223.0	267.0	603.0
FastingBS	918.0	NaN	NaN	NaN	0.233115	0.423046	0.0	0.0	0.0	0.0	1.0
RestingECG	918	3	Normal	552	NaN	NaN	NaN	NaN	NaN	NaN	NaN
MaxHR	918.0	NaN	NaN	NaN	136.809368	25.460334	60.0	120.0	138.0	156.0	202.0
ExerciseAngina	918	2	N	547	NaN	NaN	NaN	NaN	NaN	NaN	NaN
Oldpeak	918.0	NaN	NaN	NaN	0.887364	1.06657	-2.6	0.0	0.6	1.5	6.2
ST_Slope	918	3	Flat	460	NaN	NaN	NaN	NaN	NaN	NaN	NaN
HeartDisease	918.0	NaN	NaN	NaN	0.553377	0.497414	0.0	0.0	1.0	1.0	1.0

Figure 1:Quantitative Measures

The dataset includes the following variables with **Quantitative Measures** (descriptive statistics):

Age	Continuous, ranging from 28 to 77 years old, with average approximately 53.5 years. with a standard deviation of 9.43. This identifies a middle-aged demographic.
Sex	Categorical, with 2 unique values (presumably Male and Female).
ChestPainType	Categorical, with 4 unique types.
RestingBP (Resting Blood Pressure)	Continuous, with some entries as low as 0, which might indicate missing or incorrect data or noise. The mean resting blood pressure is around 132.4 mmHg, which is within normal range, however variability implies there are individuals with high blood pressure.
Cholesterol	Continuous, ranging from 0 to 603, with 0 possibly indicating missing data or noise. The average cholesterol level is about 198.8 mg/dl, but some people have high cholesterol levels (more than 240 mg/dl is considered high).
FastingBS (Fasting Blood Sugar)	Binary (0 or 1), which indicates if fasting blood sugar is above 120 mg/dl. Around 23.3% of participants have fasting blood sugar above the typical threshold of 120 mg/dl, which is a frequent diagnostic threshold for diabetes.
RestingECG (Resting Electrocardiogram results)	Categorical, with 3 unique values.
MaxHR (Maximum Heart Rate achieved)	Continuous, ranging from 60 to 202. The average maximum heart rate recorded is roughly 136.8 beats per minute. The variability reflects a range of fitness levels or heart function among people.
ExerciseAngina	Categorical, with 2 unique values (Yes or No).
Oldpeak	Continuous, ranging from -2.6 to 6.2, related to ST depression induced by exercise relative to rest. The mean value for the 'Oldpeak' variable is around 0.89, indicating that there is some exercise-induced cardiac stress in the cohort.
ST_Slope	Categorical, with 3 unique values, indicating the slope of the peak exercise ST segment.
HeartDisease	Binary (0 or 1), indicating the presence or absence of heart disease. The dataset shows a nearly equal mix of people with and without heart disease (55.3%).

Table 1: Quantitative Measures Tabel

Missing Data: There are no missing values in any of the variables. But there is noise in the data, a doctor was consulted to know the noise in the data. Box plots are a powerful tool for identifying and effectively detecting noise in data, so it used in the next dissection.

Visualizations:

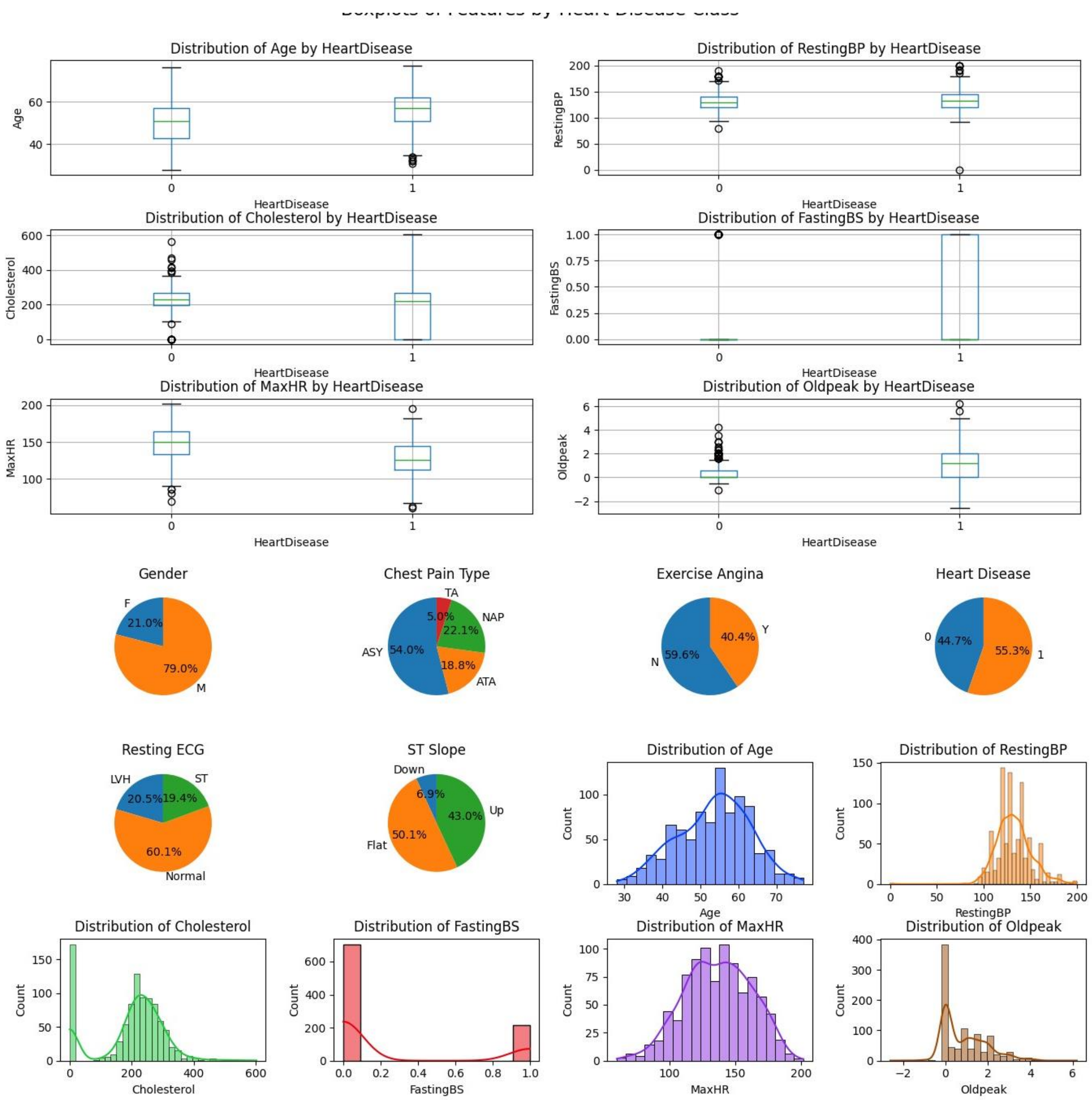


Figure 2: Visualizations

Boxplots by Heart Disease Status:

Age: Individuals with heart disease appear to have a slightly greater median age than those without. This shows that age could be a risk factor for heart disease.

RestingBP (Resting Blood Pressure): The medians for both groups are similar, but people with heart disease have a broader range of readings, indicating greater variability in blood pressure.

Cholesterol: Similar to RestingBP, cholesterol levels are more widely distributed among those with heart disease.

MaxHR (Maximum Heart Rate): People without heart disease have a higher median MaxHR and a more concentrated distribution, indicating better heart rate responses during maximum exertion.

Oldpeak: People with heart disease exhibit higher levels of ST depression (Oldpeak), implying a link between Oldpeak and heart disease.

Pie Charts:

Gender: The dataset's subjects are predominantly male, which is a common trend in heart disease investigations.

ASY (asymptomatic) is the most prevalent type of chest pain associated with heart disease, followed by non-anginal pain (NAP), atypical angina (ATA), and typical angina (TA).

Exercise Angina: A large percentage of people with heart disease have angina while exercising.

ST Slope: The flat ST slope is the most common among people with heart disease, and it is generally interpreted as an indication of an unhealthy heart.

Histograms:

Age: The age distribution is broadly bell-shaped with a little right skew, indicating that the dataset comprises a wide variety of ages but with a higher proportion of older persons.

RestingBP: The distribution of resting blood pressure is broadly normal, with some outliers that could be caused by measurement errors or individuals with extremely high or low blood pressure.

Cholesterol levels are generally distributed with a right skew, meaning that while the majority of people have normal cholesterol levels, some have extremely high levels.

FastingBS: The histogram demonstrates that the majority of people have normal fasting blood sugar levels, but a considerable percentage have increased levels, indicating a risk factor for heart disease.

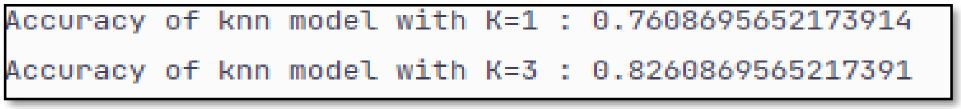
MaxHR: The distribution of maximum heart rate is somewhat left-skewed, indicating that fewer people have low MaxHR, which is generally a good marker of cardiovascular fitness.

Oldpeak: Most people have a low Oldpeak value, but there is a long tail of higher values, indicating that some people suffer from severe ST depression.

Experiments and Results

In the initial phase of the project, the baseline model was evaluated using a simple classification algorithm, namely K-Nearest Neighbors (K-NN). This evaluation was conducted twice, first with K set to 1, and then with K set to 3. The dataset, which contained approximately 918 examples, was divided into training and testing data. Specifically, 80% of the data was allocated to the training set, while the remaining 20% was considered as the testing set. The data splitting process was carried out randomly to ensure that important examples appeared in the training dataset, thereby achieving more accurate results.

Below are the accuracy results obtained when using K-NN on the testing data:

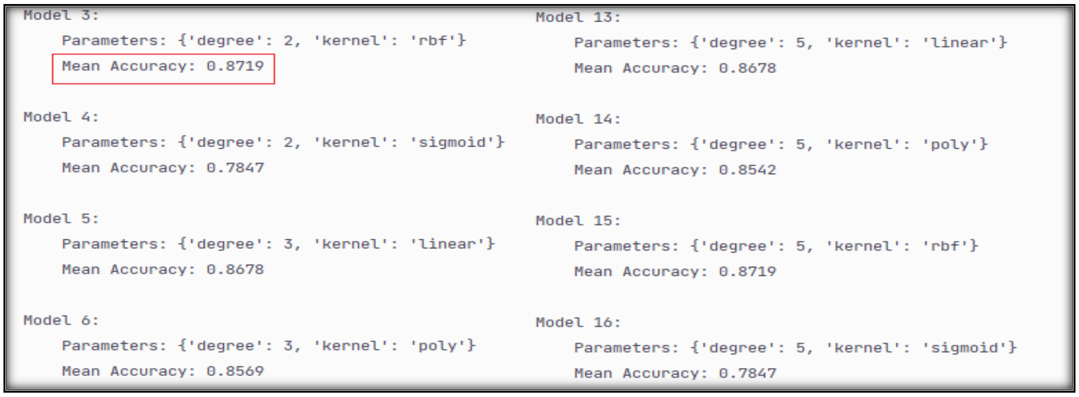


```
Accuracy of knn model with K=1 : 0.7608695652173914
Accuracy of knn model with K=3 : 0.8260869565217391
```

Figure 3: K-NN accuracy

The increase in accuracy from 76.09% (K=1) to 82.61% (K=3) suggests that the K=1 model might have been overfitting, being too sensitive to noise in the dataset. The K=3 model, with its higher accuracy, appears to generalize better, indicating a reduction in overfitting. However, it's still crucial to use additional validation techniques to confirm these findings and ensure the model's robustness.

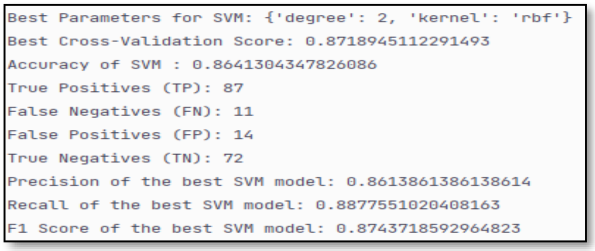
The first model employed was the Support Vector Machine (SVM), a powerful and widely used machine learning algorithm. To select the best values for its hyper parameters "kernel" and "degree" four different values were considered for each parameter: "linear," "poly," "rbf," and "sigmoid" for the kernel, and the degrees 2, 3, 4, and 5 for the "degree" As a result, 16 different models were trained from these combinations. During the training process, the data was split into cross-validation folds, totaling five folds, in order to enhance the accuracy of the trained models



```
Model 3:
    Parameters: {'degree': 2, 'kernel': 'rbf'}
    Mean Accuracy: 0.8719

Model 4:
    Parameters: {'degree': 2, 'kernel': 'sigmoid'}
    Mean Accuracy: 0.7847

Model 5:
    Parameters: {'degree': 3, 'kernel': 'linear'}
    Mean Accuracy: 0.8678

Model 6:
    Parameters: {'degree': 3, 'kernel': 'poly'}
    Mean Accuracy: 0.8569

Model 13:
    Parameters: {'degree': 5, 'kernel': 'linear'}
    Mean Accuracy: 0.8678

Model 14:
    Parameters: {'degree': 5, 'kernel': 'poly'}
    Mean Accuracy: 0.8542

Model 15:
    Parameters: {'degree': 5, 'kernel': 'rbf'}
    Mean Accuracy: 0.8719

Model 16:
    Parameters: {'degree': 5, 'kernel': 'sigmoid'}
    Mean Accuracy: 0.7847
```

Figure 4: hypered parameters selection for SVM

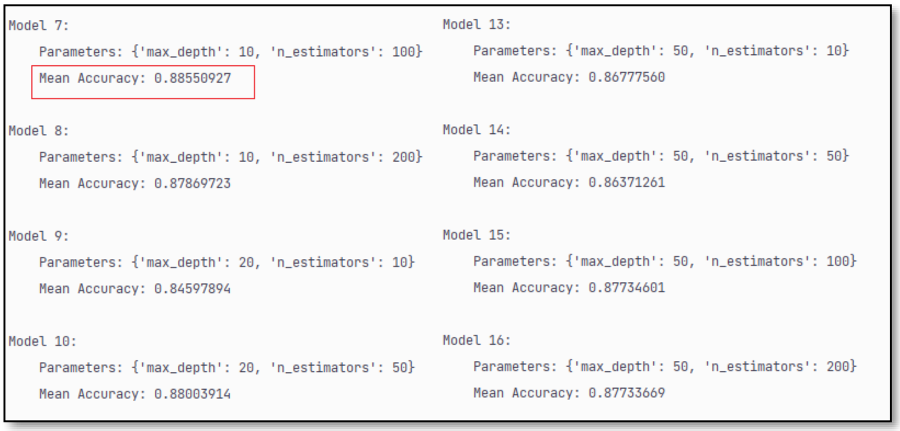
The Support Vector Machine (SVM) model, optimized with parameters {'degree': 2, 'kernel': 'rbf'}, demonstrates robust performance, achieving an accuracy of 86.41% and a cross-validation score of 87.19%. Its precision, recall, and F1 score, at 86.14%, 88.78%, and 87.44% respectively, reflect its balanced predictive capability. These metrics, derived from the confusion matrix elements (TP: 87, FN: 11, FP: 14, TN: 72), underscore the model's effectiveness in classifying instances with a high degree of reliability and precision.



```
Best Parameters for SVM: {'degree': 2, 'kernel': 'rbf'}
Best Cross-Validation Score: 0.8718945112291493
Accuracy of SVM : 0.8641304347826086
True Positives (TP): 87
False Negatives (FN): 11
False Positives (FP): 14
True Negatives (TN): 72
Precision of the best SVM model: 0.8613861386138614
Recall of the best SVM model: 0.8877551020408163
F1 Score of the best SVM model: 0.8743718592964823
```

Figure 5: SVM evaluation metrics

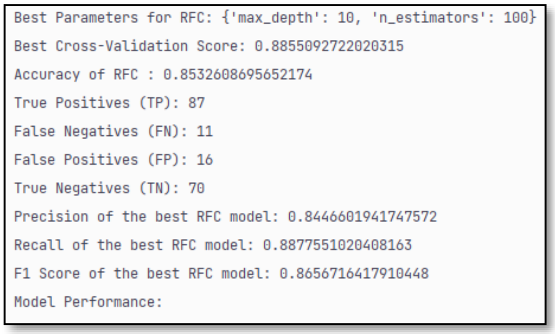
The second model employed was the Random Forest, a popular ensemble learning technique in machine learning. Random Forest combines multiple decision trees to create a robust and accurate predictive model. It mitigates overfitting and provides feature importance scores, making it suitable for various classification, to select the best values for its hyper parameters "n_estimators" and "max_depth" four different values were considered for each parameter10, 50,100 and 200 for the n_estimators, and the max depth 5,10, 20 and 50 for the "max_depth"." As a result, 16 different models were trained from these combinations. The training done in the same criteria where SVM trained “cross validation” with five floods,



```
Model 7:                                                 Model 13:
    Parameters: {'max_depth': 10, 'n_estimators': 100}       Parameters: {'max_depth': 50, 'n_estimators': 10}
    Mean Accuracy: 0.88550927                                Mean Accuracy: 0.86777560

Model 8:                                                 Model 14:
    Parameters: {'max_depth': 10, 'n_estimators': 200}       Parameters: {'max_depth': 50, 'n_estimators': 50}
    Mean Accuracy: 0.87869723                                Mean Accuracy: 0.86371261

Model 9:                                                 Model 15:
    Parameters: {'max_depth': 20, 'n_estimators': 10}        Parameters: {'max_depth': 50, 'n_estimators': 100}
    Mean Accuracy: 0.84597894                                Mean Accuracy: 0.87734601

Model 10:                                                Model 16:
    Parameters: {'max_depth': 20, 'n_estimators': 50}        Parameters: {'max_depth': 50, 'n_estimators': 200}
    Mean Accuracy: 0.88003914                                Mean Accuracy: 0.87733669
```

Figure 6: hypered parameters selection for RFC

The Random Forest Classifier (RFC), fine-tuned with the best parameters {'max_depth': 10, 'n_estimators': 100}, showcases strong predictive performance. It achieves an accuracy of 85.3% and a notable cross-validation score of 88.5%. The precision of the best RFC model is 84.4%, indicating its effectiveness in identifying positive instances. The model's recall is commendably high at 88.7%, demonstrating its strength in capturing most of the actual positive cases. The F1 score, balancing precision and recall, is impressive at 86.5%. These results, derived from confusion matrix values (TP: 87, FN: 12, FP: 16, TN: 70), underline the RFC's capability to deliver reliable and precise classifications.



```
Best Parameters for RFC: {'max_depth': 10, 'n_estimators': 100}
Best Cross-Validation Score: 0.8855092722020315
Accuracy of RFC : 0.8532608695652174
True Positives (TP): 87
False Negatives (FN): 11
False Positives (FP): 16
True Negatives (TN): 70
Precision of the best RFC model: 0.8446601941747572
Recall of the best RFC model: 0.8877551020408163
F1 Score of the best RFC model: 0.8656716417910448
Model Performance:
```

Figure 7: RFC evaluation metrics

The comparison between them and more details in analysis part.

Analysis

After assessing several models, including K-Nearest Neighbors (KNN), Support Vector Machine (SVM), and Random Forest Classifier (RFC), we found that the SVM with a polynomial kernel was the best fit for our heart disease prediction challenge.

Based on the full results in the Experiments and Results section, here's a study of the performance of the KNN, SVM, and RFC models:

K-Nearest Neighbors (KNN):	**K=1**	Achieved an accuracy of 0.7609, indicating that the training data may have been overfitted because only the nearest neighbor was considered.
	K=3	Improved accuracy to 0.8261, highlighting the importance of evaluating many neighbors in order to prevent overfitting.
Support Vector Machine (SVM):	**Best Parameters**	The best setup employed a Radial Basis Function (RBF) kernel with a degree of two.
	Cross-Validation Score	A high score of 0.8719 indicates great generalization ability.
	Accuracy	An outstanding 0.8641 on the test set, demonstrating its ability to handle nonlinear data patterns.
	Precision and Recall	Showed balanced precision (0.8614) and recall (0.8878), resulting in a high F1 score of 0.8744.
	Model Variants	We investigated numerous kernel types and degrees, and the RBF kernel consistently performed well across all degrees.
Random Forest Classifier (RFC):	**Best parameters**	A maximum depth of 10 and 100 estimators.
	Cross-validation Score	The top score was 0.8855, suggesting outstanding training performance.
	Accuracy	0.8533 on the test set, slightly lower than the cross-validation score but still reliable.
	Precision and Recall	Showed balanced precision (0.8447) and recall (0.8878), resulting in a high F1 score of 0.8657.
	Model Variants	Different depths and numbers of estimators were tested, and shallower trees with a reasonable number of estimators performed well.

Table 2: study of the performance of the KNN, SVM, and RFC models

Interesting Findings: The SVM with RBF kernel achieved a good combination of complexity and generalizability, making it especially successful for this dataset's non-linear features. RFC's performance stood out for its feature-important insights and balance of precision and recall. KNN, while simpler, performed well, particularly with a limited number of neighbors, emphasizing the dataset's underlying structure. SVM's performance constancy across kernel types and degrees demonstrates its robustness in this application.

Analysis of the Best Model: SVM

In our project, the Support Vector Machine (SVM) with an RBF kernel and a degree of 2 was the most effective model. The analysis procedure and major conclusions were as follows:

Hyper parameter tuning and model selection: We used GridSearchCV to conduct an exhaustive search for optimal hyper parameters. This technique entails experimenting with several kernel types (linear, manifold, RBF, and sigmoid) and degrees. Combining the RBF kernel with a polynomial score of 2 produced the highest cross-validation score (0.8719) as crossed validation was applied in order to test the largest number of types of data, showing a stronger ability to capture the complexity of the data set.

Evaluation metrics: On test data, the model achieved an accuracy of 0.8641, demonstrating the effectiveness of its classification. We also looked at additional metrics such as precision, recall, and F1 score (for the RFC model) to get a more complete picture of the model's performance, as a lower FP is better for medical data. The high accuracy of the SVM model during the testing phase demonstrated its flexibility and generalizability.

Performance Comparison: The investigation compared the performance of the SVM model to that of other machine learning models, specifically KNN and RFC. While KNN and RFC produced impressive results, SVM stood out, particularly when dealing with non-linear patterns, which is a prevalent difficulty in medical datasets.

Error Analysis: By studying the SVM model's confusion matrix, we found 87 true positives, 72 true negatives, 14 false positives, and 11 false negatives. This provided insights into the types of errors made by the model, as well as its strengths and shortcomings in predicting various classes.

Conclusions and Discussion:

In conclusion, the success of the SVM model in this context has been demonstrated by its suitability for classification tasks in medical datasets, which often include non-linear and complex patterns. Future work could explore the integration of more advanced feature selection techniques and the potential benefits of clustering methods to further enhance model performance.

We noted the importance of adjusting the hyper parameters, as the project emphasized the importance of carefully adjusting the hyper parameters. GridSearchCV was effective in identifying the ideal set of parameters for the SVM model and the rest of the models. The metrics and model evaluation were important in being a benchmark in various evaluations, including accuracy, precision, recall, and F1 score, a comprehensive overview of the model's performance, indicating its strengths in correctly identifying heart diseases and reducing false positives and negatives.

There are some limits on the Model and Evaluation Metrics:

Model Limitations	**KNN**	**KNN is scale-sensitive and can become computationally expensive as dataset size increases. It also difficulties with multidimensional data.**
	SVM	**While powerful, SVMs can be computationally costly, particularly with large datasets and complicated models such as those based on RBF kernels. They are also less interpretable than basic models.**
	RFC	**While Random Forests provide strong performance and feature importance insights, they are prone to overfitting, particularly with a large number of trees and deep trees. They are also less interpretable than simple linear models.**
Evaluation Metrics Limitations	**Accuracy**	**While accuracy is a valuable indicator, it can be misleading, particularly in datasets with imbalanced classes.**
	Precision, Recall, F1 Score	**These measures provide a more nuanced perspective of model performance, but they must be carefully balanced because optimizing for one can have a negative impact on the other.**

Table 3: Limitations

The project demonstrates the efficacy of machine learning models in medical prediction tasks while also emphasizing the need of knowing each model's traits and limitations. Future research could include experimenting with more advanced models, such as deep learning approaches or ensemble methods that combine the strengths of several models.

References:

[1] FEDESORIANO, "Kaggle," 10 January 2021. [Online]. Available: https://www.kaggle.com/datasets/fedesoriano/heart-failure-prediction?rvi=1. [Accessed 27 December 2023].