.


Cairo University
Faculty of Engineering
Credit Hour System

# SBEN454
# Machine Learning In Healthcare
# Project: Heart Failure Prediction

**Name:** **Sandra Adel Aziz Gebraiel**

**ID:** **1180059**

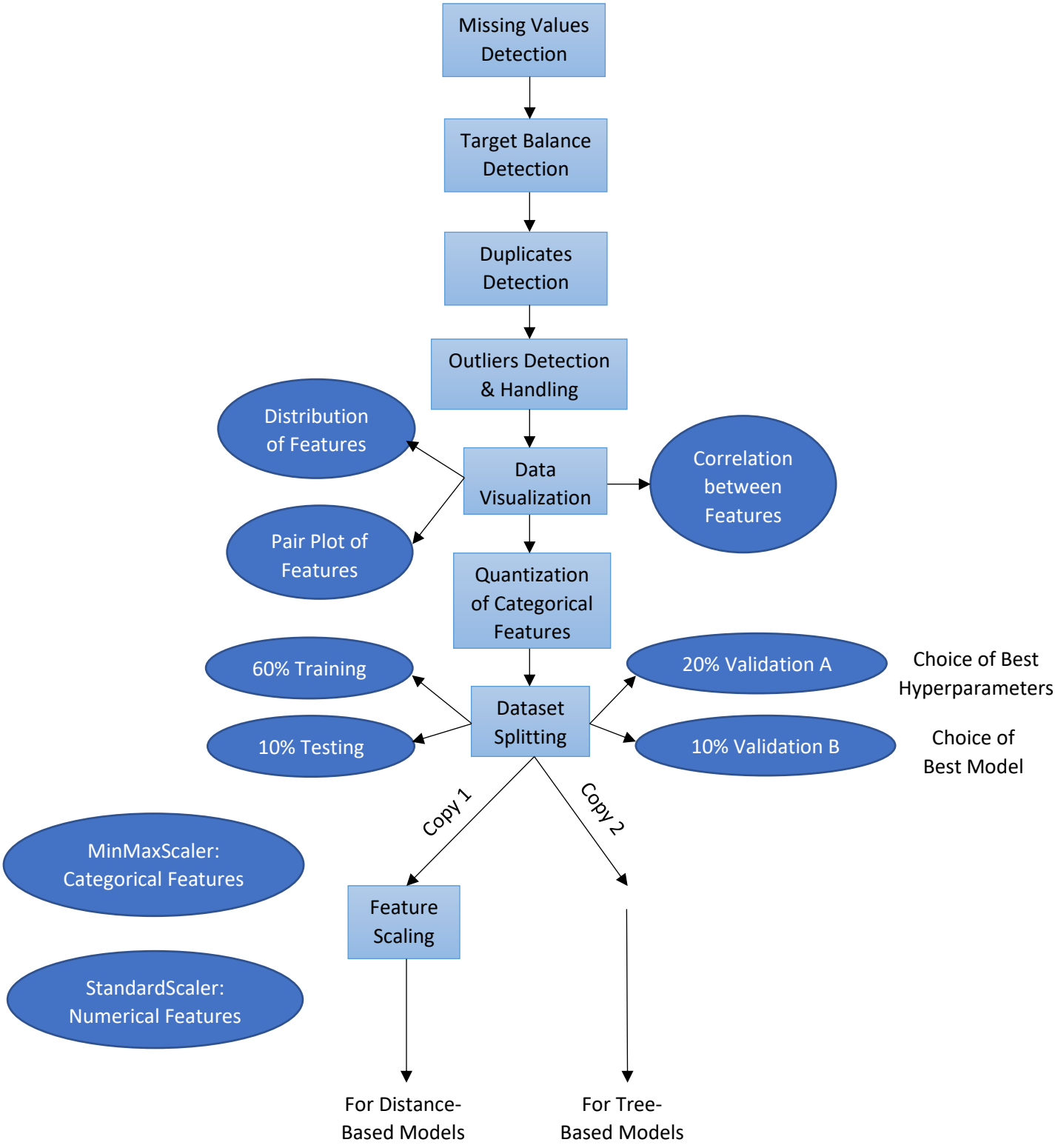
**Date:** **15/06/2022**

**Professor:** **Dr. Inas Yassine**

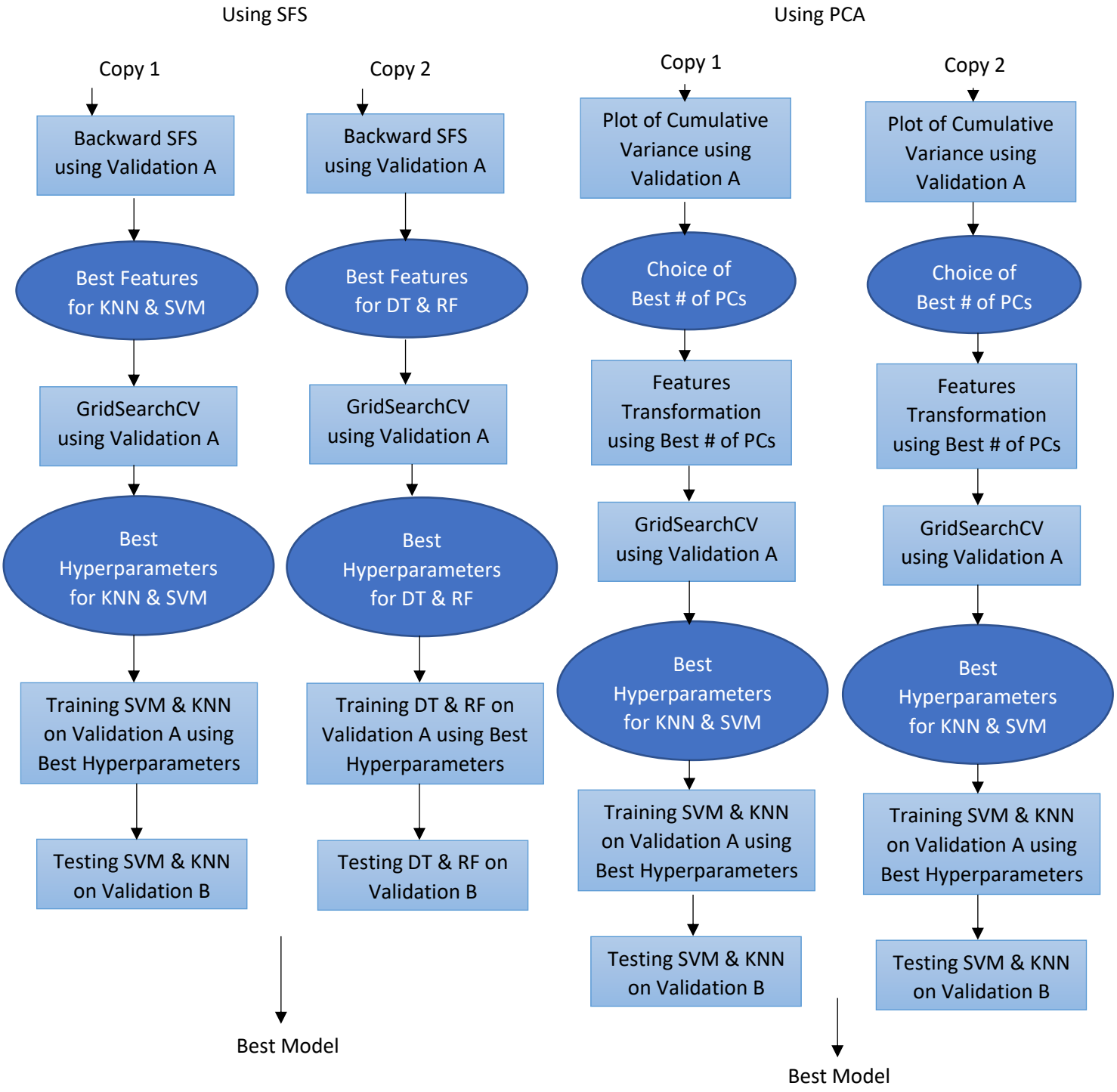
# Table of Contents:

## 1) Final Design Block Diagram:

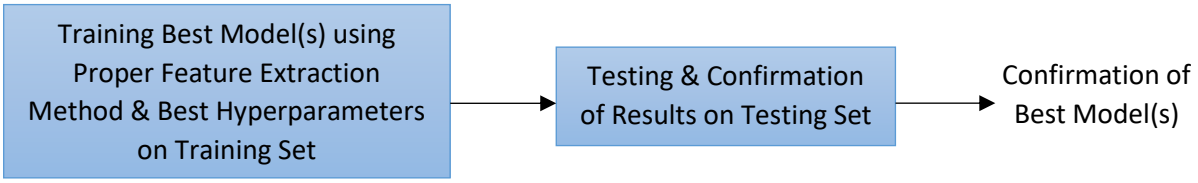
### A) Pre-processing and Visualization:



## B) Feature Extraction:




## C) Training and Testing:



## 2) Updates on Previously Decided Approaches in Phase 2:

- No discretization of features was needed as used Scikit-Learn models were able to use quantized data efficiently.
- Chi-squared filter-based technique was eliminated as feature extraction method as its inability to see relation between target and bunch of features (present in wrapped-based) was unfavoured.
- Validation dataset was split one more time from inside into 20% for choice of best hyperparameters and GridSearchCV and 10% for decision of best model.
- Two copies of dataset were handled, scaled one for distance-based models and unscaled one for tree-based models as scaling in entirely unneeded for them.

## 3) Critique of Some of Used Approaches:

### A) Dataset Split:

| 90% Training & 10% Testing | 60% Training & 30% Validation & 10% Testing **(Used)** |
|---|---|
| **Pro:** Model sees most of data to train on and optimize its performance from. | **Con:** Model could have seen more data to train on than just 60%. |
| **Con:** Less confidence in results by checking through testing set only (10%). | **Pro:** More than one method of results checking: 10% testing & 10% validation (one-third of 30%) gives more confidence in results. |

### B) Subdataset for Performing GridSearchCV:

(Noting that GridSearchCV is used for choosing hyperparameters and performs validation within its implementation)

| Training Set | Validation Set **(Used)** |
|---|---|
| **Pro:** Hyperparameters are chosen based on most of the dataset, which will definitely give better hyperparameters' choice. | **Con:** Hyperparameters are chosen based on less percentage of dataset (Here: 20%), which may not best hyperparameters. |
| **Con:** Training on dataset seen before for choosing best hyperparameters, which may give very good but deceitful results. | **Pro:** Providing more generalization to choose hyperparameters from one set and train on another (even if used models in both processes are different), giving more confidence in results. |

### C) Handling of Outliers:

**Performed:** Outliers could have been handled by either changing them into mean values of features or minimum and maximum values of features range.

After consulting some Senior-Year Medical Students, choice of minimum and maximum was preferred.

**Critique:** Maybe consulting more experienced workers in the field would have given a different preference leading to different results.

D) Meticulous Feature Scaling:

**Performed:** After observing distribution of continuous features, which seemed to be of Gaussian distribution, standard scaling was performed on them, while normalization scaling was performed on categorical features since their bars are too few to form an obvious distribution

**Critique:** Maybe this approach was too exhaustive and meticulous and performing either standard scaling or normalization scaling on all features would have led to different results, as continuous features are not perfectly Gaussian.

E) Choice of Sequential Feature Selection Direction:

**Performed:** Backward sequential feature selection was performed based on literature review as every paper we have studied which uses SFS performed it backwardly.

**Critique:** Maybe the sample of papers we studied was not representative enough of every machine learning study on heart failure prediction problem, and forward SFS would lead to different results.

F) Tried Models:

**Performed:** We tried SVM, KNN, DT and RF as a conclusion of our literature review of most models leading to the best results.

**Critique:** Again, maybe the sample of papers we studied was not representative enough of every machine learning study on heart failure prediction problem, and different models like Logistic Regression, AdaBoost, XGBoost, Naïve Bayes, ..etc would lead to different results.

## 4) Discussion of Results and Decision Making:

### A) Evaluation Metrics:

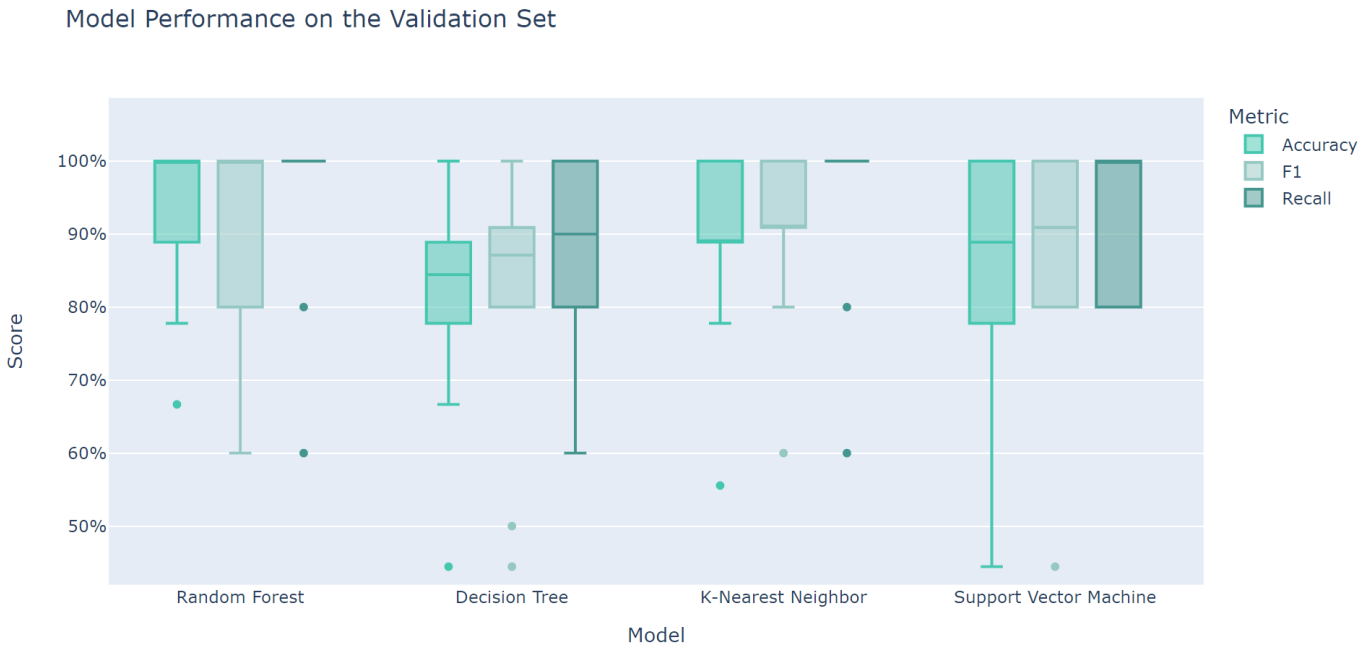
In validation stage, two parallel pipelines were followed (one using Backward Sequential Feature Selection and another using Backward Principal Component Analysis). After applying GridSearchCV to choose best hyperparameters for each model in each pipeline and training the models with them using two-thirds of validation dataset (20% of whole dataset), we test the models with one-third of validation dataset (10% of whole dataset).

The following evaluation metrics are chosen and taken into consideration:

1) Training Accuracy
2) Testing Accuracy
   (To observe amount of error in both values and difference between them for judging the quality of fitting)
3) F1-Score
4) Recall
   (In this application, false negatives are given greater importance (recall) than false positives to be reduced (precision): It is more critical to misdiagnose a heart disease patient as healthy while they need urgent care that may save their lives than to misdiagnose a healthy patient as heart-diseased and discovering they are healthy during delivering healthcare.
   For this reason, f1-score and recall are used to make sure that both false positives and negatives are small so as not to increase false positives while targeting to decrease false negatives)
5) Cross-validation Scores using Accuracy
6) Cross-validation Scores using F1-Score
7) Cross-validation Scores using Recall
   (Plotted as boxplots to observe median and standard deviation of each for estimating level of confidence of the accuracy, f1-score and recall scores calculated above)

## B) Judgement on Model Scores using Backward Sequential Feature Selection:

| | Model | Training Accuracy | Testing Accuracy | F1-Score | Recall |
|---|---|---|---|---|---|
| 0 | Support Vector Machine | 0.868132 | 0.846154 | 0.860000 | 0.860000 |
| 1 | K-Nearest Neighbor | 0.851648 | 0.868132 | 0.886792 | 0.940000 |
| 2 | Decision Tree | 0.890110 | 0.857143 | 0.865979 | 0.840000 |
| 3 | Random Forest | 0.884615 | 0.901099 | 0.910891 | 0.920000 |

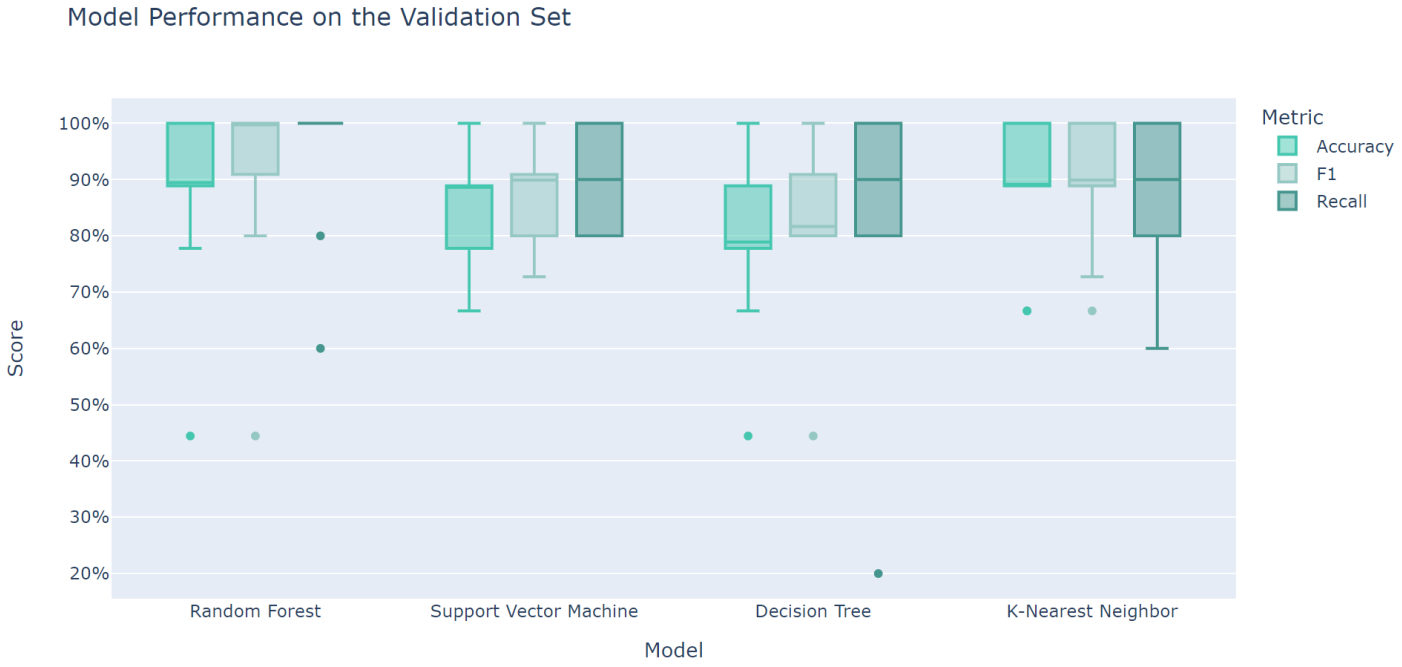



- KNN is eliminated as recall is way more than f1-score which means that false negatives are low but false positives are very high consequently.
- SVM is eliminated as it has the least testing accuracy and the standard deviations for the three scores are very high (20%) which decreases the confidence in the scores.
- Both DT and RF have high and close values of f1-score and recall, but RF has standard deviations of f1-score and recall: 20% and 2%, while DT has standard deviations of 10% and 20% for both scores. This means there is more overall confidence in both scores of for RF than DT.
  Moreover, the standard deviations of accuracy for both are equal (10%), but the median for RF is greater than DT. Finally, testing accuracy of RF is higher than that of DT in addition to difference between training and testing accuracies being less for RF than DT.

**Final Decision:**
RF is best model for this pipeline using backward sequential feature selection.

## C) Judgement on Model Scores using Principal Component Analysis:

| | Model | Training Accuracy | Testing Accuracy | F1-Score | Recall |
|---|---|---|---|---|---|
| 0 | Support Vector Machine | 0.857143 | 0.868132 | 0.882353 | 0.900000 |
| 1 | K-Nearest Neighbor | 1.000000 | 0.890110 | 0.897959 | 0.880000 |
| 2 | Decision Tree | 0.895604 | 0.868132 | 0.882353 | 0.900000 |
| 3 | Random Forest | 0.895604 | 0.901099 | 0.910891 | 0.920000 |




- SVM is eliminated because it has the least training and testing scores in comparison with rest of the models.
- DT and KNN have good f1-score and recall values but testing accuracy is higher for KNN than DT in addition to KNN having higher medians for accuracy, f1-score and recall than DT while they both have equal standard deviations for these values. However, KNN will be eliminated as its unity of training accuracy is worrisome that this good performance maybe deceitful due to validation dataset being not representative enough.
- RF shows least standard deviations and highest medians for testing accuracy, f1-score and recall scores. Furthermore, it has the highest testing accuracy value and smallest difference between training and testing accuracies.

**Final Decision:**

RF is best model for this pipeline using principal component analysis.

## D) Comparing Winner RF Models from Both Pipelines:

Since each of the two models uses different parameter values and work on different features of the data, besides their results being so close, we will move on with them to the testing phase.

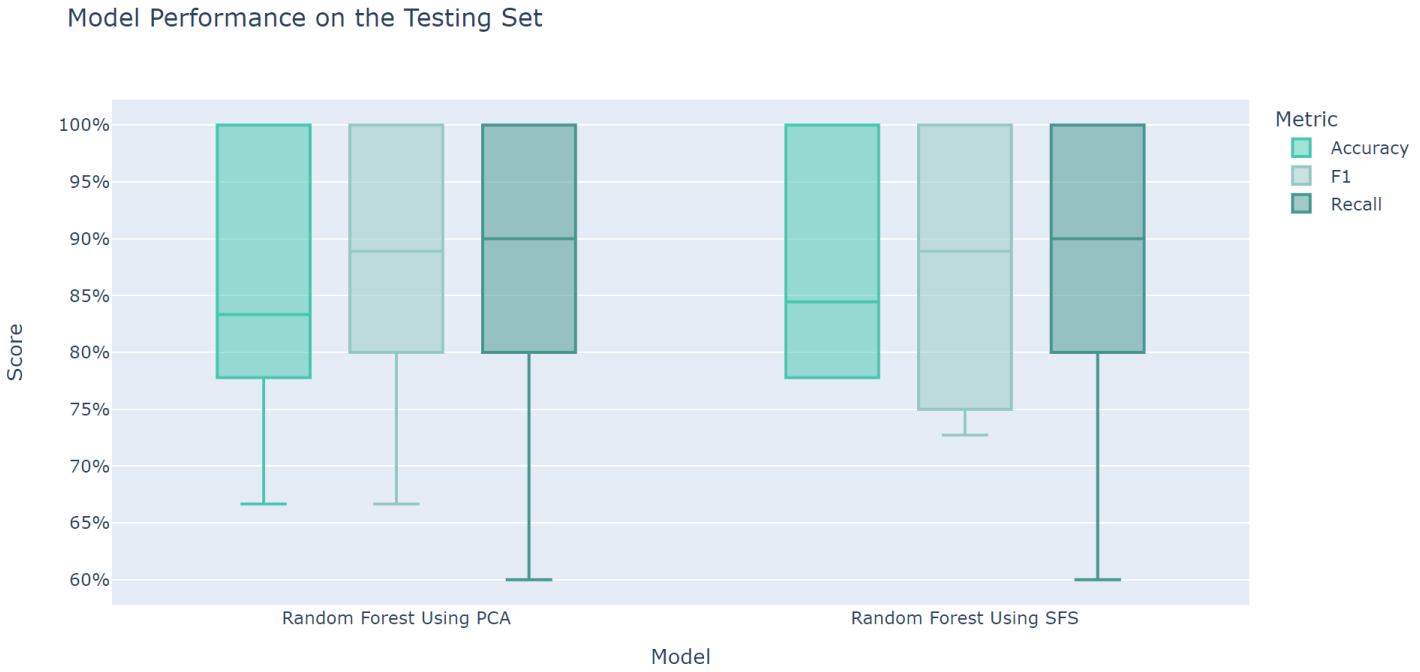
However, we will rank them first exploring their similarities and differences.

- Both models have the same testing accuracy, f1-score and recall values.
- Training accuracy score for RF using PCA is higher than that using SFS which makes difference between the training and testing accuracy values smaller.
- The standard deviations for the testing accuracy and recall value are same for both models. However, standard deviation for f1-score is less for RF using PCA (9%) than using SFS (20%) which gives more confidence that overall false negatives and false positives is less using PCA although both models have the same f1-score value, as expressed in first point.

**In conclusion**, we rank RF using PCA above RF using SFS, and we will move on to confirm our results in the testing phase.

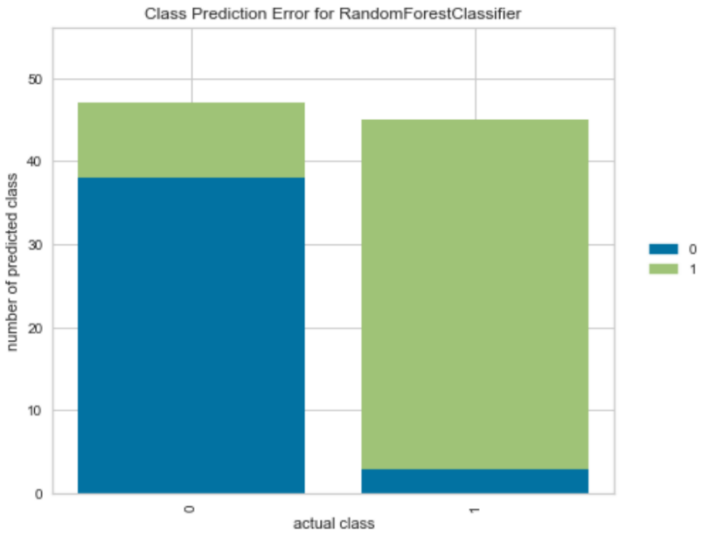
## E) Confirmation of Model Selection in Testing Phase:

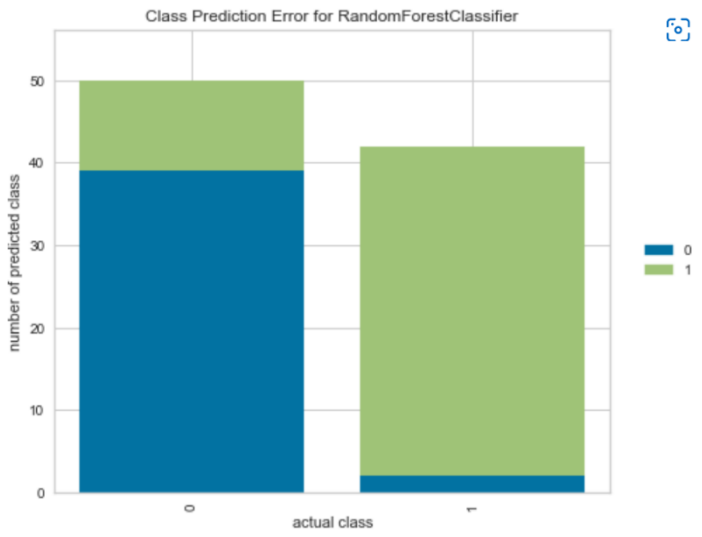
| | Model | Feature Extraction/Selection | Training Accuracy | Testing Accuracy | F1-Score | Recall | Ranking |
|---|---|---|---|---|---|---|---|
| 0 | Random Forest | Principal Component Analysis | 0.882459 | 0.869565 | 0.875000 | 0.823529 | First Place |
| 1 | Random Forest | Backward Sequential Feature Selection | 0.904159 | 0.858696 | 0.860215 | 0.784314 | Second Place |




Scores show excellency of RF using PCA over RF using SFS:

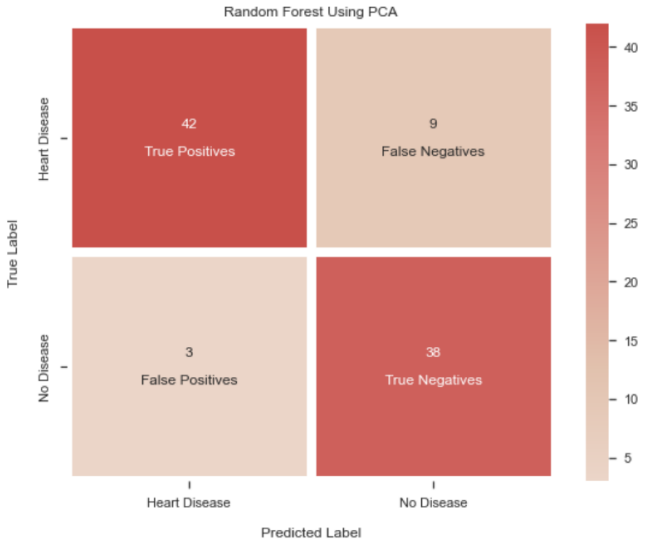
- Testing accuracy, f1-score and recall values are higher for RF using PCA than using SFS.
- Training accuracy for RF using SFS is higher than that using PCA but the difference between training and testing accuracies is less for RF using PCA than using SFS.
- The median for accuracy is higher for RF using SFS (83%) than using PCA (84%). On the other hand, the standard deviations of testing accuracy and recall is same for both models while that of f1-score is less for RF using PCA (20%) than using SFS (25%). Standard deviation difference of 5% outweighs median difference 1%.

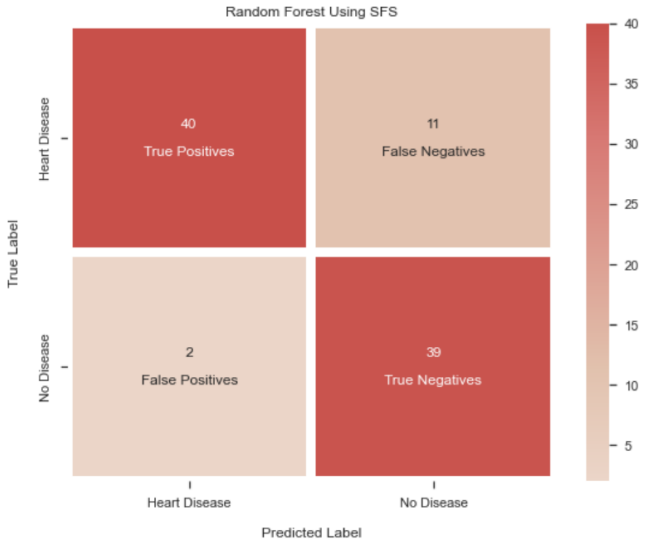



- False positives for RF using PCA is slightly higher than using SFS, but false negatives for RF using PCA is obviously much lower than using SFS.

- Less false positives for RF using SFS is still a good advantage, but we have ranked RF using PCA above it because false negatives are more important in this application.




- In RF using PCA, true positives are higher by 2 and false negatives are lesser by 2 than RF using SFS.

- In RF using SFS, true negatives are higher by 1 and false positives are lesser by 1 than RF using PCA, but again, true positives are false negatives are more targeted in this application.

## F) Final Results:

As discussed before false negatives are more important to be decreased in our application, but we have moved on with these two models not just one of them because the difference in overall performance in them is very small, and both are the best out of the rest of the previously tried models.

**Final Result:**

1) RF using PCA → Accuracy: 0.87, F1-Score: 0.88, Recall: 0.82
2) RF using SFS → Accuracy: 0.86, F1-Score: 0.86, Recall: 0.78