Early identification of PCOS with commonly known diseases: Obesity, diabetes, high blood pressure and heart disease using machine learning techniques.

Report submitted to the SASTRA Deemed to be University
as the requirement for the course

CSE300 - MINI PROJECT

Submitted by

Divya Dharshini V
(Reg. No.: 1234003086, CSE)
Sakthi Prabha S
(Reg. No.: 124003263, CSE)

June 2023

SCHOOL OF COMPUTING

THANJAVUR, TAMIL NADU, INDIA – 613 401

SASTRA

ENGINEERING · MANAGEMENT · LAW · SCIENCES · HUMANITIES · EDUCATION

DEEMED TO BE UNIVERSITY

(U/S 3 of the UGC Act,1956)

THINK MERIT | THINK TRANSPARENCY | THINK SASTRA

THANJAVUR | KUMBAKONAM | CHENNAI

SCHOOL OF COMPUTING
THANJAVUR – 613 401

Bonafide Certificate

This is to certify that the report titled "**Early identification of PCOS with commonly known diseases: Obesity, diabetes, high blood pressure and heart disease using machine learning techniques**" submitted as a requirement for the course, CSE300: **MINI PROJECT** for B.Tech. is a bonafide record of the work done by **Ms. Divya Dharshini V (Reg. No. 124003086, CSE**), **Ms. Sakthi Prabha S (Reg. No. 124003263, CSE)** during the academic year 2022-23, in the School ofComputing, under my supervision.

Signature of Project Supervisor:
Name with Affiliation: Dr. Padmakumari P, Asst. Professor-III
Date:

Mini Project *Viva voc*e held on _________________

Examiner 1 **Examiner 2**

Acknowledgements

We would like to thank our Honorable Chancellor **Prof. R. Sethuraman** for providing us with an opportunity and the necessary infrastructure for carrying out this project as a part of our curriculum.

We would like to thank our Honorable Vice-Chancellor **Dr. S. Vaidhyasubramaniam** and **Dr. S. Swaminathan,** Dean**,** Planning & Development, for the encouragement and strategic support at every step of our college life.

We extend our sincere thanks to **Dr. R. Chandramouli,** Registrar, SASTRA Deemed to be University for providing the opportunity to pursue this project.

We extend our heartfelt thanks to **Dr. A. Umamakeswari**, Dean, School of Computing, Dr. **B. Santhi**, Associate Dean, Department of Computer Application, **Dr. R. Muthaiah**, Associate Dean, Department of Information Technology and Information & Communication Technology and **Dr. V. S. Shankar Sriram**, Associate Dean, Department of Computer Science and Engineering for their motivation and support offered in materializing this project.

Our guide Dr. Padmakumari P., Assistant Professor-III, School of Computing was the driving force behind this whole idea from the start. Her deep insight in the field and invaluable suggestions helped us in making progress throughout our project work. We also thank the project review panel members for their valuable comments and insights which made this project better.

We would like to extend our gratitude to all the teaching and non-teaching faculties of the School of Computing who have either directly or indirectly helped us in the completion of the project.

We gratefully acknowledge all the contributions and encouragement from my family and friends resulting in the successful completion of this project. We thank you all for providing me an opportunity to showcase my skills through this project.

List of figures

Figure No	Title	Page No
Fig 1.1	Flow chart of Logistic Regression Model	2
Fig 4.1	SVM Algorithm (80:20) without feature selection	12
Fig 4.2	Naive Bayes Algorithm (80:20) without feature selection	12
Fig 4.3	Logistic Regression Algorithm (80:20) without feature selection	13
Fig 4.4	ROC Curve without feature selection	13
Fig 4.5	SVM Algorithm (80:20) with feature selection	14
Fig 4.6	Naive Bayes Algorithm (80:20) with feature selection	14
Fig 4.7	Logistic Regression Algorithm (80:20) with feature selection	15
Fig 4.8	ROC Curve with feature selection	15

Abbreviations

ML	Machine Learning
KNN	K- Nearest Neighbors
SVM	Support Vector Machine
LR	Logistic Regression
Corr	Correlation
STD	Standard Deviation

Abstract

Polycystic Ovary Syndrome (PCOS) is a health disorder that affects around 10 million women worldwide. If not diagnosed at an early stage, it leads to various other harmful diseases as well like diabetes, high blood pressure, obesity, heart disease etc. Thus, PCOS identification at an early stage is essential. The main objective of this project is to find the diseases which can help early identification of PCOS in a nutshell, we want to answer the question "Can we identify some commonly known diseases which taken as an indication of having PCOS". To achieve the above-mentioned goal, feature selection techniques are to be applied over the amalgamated data to reduce the number of parameters used for PCOS diagnosis. To find out the ranking of features for PCOS diagnosis, SelectKBest and chi2 feature selection methods are applied to features. Six classification algorithms of supervised learning are planned such as Decision tree, Gradient Boosting, Random Forest, Logistic regression, K-nearest neighbor and Support vector machine. Two types of datasets are used for this project: Diabetes (Pima Indians Diabetes Database) and heart disease. The developed models are compared based on their performance metrics. Depends on the features for the prediction of PCOS, the accuracies of the models differ. The focus is to improve the accuracy using the significant features for each dataset.

KEYWORDS: PCOS detection, amalgamation, Machine Learning, Decision tree, Gradient Boosting, Random Forest, Logistic regression, K-nearest neighbor, Support vector machine, Features

TABLE OF CONTENT

Title	**Page No**
Bonafide Certificate	ii
Acknowledgements	iii
List of Figures	iv
Abbreviations	vi
Abstract	vii
1. Summary of the base paper	4
2.Merits and Demerits of the base paper	7
3.Source Code	11
4.Results and Snapshots	16
5.Conclusion and Future Plans	17
6.References	18

Chapter 1: Summary of the Base Paper

1.1 Base Paper Details

Title: “”.

Authors: “”

Publication year:

Journal:

1.2 Introduction

1. PCOS is a complicated hormonal, metabolic and reproductive ailment that influences 1-in-10 women of childbearing age. PCOS is the main reason for infertility in women. Women with PCOS are at higher risk for developing type 2 diabetes and cardiovascular disease. Earlier detection can give them the chance to better manage the emotional, internal, and physical consequences of PCOS.It can also help them to prevent the outbreak of more critical illnesses associated with PCOS.

2 The objective of this project is to determine whether the Diabetes , blood pressure , obesity lead to PCOS or not. Generally women don’t go for PCOS testing but they do periodic blood pressure ,diabetes testing . In our analysis, if a woman has obesity, high blood pressure, diabetes and heart disease. We can predict that she is more likely to have PCOS. To verify, a data amalgamation is performed by merging two different datasets of heart diseases and diabetes. Data amalgamation provides a new dataset on which a feature selection technique is applied to get the eight selected attributes. Then, supervised learning classification algorithms are applied to the eight attributes to analyse the performance metrics of the model. Six classification algorithms are used Decision tree, Gradient Boosting, Random Forest, Logistic regression, K-nearest neighbor, Support vector machine. Highlights the relevance of some commonly known diseases such as obesity, heart disease, high blood pressure and diabetes for early detection of PCOS. • Created a new dataset by using different disease datasets which are further utilized for implementing supervised algorithms. • Supervised learning algorithms are applied to analyse the performance metrics of the model for all features, important features for prediction of PCOS• The findings of this project can help healthcare professionals with the early detection of PCOS by using only a few features.

2.1 Proposed Method (Architecture)

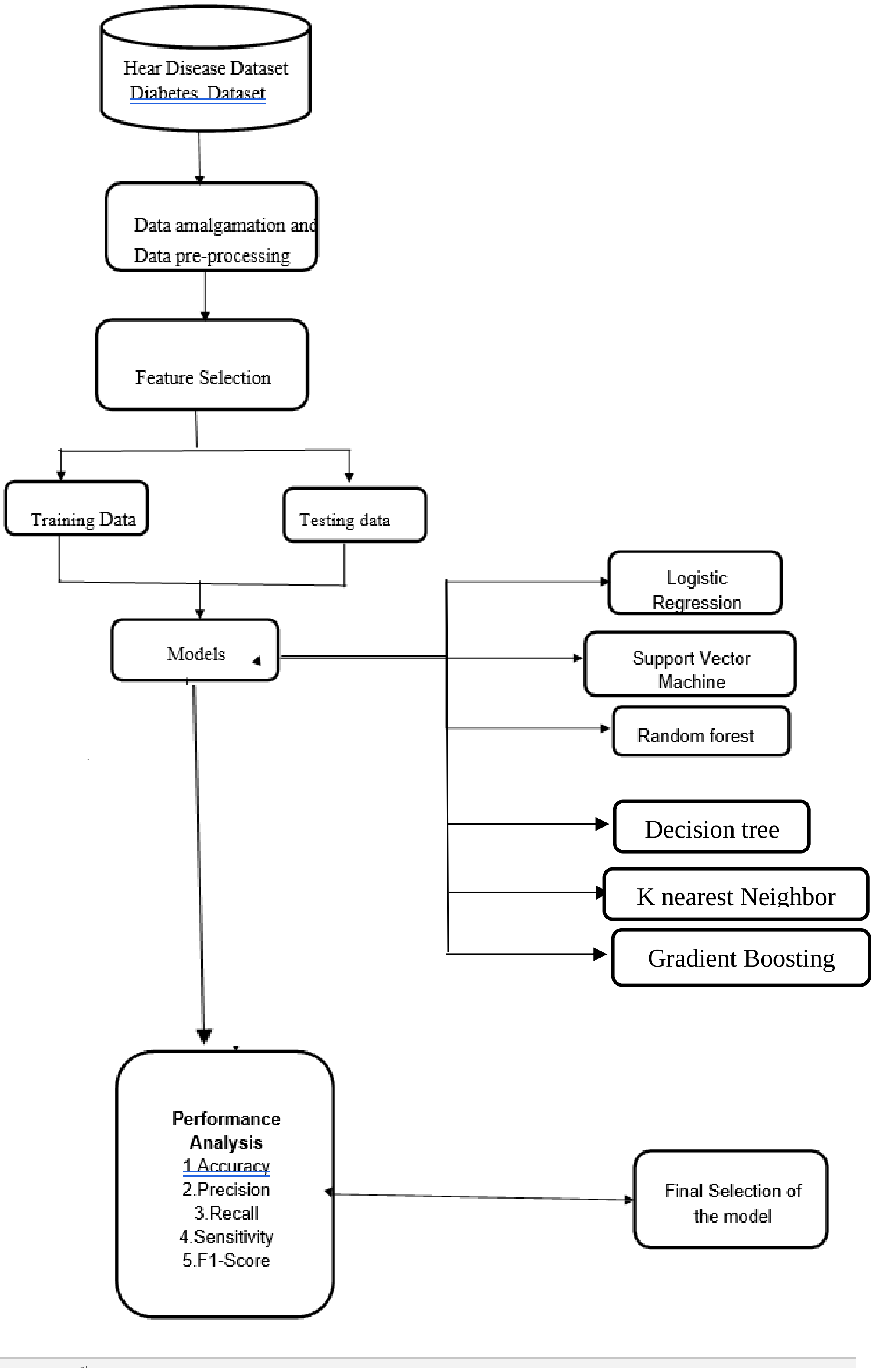


Fig 1.1

2.2 Dataset

•There are 303 records of patient details with 14 attributes describing them.

•Finally, the class either has heart-disease or is in a good condition.

Data	**Description**
age	Age of the patient
sex	If the patient is male, sex=1 If the patient is female, sex=0
cp	Chest pain type: If patient has typical angina, cp=1 If patient has atypical angina, cp=2 If patient has non-anginal pain, cp=3 If patient has asymptomatic pain, cp=4
trestbps	The individual’s resting heart rate (mm Hg on admission to the hospital)
chol	The individual's cholesterol level, expressed in mg/dl.
fbs	Fasting blood sugar>120 mg/dl If fbs=1, true If fbs=0, false

restecg	Resting electrocardiographic results
	restecg=0 (normal)
	restecg=1 (abnormal)
	restecg=2 (hyper)

thalach	The individual’s maximum heart rate achieved
exang	Exercise induced angina
oldpeak	Old peak= ST depression induced by exercise relative to rest
slope	The slope of the peak exercise ST segment
ca	No.of major vessels (0-3) coloured by fluoroscopy

thal	thal=3 (normal)
	thal=6 (fixed defect)
	thal=7 (reversible defect)

2.3 Performance Metric

In order to compare accuracy, precision, sensitivity, and specificity—which are measured using a confusion matrix, three algorithms were used 1. Support vector machine
2.Naive Bayes
3.Logistic Regression

- Accuracy is one statistic for assessing classification models, which is the percentage of accurate predictions made by our model.

Accuracy=(TP+TN)/(TP+TN+FP+FN)

- The quality of a positive prediction made by the model. Precision is calculated by dividing the total number of correct positive predictions by the number of genuine positives.

Precision = TP/(TP+FP)

- Specificity is the proportion of genuine negatives that the model correctly predicts, whereas sensitivity is the fraction of true positives that the model correctly predicts.

Sensitivity=TP/(TP+FN)
Specificity=TN/(TN+FP)

- Recall refers to the number of true positives that were recalled (found), i.e., the number of accurate hits that were also discovered.

Recall=TP/(TP+FN)

- The harmonic mean of the precision and recall is the F1 score.

F1 score = (TP/(TP+(FP+FN)/2))

Chapter 2: Merits and Demerits of Base Paper

2.1 Literature Survey

References	Methodology	Limitations
Nishadi, A. S. T. (n.d.). International Journal of Advanced Research and Publications Predicting Heart Diseases In Logistic Regression Of Machine Learning Algorithms By Python Jupyterlab.	The dataset is gathered, and the data are then pre-processed. 14 IVS and the expected value make up the data set. Given that the goal variable is categorical, binary logistic regression, one of the classification techniques, was applied.	Feature Selection is not used .
Tania Ciu, Raymond Sunardi Oetama, Logistic Regression Prediction Model for Cardiovascular Disease, International Journal of New Media Technology 7(1):33-38	The procedure begins with data retrieval, is followed by data splitting, split-data prediction using the logistic regression algorithm, and data validation at the end.	Performance analysis is not done
Sonam Nikhar , A.M. Karandikar, Prediction of Heart Disease Using Machine Learning Algorithms , International Journal of Advanced Engineering, Management and Science (IJAEMS), Infogain Publication	Dataset including patient information is gathered. The method of selecting attributes chooses the relevant attributes for heart disease prediction. The available data resources are located, then further chosen, cleansed, and transformed into the required form. To accurately forecast cardiac disease, various classification approaches will be used on preprocessed data. The accuracy of several	Here, only Naïve Bayes and Decision Tree algorithms are analyzed.

	classifiers is compared by accuracy measure.	
S. Mohan, C. Thirumalai and G. Srivastava, "Effective Heart Disease Prediction Using Hybrid Machine Learning Techniques," in IEEE Access,	They put out a cutting-edge approach that aims to identify relevant traits by utilizing machine learning techniques, thereby increasing the accuracy of cardiovascular disease prediction.	Various split ratios are not used.
M. Kavitha, G. Gnaneswar, R. Dinesh, Y. R. Sai and R S. Suraj "Heart Disease Prediction using Hybrid machine Learning Model," 2021 6th International Conference on Inventive Computation Technologies.	The machine learning model's innovative technique is created. Three machine learning algorithms—Random Forest, Decision Tree, and Hybrid Model—are used in the implementation.	Feature selection is not done

2.2 Merits

- In the base paper, proposed one machine learning algorithm, Logistic Regression for the prediction of cardiovascular disease.
- Finding the missing values is done as part of pre-processing a corpus to improve accuracy. Data is cleaned before pre-processing.
- Prediction of heart disease could be useful for early diagnosis thus it increases mortality rate.
- Proves that simple supervised algorithms can show great impact in reality.
- We have done the implementation for all split ratios i.e. 90:10, 80:20, 70:30, 60:40, 50:50.

2.3 Demerits

- Other pre-processing methods should be done accurately to get better results
- Dataset used is small, so this may lead to signs of overfitting
- In the base paper, they only implemented Logistic regression. There are other machine learning algorithms which are more accurate than this.
- The study's restriction is that it only uses the UCI dataset; however, in future work, we may try to apply it to other datasets as well.
- Feature selection is not done, it leads to increase in time and space.
- Ensemble approach can be used for improving accuracy.

Chapter 3: Source Code

K-Best

```
from sklearn.model_selection import KFold
from sklearn.metrics import classification_report,f1_score

kf = KFold(n_splits=5, shuffle=True, random_state=42)
scoring = ['f1']
cv_results = []

for train_index, test_index in kf.split(x):
    x_train_fold = x.iloc[train_index]
    y_train_fold = encoded_y_f.iloc[train_index]
    y_train_fold = np.ravel(y_train_fold)
    x_test_fold = x.iloc[test_index]
    y_test_fold = encoded_y_f.iloc[test_index]

    logreg.fit(x_train_fold, y_train_fold)
    y_pred_fold = logreg.predict(x_test_fold)

    cv_results.append({
        'train_index': train_index,
        'test_index': test_index,
        'f1_score': f1_score(y_test_fold, y_pred_fold)
    })

# Select the best fold based on the F1 score
best_fold = max(cv_results, key=lambda x: x['f1_score'])

# Train and test the model on the best fold
x_train_best_fold = x.iloc[best_fold['train_index']]
y_train_best_fold = encoded_y_f.iloc[best_fold['train_index']]
x_test_best_fold = x.iloc[best_fold['test_index']]
y_test_best_fold = encoded_y_f.iloc[best_fold['test_index']]
print(np.shape(x_train_best_fold))
np.shape(x_test_best_fold)

logreg.fit(x_train_best_fold, y_train_best_fold)
y_pred_best_fold = logreg.predict(x_test_best_fold)
# class labels
target_names = ['PCOS_not_detected', 'PCOS_detected']

# Evaluate the model on the best fold
print(classification_report(y_test_best_fold, y_pred_best_fold,target_names =target
_names))
```

Cross val score

```
from sklearn.model_selection import cross_val_score
encoded_y_fr=np.ravel(encoded_y_f)
scores = cross_val_score(logreg, x, encoded_y_fr, cv=5)
mean_score = scores.mean()
print(mean_score)
```

Confusion matrix

```
lr_conf_matrix = confusion_matrix(y_test_best_fold, y_pred_best_fold)
# Plot confusion matrix heatmap
sns.heatmap(lr_conf_matrix, annot=True, cmap="YlGnBu", fmt="d")
plt.title("Confusion Matrix")
plt.xlabel("Predicted Label")
plt.ylabel("True Label")
plt.show()
```

Logistic Regression

```
logreg = LogisticRegression()
logreg.fit(x_train, encoded_y)

import matplotlib.pyplot as plt

# Assuming your logistic regression object is called logreg
coefs = logreg.coef_[0]
features = list(x_train.columns)

# Plot the feature importances
plt.barh(features, coefs)
plt.xticks(rotation=90)
plt.xlabel('Features')
plt.ylabel('Coefficients')
plt.title('Logistic Regression Feature Importances')
plt.show()
encoded_ytest= lbl.fit_transform(y_test)
Y_pred1 = logreg.predict(x_test)
```

Classification Report

```
from sklearn.metrics import classification_report

# class labels
target_names = ['PCOS_not_detected', 'PCOS_detected']

# generate classification report
print(classification_report(encoded_ytest, Y_pred1, target_names=target_names))
```

Decision Tree

```
tree = tree= DecisionTreeClassifier()
tree.fit(x_train, encoded_y)

# Get feature importances
importances = tree.feature_importances_

# Get feature names
feature_names = x_train.columns

# Sort features by importance
sorted_idx = importances.argsort()

# Plot feature importances
plt.barh(range(len(sorted_idx)), importances[sorted_idx])
plt.yticks(range(len(sorted_idx)), [feature_names[i] for i in sorted_idx])
plt.xlabel('Importance')
plt.title('Decision tree Feature Importance')
plt.show()
```

Random Forest

```
rf= RandomForestClassifier()
rf.fit(x_train,encoded_y)

importances = rf.feature_importances_

# Get the feature names
feature_names = x_train.columns

# Create a bar plot of feature importances
plt.barh(feature_names, importances)
plt.xticks(rotation=90)
plt.xlabel('Features')
plt.ylabel('Importance')
plt.title('Random Forest Feature Importance')
plt.show()
```

KNN (K-nearest Neighbors)

```
error_rate= []
for i in range(1,40):
    knn= KNeighborsClassifier(n_neighbors=i)
    knn.fit(x_train,encoded_y)
    pred= knn.predict(x_test)
    error_rate.append(np.mean(pred != encoded_ytest))
```

graph in knn

```
plt.figure(figsize=(10,6))
plt.plot(range(1,40),error_rate,color='blue', linestyle='dashed', marker='o',
         markerfacecolor='red', markersize=10)
plt.xlabel('K Vlaue')
plt.ylabel('Error rate')
plt.title('To check the correct value of k')
plt.show()

knn= KNeighborsClassifier(n_neighbors=5)
knn.fit(x_train,encoded_y)
ypred4= knn.predict(x_test)


from sklearn.inspection import permutation_importance

# Compute permutation feature importance scores
result = permutation_importance(knn, x_train, encoded_y, n_repeats=10)

# Get feature importance scores
importance_scores = result.importances_mean

# Print feature importance scores
for i, score in enumerate(importance_scores):
    print(f"Feature {i}: {score}")

# Plot feature importance scores
plt.barh(range(x_train.shape[1]), importance_scores)
plt.xticks(range(x_train.shape[1]))
plt.xlabel("Feature index")
plt.ylabel("Importance score")
plt.title("KNN Feature Importance Scores")
plt.show()
```

SVM(Support Vector Machine)

```
from sklearn import svm
svm= svm.SVC()
svm.fit(x_train,encoded_y)

from sklearn.inspection import permutation_importance

# Compute permutation feature importance scores
result = permutation_importance(svm, x_train, encoded_y, n_repeats=10)

# Get feature importance scores
importance_scores = result.importances_mean

# Print feature importance scores
for i, score in enumerate(importance_scores):
    print(f"Feature {i}: {score}")

# Plot feature importance scores
plt.barh(range(x_train.shape[1]), importance_scores)
plt.xticks(range(x_train.shape[1]))
plt.xlabel("Feature index")
plt.ylabel("Importance score")
plt.title("SVM Feature Importance Scores")
plt.show()

ypred5= svm.predict(x_test)
```

Gradient Boosting

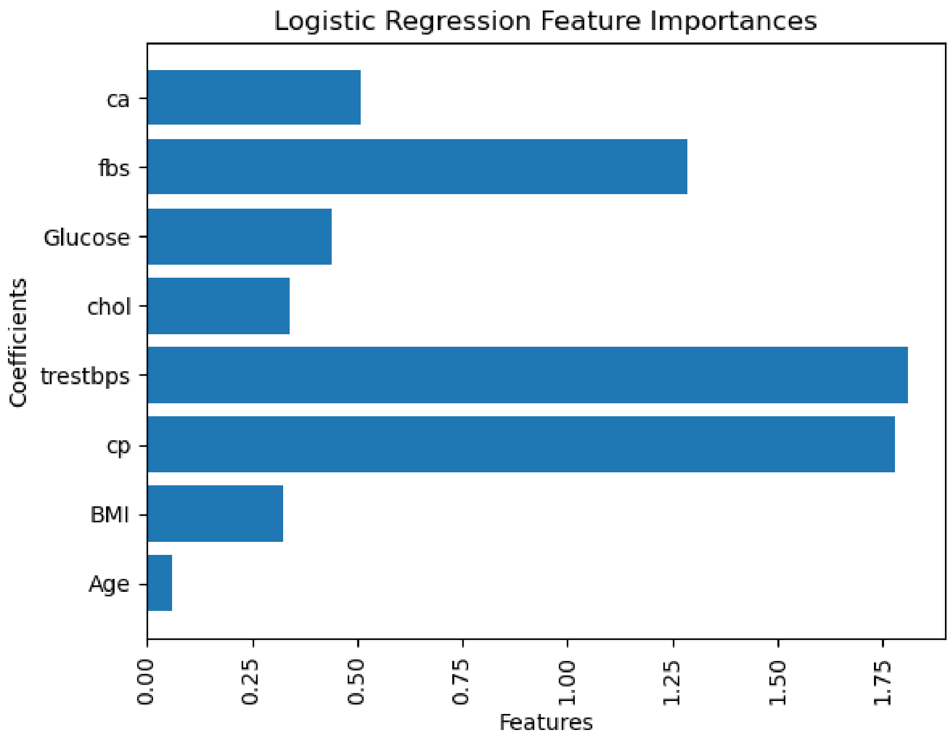
```
from sklearn.ensemble import GradientBoostingClassifier

# Create a Gradient Boosting classifier
gb = GradientBoostingClassifier()

# Train the model on the training data
gb.fit(x_train, encoded_y)

# Generate predictions on the test data
y_pred6 = gb.predict(x_test)
```

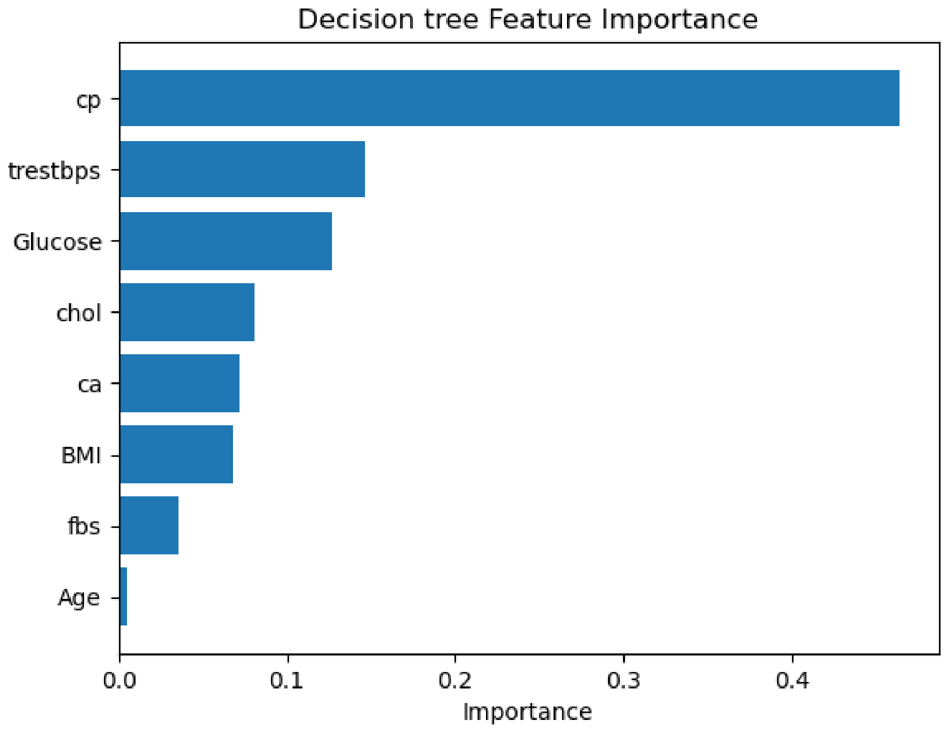
Chapter 4: Results and Snapshots

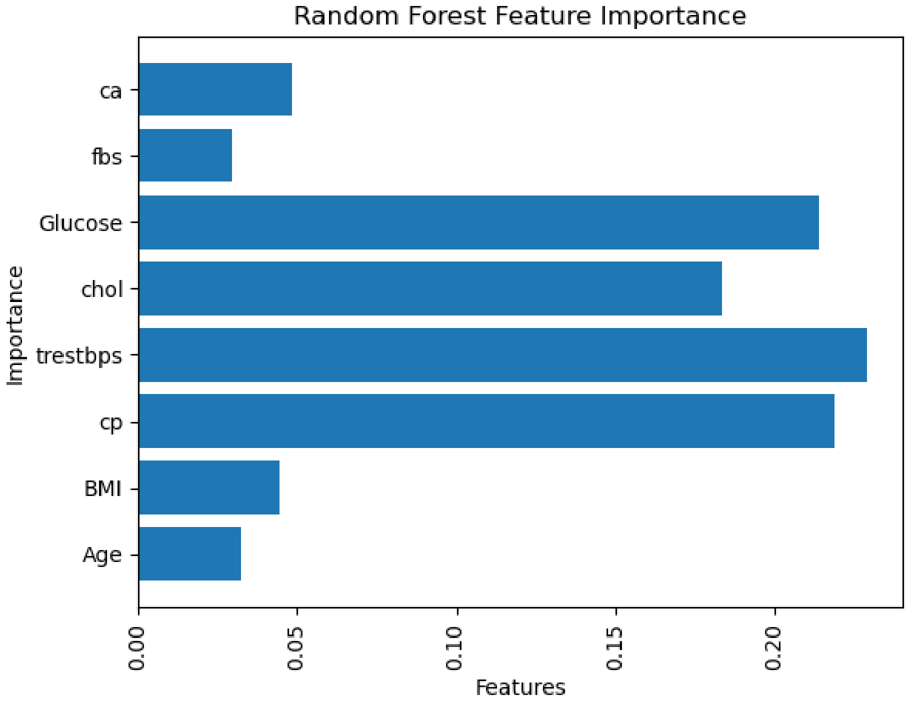
WITHOUT FEATURE SELECTION

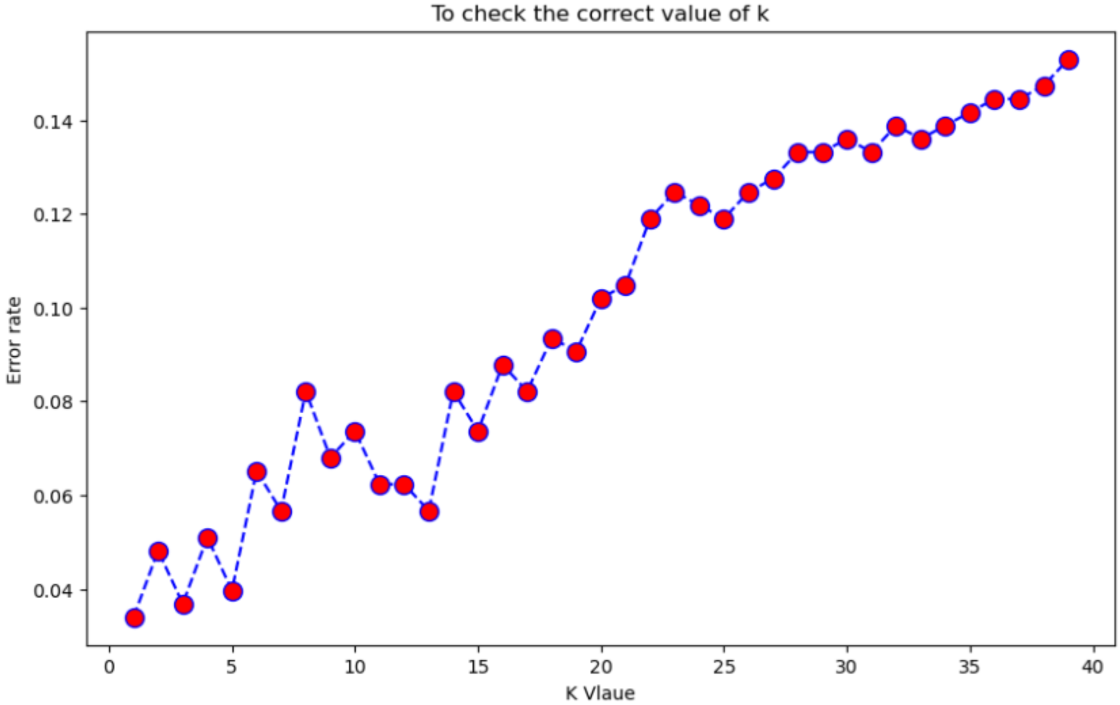
Fig 4.4

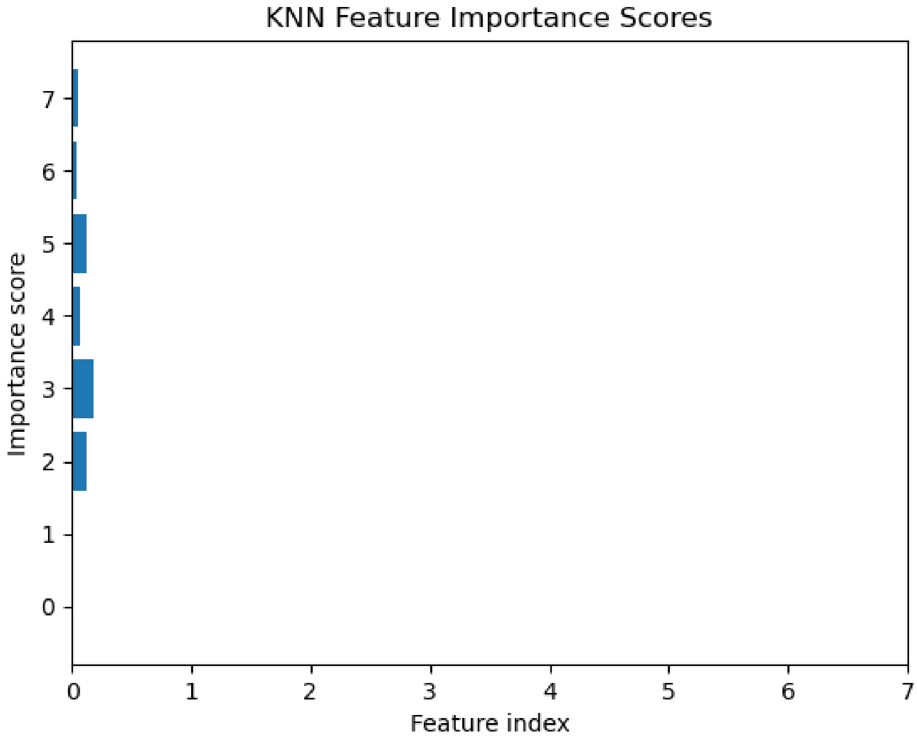
WITH FEATURE SELECTION

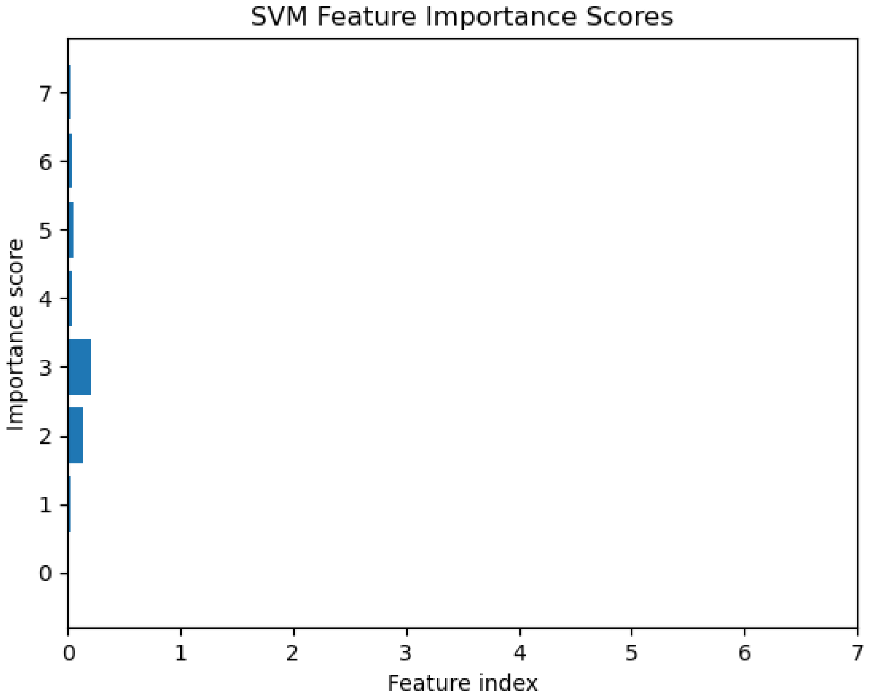





Random Forest Feature Importance
Importance
ca
fbs
Glucose
chol
trestbps
cp
BMI
Age
0.00
0.05
0.10
0.15
0.20
Features


To check the correct value of k
Error rate
0.14
0.12
0.10
0.08
0.06
0.04
0
5
10
15
20
25
30
35
40
K Vlaue


KNN Feature Importance Scores
Importance score
Feature index
0
1
2
3
4
5
6
7


SVM Feature Importance Scores
Importance score
Feature index
0
1
2
3
4
5
6
7

WITHOUT FEATURE SELECTION

Split ratio	SVM	Naive bayes	Logistic Regression
70:30	79.12	74.72	81.31
82:20	83.60	77.04	83.60

WITH FEATURE SELECTION

Split ratio	SVM	Naive bayes	Logistic Regression
70:30	76.92	71.42	78.02
80:20	78.68	73.77	78.688

Chapter 5: Conclusion and Future plans

Conclusion:

We implemented two models, one with feature selection and the other without feature selection. In those two models, we compared three algorithms performance mainly, logistic regression, support vector machine and naive bayes with two different split ratios 80:20 and 70:30. And we concluded the following points, firstly, without Feature Selection gives more accuracy than with Feature selection. Second, we compared three algorithms, support vector machine, Naive Bayes and logistic regression. From these we concluded Logistic regression gives more accuracy than other models. We used two different split ratios that are 80:20 and 70:30 and we concluded that 80:20 gives more accuracy than 70:30 for all the algorithms in the model.

Future Plans:

We will extend our project to other machine learning and deep learning algorithms and analyse the performance.

Chapter 6: References

1. World Health Organization and J. Dostupno, cardiovascular diseases: key facts, vol. 13, no. 2016, p.6,2016.https://www.who.int/en/news-room/fact-sheets/detail/cardiovascular-diseases-(cvds).

2. K. Uyar, A. Ilhan, Diagnosis of heart disease using genetic algorithm based trained recurrent fuzzy neural networks Proceed. Comput. Sci., 120 (2017), pp. 588-593

3. N. Kausar, S. Palaniappan, B.B. Samir, A. Abdullah, N. Dey, Systematic analysis of applied data mining based optimization algorithms in clinical attribute extraction and classification for diagnosis of cardiac patients in Applications of Intelligent Optimization in Biology and Medicine, Cham, Switzerland: Springer (2016), pp. 217-231

4. M. Shouman, T. Turner, R. Stocker, Integrating clustering with different data mining techniques in the diagnosis of heart disease, J. Comput. Sci. Eng., 20 (1) (2013), pp. 1-10

5. M.S. Amin, Y.K. Chiam, K.D. Varathan, Identification of significant features and data mining techniques in predicting heart disease, Telemat. Inf., 36 (2019), pp. 82-93 Mar.

6. D. Singh, J.S. Samagh, A comprehensive review of heart disease prediction using machine learning J. Crit. Rev., 7 (12) (2020), p. 2020

7. F.A. Latifah, I. Slamet, Comparison of heart disease classification with logistic regression algorithm and random forest algorithm, Proceedings of the AIP Conference (2020), p. 2296, 10.1063/5.0030579

8. Z. Khan, D.K. Mishra, V. Sharma, A. Sharma, Empirical study of various classification techniques for heart disease prediction, Proceedings of the IEEE 5th International Conference on Computing Communication and Automation (ICCCA) (2020), pp. 57-62, 10.1109/ICCCA49541.2020.9250852

9. Nishadi, A.S.T. (n.d.). International journal of advanced research and publications predicting heart diseases in logistic regression of machine learning algorithms by python jupyterlab. https://www.kaggle.com

10. M. Saw, , T. Saxena, S. Kaithwas, R. Yadav and N. Lal, Estimation of Prediction for Getting Heart Disease Using Logistic Regression Model of Machine Learning saw (Ed.), 2020 International Conference on Computer Communication and Informatics (ICCCI), IEEE, Coimbatore, India (2020), pp. 1-6 January 22-24