UTM
UNIVERSITI TEKNOLOGI MALAYSIA

UNIVERSITI TEKNOLOGI MALAYSIA

FINAL YEAR PROJECT 2 PRESENTATION

TITLE:
PREDICTION OF EARLY-STAGE CHRONIC KIDNEY DISEASE USING SUPPORT VECTOR MACHINE (SVM)

ERICA DESIRAE MAURITIUS
A20EC0032

Link:
1. Slide Presentation: https://youtu.be/JAo4fbilBf8
2. Demo Presentation: https://youtu.be/KvwO3j3wH6Q

Innovating Solutions

UTM JOHOR BAHRU

TABLE OF CONTENT

01 BACKGROUND OF STUDY

02 PROBLEM STATEMENT

03 RESEARCH OBJECTIVES AND SCOPES

04 LITERATURE REVIEW

05 METHODOLOGY

06 RESULT AND DISCUSSION

07 CONCLUSION

INTRODUCTION

BACKGROUND OF STUDY

- Chronic Kidney Disease (CKD) is a global health issue associated with increasing morbidity and mortality rates, and it is linked to other illnesses such as cardiovascular disease (Rady & Anwar, 2019; Almansour et al., 2019).
- Risk factors for CKD, including diabetes, hypertension, and poor lifestyle choices, are prevalent in the Malaysian population (Bin Abdul Ghafar et al., 2022).
- Early detection of CKD is crucial as it develops slowly and can lead to complications such as hypertension, anemia, nerve damage, and weakened immune system (Rajeshwari & Yogish, 2022).
- Machine learning algorithms, particularly Support Vector Machine (SVM), have shown effectiveness in classifying and predicting common diseases, including CKD (Rady & Anwar, 2019; Almansour et al., 2019; Bin Abdul Ghafar et al., 2022; Rajeshwari & Yogish, 2022).

BACKGROUND OF STUDY

PROBLEM BACKGROUND

Chronic Kidney Disease (CKD) lacks early clinical symptoms, leading to delayed detection and treatment (Zhou et al., 2022).

CKD imposes a significant financial burden on economies and healthcare systems, particularly with expensive and complex renal replacement therapy for End-Stage Renal Disease (ERSD) (Shanthakumari & Jayakarthik, 2021).

Previous approaches to CKD prediction using traditional methods and clinical judgment have limitations, including biases, errors, and high costs (Tekale et al., 2018).

The availability of electronic health data has spurred interest in advanced computational technologies, such as Support Vector Machine (SVM), for developing more reliable and sophisticated CKD prediction models (Sharma & Kaur, 2022).

PROBLEM STATEMENT

ISSUES

- Investigating the causes of early-stage Chronic Kidney Disease (CKD) and evaluating the effectiveness of machine learning in CKD prediction.
- Ensuring early detection and treatment of CKD to prevent the disease and improve patient outcomes

PROPOSED SOLUTIONS

- Utilizing the Support Vector Machine (SVM) algorithm to accurately predict the factors associated with early-stage CKD and determine if SVM can be a reliable tool for early-stage CKD prediction.

PROBLEM

- Improving the accuracy and assessment of the proposed SVM-based method by incorporating evaluation metrics such as Accuracy, Recall, Precision, and F1 Score. This ensures a comprehensive evaluation of the model's performance in predicting early-stage CKD.

RESEARCH OBJECTIVES

R01 To study and identify the important features of early-stage Chronic Kidney Disease (CKD) and the performance of machine learning in the prediction of Chronic Kidney Disease (CKD).

R02 To develop model for Chronic Kidney Disease (CKD) using Support Vector Machine (SVM).

R03 To evaluate the early-stage Chronic Kidney Disease in Support Vector Machine (SVM) based on Accuracy, F1-Score, Correlation Coefficients and Mean Absolute Error (MAE).

RESEARCH SCOPE

MACHINE LEARNING CLASSIFIER

Support Vector Machine (SVM)

DATASET FEATURES

consists of 25 features for 400 people in which 11 and 14 features are numerical and categorical respectively.

PROGRAMMING LANGUAGE

Python Programming language will be utilized to design and develop the algorithm.

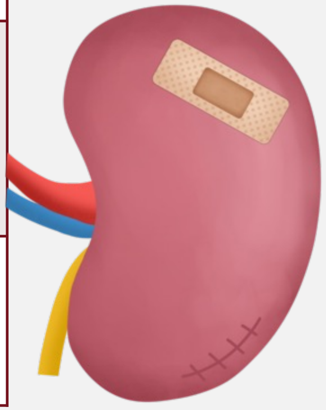
SOURCES

Kaagle "Chronic Kidney Disease Dataset" and "Chronic Kidney Diseases Prediction"

LITERATURE REVIEW

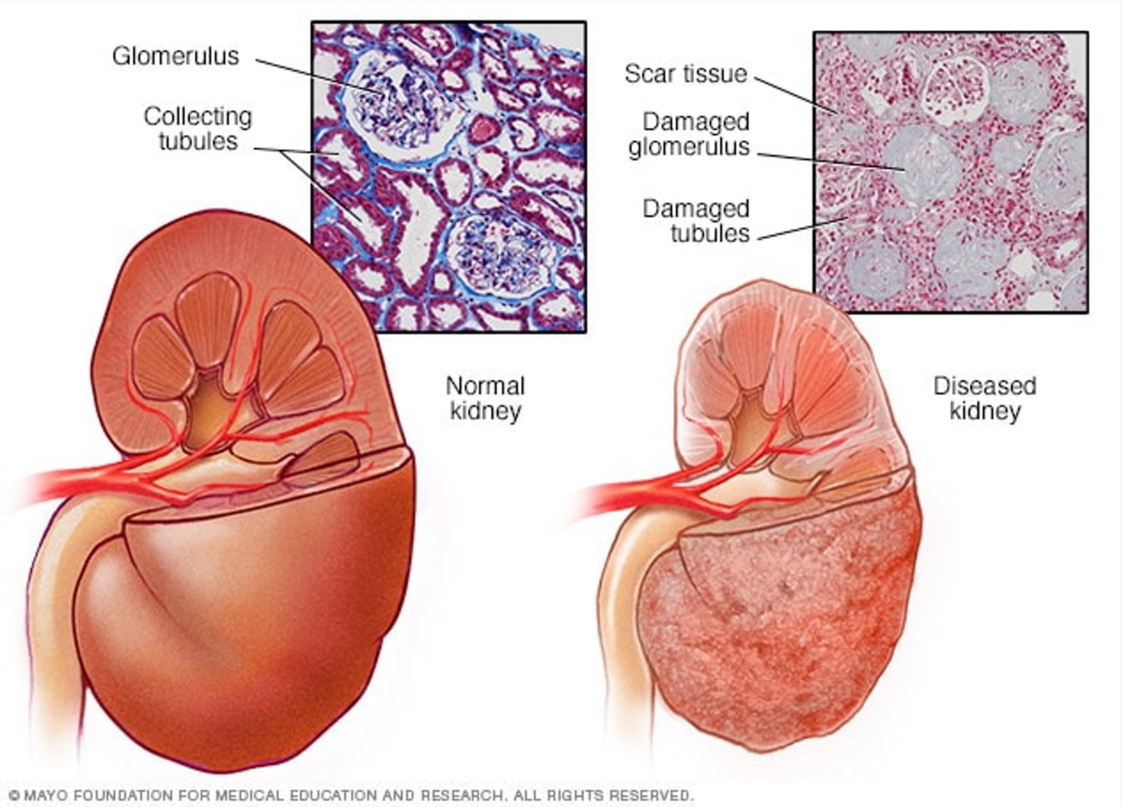
Table 1: Comparative performance between related works

Related Works	Key Findings
Gupta et al. (2020)	Decision Trees, Logistic Regression, SVM, and Random Forest achieved accuracies consistently above 90%, except for KNN.
Pankaj Chittora et al. (2021)	Linear SVM (LSVM) achieved highest accuracy of 98.46% among multiple classifiers for CKD prediction.
Rajeshwari and Yogish (2022)	Compared SVM, Random Forest, Decision Tree, and Naïve Bayes; Random Forest attained highest accuracy of 98.75% on a dataset with 14 columns and 400 rows.
Ghafar et al. (2022)	SVM achieved 93.5% accuracy in CKD prediction; suggested potential accuracy improvement through dataset augmentation.



LITERATURE REVIEW

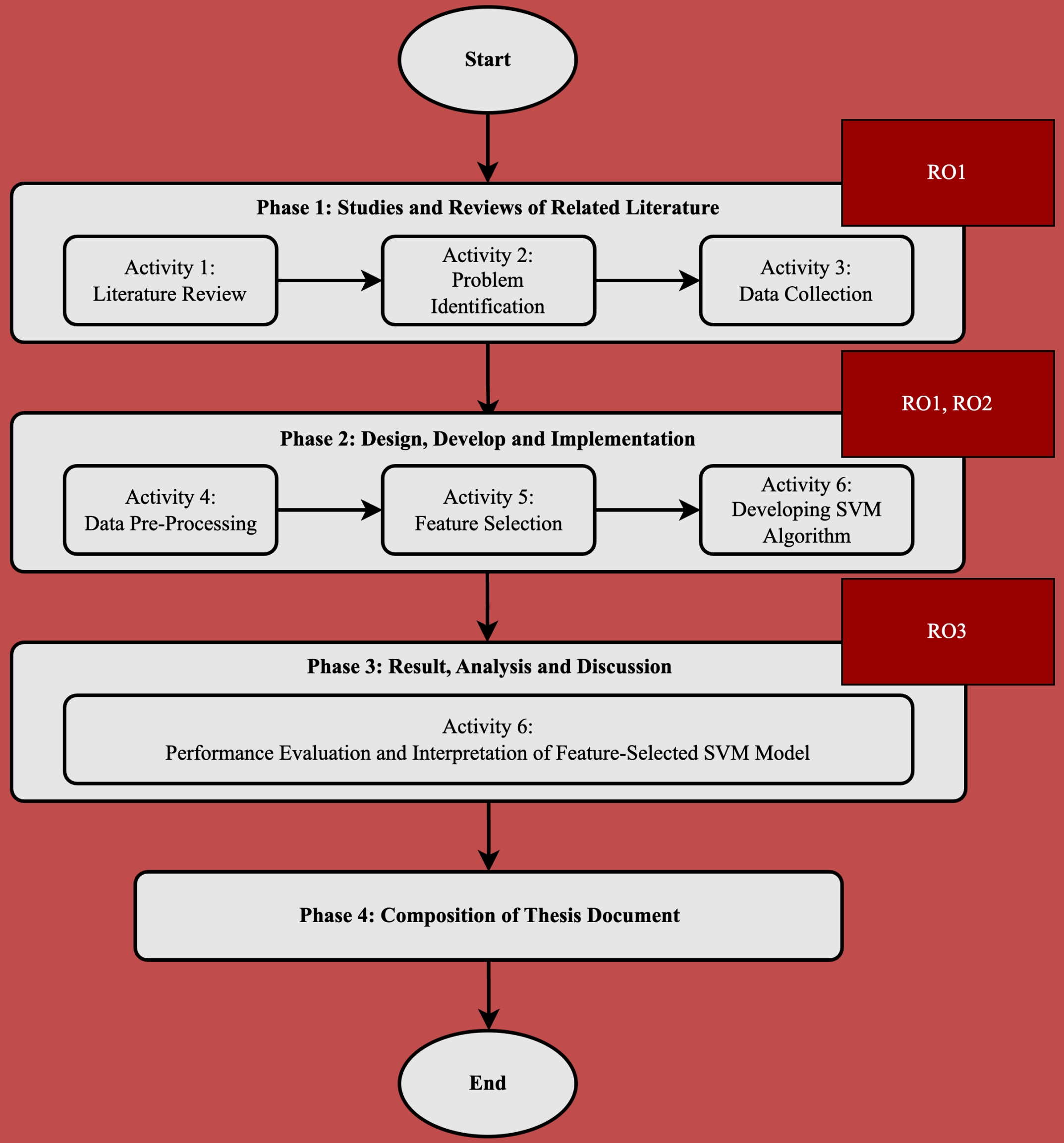
- Almansour et al. (2019) stated that the accuracy decreased due to fewer features where initially SVM produced a greater accuracy but ANN outperformed SVM with just two features.
- Therefore, this paper will compare various feature selection techniques based on the number of features and different train-test splits to determine which technique achieves better accuracy with SVM.
- There will also be an identification of important features, aiding in earlier detection of its often asymptomatic symptoms.



© MAYO FOUNDATION FOR MEDICAL EDUCATION AND RESEARCH. ALL RIGHTS RESERVED.

METHODOLOGY

RESEARCH FRAMEWORK



Start

Phase 1: Studies and Reviews of Related Literature

RO1

Activity 1: Literature Review

Activity 2: Problem Identification

Activity 3: Data Collection

Phase 2: Design, Develop and Implementation

RO1, RO2

Activity 4: Data Pre-Processing

Activity 5: Feature Selection

Activity 6: Developing SVM Algorithm

Phase 3: Result, Analysis and Discussion

RO3

Activity 6: Performance Evaluation and Interpretation of Feature-Selected SVM Model

Phase 4: Composition of Thesis Document

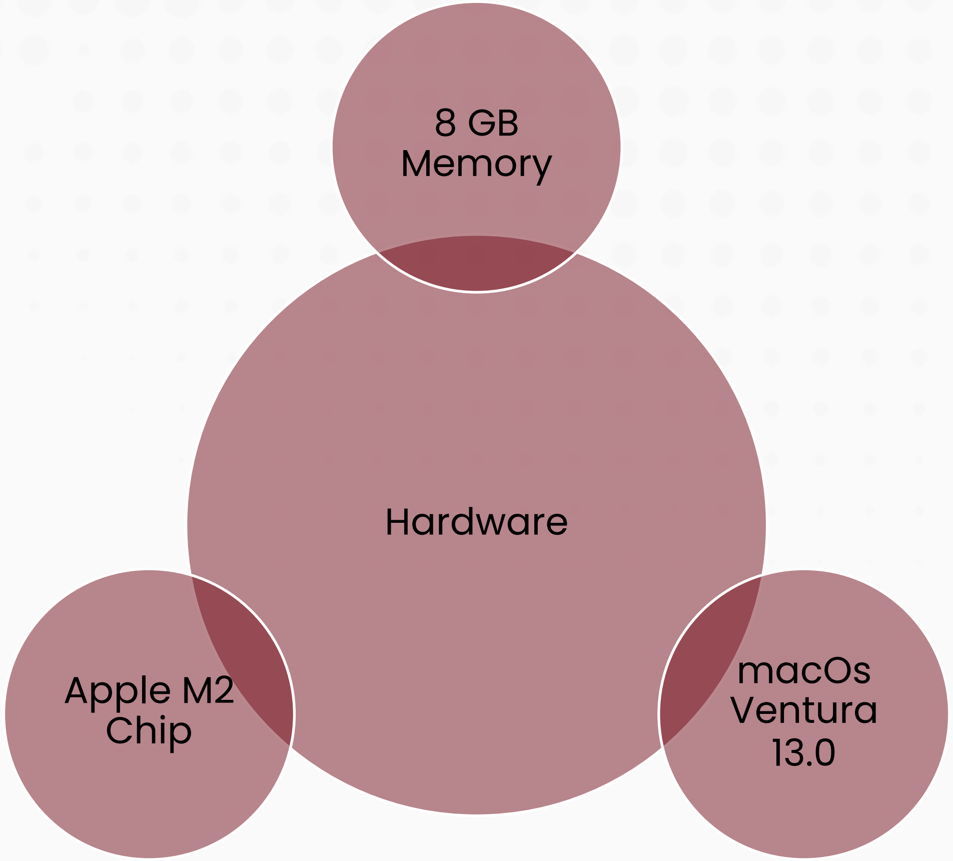
End

METHODOLOGY (DATASET)

Table 2 Criteria of Dataset

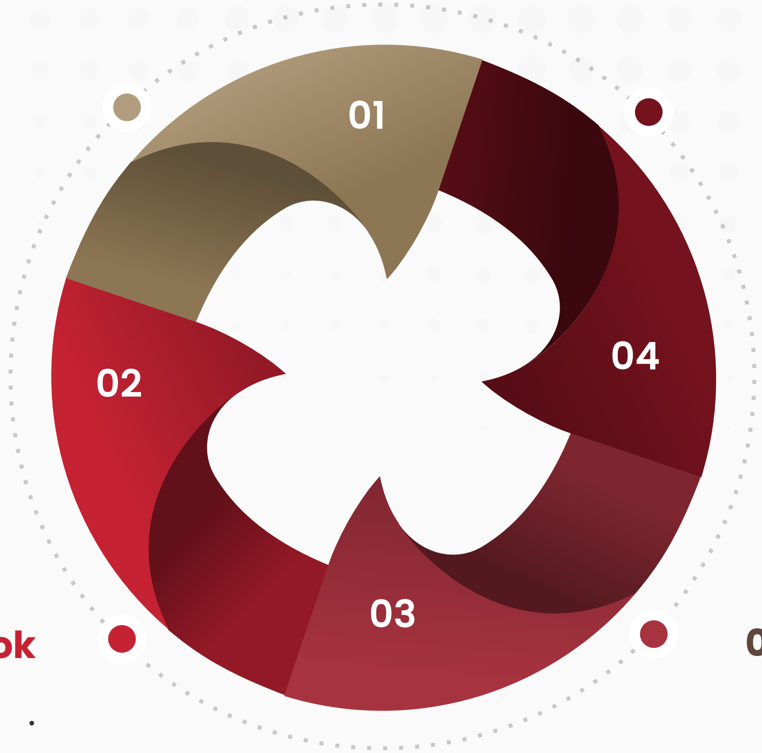
CRITERIA	VALUE
Dataset Size	400 instances/samples
Patient Type	• Chronic Kidney Disease Patient (250 instances) • Unaffected Patient (150 instances)
Attributes	25 (24 numerical and 1 class attribute)
Data Collection sources	Measurement data from blood and urine tests, as well as survey responses

METHODOLOGY



METHODOLOGY

Software



01. Visual Studio Code

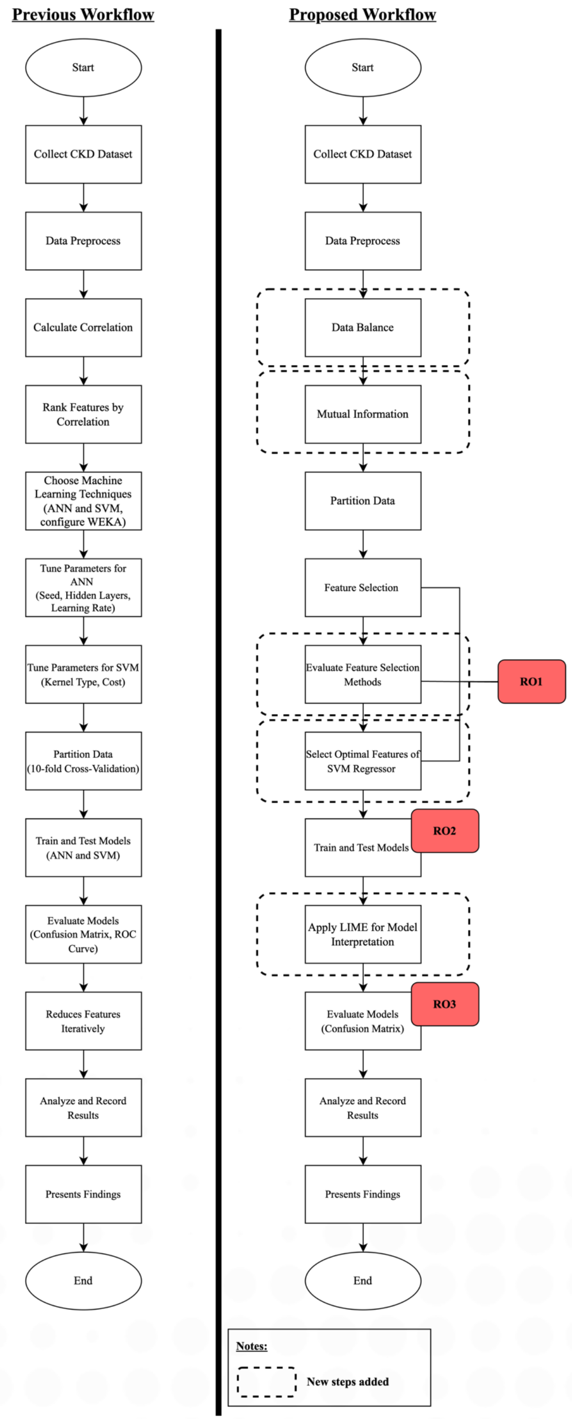
02. Jupyter Notebook

03. Microsoft Excel 2019

04. Microsoft Word 2019

METHODOLOGY

EXPERIMENTAL SETUP AND DESIGN



Previous Workflow

Start

Collect CKD Dataset

Data Preprocess

Calculate Correlation

Rank Features by Correlation

Choose Machine Learning Techniques (ANN and SVM, configure WEKA)

Tune Parameters for ANN (Seed, Hidden Layers, Learning Rate)

Tune Parameters for SVM (Kernel Type, Cost)

Partition Data (10-fold Cross-Validation)

Train and Test Models (ANN and SVM)

Evaluate Models (Confusion Matrix, ROC Curve)

Reduces Features Iteratively

Analyze and Record Results

Presents Findings

End

Proposed Workflow

Start

Collect CKD Dataset

Data Preprocess

Data Balance

Mutual Information

Partition Data

Feature Selection

Evaluate Feature Selection Methods

RO1

Select Optimal Features of SVM Regressor

Train and Test Models

RO2

Apply LIME for Model Interpretation

Evaluate Models (Confusion Matrix)

RO3

Analyze and Record Results

Presents Findings

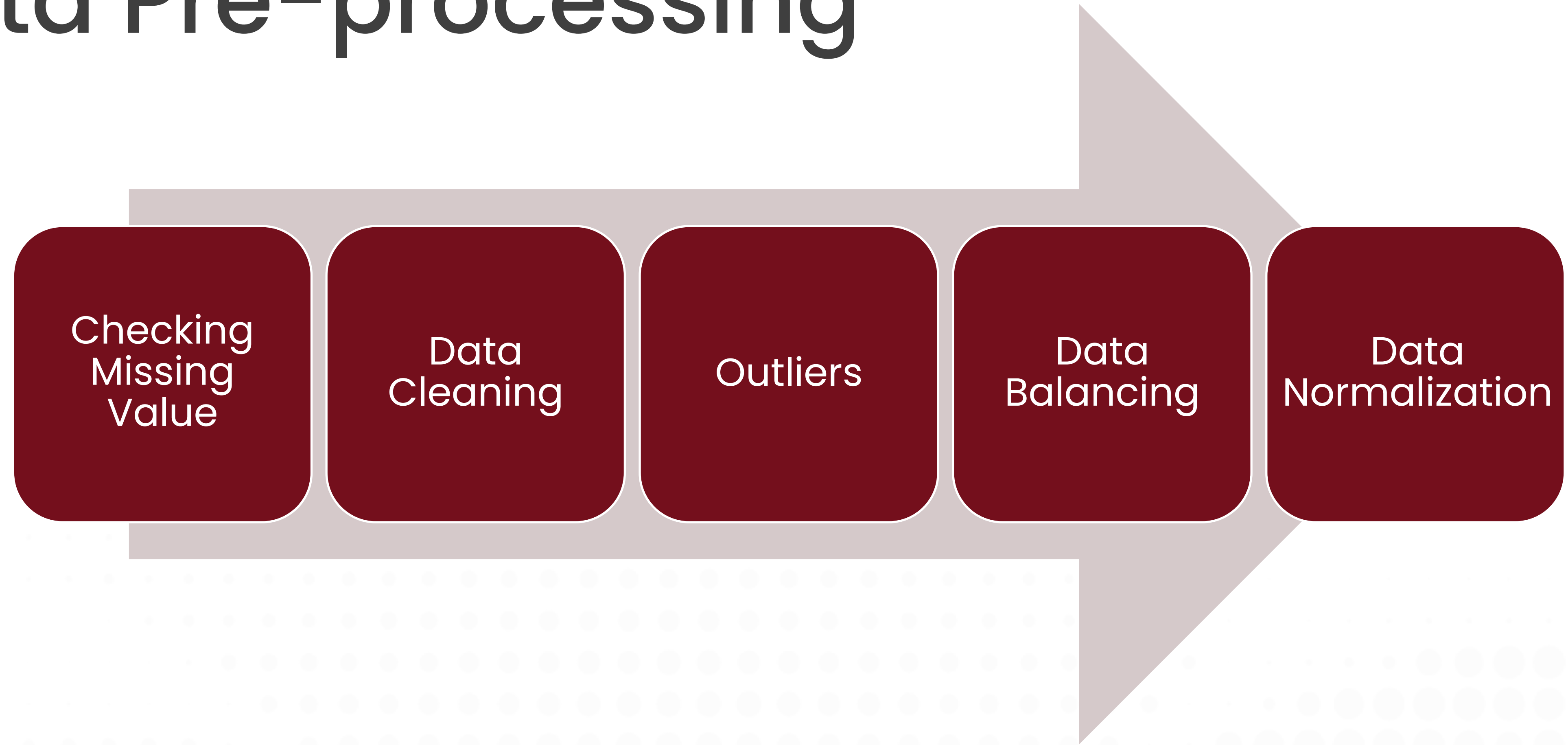
End

Notes:

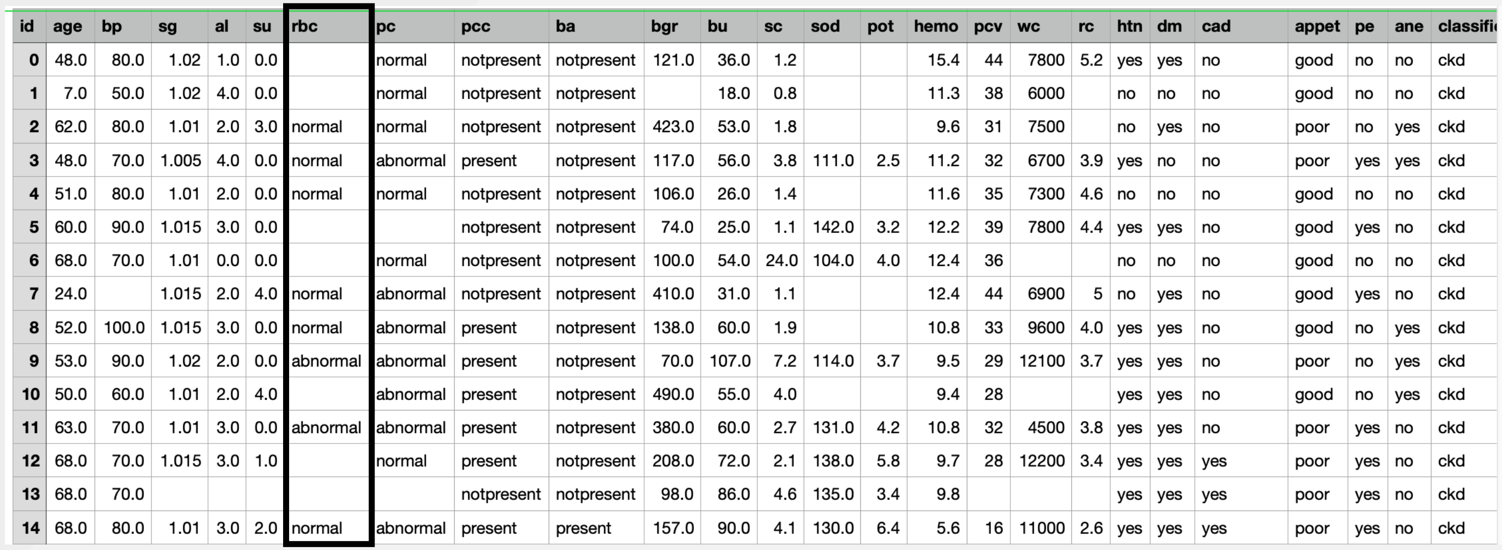
New steps added

METHODOLOGY

Data Pre-processing

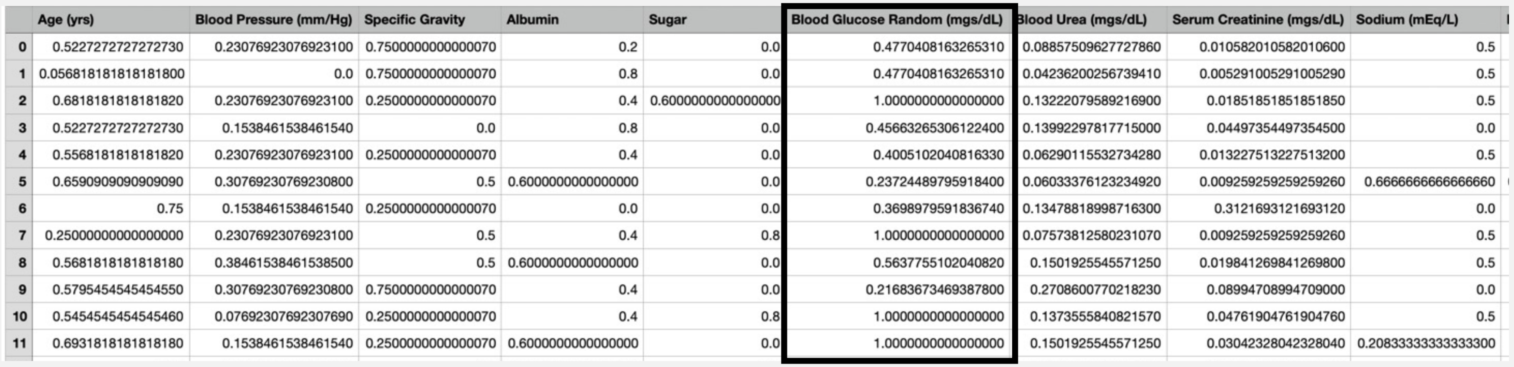


Before Data Pre-processing



id	age	bp	sg	al	su	rbc	pc	pcc	ba	bgr	bu	sc	sod	pot	hemo	pcv	wc	rc	htn	dm	cad	appet	pe	ane	classifi
0	48.0	80.0	1.02	1.0	0.0		normal	notpresent	notpresent	121.0	36.0	1.2			15.4	44	7800	5.2	yes	yes	no	good	no	no	ckd
1	7.0	50.0	1.02	4.0	0.0		normal	notpresent	notpresent		18.0	0.8			11.3	38	6000		no	no	no	good	no	no	ckd
2	62.0	80.0	1.01	2.0	3.0	normal	normal	notpresent	notpresent	423.0	53.0	1.8			9.6	31	7500		no	yes	no	poor	no	yes	ckd
3	48.0	70.0	1.005	4.0	0.0	normal	abnormal	present	notpresent	117.0	56.0	3.8	111.0	2.5	11.2	32	6700	3.9	yes	no	no	poor	yes	yes	ckd
4	51.0	80.0	1.01	2.0	0.0	normal	normal	notpresent	notpresent	106.0	26.0	1.4			11.6	35	7300	4.6	no	no	no	good	no	no	ckd
5	60.0	90.0	1.015	3.0	0.0			notpresent	notpresent	74.0	25.0	1.1	142.0	3.2	12.2	39	7800	4.4	yes	yes	no	good	yes	no	ckd
6	68.0	70.0	1.01	0.0	0.0		normal	notpresent	notpresent	100.0	54.0	24.0	104.0	4.0	12.4	36			no	no	no	good	no	no	ckd
7	24.0		1.015	2.0	4.0	normal	abnormal	notpresent	notpresent	410.0	31.0	1.1			12.4	44	6900	5	no	yes	no	good	yes	no	ckd
8	52.0	100.0	1.015	3.0	0.0	normal	abnormal	present	notpresent	138.0	60.0	1.9			10.8	33	9600	4.0	yes	yes	no	good	no	yes	ckd
9	53.0	90.0	1.02	2.0	0.0	abnormal	abnormal	present	notpresent	70.0	107.0	7.2	114.0	3.7	9.5	29	12100	3.7	yes	yes	no	poor	no	yes	ckd
10	50.0	60.0	1.01	2.0	4.0		abnormal	present	notpresent	490.0	55.0	4.0			9.4	28			yes	yes	no	good	no	yes	ckd
11	63.0	70.0	1.01	3.0	0.0	abnormal	abnormal	present	notpresent	380.0	60.0	2.7	131.0	4.2	10.8	32	4500	3.8	yes	yes	no	poor	yes	no	ckd
12	68.0	70.0	1.015	3.0	1.0		normal	present	notpresent	208.0	72.0	2.1	138.0	5.8	9.7	28	12200	3.4	yes	yes	yes	poor	yes	no	ckd
13	68.0	70.0						notpresent	notpresent	98.0	86.0	4.6	135.0	3.4	9.8				yes	yes	yes	poor	yes	no	ckd
14	68.0	80.0	1.01	3.0	2.0	normal	abnormal	present	present	157.0	90.0	4.1	130.0	6.4	5.6	16	11000	2.6	yes	yes	yes	poor	yes	no	ckd

After Data Pre-processing



	Age (yrs)	Blood Pressure (mm/Hg)	Specific Gravity	Albumin	Sugar	Blood Glucose Random (mgs/dL)	Blood Urea (mgs/dL)	Serum Creatinine (mgs/dL)	Sodium (mEq/L)
0	0.5227272727272730	0.2307692307692310	0.7500000000000070	0.2	0.0	0.4770408163265310	0.0885750962772786	0.0105820105820106	0.5
1	0.0568181818181800	0.0	0.7500000000000070	0.8	0.0	0.4770408163265310	0.0423620025673941	0.0052910052910053	0.5
2	0.6818181818181820	0.2307692307692310	0.2500000000000070	0.4	0.6000000000000000	1.0000000000000000	0.1322207958921690	0.0185185185185185	0.5
3	0.5227272727272730	0.1538461538461540	0.0	0.8	0.0	0.4566326530612240	0.1399229781771500	0.0449735449735450	0.0
4	0.5568181818181820	0.2307692307692310	0.2500000000000070	0.4	0.0	0.4005102040816330	0.0629011553273428	0.0132275132275132	0.5
5	0.6590909090909090	0.3076923076923080	0.5	0.6000000000000000	0.0	0.2372448979591840	0.0603337612323492	0.0092592592592593	0.6666666666666660
6	0.75	0.1538461538461540	0.2500000000000070	0.0	0.0	0.3698979591836740	0.1347881899871630	0.3121693121693120	0.0
7	0.2500000000000000	0.2307692307692310	0.5	0.4	0.8	1.0000000000000000	0.0757381258023107	0.0092592592592593	0.5
8	0.5681818181818180	0.3846153846153850	0.5	0.6000000000000000	0.0	0.5637755102040820	0.1501925545571250	0.0198412698412698	0.5
9	0.5795454545454550	0.3076923076923080	0.7500000000000070	0.4	0.0	0.2168367346938780	0.2708600770218230	0.0899470899470900	0.0
10	0.5454545454545460	0.0769230769230769	0.2500000000000070	0.4	0.8	1.0000000000000000	0.1373555840821570	0.0476190476190476	0.5
11	0.6931818181818180	0.1538461538461540	0.2500000000000070	0.6000000000000000	0.0	1.0000000000000000	0.1501925545571250	0.0304232804232804	0.2083333333333330

METHODOLOGY

Mutual Information

- Mutual information measures information shared between variables, capturing linear and nonlinear associations.
- It considers joint and marginal distributions, providing a comprehensive understanding of variable relationships.
- Mutual information aids in feature selection, identifying important features for prediction or classification tasks.

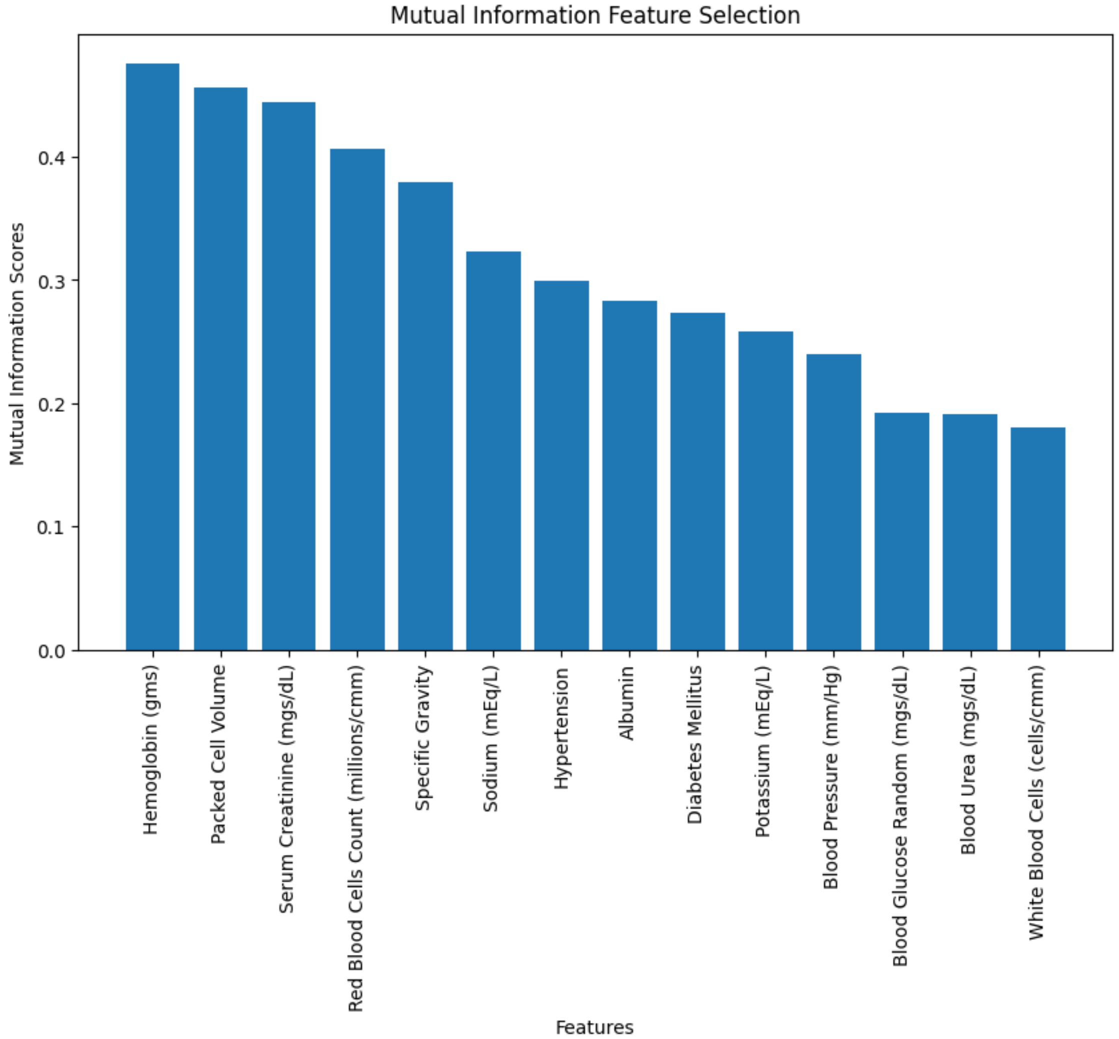


Figure 4.2: Bar Graph of the Mutual Information Score of various features

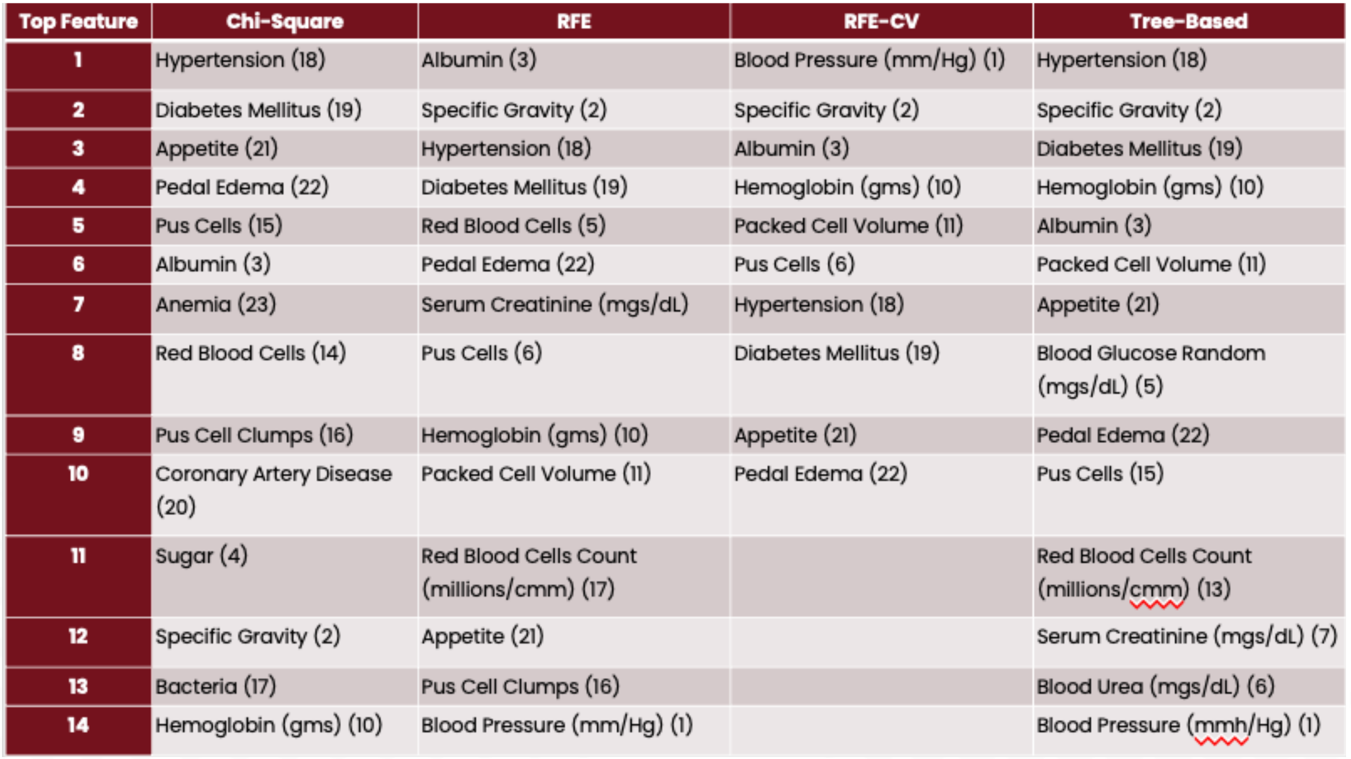
METHODOLOGY

Feature Selection

- Feature selection is a method used to choose relevant traits for a classification task.
- There are 4 feature selection methods that have been applied in this study to identify the important feature

1. Chi-Square
2. RFE
3. RFE-CV
4. Tree-Based

Table 5.1 Top Feature selected by different feature selection methods

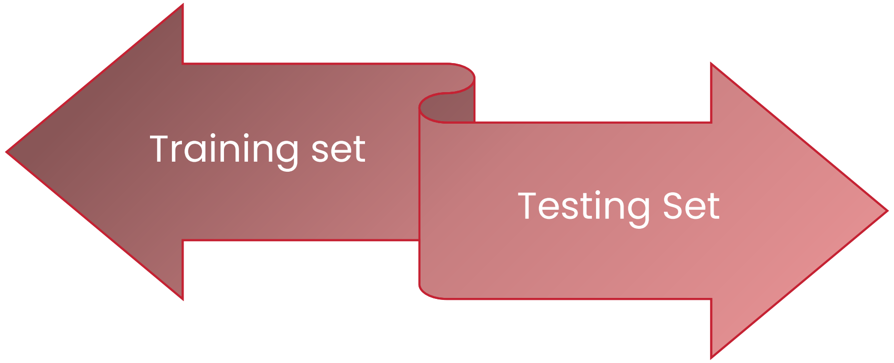


Top Feature	Chi-Square	RFE	RFE-CV	Tree-Based
1	Hypertension (18)	Albumin (3)	Blood Pressure (mm/Hg) (1)	Hypertension (18)
2	Diabetes Mellitus (19)	Specific Gravity (2)	Specific Gravity (2)	Specific Gravity (2)
3	Appetite (21)	Hypertension (18)	Albumin (3)	Diabetes Mellitus (19)
4	Pedal Edema (22)	Diabetes Mellitus (19)	Hemoglobin (gms) (10)	Hemoglobin (gms) (10)
5	Pus Cells (15)	Red Blood Cells (5)	Packed Cell Volume (11)	Albumin (3)
6	Albumin (3)	Pedal Edema (22)	Pus Cells (6)	Packed Cell Volume (11)
7	Anemia (23)	Serum Creatinine (mgs/dL)	Hypertension (18)	Appetite (21)
8	Red Blood Cells (14)	Pus Cells (6)	Diabetes Mellitus (19)	Blood Glucose Random (mgs/dL) (5)
9	Pus Cell Clumps (16)	Hemoglobin (gms) (10)	Appetite (21)	Pedal Edema (22)
10	Coronary Artery Disease (20)	Packed Cell Volume (11)	Pedal Edema (22)	Pus Cells (15)
11	Sugar (4)	Red Blood Cells Count (millions/cmm) (17)		Red Blood Cells Count (millions/cmm) (13)
12	Specific Gravity (2)	Appetite (21)		Serum Creatinine (mgs/dL) (7)
13	Bacteria (17)	Pus Cell Clumps (16)		Blood Urea (mgs/dL) (6)
14	Hemoglobin (gms) (10)	Blood Pressure (mm/Hg) (1)		Blood Pressure (mmh/Hg) (1)

METHODOLOGY

Data Splitting

- Data splitting involves dividing a dataset into training and test sets for machine learning.
- The training set is used to train models, while the test set evaluates model performance on unseen data.
- The dataset is split into X_train, y_train, X_test, and y_test subsets before feature selection techniques are applied.



set to 0.9 or 90%

Training set

Testing Set

set to 0.1 or 10%

SVM Classifier (Linear SVM)

Parameter	Settings
Model	SVC
Kernel	Linear
Random State	42

METHODOLOGY

Support Vector Machine

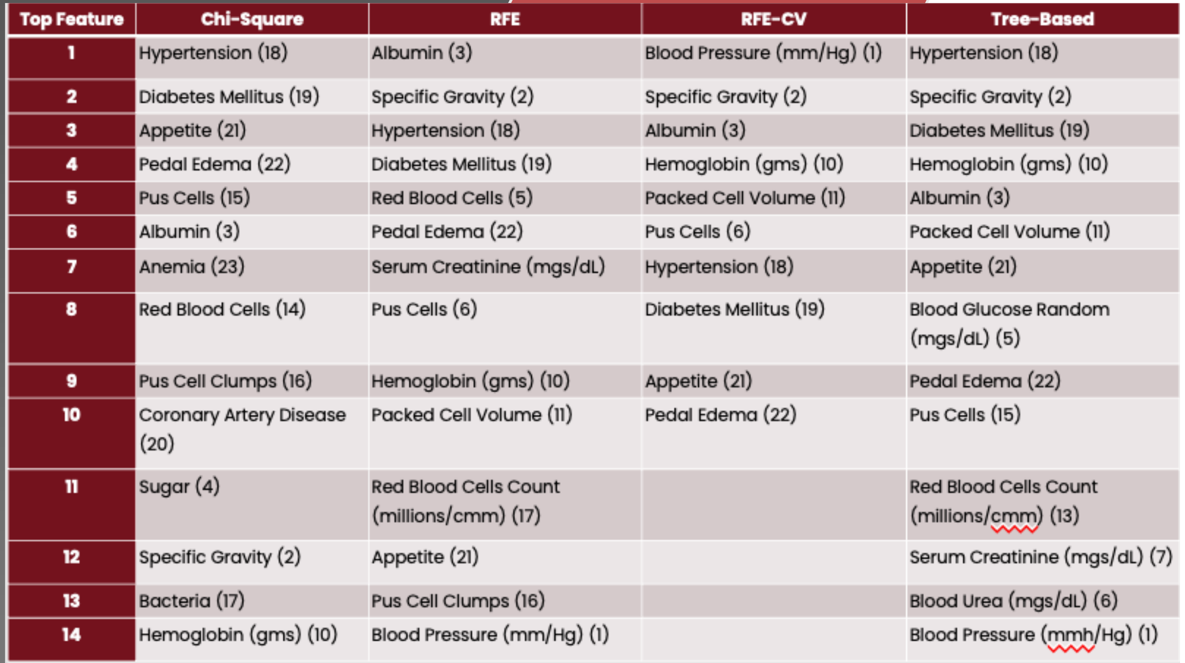
- Support Vector Machines (SVM) is a popular supervised learning approach used for classification and regression tasks.
- It is effective in handling categorization problems, works well in high-dimensional spaces, has efficient memory usage, and can utilize custom kernels for non-linear data.

Table 4.12 Past Research and Preliminary Result using top 12 features

Research	Prediction Model	Result			
		Accuracy	Precision	Recall	F1-Score
(AlMansour et al., 2019)	SVM	0.9775	0.982	0.964	1.000
Preliminary Result using parameter setting AlMansour et al., 2019	SVM	0.700	0.700	1.000	0.824

RESULTS AND DISCUSSION

Table 5.1 Top Feature selected by different feature selection methods



Top Feature	Chi-Square	RFE	RFE-CV	Tree-Based
1	Hypertension (18)	Albumin (3)	Blood Pressure (mm/Hg) (1)	Hypertension (18)
2	Diabetes Mellitus (19)	Specific Gravity (2)	Specific Gravity (2)	Specific Gravity (2)
3	Appetite (21)	Hypertension (18)	Albumin (3)	Diabetes Mellitus (19)
4	Pedal Edema (22)	Diabetes Mellitus (19)	Hemoglobin (gms) (10)	Hemoglobin (gms) (10)
5	Pus Cells (15)	Red Blood Cells (5)	Packed Cell Volume (11)	Albumin (3)
6	Albumin (3)	Pedal Edema (22)	Pus Cells (6)	Packed Cell Volume (11)
7	Anemia (23)	Serum Creatinine (mgs/dL)	Hypertension (18)	Appetite (21)
8	Red Blood Cells (14)	Pus Cells (6)	Diabetes Mellitus (19)	Blood Glucose Random (mgs/dL) (5)
9	Pus Cell Clumps (16)	Hemoglobin (gms) (10)	Appetite (21)	Pedal Edema (22)
10	Coronary Artery Disease (20)	Packed Cell Volume (11)	Pedal Edema (22)	Pus Cells (15)
11	Sugar (4)	Red Blood Cells Count (millions/cmm) (17)		Red Blood Cells Count (millions/cmm) (13)
12	Specific Gravity (2)	Appetite (21)		Serum Creatinine (mgs/dL) (7)
13	Bacteria (17)	Pus Cell Clumps (16)		Blood Urea (mgs/dL) (6)
14	Hemoglobin (gms) (10)	Blood Pressure (mm/Hg) (1)		Blood Pressure (mmh/Hg) (1)

- Njoud Abdullah Almansour *et al.* (2019) found that using the top 12 features achieved the highest accuracy of 97.75% for CKD prediction.
- This research compares subsets of 6, 10, 12, and 14 features using various feature selection methods.
- RFE-CV identified 10 optimal features through iterative cross-validation, aiming to enhance model accuracy and mitigate overfitting.
- Selected features like Hypertension, Diabetes Mellitus, Albumin, and Specific Gravity consistently emerged as critical predictors across Chi-Square, RFE, RFE-CV, and Tree-Based methods.

RESULTS AND DISCUSSION

Results of SVM Regression

- The best approach is Recursive Feature Elimination (RFE) with six features,
  - ❑ the lowest Mean Absolute Error (MAE) of 0.2181
  - ❑ a high Pearson correlation value of 0.8333,
  - ❑ the highest Spearman rank-order correlation coefficient of 0.863.

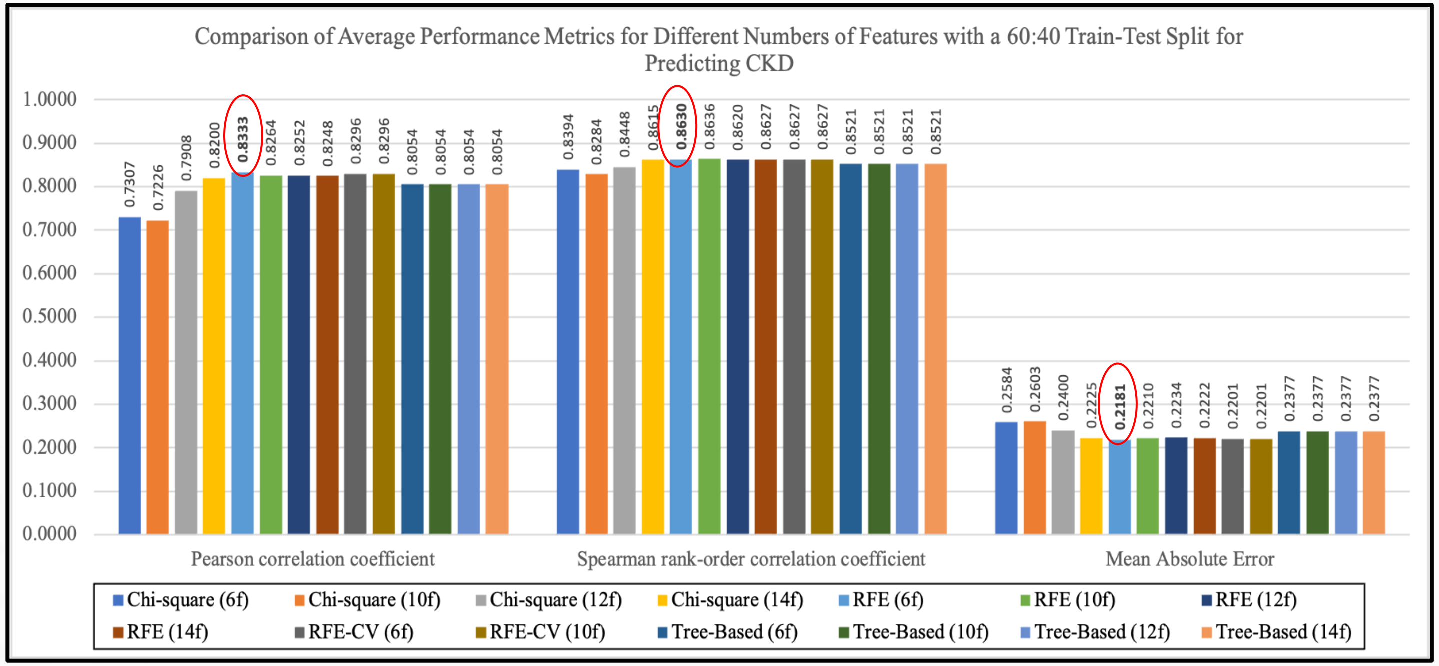


Figure 5.1 Comparison of Average Performance Metrics for Different Numbers of Features with a 60:40 Train-Test Split for Predicting CKD.

RESULTS AND DISCUSSION

Results of SVM Regression

- The best approach is Recursive Feature Elimination (RFE) with six features,
  - ❑ Mean Absolute Error (MAE): 0.2211
  - ❑ Pearson Correlation Coefficient: 0.8254
  - ❑ Spearman Rank-Order Correlation: 0.8635

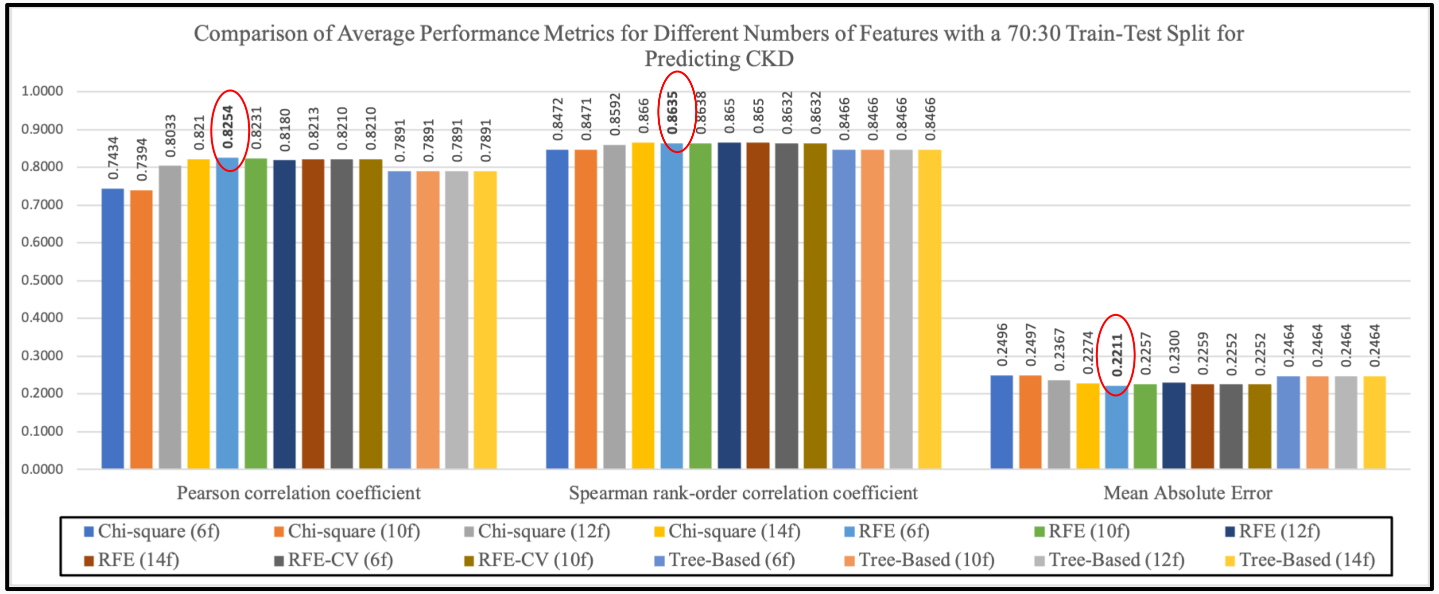


Figure 5.2 Comparison of Average Performance Metrics for Different Numbers of Features with a 70:30 Train-Test Split for Predicting CKD.

RESULTS AND DISCUSSION

- The best approach is Recursive Feature Elimination (RFE) with **six** features,
  - ❑ Mean Absolute Error (MAE): 0.2137 (lowest value)
  - ❑ Pearson Correlation Coefficient: 0.8388
  - ❑ Spearman Rank-Order Correlation: 0.8612

Results of SVM Regression

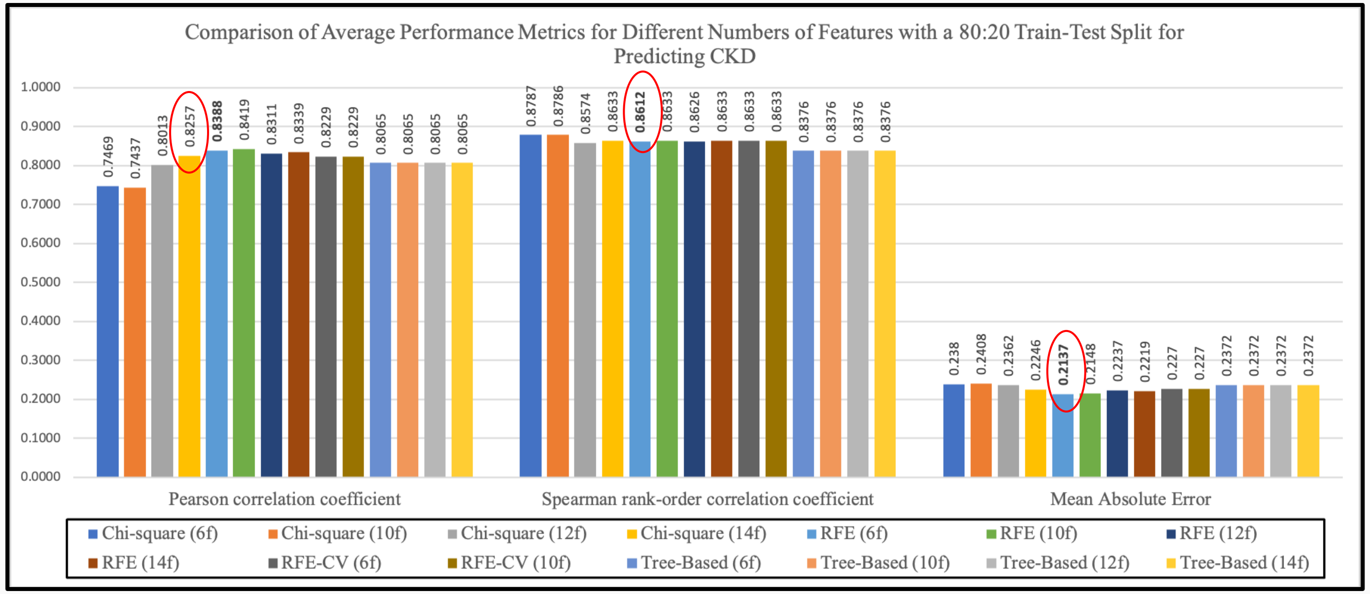


Figure 5.3 Comparison of Average Performance Metrics for Different Numbers of Features with a 80:20 Train-Test Split for Predicting CKD.

RESULTS AND DISCUSSION

Results of SVM Regression

- The best approach is Recursive Feature Elimination (RFE) with **ten** features,
  - ❑ Mean Absolute Error (MAE): 0.2151
  - ❑ Pearson Correlation Coefficient: 0.8504
  - ❑ Spearman Rank-Order Correlation: 0.8655

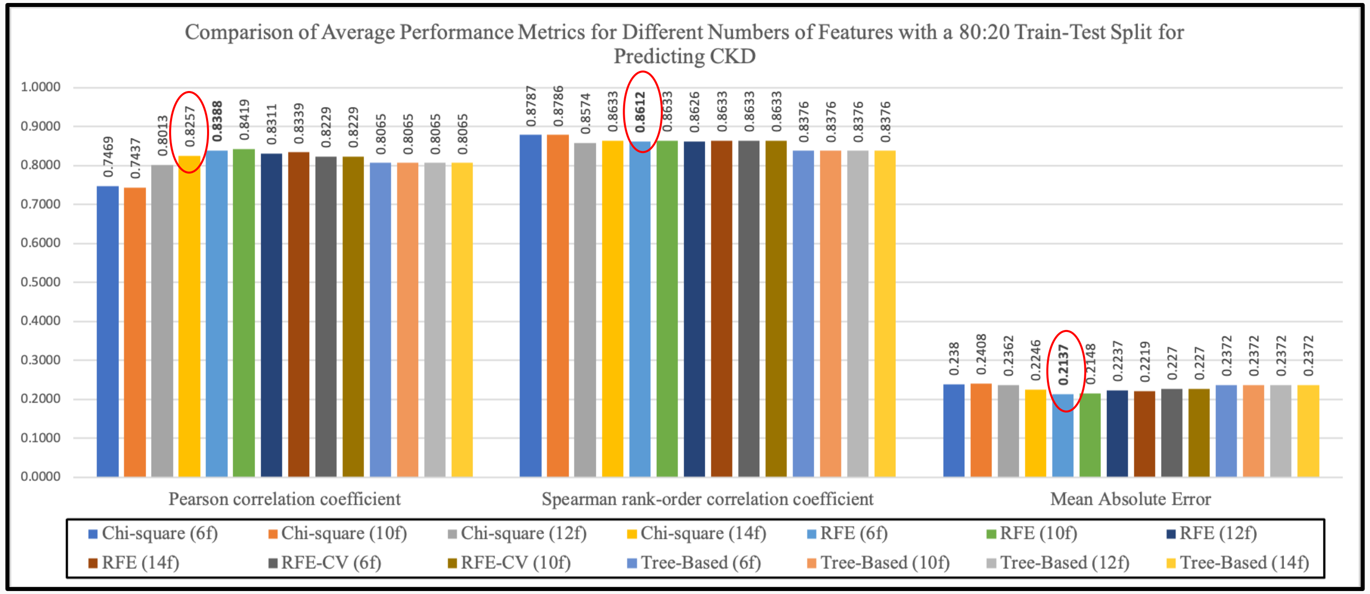


Figure 5.4 Comparison of Average Performance Metrics for Different Numbers of Features with a 90:10 Train-Test Split for Predicting CKD.

RESULTS AND DISCUSSION

Results of SVM Regression

Table Summary Comparison of Average Performance Metrics for Different Numbers of Features with different Train-Test Split for Predicting CKD.

Methods	Features	Train-Test Split	Pearson correlation coefficient	Spearman rank-order correlation coefficient	Mean Absolute Error
RFE	6	60:40	0.8333	0.8630	0.2181
RFE	6	70:30	0.8254	0.8635	0.2211
RFE	6	80:20	0.8388	0.8612	0.2137
RFE	10	90:10	0.8504	0.8655	0.2151

- RFE consistently performs well across different train-test splits and feature counts.
- The optimal configuration varies slightly with the split ratio, with RFE (10 features) performing best for the 90:10 split in terms of Pearson and Spearman correlations and MAE

RESULTS AND DISCUSSION

Results of SVM Prediction Model

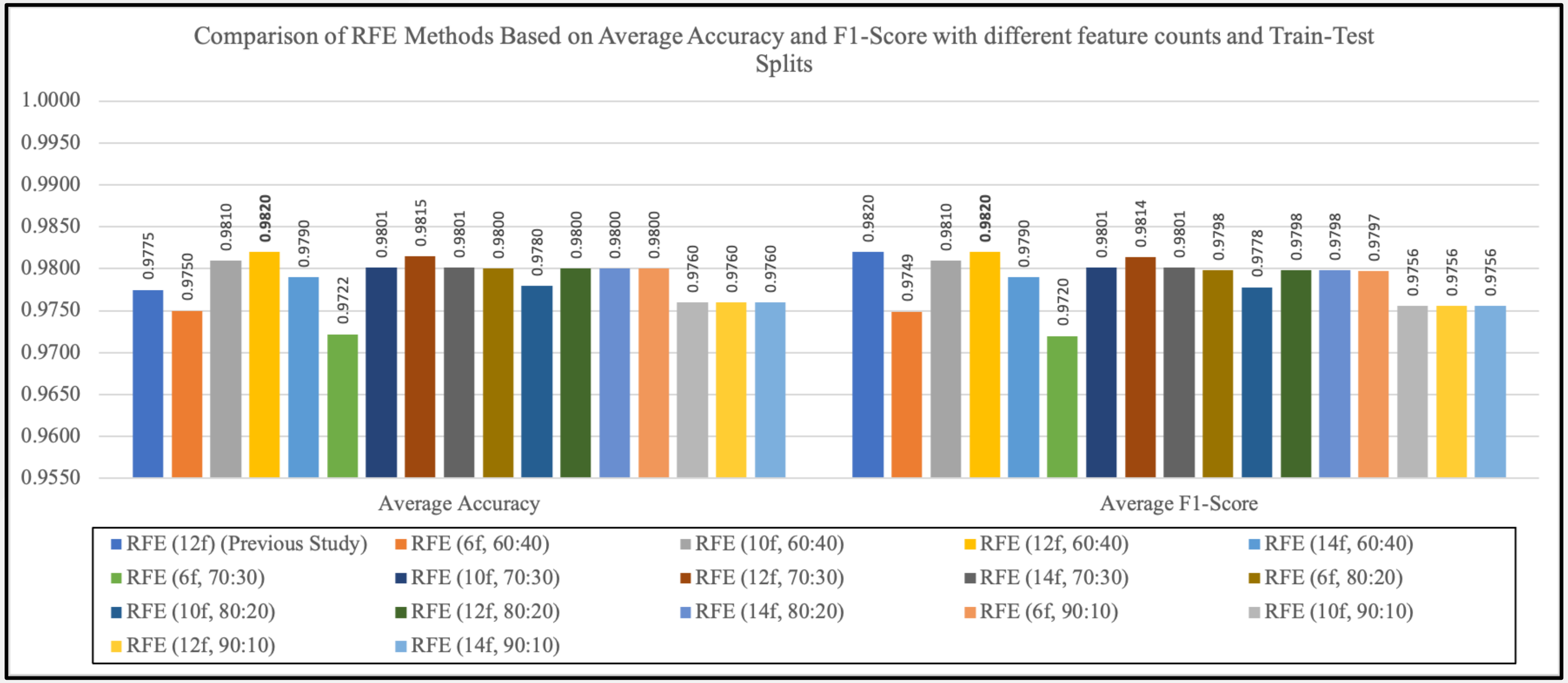


Figure 5.6 Comparison of RFE Methods Based on Average Accuracy and F1-Score with different feature counts and Train-Test Splits.

RESULTS AND DISCUSSION

Results of SVM Prediction Model

Table Comparisons of SVM Result between Previous and Current Study

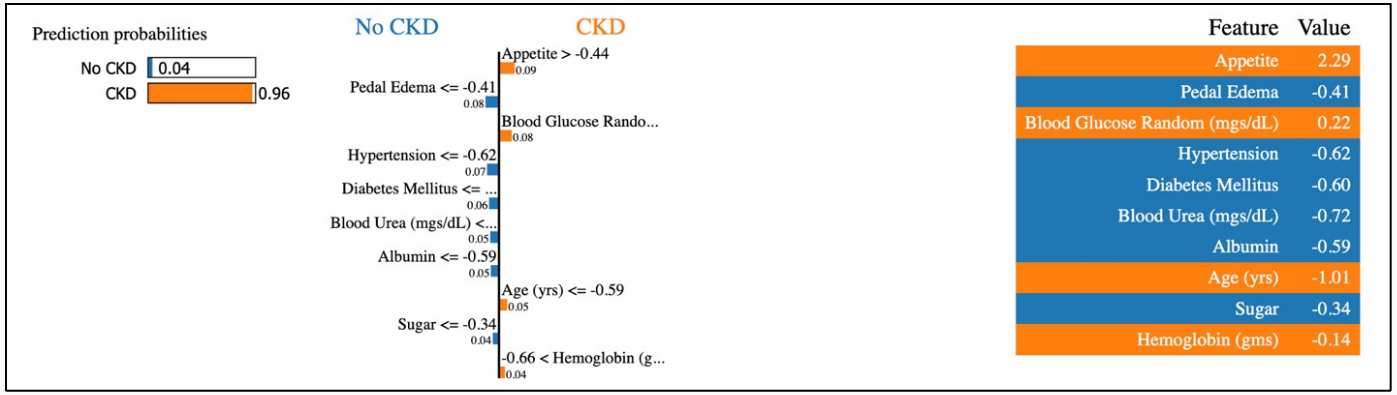
Method	Features	Train Test Split	Accuracy	F1-Score
RFE (Previous Study)	12	90:10	0.9775	0.9820
RFE	12	60:40	0.9820	0.9820

- Previous Study (Njoud Abdullah Almansour *et al.* (2019)): SVM achieved 97.75% accuracy with 12 features using a 90:10 split and 10-fold cross-validation.
- Current Study:
    - Achieved 98.20% accuracy using RFE with the top 12 features in a 60:40 split.
    - Achieved 97.60% accuracy with the same 90:10 split and 12 features, slightly lower than the previous study.
- Highlighted the impact of train-test split selection on performance, with the 60:40 split showing the greatest improvement.

RESULTS AND DISCUSSION

- Prediction: SVM model predicts 'CKD' class with 96% confidence.
- Positive Contributions:
  - High appetite value: 2.29
  - Moderate blood glucose random value: 0.22
- Negative Contributions:
  - Absence of pus cell clumps: <= -0.31
  - Absence of diabetes mellitus: <= -0.61
  - Absence of hypertension: <= -0.62
  - Low albumin level: <= -0.58
  - Younger age: <= -1.03
  - Slightly lower blood pressure: <= -0.46
  - Pedal edema: <= -0.4

Local Interpretable Model-agnostic Explanation (LIME) Interpretation



Feature	Value
Appetite	2.29
Pedal Edema	-0.41
Blood Glucose Random (mgs/dL)	0.22
Hypertension	-0.62
Diabetes Mellitus	-0.60
Blood Urea (mgs/dL)	-0.72
Albumin	-0.59
Age (yrs)	-1.01
Sugar	-0.34
Hemoglobin (gms)	-0.14

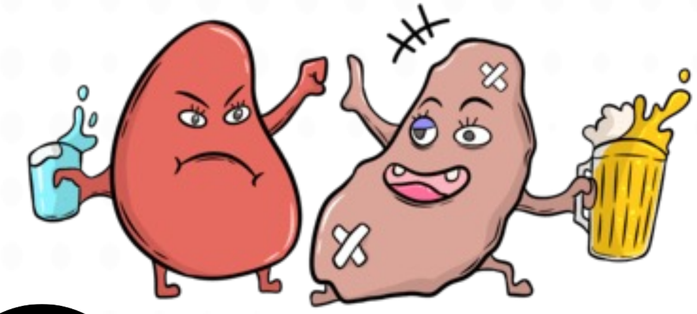
Figure 5.7 The Outcomes of LIME.

RESULTS AND DISCUSSION

Comparison between Local Interpretable Model-agnostic Explanation (LIME) and other feature selections

- There are significant overlaps in features identified by different feature selection techniques.
- Hypertension is an important feature in LIME and is also top-ranked by both Chi-Square and Tree-Based methods.
- LIME highlights low albumin values, which is also top-ranked by RFE.
- Blood pressure is ranked highly by RFE-CV and is significant in LIME's findings.
- LIME emphasizes features like age and blood glucose random, which do not frequently appear in the top ranks of other feature selection methods.
- Overall, LIME and traditional feature selection techniques show significant compatibility on features such as diabetes mellitus, pedal edema, blood pressure, albumin, and hypertension.

CONCLUSION



✓ RO1

- Reviewed sources like journals and articles from platforms such as ResearchGate and IEEE.
- Applied four feature selection methods: tree-based selection, chi-square analysis, RFE, and RFE with cross-validation.

✓ RO2

- Selected the optimal feature set for the SVM model.
- Trained and tested the model to accurately identify CKD cases.

✓ RO3

- Used correlation coefficients, regression analysis, and confusion matrix to assess performance.
- Explained accuracy and F1-score with the confusion matrix.

CONCLUSION

Contributions

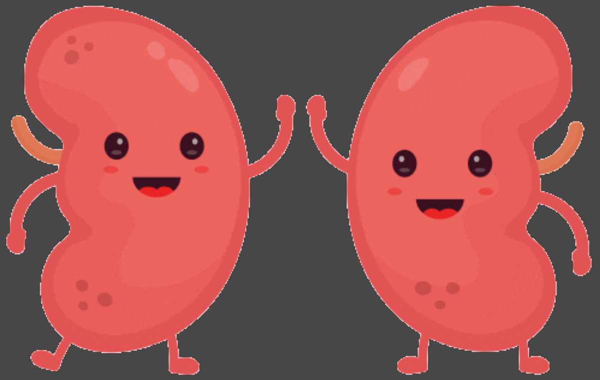
- Data imbalance issues were addressed by implementing SMOTE oversampling, significantly improving the model's sensitivity to CKD cases.
- LIME was used for interpreting individual predictions, which improved understanding of the model’s decision-making process.

Limitation

The study was constrained by a small dataset size, potentially limiting the robustness and generalizability of findings.

Future works

- Increase dataset size to improve model robustness and generalizability.
- Validate the findings, especially the perfect F1-score observed, across multiple datasets to ensure consistency and robustness in different circumstances.
- Investigate advanced machine learning methods to enhance model accuracy and reliability for CKD prediction



THANK YOU

UTM
UNIVERSITI TEKNOLOGI MALAYSIA

In the Name of God for Mankind

utm.my univteknologimalaysia utmofficial