

CONTENTS

TITLE PAGE	i
APPROVAL PAGE	ii
DECLARATION PAGE	iii
MOTTO PAGE	iv
PREFACE	v
CONTENTS	vi
LIST OF TABLES	ix
LIST OF FIGURES	xi
ABSTRACT	xii
ABSTRACT	xiii
I INTRODUCTION	1
1.1 Research Background	1
1.2 Research Problem	4
1.3 Research Scope	4
1.4 Research Objective	4
1.5 Research Benefit	5
1.6 Research Methodology	5
1.7 Writing Systematics	7
II LITERATURE REVIEW	8
III Theoretical Framework	12
3.1 Noise and Outlier	12
3.2 DBSCAN	14
3.3 K-Means Clustering	17
3.4 Naive Bayes	19

3.5	Evaluation Metrics	22
3.5.1	Confusion Matrix	22
3.5.2	Accuracy	23
3.5.3	Precision	23
3.5.4	Recall	23
3.5.5	F1-Score	23
3.6	Diabetes Mellitus	24
IV	RESEARCH METHODOLOGY	26
4.1	Research Description	26
4.2	Tools and Materials	27
4.3	Research Steps	28
4.3.1	Dataset Gathering	28
4.3.2	System Design	29
4.3.3	Replacing NaN and Impossible Values	30
4.3.4	Obtain Optimum Values of DBSCAN Parameters	31
4.3.5	Outlier Removal with DBSCAN	33
4.3.6	Noise Removal with K-Means Clustering	33
4.3.7	Classification with Naive Bayes	34
4.4	Evaluation Design	35
4.4.1	K-Fold Cross Validation	35
4.4.2	Comparison with Past Researches	36
V	IMPLEMENTATION	38
5.1	Importing Necessary Libraries	38
5.2	Replacing NaN and Impossible Values	39
5.3	Outlier Removal	40
5.3.1	DBSCAN Implmentation	40
5.3.2	Obtaining Optimum Epsilon	43
5.3.3	Outlier Removal Process	44
5.4	Noise Removal	45
5.4.1	K-Means Clustering Implementation	45
5.4.2	Noise Removal Process	48
5.5	Naive Bayes Classification and Evaluation	50

VI RESULT AND DISCUSSION	51
6.1 Data Preprocessing Result	51
6.2 Outlier Removal Result	53
6.2.1 Obtaining Optimum Epsilon Result	53
6.2.2 DBSCAN Clustering Result	54
6.2.3 Outlier Removal Process Result	56
6.3 Noise Removal Result	57
6.3.1 K-Means Clustering Result	57
6.3.2 Noise Removal Process Result	59
6.4 Evaluation Result	61
6.4.1 Confusion Matrix Result	61
6.4.2 Performance Evaluation Result	63
6.5 Comparison with Past Researches	67
VII CONCLUSION AND FUTURE WORKS	70
7.1 Conclusion	70
7.2 Suggestions	70
REFERENCES	72

LIST OF TABLES

2.1	Literature Review Comparison Table	10
3.1	Noise and Outlier Comparison	14
3.2	Confusion Matrix	22
3.3	Diabetes Risk Factors	25
4.1	PIDD Dataset Attributes	29
4.2	Past Researches for Comparison	37
6.1	Confusion Matrix Values of Dataset 1	62
6.2	Confusion Matrix Values of Dataset 2	62
6.3	Confusion Matrix Values of Dataset 3	63
6.4	Confusion Matrix Values of Dataset 4	63
6.5	Classification Report for Dataset 1	64
6.6	Classification Report for Dataset 2	64
6.7	Classification Report for Dataset 3	65
6.8	Classification Report for Dataset 4	65
6.9	Result Comparison for All Dataset	66
6.10	Comparison of Accuracies with Past Researches	69

LIST OF FIGURES

3.1	DBSCAN Clustering Flowchart	17
3.2	K-Means Clustering Flowchart	18
3.3	Naive Bayes Flowchart	21
4.1	Research Description Flowchart	26
4.2	System Design Flowchart	30
4.3	Replacing NaN and Impossible Values Flowchart	31
4.4	Obtaining Optimum Epsilon Flowchart	32
4.5	Distance Graph to Obtain Knee	32
4.6	Outlier Removal Process Flowchart	33
4.7	Noise Removal Process Flowchart	34
4.8	K-Fold Cross Validation Evaluation Process Flowchart	36
5.1	Installing and Importing Necessary Libraries	39
5.2	Import Dataset from Drive Using Pandas	39
5.3	Change Impossible Values to NaN and Replace with Mean	40
5.4	Function to Get Neighbouring Data	41
5.5	Function to Determine Core Point	41
5.6	Function to Visit Neighbouring Data	42
5.7	Main Runner Function	42
5.8	Full DBSCAN Class	43
5.9	Acquiring Optimum Epsilon using Knee Method	44
5.10	Apply DBSCAN on Data and Remove Detected Outlier	44
5.11	Full Function for Outlier Removal Process	45
5.12	Function to Assign Cluster on Each Data Points	46
5.13	Function to Move Centroids to Current Mean on New Iterations	46
5.14	Main Runner Function for K-Means Clustering	47
5.15	Full K-Means Clustering Class	48
5.16	Apply K-Means Clustering to Data	49
5.17	Remove Noise Based on Formed Clusters	49
5.18	Full Function for Noise Removal Process	49
5.19	Features and Outcomes Separation and Instances Initialization	50
5.20	10-Fold Cross Validation on Naive Bayes Classifier	50

6.1	Imported PIDD Dataset in data frame	51
6.2	PIDD Data Frame After Mean Replacement Process	52
6.3	Two Dimensional Scatter Plot of PIDD Dataset	52
6.4	Resulting Distance Graph	53
6.5	Distance Graph with Obtained Knee	54
6.6	Obtained Optimum Epsilon from Distance Index	54
6.7	Two Dimensional Scatter Plot of PIDD Dataset Before DBSCAN	55
6.8	Two Dimensional Scatter Plot of PIDD Dataset with DBSCAN Result	55
6.9	PIDD Data Frame with DBSCAN Labels Column	56
6.10	Two Dimensional Scatter Plot of PIDD Dataset after Outlier Removal Process	56
6.11	PIDD Data Frame after Outlier Removal Process	57
6.12	Two Dimensional Scatter Plot of PIDD Dataset Before K-Means Clus- tering	58
6.13	Two Dimensional Scatter Plot of PIDD Dataset with K-Means Clus- tering Result	58
6.14	PIDD Data Frame with K-Means Cluster Labels Column	59
6.15	Two Dimensional Scatter Plot of PIDD Dataset After Noise Removal Process	60
6.16	PIDD Data Frame After Noise Removal Process	60
6.17	Comparison for All Dataset Bar Chart	66