

PREDICTION OF DIABETES USING MACHINE LEARNING CLASSIFIERS

BY

SAFIA AHMED

ID: 181-15-1786

This Report Presented in Partial Fulfillment of the Requirements for the Degree of Bachelor of Science in Computer Science and Engineering

Supervised By

Tasfia Anika Bushra

Lecturer

Department of CSE

Daffodil International University

Co-Supervised By

Mohammad Jahangir Alam

Lecturer

Department of CSE

Daffodil International University



DAFFODIL INTERNATIONAL UNIVERSITY

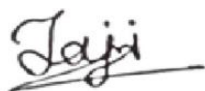
DHAKA, BANGLADESH

JANUARY 2022

APPROVAL

This Project titled “**PREDICTION OF DIABETES USING MACHINE LEARNING CLASSIFIERS**”, submitted by Safia Ahmed to the Department of Computer Science and Engineering, Daffodil International University, has been accepted as satisfactory for the partial fulfillment of the requirements for the degree of B.Sc. in Computer Science and Engineering and approved as to its style and contents. The presentation has been held on 18 January 2022.

BOARD OF EXAMINERS



Tajim Md. Niamat Ullah Akhund
Lecturer

Department of Computer Science and Engineering
Daffodil International University

Internal Examiner



Mohammad Jahangir Alam
Lecturer

Department of Computer Science and Engineering
Daffodil International University

Internal Examiner



Dr. Dewan Md. Farid
Associate Professor

Department of Computer Science and Engineering
United International University, Bangladesh

External Examiner

DECLARATION

I hereby declare that, this project has been done by me under the supervision of **Tasfia Anika Bushra, Lecturer, Department of CSE** Daffodil International University. I also declare that neither this project nor any part of this project has been submitted elsewhere for award of any degree or diploma.

Supervised by:



Tasfia Anika Bushra

Lecturer
Department of CSE
Daffodil International University

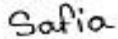
Co-Supervised by:



Mohammad Jahangir Alam

Lecturer
Department of CSE
Daffodil International University

Submitted by:



Safia Ahmed

ID: 181-15-1786
Department of CSE
Daffodil International University

ACKNOWLEDGEMENT

First I express my heartiest thanks and gratefulness to almighty God for His divine blessing makes me possible to complete the final year project successfully.

I really grateful and wish my profound my indebtedness to **Tasfia Anika Bushra, Lecturer**, Department of CSE, Daffodil International University, Dhaka. Deep Knowledge & keen interest of my supervisor in the field of “*Machine Learning*” to carry out this project. Her endless patience, scholarly guidance, continual encouragement, constant and energetic supervision, constructive criticism, valuable advice, reading many inferior draft and correcting them at all stage have made it possible to complete this project.

I would like to express my heartiest gratitude to **Dr. Touhid Bhuiyan**, Head, Department of CSE, for his kind help and also to other faculty member and the staff of CSE department of Daffodil International University.

I would like to thank my entire course mate in Daffodil International University, who took part in this discuss while completing the course work.

Finally, I must acknowledge with due respect the constant support and patients of my parents.

ABSTRACT

Diabetes is a long-term condition that affect our body's ability to convert food into energy. It is a condition in which the blood glucose level is high. Insulin is a hormone which help glucose to move to cell and produce energy. In the body of diabetes patient this procedure does not work properly. Diabetes causes a plethora of problems in our body. Our kidneys, eyes, heart, and other organs are all affected. However, with the advancement of data mining and machine learning technology, a solution to this problem has been discovered. This paper details our research into using machine learning classifiers to predict diabetes in women at an early stage. When it comes to accuracy, the Gradient Boosting algorithm is at the top of the list. It has 88.74 percent accuracy rate, which is significantly higher than the other algorithms. Pregnancies, glucose, blood pressure, BMI, insulin, age, and other common factors linked to chronic disease are being investigated in this study.

TABLE OF CONTENTS

CONTENTS	PAGE
Board of examiners	ii
Declaration	iii
Acknowledgements	iv
Abstract	V

CHAPTER

CHAPTER 1: Introduction 1-4

1.1 Introduction	1-2
1.2 Motivations	2
1.3 Research Questions	3
1.4 Expected Outcome	3
1.5 Report Layout	4

Chapter 2: Background 5-9

2.1 Preliminaries	5
2.2 Related Works	5-8
2.3 Research summary	9
2.4 Scope & Challenges	9

Chapter 3: Research Methodology	10-22
3.1 Introduction	10
3.2 Research Subject and Instrumentation	11
3.3 Data set	11-12
3.4 Data Visualization	12-16
3.5 Data preprocessing	17-18
3.5.1 Missing value identification	17
3.5.2 Missing value handling	17
3.5.3 Feature selection	18
3.5.4 Data standardization	18
3.6 Dataset train and test method	18
3.7 Detail Working Flowchart	19
3.8 Algorithms implementation	20-23
3.8.4 Logistic Regression	20
3.8.1 Support Vector Classifier	20-21
3.8.3 Decision Tree	21
3.8.2 Naive Bayes	21
3.8.5 Random Forest Machine Learning	22
3.8.6 K Nearest Neighbors	22
3.8.6 Gradient Boosting	23
CHAPTER 4 : Experiment result and discuss	24-33
4.1 Experimental Results & Analysis	24

4.1.1 Accuracy of models	25-26
4.1.2 Confusion Matrix and Heat Map of Support Vector	27
4.1.3 Confusion Matrix and Heat Map of Logistic Regression	28
4.1.4 Confusion Matrix and Heat Map of Naive Bayes	29
4.1.5 Confusion Matrix and Heat Map of Decision Tree	30
4.1.7 Confusion Matrix and Heat Map of K Nearest Neighbors	31
4.1.6 Confusion Matrix and Heat Map of Random Forest	32
4.1.8 Confusion Matrix and Heat Map of Gradient Boosting	33
4.2 Evaluation measures	34
Chapter 5: Impact on Society, Environment and Sustainability	35-36
5.1 Impact On Society	35
5.2 Impact On Environment	35
5.3 Ethical Expects	36
5.4 Sustainability Plan	36
Chapter 6: Summary, Conclusion, Recommendation and Implication for Future Research	37-38
6.1 Conclusion	37
6.2 Future Work	38
REFERENCES	39-40

LIST OF FIGURES

FIGURES	PAGE NO
Figure 3.1: Model Diagram	10
Figure 3.4.1: Outcome Cases	13
Figure 3.4.2: Histogram relation between Single attribute and outcome	14
Figure 3.4.3: Density plot to represent numeric variable's distribution	15
Figure 3.4.4: Heat map to represent the co-relation	16
Figure 3.5.2: Total missing values in each attributes	17
Figure 3.7: Detail workflow	19
Figure 4.1.1.1: Models accuracy	25
Figure 4.1.1.2: Models accuracy diagram	26
Figure 4.1.2.1: Confusion Matrix of SVC	27
Figure 4.1.2.2: Heat map of SVC	27
Figure 4.1.3.1: Confusion Matrix of LR	28
Figure 4.1.3.2: Heat map of LR	28
Figure 4.1.4.1: Confusion Matrix of NB	29
Figure 4.1.4.2: Heat map of NB	29
Figure 4.1.5.1.: Confusion Matrix of DT	30
Figure 4.1.5.2: Heat map of DT	30
Figure 4.1.6.1: Confusion Matrix of KNN	31
Figure 4.1.6.2: Heat map of KNN	31

Figure 4.1.7.1: Confusion Matrix of RF	32
Figure 4.1.7.2: Heat map of RF	32
Figure 4.1.8.1: Confusion Matrix of GB	33
Figure 4.1.8.2: Heat map of GB	33

LIST OF TABLES

TABLES	PAGE NO
Table 2.2: Related studies on Diabetes Prediction	6-8
Table 3.3.1 : PIDD Dataset Description	11
Table 3.3.2 :The attributes in PIDD	12
Table 4.1: Train test dataset ratio	24
Table 6.1: Results	37

CHAPTER 1

Introduction

1.1 Introduction

Diabetes is one of the world's fastest-growing and deadly diseases. It is a collection of metabolic illnesses in which a person's blood glucose levels are elevated either because the body generates insufficient insulin or because the body's insulin is not adequately absorbed by the body's cells. Diabetic patients are likely to be more impotent in the face of a higher risk of micro vascular damage, and as a result, long-term complications of cardio-vascular diseases are the primary cause of death. There are Three type of Diabetes Mellitus; Type 1 is known as Diabetes Mellitus, in this type human body cannot produce insulin. Patient with Diabetes Mellitus needs to produce insulin.

Type 2 Diabetes Mellitus is different from type 1, it does not depends on production of insulin. It happens when body produce insulin but it cannot be utilized. Type-3 Gestational Diabetes is characterized by an increase in blood sugar levels in pregnant women who have diabetes that has not been recognized previously. Diabetes mellitus has long-term consequences. A diabetic also faces a high chance of developing a variety of health concerns. According to the International Diabetes Federation, the number of individuals living with diabetes rose to 382 million in 2013[1].Diabetic disease is anticipated to rise from 376 billion to 490 billion by 2030, according to world healthcare medical data[5].

Early detection of such disorders allows for disease control and the preservation of human life. This study focuses on the early detection of diabetes using machine learning approaches in order to attain this goal.

Several diabetes prediction algorithms have been proposed and published in recent years. They conducted thorough tests on outlier rejection and filling missing variables in order to improve the ML model's performance. They shows, the adoption of Machine Learning

classification-based techniques in illness diagnosis and treatment can reduce medical errors and human costs dramatically. According to the findings of the study, When compared to other methods for data classification, machine learning-based classification techniques offer a promising performance in prediction accuracy[2].

This research focuses on developing a diabetes prediction model utilizing machine learning algorithms and data mining techniques. The National Institute of Diabetes and Digestive and Kidney Diseases' PIMA Indians Diabetes Dataset, which contains information on female diabetic patients, was used in this study. In our proposed approach, different ML classifiers LR, KNN, NB, SVC, DT, RF, GB were used. Grid search technique is applied to improve accuracy. Various approaches of machine learning models are carried out in order to find the best classifier, which employs the best performing preprocessing from prior experiments.

1.2 Motivations

Diabetes is a serious disease that no one should neglect. Diabetes is a category of metabolic disorders characterized by a persistently high blood sugar level. Many types of Complications in major organs of the body as a result of diabetes. Recognizing at early stage and be conscious about health is the only way to avoid this serious problem. Women are disproportionately affected by diabetes because they are not accustomed to enough physical activity and are unaware of their health status. They always try to conceal health issues since they don't want to go through the extra effort of medical care as a result, various studies, tests and researches are carried out throughout this time period This research work is focused on female patients .This study is also aims to determine which of the seven data mining algorithms that are used has the greatest accuracy rate using machine learning approaches. Since seven algorithms are this study, It will aid in the selection of the most appropriate algorithm for this goal.

1.3 Research Questions

- What is Diabetes, and how does it affect you?
- What are the most serious consequences of diabetes?
- Is it possible to tackle the diabetes prediction problem with machine learning?
- Which machine-learning approach provides the most accurate predictions?
- Is data mining playing a significant role in diabetes forecasting?

1.4 Expected outcome

Machine-learning algorithms have been shown to be more effective in diagnosing various diseases in studies. The capacity to manage a vast amount of data, merge data from multiple sources, and integrate background information in the study gives Data Mining and Machine learning algorithms their power. Compared to an individual classifier, classification techniques are widely applied in the medical field for classifying data into different classes based on specified constraints. Machine-learning algorithms have been shown to be more effective in diagnosing various diseases in studies. The strategies are also extremely simple to apply. K Nearest Neighbors, Decision Tree, Logistic Regression, Support Vector Machine, Nave Bayes, Random Forest, and Gradient Boosting classifier were applied in this work. We also used the component of confusion matrix to determine the best classifier. The statistical classification is known as the confusion matrix. The confusion matrix is used as a performance visualization of algorithms in supervised learning and as a matching matrix in unsupervised learning.

1.5 Report Layout

We need an introduction and motivation before a work begin and in Chapter 1 of this paper, the introduction and motivation for this topic is addressed, Diabetes prediction, and why it is chosen. The associated working area is covered in Chapter 2 of this study. After finishing introduction, the study concentrated on people who worked on projects relating to this topics and gathered their useful knowledge. Basically, our 2 chapter normally discussed about related work. Consequence in chapter 3 we discuss about implement algorithm, we chose famous 6 algorithms' (K Nearest Neighbors, Decision Tree, Logistic Regression, Support Vector Classifier, Naïve Bayes and Random Forest) and applied those algorithms in our dataset to searching best one and it's our methodology part. Chapter 4 we talk about all of our algorithm's outcome and evaluation. Then in Chapter 5 we talk about the impact on society, environment. In general, Chapter is about literature review. As a result, in Chapter 3, it is about how to implement algorithms. Here, seven six well-known algorithms (K Nearest Neighbors, Decision Tree, Logistic Regression, Support Vector Classifier, Nave Bayes, and Random Forest, Gradient Boosting) are taken and applied them to dataset to find the best one. In Chapter 4, the results and evaluation of algorithm is discussed. The influence on society, the environment, and the sustainability plan are discussed in Chapter 5. In Chapter 6, recieving distinct outcome, which might be described as the comparative best. It's time to wrap things up. Finally, the sources from which gathered data for this research is presented under references.

CHAPTER 2

Background

2.1 Preliminaries

Although there is no cure for diabetes, persons with well-managed diabetes can live long and healthy lives. Diabetes identification at an early stage is critical in this regard. Machine learning and data mining, like other fields of medical science, can play an essential role in reducing diabetes-related premature mortality. There are also several studies and tests on female patients with the primary purpose of predicting diabetes sooner. Charge effective, efficacious, and speedier strategies for diabetes prediction can be made employing computer expertise, algorithms of machine learning, and data mining, as shown by collaboration researches. Many academics have used data mining to construct different prediction models to predict diabetes. Data mining and machine learning currently a viable technology capable of handling both simple and complex data.

2.2 Literature review

Using the Pima Indian Diabetes dataset, several researchers employed the machine learning technology to predict diabetes (PIDD). There are a variety of data mining approaches that can link unstructured and structured data together. That is why these technologies are so widely applied in diabetes prediction and in research. This paper metering some similar papers on this issue in the table 2.2:

Table 2.2: Related studies on Diabetes Prediction

Reference	Used Dataset	Model Details and Accuracy
Sisodia et al.[8]	Pima Indian Dataset	On PIDD, it was discovered that the NB classifier outperforms the SVM, NB, and DT techniques.
Kavakiotis et al[3]	Pima Indian Dataset	10 fold cross validation was utilized to evaluate three distinct algorithms: logistic regression, Naive Bayes, and SVM, with SVM providing the best performance and accuracy of 84 percent.
Zheng et al[6]	Pima Indian Dataset	To predict diabetes mellitus at an early stage, researchers used Random Forest, KNN, SVM, Naive Bayes, decision trees, and logistic regression.
Swarupa et al.[9]	Pima Indian Dataset	With an accuracy value of 77.01 percent, Naive Bayes (NBs) provided good accuracy.
Asma [4]	Pima Indian Dataset	The accuracy was shown to be 78.1768 percent with Decision tree
Dr. D. Asir Antony Gnana Singh et al.[11]	Dataset from their own medical history	NB, ML, and PRF were used. NB was more

		accurate.
Thirumal et al.[10]	Pima Indian Dataset	C4.5, SVM, KNN, and Nave Bayes were demonstrated to have better accuracy than the others, with an accuracy value of 78.2552 percent.
Aishwarya Mujumdar and Dr. Vaidehi V [6]	A collection of data contains 800 records and 10 attributes	AB, GBC, LR, NB, LDA, KNN, SVC, RF, DT ETC, Perceptron, AB, GBC, LR, NB, LDA, KNN The maximum accuracy was achieved by LR, which was around 96 percent.
M. K. Hasan et al.[15]	Pima Indian Dataset	To improve accuracy with ensembling classifiers, different ML classifiers were used and the grid search technique was used.
Cut Fiarni et al. [12]	Tested on three data sets (Sri Pamela Hospital and Kumpulan Pane Hospital, Tebing Tinggi, and Dolok Community Health Center)	K-means (68 percent) had the highest accuracy among K-means NB, C4.5 DT.
Veena and Anjali[14]	Pima Indian Dataset	SVM, Decision Stump,NB, and decision tree was used. Decision stump rated 80.72 percent accuracy as superior.
Nongyao and Rungruttikarn[13]	Pima Indian Dataset	The algorithms LR, Boosting, Nave Bayes, ANN, Bagging, and Decision Tree were used.This Random Forest

		<p>approach yielded an accuracy of 85.558 percent.</p> <p>It is noted as having a higher level of precision.</p>
Sajida et al.[17]	A dataset from CPCSSN (2003 to 2013)	The algorithms Adaboost, j48, and Bagging were used. Adaboost showed to be more accurate than previous methods.
Xue-Hui Meng et al. [16]	A collection of data from random people of China and Guangzhou	LR (76.13%), ANN (73.23%), DT C5.0 (77.87%)

2.3 Research summary

Going through these related works it can be realized that many machine learning approaches and data mining techniques are used to predict diabetes. Pima Indian data set is mostly used in this purpose. Researchers apply algorithms to predict this deadly disease at early stage. Some of the researchers worked in WEKA, while others used the Python programming language. WEKA is a knowledge analysis environment that was created in New Zealand. WEKA is one of the most popular and widely used tools for this purpose. In this case, machine learning approaches are also used, and in this day and age, machine learning is a leading methodology for any type of experiment because it is simple to use and speedier. Many research papers have gone through but any research couldn't be discovered that compares seven data mining strategies with machine learning classifiers utilizing the Python programming language. The majority of studies have not attempted to compare algorithms on a single dataset in this way. In this study, I have applied seven algorithms (Logistic Regression, K Nearest Neighbors, Decision Tree, Support Vector Classifier, Naïve Bayes and Random Forest, Gradient Boosting) and also get a better accuracy of 88.74% for Gradient Boosting classifier. It is a comparison based study and algorithm by which I have found the best algorithm for predicting diabetes.

2.4 Scope & Challenges

After collecting Pima Indian Dataset initial objective was to manually check any missing, zeroed, or impertinent data. There were a lots of zero values which I handled by preprocessing which was a challenging part. The key focus of this research is that no research paper has combined all these seven algorithms on the same dataset with a better accuracy.

CHAPTER 3

Research Methodology

3.1 Introduction and process model

The methodology is separated into several portion to gain the study goal, including collection of data, processing the collected data, application of various machine learning approaches, and associated prediction analysis. The procedure which is followed is summarized below as Model Diagram that shows the working flow of the research work In short:

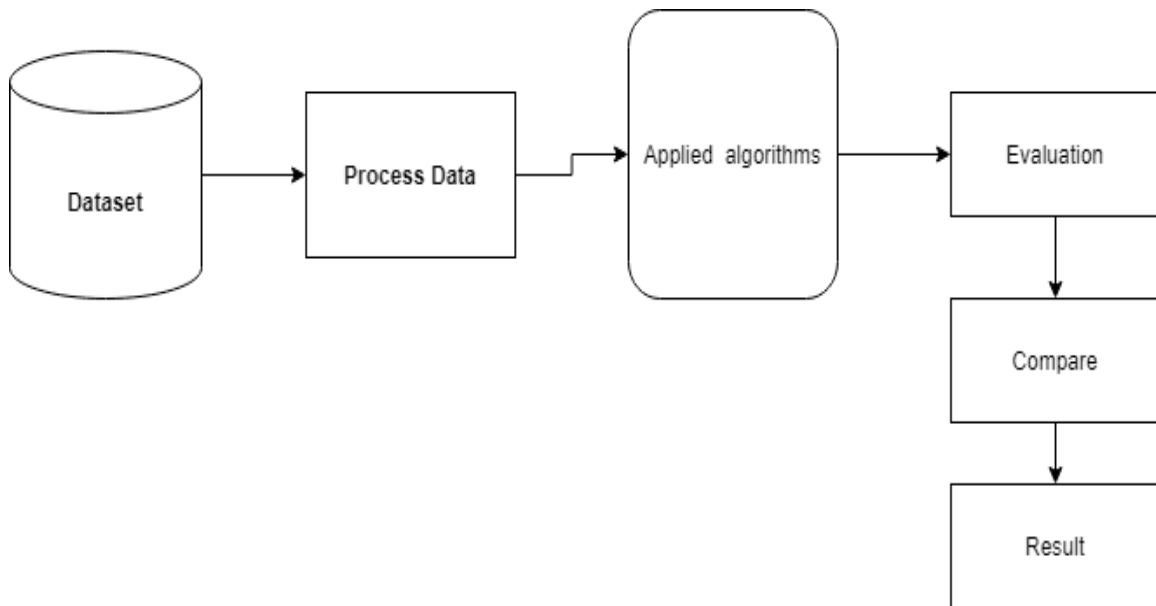


Figure 3.1: Model Diagram

3.2 Research Subject and Instrumentation

In this study basically ML classification algorithms are used which are under supervised learning. Algorithms learn from labeled data in supervised learning. The algorithm determines which label should be given to new data after analyzing the data by associating patterns with the unlabeled new data. Python programming language is used to apply the algorithm because it is easy to learn and quite similar to human language. For coding Jupyter notebook is used which is not only work as a IDE but also a beautiful educational tool. The data set is in csv format so that Microsoft office excel is also used.

3.3 Dataset

Pima Indian diabetes dataset (PIDD) is used in this research which is collected from UCI Machine Learning Repository. This dataset is mainly originated from National Institute of Diabetes and Digestive and Kidney diseases. PIDD is focused on female patients. All of them are at least 21 years old. This data set is containing the information of 768 female patients. There is nine attributes. In the table 3.3.1 and 3.3.2 below details about PIDD is shown.

Table 3.3.1: PIDD Dataset Description

Dataset	No. of Instance	No. of Attributes
PIDD	768	9

There are nine attributes in PIDD, all of them are numerical. The features are below in table 3.3.2.

Table 3.3.2: The attributes in PIDD

Attribute	type	Average Mean
Pregnancies	Numeric	3.845052
Glucose	Numeric	121.69
BloodPressure	Numeric	72.41
SkinThickness	Numeric	29.15
Insulin	Numeric	155.55
BMI	Numeric	32.45
DiabetesPedigreeFunction	Numeric	0.47
Age	Numeric	33.24
Outcome	Numeric	0.34

3.4 Data Visualization

Data visualization is important to have a better idea about the data. It is visual context of data information. It helps to understand the pattern of data. For achieving goal it is necessary to interpret big and complex dataset, data visualization will a great role in this area. As the purpose of data analysis to gain insights, data visualization is important because also has a great impact on human body.

Outcome can be plotted to get a better view. As outcome column has two type values 0 and 1. In PIDD, There is 500 healthy (when outcome=0) and 268 (when the outcome=1) diabetic cases.

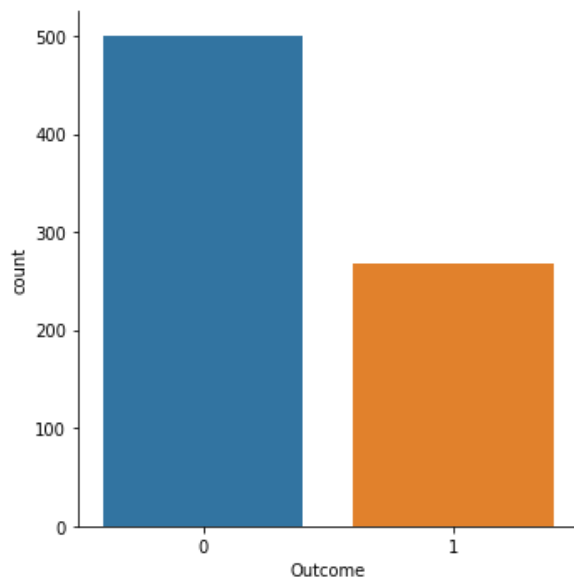


Figure3.4.1: Outcome Cases

Histogram will help us to visualize relations between a single variable and the outcome. Below, we'll see the relation between every parameter and outcome. It is the most used graph for displaying frequency distributions. It has the appearance of a bar chart.

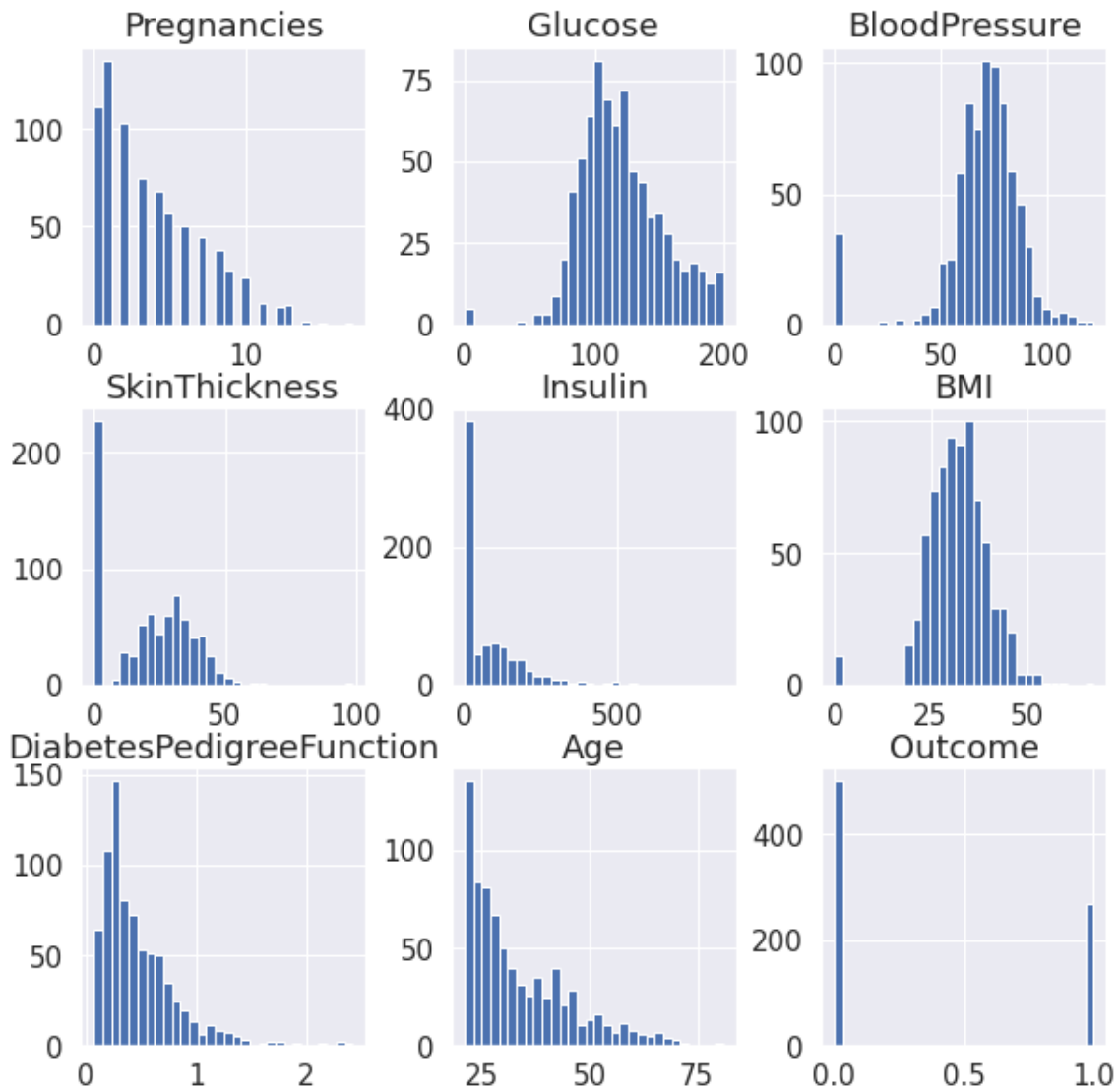


Figure 3.4.2: Histogram relation between Single attribute and outcome

Density plot can also be used to examine this relationship. A density plot is a visual representation of a numeric variable's distribution

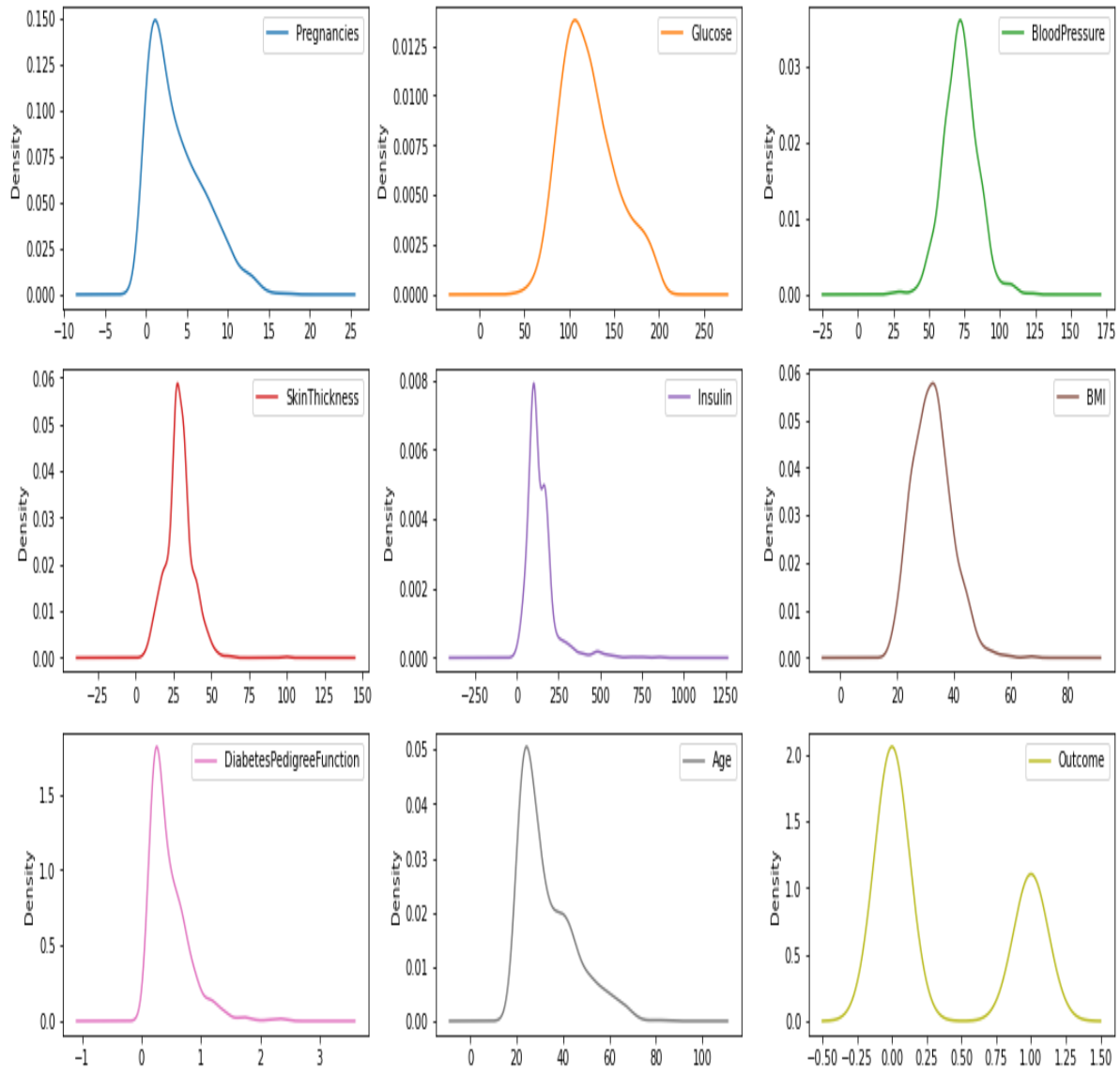


Figure 3.4.3: Density plot to represent numeric variable's distribution

Visualizing heat map co-relation between every column can be understood

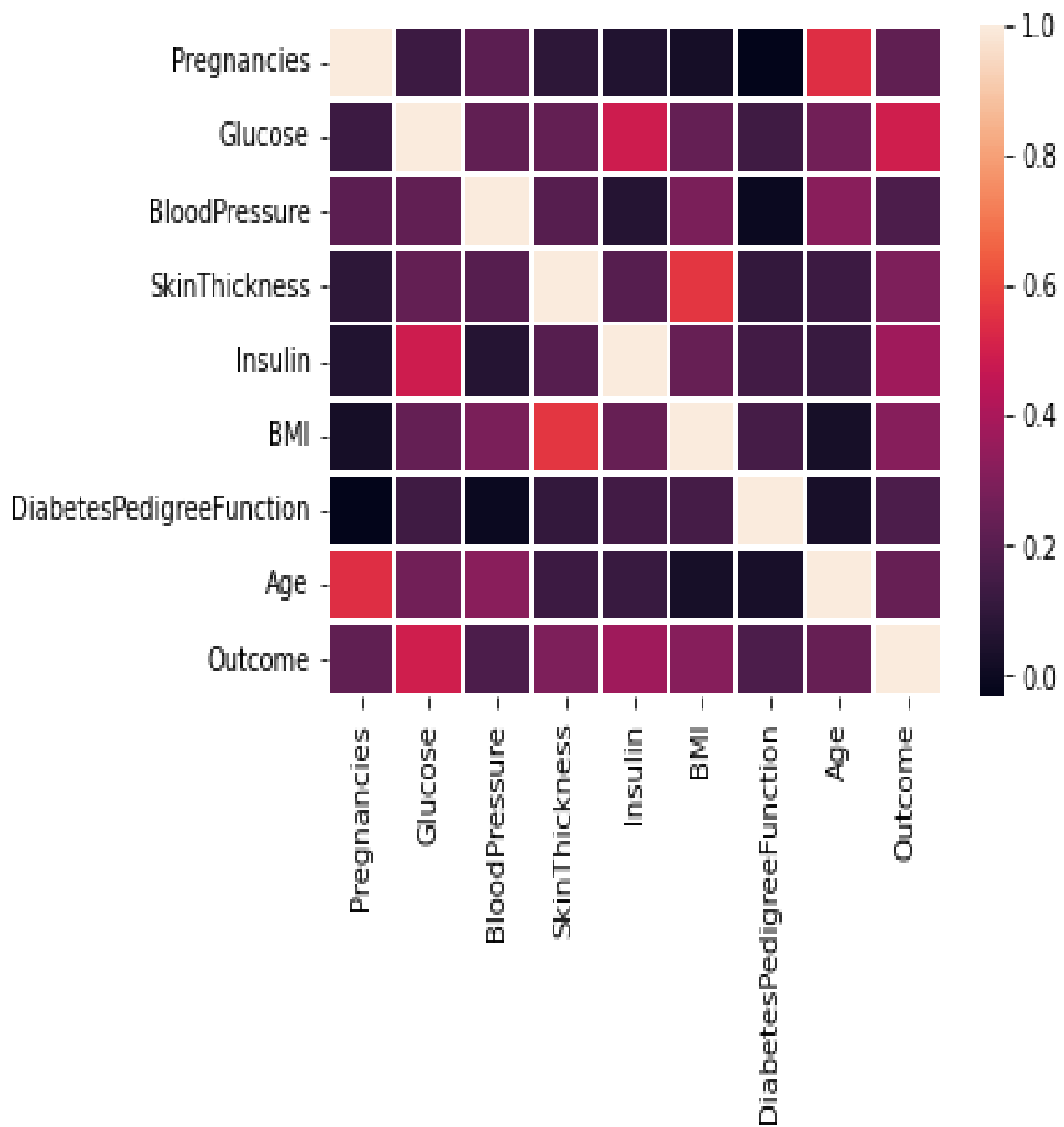


Figure 3.4.4: Heat map to represent the co-relation

3.5 Data processing

Data processing is an important part for improving data quality to get better accuracy. Data processing includes data cleaning, remove, handling missing values. Using these steps we get a clean data which helps to gain our goal of getting better accuracy.

3.5.1 Missing value identification

In PIDD, directly there is no missing or null value but there are many zero value which are illogical for example Glucose, Blood Pressure, Skin Thickness, Insulin, BMI of a human cannot be zero. So there are data gap which need to be handled to gain a better accuracy and achieve the research goal.

3.5.2 Missing value handling

To handle it first this values are replaced with NaN values. Then using data `isnull().sum()` all the null value can be shown which is shown in figure 2 below. Then these values are handled by replacing corresponding mean values.

```
[ ] df.isnull().sum()

Pregnancies          0
Glucose              5
BloodPressure        35
SkinThickness        227
Insulin              374
BMI                  11
DiabetesPedigreeFunction  0
Age                  0
Outcome              0
dtype: int64
```

Figure 3.5.2: Total missing values in each attributes

3.5.3 Feature Selection

In order to creating a predictive model, feature selection is the process of minimizing the number of input variables. Reducing the numbers feature can make huge effect in model for some datasets that's why it is used by many people to for increasing speed, avoiding over fitting, and achieving the highest categorization results. In this data set it didn't effect a lot so in this study all nine attributes have been used and the goal class is 'Outcome,' where a value of '0' indicates diabetes negative and a value of '1'.

3.5.4 Data standardization

Data standardization is a part of data preparation. It is a process of converting data in a common format. It aids in the establishment of clearly defined elements and attributes so that frequent utilization of data standardization is seen in machine learning algorithm. The features will be rescaled as a result of standardization to ensure that the mean and standard deviation are 0 and 1, respectively. The equation is-

$$X_{stand} = \frac{X - Mean(x)}{standard\ deviation(x)}$$

3.6 Dataset train and test method

After preprocessing data set, the prepared Pima Indian Diabetes Dataset is ready to use and applying various machine learning algorithms. Train test split is used which is a easy technique to evaluate the performance of machine learning algorithms. The data set divided to two different portions. One of them is called training data which is used to train the model and other one is known as test dataset which is used to test the model and evaluate model performances. Here in this study the dataset is divided into 70% is used to train the model and 30% used to test the model.

3.7 Detail work flow:

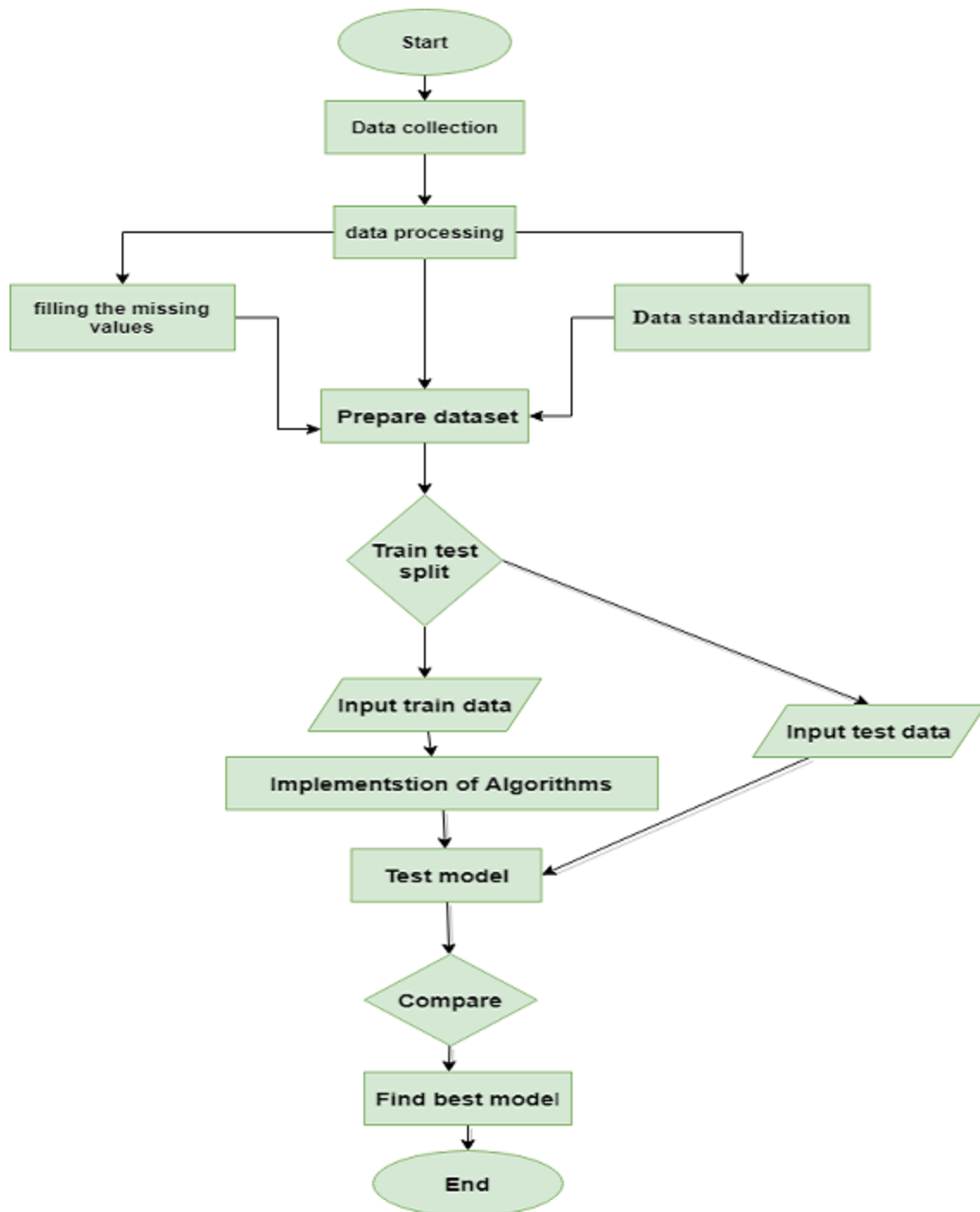


Figure 3.7: Detail workflow

3.8 Algorithms implementation

In this study, machine learning supervised technique is used. Using the training data classification model is construct with Logistic Regression, Support Vector Machine, Decision Tree, Naïve Bayes, k-nearest neighbor, Random Forest, Gradient Boosting Classifier. Then evaluate and compare all the models and select the best model.

3.8.1 Logistic Regression

Logistic Regression is a simple and efficient algorithm. In the early twentieth century, logistic regression was mostly employed in biological research and applications [19]. It measures the relationship of independent and dependent variables. The sigmoid function, $S(x) = \frac{1}{1+e^{-x}}$ replace a certain continues value or range to a discrete number. It is a fundamental model that describes dichotomous output variables and can be used to predict disease classification [18]. Mainly it is used when the output or dependent variable. There are always two possible outcome that's why it is also known as binary logistic regression. It is an algorithm which is used in machine learning deals with probability to measure the relationship between dependent and independent variable. In logistic regression binary is used as a dependent variable and its limited number of possible values are 0 to 1.

3.8.2 Support Vector Classifier

Support vector classifier is one of the popular machine learning algorithms. It can solve both classification and regression problems but as a classification it is used mostly. SVM divide the data set into classes using hyper plane. Hyper plane is situated maximizing the margin between classes. Hyper plane should not be closer to data points belonging to the other class for better generalization. Supervised machine learning follows kernel trick

where kernel takes non separable problems into separable problems which makes supervised machine learning more powerful, flexible and exact. Supervised machine learning find hyper plane using support vector and margins. A hyper plane that is far from the data points in each category should be chosen. The support vectors are the spots closest to the classifier's margin[20]. The support vectors are the spots closest to the classifier's margin.

3.8.3 Decision Tree

Decision tree is Algorithm which follow supervised machine learning techniques. It is an approach that divides a large dataset into smaller sections. It follows tree structure. Decision tree selects each node at each stage by weighing the maximum information gain across all attributes [21]. Entropy is a metric that estimates the number of errors in a data set and is measured between 0 and 1. It is, however, employed for classification because it is simple to explain to the technical team and is an automatic process. The previous stages are repeated for each subset or branch obtained in the previous step, building a decision tree for each partitioned sample. The term "tree-based knowledge" refers to the gathering of information for the purpose of reaching a decision.

3.8.4 Naive Bayes

Naïve Bayes is an efficient algorithm. It mainly deals with probability. Classification problem is solved by this algorithm. This approach needs binning to manage data. A class's highlighted features are unaffected by another feature. All of the feathers are unique and unrelated to one another. That is, applies strong independent assumption. Presence of one feature doesn't affect others. Bayes Theorem:

$$P(A|B) = \frac{P(B|A)P(A)}{P(B)}$$

That is A occurring if B has already occurred using Bayes' theorem. The evidence is B, and the hypothesis is A.

3.8.5 Random Forest

Random forest is an algorithm which is easy to use and efficient. Most of the time we can get higher accuracy using this algorithm even there no need to use hyper-parameter tuning. In this approach to predict the output or class RF combines multiple trees, here not all the trees predict the correct output but combining all the trees output it predicts the correct output. The random forest method can maintain accuracy even when a high percentage of data is missed. Finally, random forest is employed in both regression and classification applications. This random forest classifier works at random and improves the model's accuracy. Random forest predict with higher accuracy taking less training time compared to other algorithms.

3.8.6 K-Nearest Neighbors

K-Nearest Neighbors is used to solve both classification and regression problems. The sign 'K' represents the number of nearest neighbors to a new unknown variable that must be predicted or categorized. The blending of numerical and categorical values in datasets is achieved using the K nearest neighbors technique, which is a simple and soft approach. The issue of numerical variables between 0 and 1 is likewise raised by K nearest neighbors. To classify other data, it can be used the majority number of votes for the label to locate the element's neighbors. It estimates the distance between classes, finds neighbors, and votes on labels after initializing the data. k nearest neighbors uses a categorized machine learning method to count all of the training samples and determine the number of nearest neighbors.

3.8.7 Gradient Boosting

A weak learners can be turned into strong learners with the boosting algorithm. It considers a series of weak learners which provide slightly better predictions. After that, the forecasts are merged using a majority vote. Gradient boosting technique is very attractive because of its fast prediction speed and higher accuracy even if the data set is complex. It is one of the most powerful algorithms in the machine learning field. It can not only work as a classifier but also solve regression problems. Using GridSearchCv, the best value of `n_estimators` can also be found.

CHAPTER 4

Experimental Results and Discuss

4.1 Experimental Results & Analysis

In this study, accuracy score matrix is used to performance evaluation. In PIDD Pregnancies, Glucose, Blood Pressure, Skin Thickness, Insulin, BMI, Diabetes Pedigree Function and age were the independent variable and target column which is named as outcome was the dependent variable. Data set was split in 7:3 ratio that is 70% of data was used as training data and remaining 30% as test data. Seven algorithms are implemented. Gradient Boosting classifier predict with higher accuracy of 88.74%

Table 4.1: Train test dataset ratio

Total No. of Instance (%)	No. of Training Instance (%)	No. of Testing Instance (%)
100	70	30

4.1.1 Accuracy of Models

Seven classifiers K Nearest Neighbors, Decision Tree, Logistic Regression, Support Vector Classifier, Naïve Bayes Random Forest, Gradient Boosting classifier are used to predict the best model. The accuracy of models are below in figure 4.1.1.1 and 4.1.1.2;

```
-----  
LogisticRegression:  
Accuracy: 75.7576%  
-----  
KNeighborsClassifier:  
Accuracy: 77.4892%  
-----  
DecisionTreeClassifier:  
Accuracy: 84.8485%  
-----  
GaussianNB:  
Accuracy: 74.4589%  
-----  
SVC:  
Accuracy: 75.7576%  
-----  
RandomForestClassifier:  
Accuracy: 86.1472%  
-----  
GradientBoostingClassifier:  
Accuracy: 88.7446%
```

Figure 4.1.1.1: Models accuracy

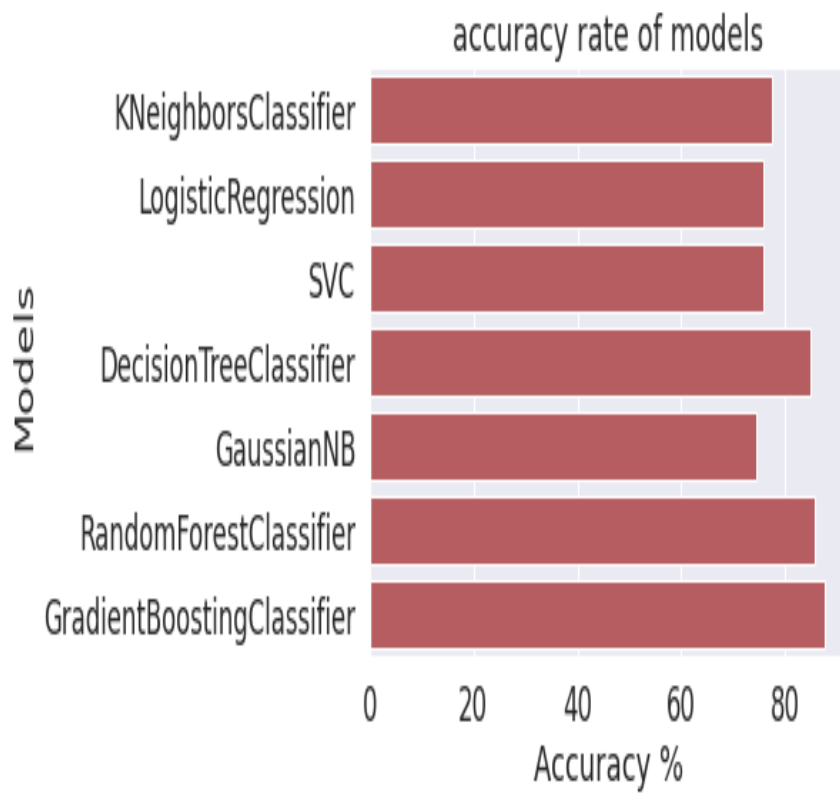


Figure 4.1.1.2: Models accuracy diagram

4.1.2 Confusion Matrix and Heat Map of Support Vector

	precision	recall	f1-score	support
0	0.81	0.83	0.82	151
1	0.66	0.62	0.64	80
accuracy			0.76	231
macro avg	0.73	0.73	0.73	231
weighted avg	0.76	0.76	0.76	231

Figure 4.1.2.1: Confusion Matrix of SVC

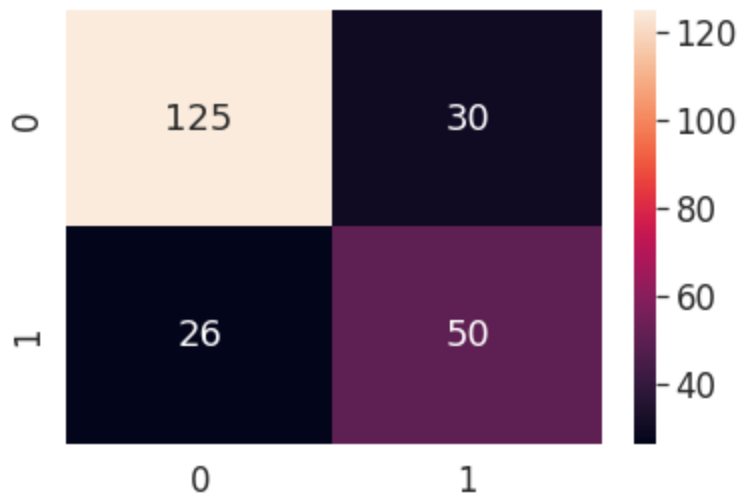


Figure 4.1.2.2: Heat map of SVC

4.1.3 Confusion Matrix and Heat Map of Logistic Regression

	precision	recall	f1-score	support
0	0.80	0.84	0.82	151
1	0.67	0.60	0.63	80
accuracy			0.76	231
macro avg	0.73	0.72	0.73	231
weighted avg	0.75	0.76	0.75	231

Figure 4.1.3.1: Confusion Matrix of LR

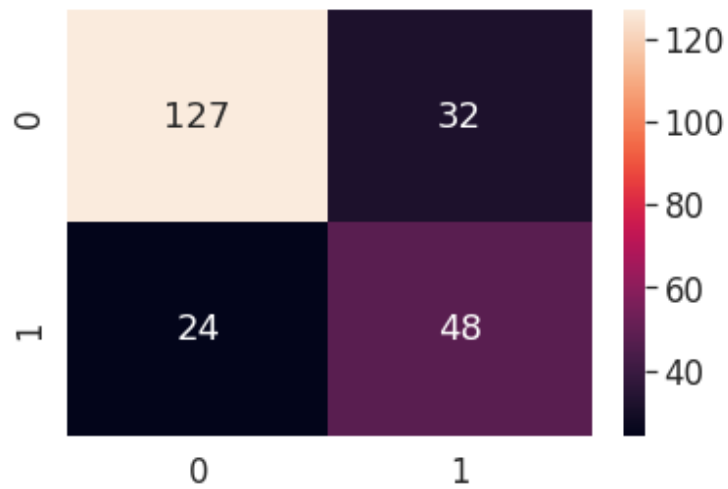


Figure 4.1.3.2: Heat map of LR

4.1.4 Confusion Matrix and Heat Map of Naïve Bayes

	precision	recall	f1-score	support
0	0.81	0.79	0.80	151
1	0.63	0.65	0.64	80
accuracy			0.74	231
macro avg	0.72	0.72	0.72	231
weighted avg	0.75	0.74	0.75	231

Figure 4.1.4.1: Confusion Matrix of NB

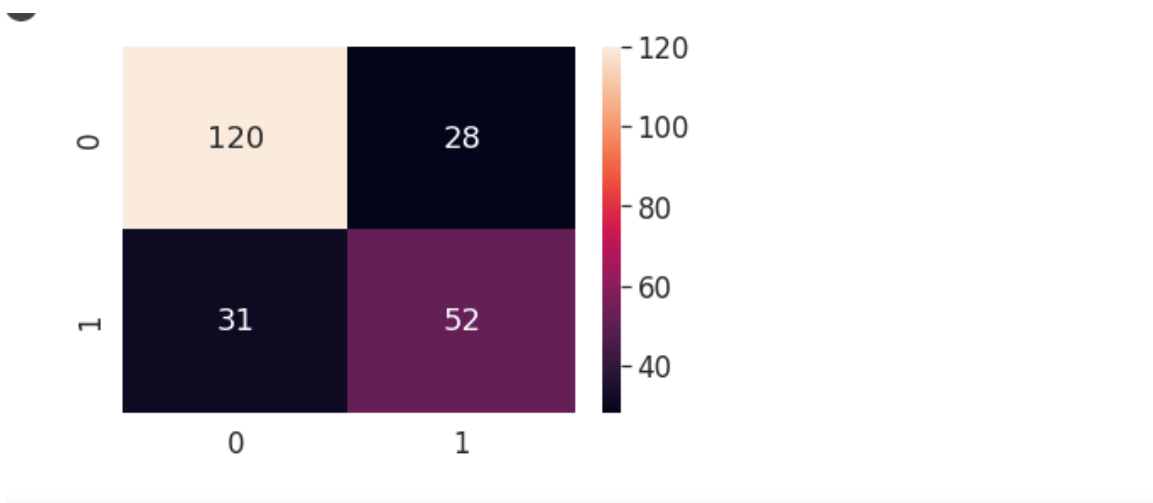


Figure 4.1.4.2: Heat map of NB

4.1.5 Confusion Matrix and Heat Map of Decision Tree

	precision	recall	f1-score	support
0	0.90	0.86	0.88	151
1	0.76	0.82	0.79	80
accuracy			0.85	231
macro avg	0.83	0.84	0.84	231
weighted avg	0.85	0.85	0.85	231

Figure 4.1.5.1: Confusion Matrix of DT

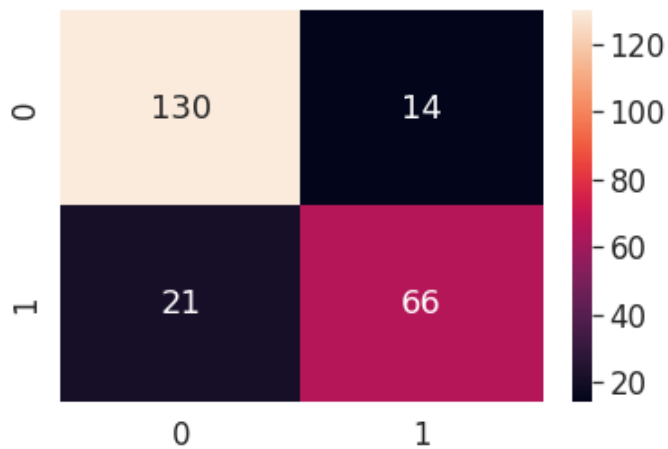
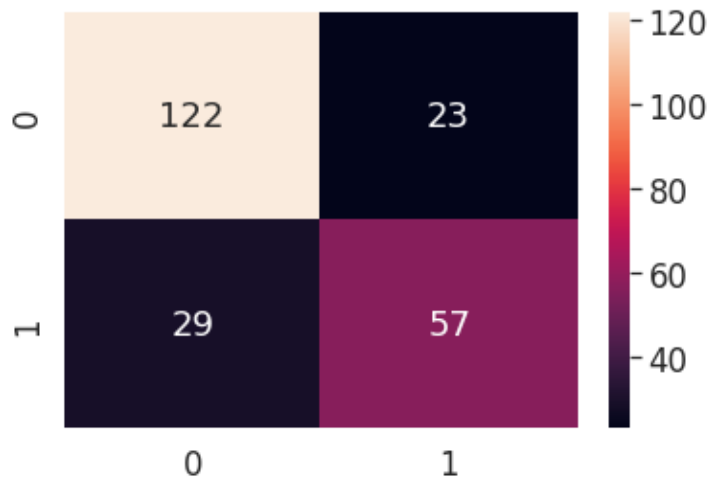


Figure 4.1.5.2: Heat map of DT

4.1.6 Confusion Matrix and Heat Map of K-Nearest Neighbors

	precision	recall	f1-score	support
0	0.84	0.81	0.82	151
1	0.66	0.71	0.69	80
accuracy			0.77	231
macro avg	0.75	0.76	0.76	231
weighted avg	0.78	0.77	0.78	231

Figure 4.1.6.1: Confusion Matrix of KNN



+ Co

Figure 4.1.6.2: Heat map of KNN

4.1.7 Confusion Matrix and Heat Map of Random Forest

	precision	recall	f1-score	support
0	0.90	0.88	0.89	151
1	0.79	0.82	0.80	80
accuracy			0.86	231
macro avg	0.85	0.85	0.85	231
weighted avg	0.86	0.86	0.86	231

+ Code

Figure 4.1.7.1: Confusion Matrix of RF

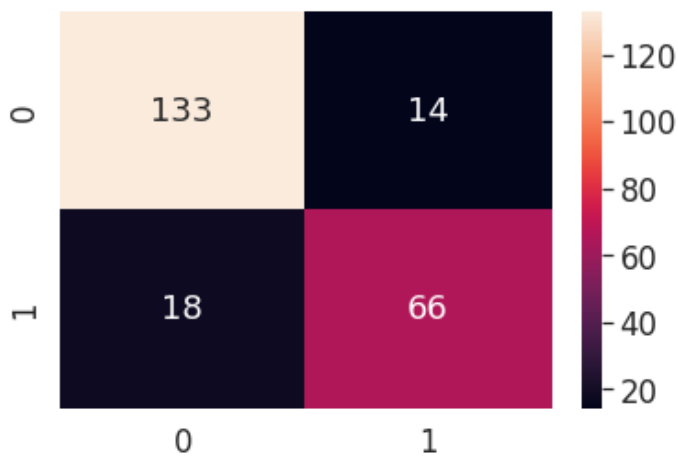


Figure 4.1.7.2: Heat map of RF

4.1.8 Confusion Matrix and Heat Map of Gradient Boosting

	precision	recall	f1-score	support
0	0.91	0.91	0.91	151
1	0.84	0.84	0.84	80
accuracy			0.89	231
macro avg	0.88	0.88	0.88	231
weighted avg	0.89	0.89	0.89	231

Figure 4.1.8.1: Confusion Matrix of RF



Figure 4.1.8.2: Heat map of GB

4.2 Evaluation measures

The performance of the classifiers has been discovered in Diabetes prediction for various plans with their suitable parameters, such as the ratios of accuracy, recall, precision, F1-score, and report that are received using a confusion matrix. True-Positive (TP) and False-Positive (FP) cases are explained, whereas True-Negative (TN) and False-Negative (FN) cases are explained (FN).

$$\text{Precision} = \frac{TP}{(FP + TP)} \quad [\text{Ratio of the anticipated positive cases}]$$

$$\text{Recall} = \frac{TP}{(FN + TP)} \quad [\text{Ratio of positive cases}]$$

$$\text{Accuracy} = \frac{(TP + TN)}{(FP + TN + FN + TP)} \quad [\text{Ratio of the total number of predictions}]$$

$$\text{F-Measure} = \frac{2 \times \text{Precision}}{(\text{recall} + \text{precision})} \times \text{Recall} \quad [\text{Count Balance between recall and precision}]$$

CHAPTER 5

Impact on Society, Environment and Sustainability

5.1 Impact on Society

This study is entirely based on the use of women's data to forecast diabetes. Most people in our society are unable to keep their pre-work or personal boundaries in order to stay healthy, resulting in a variety of diseases. One of these diseases is diabetes. Many women are impacted with diabetes as a result of their unconsciousness, which has societal implications. As a result, this thesis paper will guarantee that our diagnostic is accurate and competent. The entire state should be covered by such an application. It is possible to be conscious of one's personal health and well-being by employing this.

5.2 Impact on Environment

In medicine, doctors have conducted blood sugar tests and hemoglobin tests to forecast diabetes and utilized toxic chemicals or other substances to perform these tests. However, we can simply forecast our diabetes level if we employ this thesis type solution. This thesis was entirely reliant on software. Computer systems process all data using software. So, utilizing software to forecast diabetes is not dangerous; in fact, it is a simple and rapid process to identify. This procedure has no negative effects on our air, land, or water, and it does not harm our immune cells. As a result, it's a safe and quick procedure to follow.

5.3 Ethical Expects

Using a diabetes prediction system is completely ethical in this paper. To continue the diabetes prediction algorithm in this work, it required to employ a variety of data sets to predict diabetes. As a result, in order to gather this information from people, it must be obtained their consent and ensure that the information is accurate. People are not harmed in any way during the data collection procedure. As a result, this data collection procedure is completely ethical.

5.4 Sustainability Plan

Passion, devotion, and excitement for the task are essential components of a long-term diabetic self-management program. At current time, both software and hardware are used to provide successful and satisfied diabetes prediction results. The diabetic initiative proved that there is a key method that can boost the program's long-term viability. The diabetes prediction algorithms are long-lasting, which means that individuals can use them for a long time and that they are extremely beneficial to everyone.

CHAPTER 6

Summary, Conclusion, Recommendation and Implication for Future Research

6.1 Conclusion

Diabetes Dataset for Pima Indians In the case of diabetes prediction, feature engineering and data analysis are critical for improving the performance of the classifier for patients. Many researchers are considering how to diagnosis other prevalent diseases. To accurately anticipate Diabetes, we should employ algorithms that provide better representation. In this research paper, we have discussed about machine-learning techniques (SVC, DT, LR, KNN, RF, NB, GB) and also try to find out the suitable method and there 75.76% in LR, 77.48% in KNN, 74.46% in NB, 75.75% in SVC, 84.84% in DT and 86.15% in RF, 88.74% in GB outperformed the other classifiers found in the empirical evaluation.. RF is also gives better accuracy very near to GB .

Table 6.1: Results

Algorithms Name	Accuracy (%)
SVC	75.75
NB	74.46
LR	75.76
DT	84.84
KNN	77.49
RF	86.15
GB	88.74

6.2 Future Work

In our study, seven machine learning algorithms are implemented. More algorithms can be implemented to this data set and may get a better solution. Also this data set is about women information. So this work can be extended adding men information, there also can be software based research in future. So that a patient does not need to go to the hospital. A patient's information can be entered into a website, online application, or software, which will be able to predict diabetes in that patient. After a few years, it will be more appealing, efficient, and acceptable to people.

References

- [1] V., A. K. and R., C. 2013. Classification of Diabetes Disease Using Support Vector Machine. *International Journal of Engineering Research and Applications*. 3, (April. 2013), 1797-1801
- [2] Mathers, C.D. and Loncar, D. 2006. Projections of Global Mortality and Burden of Disease from 2002 to 2030. *PLoS Medicine*. 3, 11 (Nov. 2006), e442.
- [3] Kavakiotis, Ioannis, Olga Tsave, Athanasios Salifoglou, Nicos Maglaveras, Ioannis Vlahavas, and Ioanna Chouvarda. "Machine learning and data mining methods in diabetes research." *Computational and structural biotechnology journal* (2017).
- [4] Al Jarullah, Asma A. "Decision tree discovery for the diagnosis of type II diabetes." In *Innovations in Information Technology (IIT)*, 2011 International Conference on, pp. 303-307. IEEE, 2011.
- [5] Carlo, B G., Valeria, M. and Jesús, D. C. 2011. The impact of diabetes mellitus on healthcare costs in Italy. *Expert review of pharmacoeconomics & outcomes research*. 11, (Dec. 2011), 709-19.
- [6] Zheng, Tao et al. "A machine learning-based framework to identify type 2 diabetes through electronic health records." *International journal of medical informatics* 97 (2017): 120- 127.
- [7] Aishwarya Mujumbara, Dr. Vaidehi V. Diabetes Prediction using Machine Learning Algorithms. ScienceDirect. Kodaikanal, India. Vol 165. pp 292–299. 2019.
- [8] D. Sisodia, D.S. Sisodia Prediction of diabetes using classification algorithms, *Procedia Comput. Sci*. 132(2018) 1578-1585.
- [9] Rani, A. Swarupa, and S. Jyothi. "Performance analysis of classification algorithms under different datasets." In *Computing for Sustainable Global Development (INDIACom)*, 2016 3rd International Conference on, pp. 1584-1589. IEEE, 2016.
- [10] Thirumal, P. C., and N. Nagarajan. "Utilization of data mining techniques for diagnosis of diabetes mellitus-a case study." *ARPN Journal of Engineering and Applied Science* 10, no. 1 (2015): 8-13.
- [11] Aishwarya Mujumbara, Dr. Vaidehi V. Diabetes Prediction using Machine Learning Algorithms. ScienceDirect. Kodaikanal, India. Vol 165. pp 292–299. 2019
- [12] Cut Fiarni, Evasaria M. Sipayung, Siti Maemunah. Analysis and Prediction of Diabetes Complication Disease using Data Mining Algorithm. ScienceDirect. Bandung 4132, Indonesia. Vol 161. Pp 449–457. 2019. Machine Learning Algorithms. ScienceDirect. Kodaikanal, India. Vol 165. pp 292–299. 2019.
- [13] Nai-arun, Nongyao, and Rungruttikarn Moungrmai. "Comparison of classifiers for the risk of diabetes prediction." *Procedia Computer Science* 69 (2015): 132-142.

- [14] Vijayan, V. Veena, and C. Anjali. "Prediction and diagnosis of diabetes mellitus—A machine learning approach." In *Intelligent Computational Systems (RAICS), 2015 IEEE Recent Advances in*, pp. 122-127. IEEE, 2015.
- [15] Hasan, M.K., Alam, M.A., Das, D., Hossain, E. and Hasan, M., 2020. Diabetes prediction using ensembling of different machine learning classifiers. *IEEE Access*, 8, pp.76516-76531.
- [16] Xue-Hui Meng, Yi-Xiang Huang, Dong-Ping Rao, Qiu Zhang, Qing Liu. Comparison of three data mining models for predicting diabetes or prediabetes by risk factors. *ScienceDirect. Cancer Center of Sun Yat-Sen University, People's Republic of China. Vol 29. Pp 93-99. 12 March 2012.*
- [17] Perveen, Sajida, Muhammad Shahbaz, Aziz Guergachi, and Karim Keshavjee. "Performance analysis of data mining classification techniques to predict diabetes." *Procedia Computer Science* 82 (2016): 115-121. [
- [18] M. S. Klein and J. Shearer, "Metabolomics and Type 2 Diabetes: Translating Basic Research into Clinical, Application", Hindawi Publishing Corporation *Journal of Diabetes Research*, 2015.
- [19] Safavian, S.R. and Landgrebe, D. 1991. A survey of decision tree classifier methodology. *IEEE Transactions on Systems, Man, and Cybernetics*. 21, 3 (1991)
- [20] Sisodia, D., Singh, L., Sisodia, S., 2014. Fast and Accurate Face Recognition Using SVM and DCT, in: *Proceedings of the Second International Conference on Soft Computing for Problem Solving (SocProS 2012)*, December 28-30, 2012, Springer. pp. 1027–1038.
- [21] Iyer, A., S, J., Sumbaly, R., 2015. Diagnosis of Diabetes Using Classification Mining Techniques. *International Journal of Data Mining & Knowledge Management Process* 5, 1–14. doi:10.5121/ijdkp.2015.5101, arXiv:1502.03774.

Plagiarism report

Student Paper Diabetes

ORIGINALITY REPORT

21 %	14 %	9 %	15 %
SIMILARITY INDEX	INTERNET SOURCES	PUBLICATIONS	STUDENT PAPERS

PRIMARY SOURCES

1	Submitted to Daffodil International University Student Paper	7 %
2	dspace.daffodilvarsity.edu.bd:8080 Internet Source	2 %
3	Submitted to University of Chichester Student Paper	1 %
4	Submitted to Mar Baselios Engineering College Student Paper	1 %
5	Jobeda Jamal Khanam, Simon Y. Foo. "A comparison of machine learning algorithms for diabetes prediction", ICT Express, 2021 Publication	1 %
6	Submitted to Liverpool John Moores University Student Paper	1 %
7	Submitted to Birkbeck College Student Paper	<1 %
8	ijece.iaescore.com Internet Source	<1 %

24	Cut Fiarni, Evasaria M. Sipayung, Siti Maemunah. "Analysis and Prediction of Diabetes Complication Disease using Data Mining Algorithm", Procedia Computer Science, 2019 Publication	<1 %
25	www.kdnuggets.com Internet Source	<1 %
26	Submitted to Asia Pacific University College of Technology and Innovation (UCTI) Student Paper	<1 %
27	link.springer.com Internet Source	<1 %
28	Md. Faisal Faruque, Asaduzzaman, Iqbal H. Sarker. "Performance Analysis of Machine Learning Techniques to Predict Diabetes Mellitus", 2019 International Conference on Electrical, Computer and Communication Engineering (ECCE), 2019 Publication	<1 %
29	Submitted to University of East London Student Paper	<1 %
30	bbrc.in Internet Source	<1 %
31	ijirset.com Internet Source	<1 %

17	Md. Fahad Hossain, Md. Mehedi Hasan, Hasmot Ali, Md Rahmatul Kabir Rasel Sarker, Md. Toukirul Hassan. "A Machine Learning Approach to Recognize Speakers Region of the United Kingdom from Continuous Speech Based on Accent Classification", 2020 11th International Conference on Electrical and Computer Engineering (ICECE), 2020 Publication	<1 %
18	Wangshu Zhang, Liping Ma. "Research and application of second-hand commodity price evaluation methods on B2C platform: take the used car platform as an example", Annals of Operations Research, 2021 Publication	<1 %
19	acadpubl.eu Internet Source	<1 %
20	dspace.library.uvic.ca Internet Source	<1 %
21	github.com Internet Source	<1 %
22	www.ijstr.org Internet Source	<1 %
23	"Soft Computing: Theories and Applications", Springer Science and Business Media LLC, 2020 Publication	<1 %

9	Preeti Grover, Sanjeev Prasad. "A Review on Block chain and Data Mining Based Data Security Methods", 2021 2nd International Conference on Big Data Analytics and Practices (IBDAP), 2021 Publication	<1 %
10	Submitted to University of Newcastle Student Paper	<1 %
11	Andres Robles-Durazno, Naghmeh Moradpoor, James McWhinnie, Gordon Russell, Zhiyuan Tan. "Newly engineered energy-based features for supervised anomaly detection in a physical model of a water supply system", Ad Hoc Networks, 2021 Publication	<1 %
12	researchspace.ukzn.ac.za Internet Source	<1 %
13	Submitted to Lebanese University Student Paper	<1 %
14	Submitted to University of Surrey Student Paper	<1 %
15	www.commonlounge.com Internet Source	<1 %
16	"Proceedings of International Joint Conference on Computational Intelligence", Springer Science and Business Media LLC, 2020 Publication	<1 %

32	www.coursehero.com Internet Source	<1 %
33	www.ijeat.org Internet Source	<1 %
34	www.mdpi.com Internet Source	<1 %
35	www.ijana.in Internet Source	<1 %
36	Gaurav Shetty, Vijay Katkar. "Type-II Diabetes detection using Decision-tree based Ensemble of Classifiers", 2019 5th International Conference On Computing, Communication, Control And Automation (ICCUBEA), 2019 Publication	<1 %
37	Xuehu Wang, Tianqi Wang, Yongchang Zheng, Xiaoping Yin. "Hyperspectral-attention mechanism-based improvement of radiomics prediction method for primary liver cancer", Photodiagnosis and Photodynamic Therapy, 2021 Publication	<1 %
38	serisc.org Internet Source	<1 %
39	www.ijesrt.com Internet Source	<1 %