International Journal of Biometrics, Image Processing and Computing Research
ISSN: 2997-688X |
Volume 11 Issue 1, January-March, 2023
Journal Homepage: https://ethanpub.online/Journals/index.php/E23
Official Journal of Ethan Publication

PREDICTING DIABETES: IMPROVING SUPPORT VECTOR MACHINES AND LOGISTIC REGRESSION

Li Wei and Sofia Martinez
Computer Science Department, University of São Paulo, São Paulo, Brazil

Abstract:
The increasing prevalence of Diabetes is a global health concern, particularly in Sub Saharan Africa, where the mortality rates are alarming. Projected statistics indicate that by 2045, approximately 12.2% of the world population will be affected by Diabetes. This distressing reality demands collective efforts to combat the disease. Recognizing this, as a manufacturer, I identified an opportunity to address this health crisis while contributing to humanity's well-being.
We collaborated with traditional herbalists in Kano State, Nigeria, to obtain local formulations used in the treatment of Diabetes. Subsequently, we conducted a detailed analysis in China, identifying two effective formulations out of nine samples. However, as a non-medical practitioner and a student of Computer Science, I contemplated how I could contribute to this field.
A feasibility study revealed that machine learning, known for its prowess in making complex predictions, could be harnessed for Diabetes prediction. This approach would empower doctors to detect the disease at an early stage with heightened accuracy, offering patients a better chance of managing the condition effectively before it becomes life-threatening.

Keywords: Diabetes, machine learning, early detection, global health, traditional medicine.

1. INTRODUCTION

The rate at which people are dying as a result of Diabetes especially in Sub Saharan Africa and other part of the world is very alarming, it is even worse when you know the statistics of people fighting with Diabetes disease; it is projected that in the next two decades by year 2045about 12.2 % of the world population will be diabetic. These are very worrisome, devastating and disturbing figures anyone with power to do something that can lower this figures is challenged to do so, that is why it occurred to me as a manufacturer that this will be a very good market opportunity to make money and help humanity. We gather a lot of professional herbalist in Kano State Nigeria and got traditional medicines formulations from them that are locally used and sent the samples to China for detailed analysis and two proved to be effective among the 9 samples sent but as a non-medical practitioner and student studying Computer Science start thinking on what contribution can I make in terms of computing aspect since inventing the medicine is not a contribution in Computing field.

Based on the feasibility study conducted I discovered that machine learning has been used to make complex predictions and proved good, then it can also be used in diabetes prediction to help doctors identify the disease at an early stage with a higher degree of accuracy and by doing so the patient might have a better chance of managing the disease before it become late and fatal.

Technological advancement has bring about breakthrough in many fields and aspect of life nowadays cutting across every aspect of life from military, entertainment, education and healthcare and so on, but

the development is continuous which is why it opens so many doors to many untapped research opportunities in various sectors. Machine learning is a branch of Artificial intelligence with focus on training computer models in making accurate predictions, classifications, pattern recognition and mimicking human cognition in solving problems that are very challenging and difficult to solve by humans. Machine learning has provided greatest support in diseases diagnosis with absolute degree of accuracy given enough training and testing dataset.

Our Healthcare givers can leverage great support of machine learning in diagnosing diseases that are difficult or complex to diagnose in an early stage such as diabetes. Diabetes is a common chronic disease as a result of blood glucose is higher than the normal level, which is caused by defective insulin secretion or its impaired biological effects, or both Lonappan et al., (2007). It is widely known that there is no cure to diabetes but early detection can help in preventing the permanent damages that are associated with the disease such as dysfunction of various tissues and organs, especially eyes, kidneys, heart, blood vessels and nerves which are sometimes lead to fatality. The disease poses great threat to humanity due to the level of complications, failure to detect the disease early lead to increase in fatal rate per annum with number of patients skyrocketing. According to CDA (Canadian Diabetes Association) between 2010 and 2020, the expected increase of the disease would be from 2.5 million to 3.7 million and 643 million (11.3%) by 2030 and to 783 million (12.2%) by 2045.

In an effort to provide a novel solution in predicting early diabetes using machine learning techniques, this research tries to optimized some of the most prominent machine learning classification algorithm Support vector machine algorithm (SVM) and Logistic Regression algorithm (LR) using particle swarm optimization technique (PSO) for prediction of diabetes in order to improve their performances and yield an accuracy that is higher than the benchmark. The performances of the two algorithms were compared to determine the best performing algorithm in terms of accuracy and also compare with other related works.

Diabetes is one of the most current health problems that threaten a lot of people in both developed and developing nations. Diabetes is a disease whereby blood sugar (glucose) is not metabolized in the body. This increases the glucose in the blood to alarmingly high levels. This is known by the name hyperglycemia. In this condition, body is unable to produce sufficient insulin. The other possibility is that body cannot respond to the produced insulin. Diabetes is incurable; it has to be controlled. A diabetic person can develop severe complications like nervedamage, heart attack, kidney failure and stroke. According to statistics in 2017, an estimated 8.8% of global population has Diabetes in 2015 and is predicted to rise to 11.3% by 2030 and to 12.2% by 2045. Hyperglycemia caused by Diabetes, create abnormalities in the cardiovascular system independent of the possible presence of dyslipidemia, arterial hypertension etc.

The need for effective healthcare in order to manage Diabetes is like never before, as the disease has no known cure and the number of people affected is on the rise, the challenge is even worse in developing nation where proper healthcare is wanting. One of the best ways for now to manage this disease is to detect it at an early stage, which is why we want to leverage on current technological development to enhance the healthcare system in diabetes diagnosis by providing a machine learning approach to Diabetic diagnosis.

Diabetes no doubt is chronic disease that has no known cure; the disease can cause organ damage and lead to fatality if it was not diagnosed on time. When diabetes is detected on time the disease can be manage or even avoided in situation where early diagnosis is observed. The problem of early diabetes diagnosis is a challenge to healthcare practitioners all over the world in the sense that they have to make so many wild educated hunches and tests before they come up with the diagnosis. The physicians combine those tests

results together with medical history of the patients in order to decide the diagnosis; the process is time consuming and complicated which left the patient helplessly in pain and lead to many damages. The physician's tries to put all those information together and guess the diagnosis based on heuristics and degree of proximity between the symptoms and the disease which sometime becomes more confusing with other diseases with similar symptoms.

The proposed optimized algorithms will try to improve the performance of the ordinary machine learning algorithms and helps provides faster methods of diagnosis, more accurate than the ordinary models used before in diabetes diagnosis and will simply aid in the task of decision making made by physicians.

The aim of this research isto build an optimized model of machine learning for diabetes prediction, by optimizing the parameters of Support Vector Machine (SVM) and Logistic Regression algorithms (LR) using Particle Swarm Optimization (PSO) to reduced dimensionality of features and selectthe best features that will boost the model accuracies.

The importance of computerized diabetes detection and prediction cannot be overemphasized, but not just computerized, using artificial intelligence technology; a machine learning approach in which the system will not limit itself to its knowledge base as in case of expert systems, but a more robust and accurate machine learning model which will continue to learn and improve upon its accuracy. Machine learning finds its applications in many automated systems such as Airline industries, medical diagnosis etc. In most cases, these applications were proved to be worthy of use as it helped in improving efficiency, accuracy and reliability of systems. (E.g. automatic grading system, object recognition, license plate recognition systems, etc.). In this research, we intend to optimized and implement some of the most popular classification algorithms in diabetes prediction and evaluate their performances. Optimized Logistic regression (LR) and optimized Support Vector Machine (SVM) will be applied to the PIMA Indian Diabetes dataset for diabetes diagnosis and their performance will be evaluated. The algorithms will be compared in terms of accuracy of the prediction in order to identify the best algorithm with high performance.

2. Literature Review

This chapter discusses the existing literatures reviewing previous works and researches carried out in the field of optimization of machine learning algorithms, medical diagnosis and other related researches carried out by experts from various fields that will add more light to the current research in view.

There is no doubt that the alarming figures of diabetes patients needs great attention of several experts even though many researches have been carried out on how to aid in diabetes diagnosis from the doctors knowledge, expert systems to machine learning approaches in the past, but still there is much that needs to be done to improve the diagnosis and improved prediction accuracy because is a matter of life and death. Some of the researchers who tried to predict diabetes using machine learning algorithms with a success stories include: Zou*et al,* (2018) in their research they used Machine learning algorithms to predict diabetes condition, the dataset used in the research consist of 151,598 instances of real diabetic patients and 69,082 of non-diabetic patients used for training and testing of the models. The datasets were gotten from hospital physical examination data in Luzhou, China. Specifically they used three algorithms namely Decision tree, Neural network and random forest in their research and were able to achieve 80% accuracy been the highest accuracy attained at that time.

Iyer A, Jeyalatha S, and Sumbaly, R., (2015) among the earliest time since 2015 have tried to use ML technology in diabetes prediction. The study of Iyer A, Jeyalatha S., Sumbaly, R. proposed the use of the Naïve Bayes algorithm to predict the onset of diabetes. The study gave an accuracy result of 79.56%.

Furthermore, Jakka, A. and Vakula, R., (2019) took an in depth performance evaluation of about six algorithms in diabetes prediction. The major aim of their study was to investigate the performance of various classification models that can anticipate the probability of disease in patients with the greatest accuracy and precision. Among the algorithms evaluated were K Nearest Neighbor (KNN), Decision Tree (DT), Naive Bayes (NB), Support Vector Machine (SVM), Logistic Regression (LR) and Random Forest (RF). The algorithms were used on Pima Indian diabetes datasets available online at UCI repository, the results proved that Logistic Regression (LR) performs better with the accuracy of 77.6 % in comparison to other algorithms.

Looking at the trends of the works and studies carried out in the field of diabetes prediction from 2010 to date, we can see the performances of the models are increasing accordingly but slowly and this is a matter of life and death as we saw other models attaining such accuracy in different application arrears for example email classification (Emmanuel et al., 2019). It is clear that ordinary algorithms cannot achieve such optimum accuracy alone, which is why the state of the art researchers employed several techniques on how to improve the accuracies of the algorithms using optimization and hybridization.

Among the methods tried in an effort to improve the accuracy of diabetes diagnosis models using machine learning were a hybrid methods in which more than one algorithm were applied to same dataset for better results as we have seen in the work of Mahsa, A*., et al*(2019). Another method was to use principal component analysis with one or two algorithms combined together as in the case of Zhou, C., (2019) and many other researches but nonetheless the results shows a significant improvement from the models that uses a single algorithms alone but they were not able to attained that optimum accuracy.

The accuracies of all the models in papers reviewed both ordinary and optimized algorithms were less than 90% altogether, until Tarun J., and Pawan, K.M., (2014) tried improving support vector machine (SVM) algorithm using Principal component analysis (PCA) for the classification of diabetic patients. The experimental result from the study showed that the algorithm can be improved upon when optimized with PCA as they did and obtained an accuracy of 93.66%.

Suyash, S., *et al,* (2019) viewed that the problem of diabetes prediction in machine learning can best be tackled with a neural network as it has high classification accuracy compared to other algorithms, in their work they build an artificial neural network (ANN) model for diabetes classification. The model was train and tested on Pima Indian diabetes datasets and was ideal for predicting the possibility of diabetes and achieved 92% testing accuracy. They recommend that the model can achieve more accuracy if it trains with large sample training data in future claiming that the instances were simply not enough for the model and its known fact that ANN models improved with addition of datasets.

It can be seen that even though Suyash*et al,* (2019) claim that neural network can perform better than the rest of ML algorithms, which lead to the development of their ANN model for diabetes prediction the accuracy obtained by their model is less than that of Tarun J., Pawan, K.M., (2014) who optimized the ordinary algorithm using PCA in their works. It can be inferred that an optimized algorithm will outperformed all other single or ordinary algorithms in term of accuracy even the artificial neural network as the above comparison shows between the work of Suyash and Tarun which give us more reason as to why the proposed or intended work of optimizing SVM and LR should be a success since accuracy improvement is one of the objective of the work as the literature suggest.

The above discussed research works related to Machine learning in diabetes diagnosis shows that the accuracy is still not optimal as we have seen how Machine leaning algorithms were able to achieved 99%

accuracies in other fields (Emmanuel et al., 2019). With this in mind researchers begin to optimized ordinary machine learning algorithms to improve upon it and increases the accuracies of their models and these are what caught the attention of Yuxiang S,.(2010) of the earliest people to have optimized an ML algorithm using particle swarm optimization in classification problem just to see if the technique will work or not, whether the performance will improve or not. Yuxiang S., Qing C., andHong J., (2010) were the earliest set of people to have optimized a machine learning algorithm using particle swarm optimization (PSO) in order to improve its performance. Their paper attempt to develop a Radial Basis function (RBF) neural network based on particle swarm optimization (PSO) algorithm. RBF parameters were optimized including clustering centers, variances of Radial Basis Function and weight. The work simplified the structure of RBF neural network and also enhanced training speed and mapping accuracy. The performance and effectiveness of the new method were evaluated by using function simulation and compared with ordinary RBF neural network. The result shows that the optimized RBF neural network has significant advantages in terms of fast convergence speed, good generalization ability and not easy to yield minimal local results.

Ming, Y, C. and Thi, T., H., (2017) also optimized SVM using PSO in classification of faults in power distribution systems. Their model PSO-SVM classifier is able to select appropriate input features and optimize SVM parameters to increase classification accuracy. The technique has been tested on a typical datasets containing ten different types of faults with 12 given input features. They were able to achieve a success rate with accuracy of 97%, which demonstrates the effectiveness and high efficiency of the developed model.

Harleen, K., and Vinita, K., (2018) used multifactor dimensionality reduction (MDR) to optimized neural networks in prediction of diabetes and compared the performance of the optimized algorithm with other three machine learning algorithms specifically SVM-linear, KNN, RBF in order to validate their model. According to Harleen, K., and Vinita, K., (2018) "Multifactor dimensionality reduction is an approach for finding and representing the consolidation of independent variables that can somehow influence the dependent variables. It is basically designed to find out the interactions between the variables that can affect the output of the system."

Put together with ANN the system were able to achieve accuracy of 83% less than the rest of the models where KNN and RBF achieved 86% and 89% respectively. It therefore shows that not all optimization are proved to be worthy of implementation or capable of improving an algorithms or outperforming native algorithms as the case may be.

Moreover researches are still carrying out many works on how to achieve the optimum accuracy in diabetes prediction which is far from reach even with current optimized algorithms we have seen so far, which is why Changshen, Z., Christian., U.I., Wenfang, F., (2019) proposed a multi- algorithms classifier in diabetes prediction. In their work they primarily used logistic regression which was optimized by K-means algorithm and also optimized by PCA they called it "PCA+K-means+ logistic regression" technique. They first use PCA to transform the initial set of features thereby solving the problem of correlation which makes it difficult for the classification algorithm to find the relationship among the data. PCA helps them filter the irrelevant features and concentrate or work only on the relevant ones, thereby lowering training time and increased performance. The result from the PCA was then passed to a K-means clustering algorithm so that outlier's issues will be addressed. The clean result from the K-means was finally passed to logistic regression for classification. The above hybrid multi-algorithms technique of optimization was able to

achieve an accuracy of 97.40%in diabetes prediction, but nonetheless it still suffered from many drawbacks such as time taken to build the model was significantly high. The model was computationally expensive as it uses three algorithms simultaneously. They concentrate on the accuracy and discard time and resource constrain. Yet despite all the complexities, resources and time consuming nature of the model it failed to obtained optimum accuracy.

However, almost all the literatures discussed above almost none have tried to describe how SVM or LR can be optimized using PSO on diabetes prediction, which shows the viability of this research work and they have been optimized before with different optimization technology and in different application area, which clear the road for our intended study. It can also be seen from the discussion that not all optimization techniques are capable of improving a machine learning algorithm, which is why we selected the PSO optimization technique as it provides success stories from the consulted literature in combination with different algorithms which will serve as a justification for using PSO. Justification for using SVM and LR was because the work had never been carried out in relation to diabetes prediction with those algorithms.

The discussion of the literature discusses numerous paper specifically several articles were consulted in which some were cited, three textbooks and a couple of published conference proceedings. The discussion provides us with clear understanding of the problem we intended to solve as many similar problems were consulted and also provides a kind of roadmap or base on which our proposed work will fit as the new study that will contribute to the realm of knowledge when conducted as such giving us a confidence and basis for the research.

The literature review showcase a research gaps that needs to fill with relation to diabetes diagnosis using machine learning technology, it shows that many works have been conducted in past on diabetes prediction but the works suffered with lesser accuracy in terms of performance, we also saw how some people tried to improve the accuracy by trying several methods of improvement, hybridization and optimization but none was capable of achieving the optimum classification accuracy.

3. Methodology

The research methodology employed in this research work is hybrid both qualitative and quantitative methods were combined, where an interview was conducted with various doctors in hospitals in order to obtained a first class information about the disease under investigation, to enabled us understand the dynamics, symptoms and complexities involved in diagnosis of Diabetes. Quantitative method is where the datasets was obtained. It was shown before in previous researches that the success rate of classification algorithms depend on the size of training sets, without rich datasets the models will not learn enough to make accurate predictions, which is why large quantity datasets was required from various hospitals to allow our model learn and predict with high degree of accuracy. Nonetheless the number of dataset we were able to get locally was highly insufficient as such we improvised and used open source diabetes datasets from India and the fact that it is the benchmark dataset and used in many researches in diabetes domain. Literature were reviewed about previous worked carried out in areas of machine learning and disease diagnosis in which many journals and articles were reviewed, to make sure enough information was gathered and guarantee the success of this research.

3.1 Dataset Description

The dataset which was obtained from PIMA Indian diabetes database in an open source diabetes datasets which consist of 768 instances each tested for diabetes. The dataset has a total of 9 attributes that was used in diabetes diagnosis plus one target class attribute which represents the status of each tested

International Journal of Biometrics, Image Processing and Computing Research
ISSN: 2997-688X |
Volume 11 Issue 1, January-March, 2023
Journal Homepage: https://ethanpub.online/Journals/index.php/E23
Official Journal of Ethan Publication

individual either positive or negative. The dataset contains a total of 268 positive instances and 500 negative instances. Below are the attributes available in the in the dataset:

1. Number of times pregnant (Pregnancies)
2. Plasma Glucose concentration at 2hr in an oral glucose tolerance test (Glucose)
3. Diastolic Blood pressure (BloodPressure)
4. Triceps skin fold thickness (Skin Thickness)
5. 2-hr serum insulin (Insulin)
6. Body mass index (BMI)
7. Diabetes pedigree function (Pedigree)
8. Age (Age)
9. Target Variable (Outcome)

Index	Pregnancies	Glucose	BloodPressure	SkinThickness	Insulin	BMI	etesPedigreeFunc	Age	Outcome
0	6	148	72	35	0	33.6	0.627	50	1
1	1	85	66	29	0	26.6	0.351	31	0
2	8	183	64	0	0	23.3	0.672	32	1
3	1	89	66	23	94	28.1	0.167	21	0
4	0	137	40	35	168	43.1	2.288	33	1
5	5	116	74	0	0	25.6	0.201	30	0
6	3	78	50	32	88	31	0.248	26	1
7	10	115	0	0	0	35.3	0.134	29	0
8	2	197	70	45	543	30.5	0.158	53	1
9	8	125	96	0	0	0	0.232	54	1
10	4	110	92	0	0	37.6	0.191	30	0
11	10	168	74	0	0	38	0.537	34	1
12	10	139	80	0	0	27.1	1.441	57	0
13	1	189	60	23	846	30.1	0.398	59	1
14	5	166	72	19	175	25.8	0.587	51	1
15	7	100	0	0	0	30	0.484	32	1

Fig 3.1: Screenshot of first 16 instances in the datasets

3.2 Data Dictionary

Each column in the datasets or an attribute represents a distinct character attributed to the patient under investigation which when put collectively determined the condition of the patient.

Index: it is the first column in the dataset which describe the number of instances in the dataset from 0 to the last number of the instance 767.

- **Pregnancy**: it's the second column which indicates the number of times the patient has been pregnant for female patient as it's very important in diagnosis of diabetes in females. It's simply represented by 0 for a male patient.
- **Glucose:** It's the third column which contains the number of glucose level in blood plasma in mg/dL. The numbers were obtained by testing the patient glucose and obtained the numbers using glucometer.
- **Diastolic Blood Pressure**: the column contains diastolic reading or the bottom number, is the pressure in the arteries when the heart rest between beats. A time when the heart fills with blood and gets oxygen. A normal diastolic pressure is lower than 80.
- **Triceps Skin fold thickness**: it contains numbers from the skinfold measuring, the triceps skinfold is necessary for calculating the upper arm muscle circumference which gives information about fat reserve of the body. The normal reading varies between ages and gender of the individual.
- **Serum Insulin:** the column contains numbers that represent the amount of insulin in the bloodstream. The figures are derived through an insulin test. Insulin is a hormone that helps move blood sugar known as glucose, low insulin means glucoses can't get into cells it stays at bloodstream and causes problems. The normal insulin level should be around 111-1153pmol/L
- BMI: the column contains the measures of Body Mass Index (BMI) of the patients. BMI is a value derived by dividing the body mass by square of the height of an individual which is expressed in Kg/M^z. The normal BMI is between 18.5 to 24.5 depending on the gender and age also.
- **Diabetes Pedigree Function**: Contains figures that indicate the likelihood of diabetes based on family history.
- **Age**: contains the numbers representing the ages of the people samples in the dataset.
- **Outcome**: It contains the numbers which indicate the prediction result of the individual; it has a Boolean data type with 1 indicating positive for diabetes and 0 indicating a negative case.

It can be seen that all the columns contains numerical values that are very significant when it comes to diabetes prediction be it manual or with computers.
The first 6 columns attribute have integer data type property, the seventh and eighth have float data type, with the last attribute outcome been a Boolean type.

3.3 Data Preprocessing

3.3.1 Data Cleaning

According to *Zhou et al,* (2019). Today's real world datasets are highly susceptible to noisy, missing values, and inconsistent data due to their typically huge sizes and their likely origin from multiple, heterogeneous sources. Data quality is an important factor in the data mining process for disease prediction and diagnosis, because low quality data may lead to inaccurate or low prediction result.
From the above dataset described, it is quite obvious that our datasets contains such inconsistencies such as noise, missing values, for example 0 values in an insulin column is considered as a missing value since no matter how low the insulin reading is cannot be 0, to handle such a problems in our dataset and make our original dataset more productive and applicable for predicting diabetes, we applied several preprocessing techniques using various packages offered within the Anaconda integrated development environment. The dataset was cleaned and all unnecessary details were removed and it was made sure that the dataset is of right format compatible to our toolkits, the cleaning process was necessary in order

to ensure that the dataset is consistent and free of missing values. The method employed in handling the missing value is by calculating the average mean of all the values of a given attribute and the replacing the missing values with the mean value as t will increase the accuracy of the prediction compare to the one with missing values.

S/N	Attributes	Average Mean Values
1	Insulin	80
2	Blood Pressure	69
3	BMI	32
4	AGE	33
5	Skin Thickness	21

Table 3.1: Mean attributes values

For all the attributes with missing values we are going to replace them with the mean values calculated from the rest of the values available in order to boost the performance of our model.

3.4 Tools used

In this research work python programming language was used to implement the model using anaconda toolkit in combination with other data science related packages.

Matlab was also used to plot some of the graphs as it provided more graphical library then python, three separate notebooks were created for each model there is a dedicate Jupyter notebook for it and another one for comparison and validation.

Many python libraries were used over the course of the model implementation some of which are: Pandas, Numpy, Matplotlib, Keras, Sklearn and so on.

3.4.1 HARDWARE REQUIREMENT

The model was developed in Google cloud platform with high Processing CPU's and GPU to aid in speedy development and minimizes resource utilization. The typical system setup of the model is as follows:

3.5.2 Machine Setup

SYSTEM REQUIREMENT

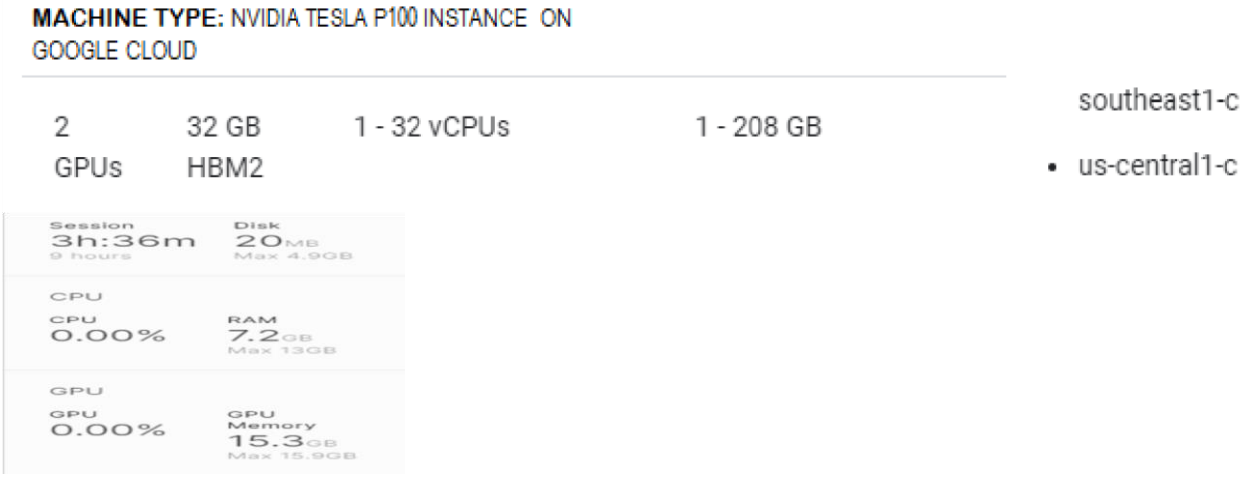


Fig 3.3: Showing the available resources on the machine used, time spent to build and train the model.

3.4.2 THE PROPOSED FRAMEWORK

The proposed framework for the models gives a detailed flow of how the tasks will be performed, describing where will be the beginning and how the task can proceed from a current state to subsequent ones, and it is diagrammatically represented in the figure 3.4.

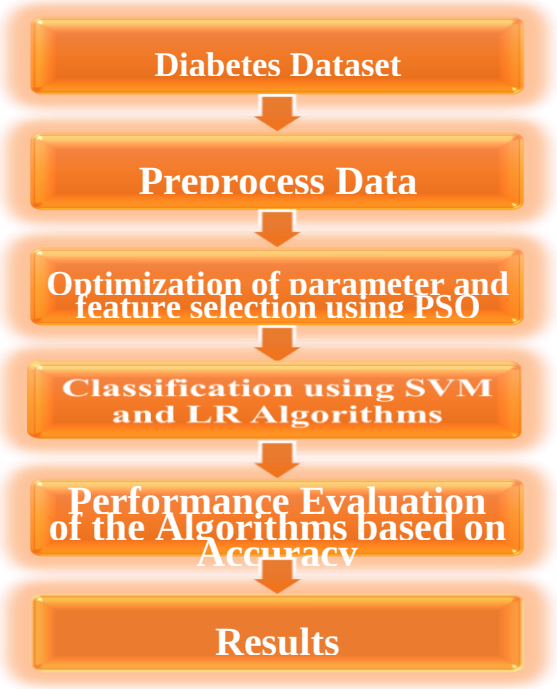


Fig 3.4: Proposed Framework Description

The above framework presents the sequence and flow of the entire experimental process that will be followed in the research work.

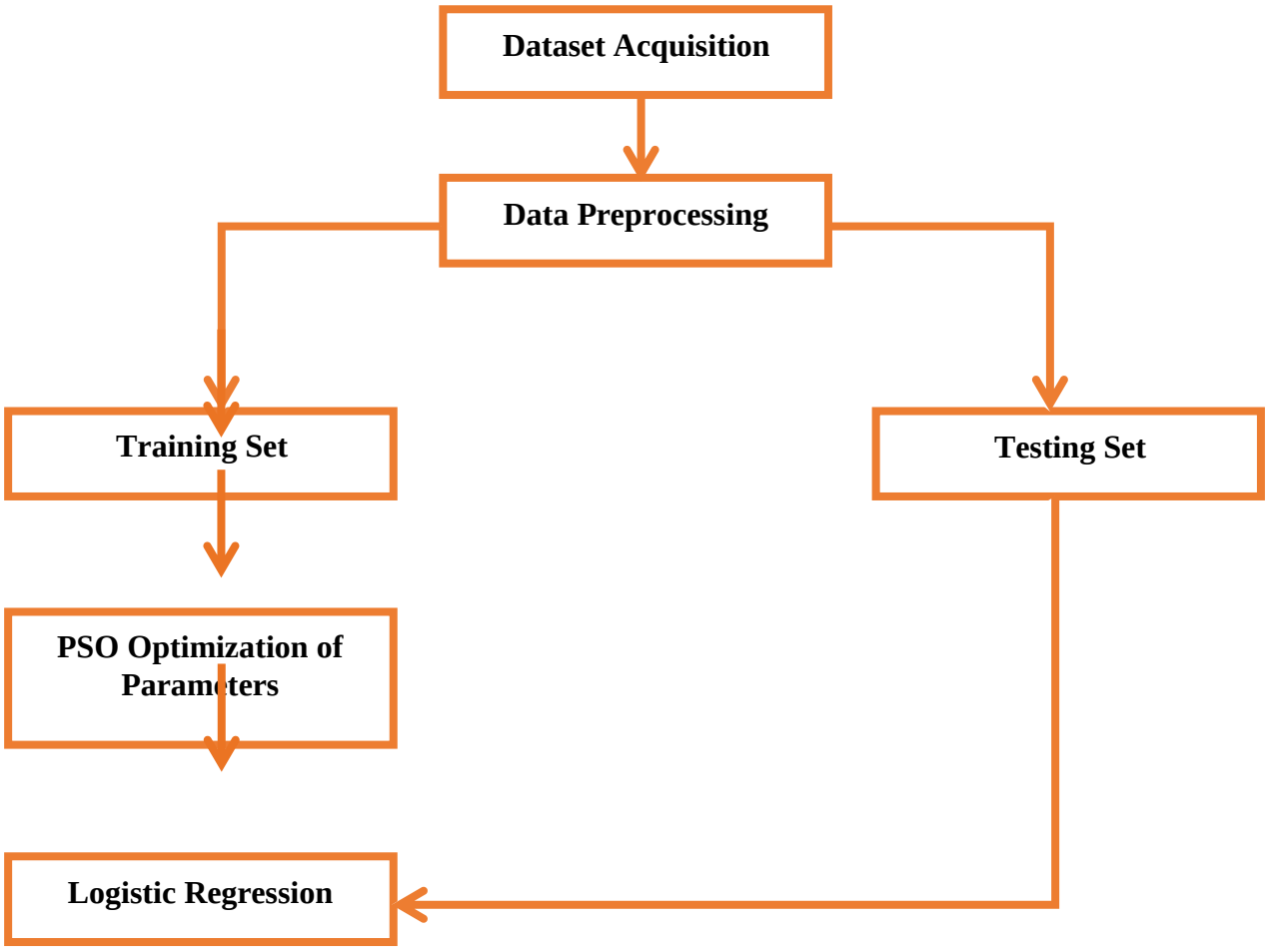
3.5 Optimization of Logistic Regression Model

The application of the Logistic regression model in many domains learning and real life such as the biological sciences, weather stations, stock exchange etc. The Logistic regression algorithm is best used when the objective is to classify data items into categories. Usually in Logistic Regression the target variable is binary, which means that it only contains data classified as 1 or 0 which in our case positive or negative for diabetes. The purpose of our logistic regression algorithm is to find the best fit that is diagnostically reasonable to describe the relationship between our target variable and the predictor variables.

Logistic Regression was optimized with the aid of PSO algorithm, first a principal component analysis (PCA) was done on the model for feature selection then optimization of the parameters of eigenvalues and vectors using PSO and train out model using this setup, it can be clearly describe using algorithm below:

Step 1: Get the Datasets
Step 2: Prepare the Datasets
Step 3: Split the Data into Training sets and Testing Sets

Step 4: Read Training Set
Step 5: Find the mean
Step 6: Calculate variance
Step 7: Calculate covariance
Step 8: Find the eigenvalues and the eigenvectors
Step 9: Choose Parameters within the range of eigenvalues and eigenvectors
Step 10: Form Logistic regression model using the Training sets
Step 11: Evaluate fitness of each parameter and select the best parameters
Step 12: Loop through parameters
Step 13: If max parameter then go to step (7); else go to step (14)
Step 14: Optimum solution obtained
Step 15: Retrain with optimum parameters; then identify unknown samples on testing sets. Step 16: End



International Journal of Biometrics, Image Processing and Computing Research
ISSN: 2997-688X |
Volume 11 Issue 1, January-March, 2023
Journal Homepage: https://ethanpub.online/Journals/index.php/E23
Official Journal of Ethan Publication

Output (PSO_LR)

Fig 3.5: PSO-LR Flowchart

3.5.1 Model Description

The logistic regression model was based on linear regression and is described by equation below:

$y = h(x) = \theta Tx$ ----------Equation 1 (Linear Regression)

Then by adding a sigmoid function of:

$\sigma(t) = \frac{1}{1+e^{-t}}$---------- Equation 2 (Sigmoid function)

Now putting the two Equations together, we have:

$y = h(x) = \sigma(\theta Tx)$ ---------- Equation 3 (Logistic Regression)

By performing PCA to obtain necessary parameters that have more relevance in the prediction accuracy and the one that will be optimized by Particle Swarm Optimization Algorithm (PSO).

Assume Dataset =X

Dimension= P

Where Y is the objective function

$\therefore Y = pc(X)$---------- Equation 4

X has n vectors $X = [x1, x2, x3 \ldots xn]$

Where $x1, x2 \ldots xn$ single instances in the dataset are

Next is to compute the mean value using:

$Mean(X)\ \{\sum_{i=1}^{n} Xi\}/\mathrm{n}$ ----------Equation 5

Now let's compute the covariance from which eigenvalues and eigenvectors will be used to derived as the parameters

$Xn \times n = (x, x = cov\,(Dimi, Dimj))$ ----------Equation 6 (Covariance)

Where $X\ n \times n$ is the data matrix and $Dimi$ is the $_{ith}$dimension

Now the Eigenvalues and Eigenvectors of the covariance will determine the direction and the magnitude of the new features space respectively otherwise it refers as parameters and can be obtained as follows:

Let Abe $n \times n$matrix, then a nonzero vector y in $R^n$is called an eigenvector of A. If $Ax$is a scalar multipleof x; that is,

$Ax = \lambda x$ ----------Equation 7

For some scalar λ. The scalar λ is called an eigenvalue of A and x is said to be an eigenvector corresponding to λ. Since the eigenvectors corresponding to an eigenvalue of a matrix A are the nonzero vectors that satisfy the equation:

$(\lambda I - A)\, x = 0$----------Equation 8

Then we define the set E to be all vectors x that satisfy equation (8) as our corresponding Eigen space which can be mathematically express as:

$E = \{x: (A - \lambda I)\, x\} = 0$ ----------Equation 9

Where E is space containing the new parameters obtained from PCA then we sort the sets from highest to lowest where highest are the principal components needed for optimization.

To apply PSO to the new parameters E we made some simple modifications and assumption as follows:

We feed E to PSO as input or initial population called swarm with Size n and dimension Dim then

$E = [E1, E2, E3 \ \ldots \ En]^T$ ----------Equation 10 (where T is transpose operator)
Initial velocity of the swarm $V = [V1, V2, V3 \ldots VDim]$
Let $p = 1, 2, 3 \ldots n$
Let $q = 1, 2, 3 \ldots . Dim$
Let r_1, r_2 be two random numbers between 0-1 Let c_1, c_2 be two accelerating factors Then the model will be:
Vp, qk+1 +$w \times V^k{}_{p,q} + c\ r1 \quad 1\ (Pbest^k{}_p, - E^k p,) + c2 r2\ (Gbest^k{}_q - E^k{}_p), E^k{+}1p, = E^k{}_p, + V^k{+}1{}_p,$ ----- Equation 11
Where: w is an inertia factor
The initial Pbest is the first position of each particle and Gbest is the initial best position among the random swarms and they are updated as follows:
If Enext<Pbestcurrent then Pbestnext = E current else Pbestnext = pbestcurrent

If f (E)< f(Gbest) then Gbest= E_{curent} else $Gbest_{next} = Gbest_{current}$
Where $F(E)$ is the new objective function of Equation(3)
Replacing x with E Then
$y = h(E) = \sigma(\theta TE)$---------- Equation 12 (Optimized Model of LR)

3.6 Optimization of SVM Model Using PSO

To be able to give enough insight of what is to be done when optimizing the parameters of SVM using PSO below is a step by step algorithm of the experimental procedure:
Step 1: Get the Datasets
Step 2: Prepare the Datasets
Step 3: Split the Data into Training sets and Testing Sets
Step 4: Read the p Training Set and set PSO parameters
Step 5: Initialized sets of SVM Parameters W, C_1, C_2within the range of position and Velocity
Step 6: Form SVM using Training Datasets
Step 7: Evaluate fitness of each particle and select the best particle
Step 8: Set iteration count k=1
Step 9: If k max iteration then k=k+1 and go to step (7); else go to step (10)
Step 10: Optimum solution obtained: print the results of optimum generation as best
Step 11: Retrain SVM with optimum features and parameters; then identify unknown samples on testing dataset. Step 12: End

International Journal of Biometrics, Image Processing and Computing Research
ISSN: 2997-688X |
Volume 11 Issue 1, January-March, 2023
Journal Homepage: https://ethanpub.online/Journals/index.php/E23
Official Journal of Ethan Publication

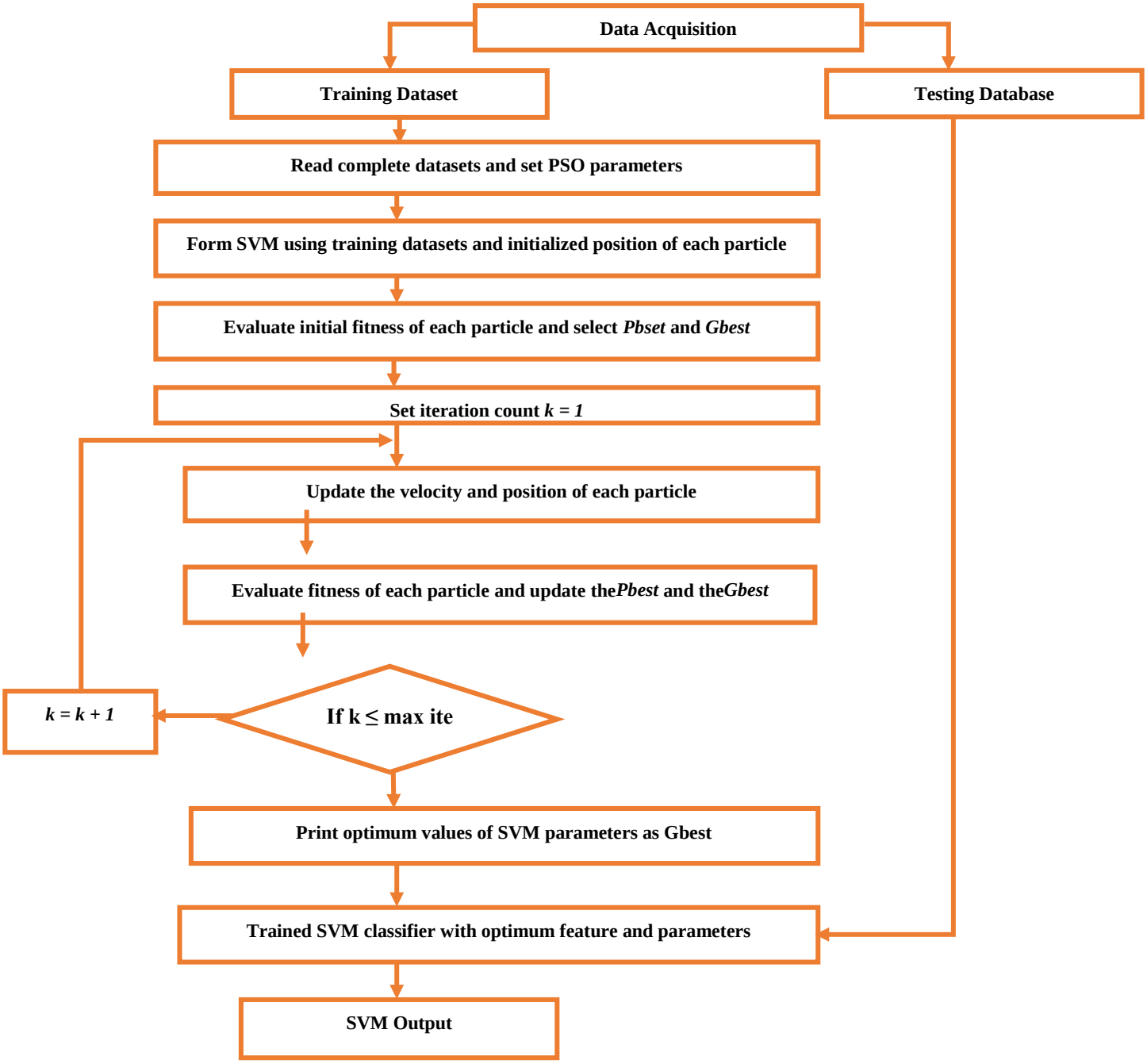


Fig 3.6: PSO SVM Flowchart

3.6.1 Model Description

For SVM we don't need to repeat the procedure on how PSO traverse the population sets and no PCA is required therefore the model will be described as follows:

International Journal of Biometrics, Image Processing and Computing Research
ISSN: 2997-688X |
Volume 11 Issue 1, January-March, 2023
Journal Homepage: https://ethanpub.online/Journals/index.php/E23
Official Journal of Ethan Publication

$y = a \times x + b$ ----------Equation 13
Let $X = (x, y)$ and $W = (a, -1)$ ----------Equation 14
$\therefore W.X + b = 0$ ----------Equation 15
Then
$yi(w.xi + b) - 1 \geq 0 \quad for\ yi = +1, -1$ ----------Equation 16 Therefore the optimization problem can be presented as follows
$Min: W| + C \sum_{i=1}^{m} \varepsilon i$ ----------Equation 17
Subject to: $yi\ (w \times xi) + b \geq 1 - \xi i$ ----------Equation 18
Where C is regularization parameter and εi is relaxation parameter
$\varepsilon i > 0, \quad i = 1 \ldots m$
$w \times x\, i + b \geq +1\ for$
$yi = +1$ –---------- Equation 19
$w \times x\, i + b \geq -1\ for\ yi = -1$---------- Equation 20

Therefore the required objective function will then be:
$f\ (x) = sign\ (m\Sigma i = 1 \alpha i * \times\ yi \times (x\, i, yi) + b *)$---------- Equation 21

Equation 21 is the required objective function equation

To apply PSO we simply check our swarms for local and global best parameter using the PSO algorithm discussed in model 1, the selection process can be achieved with an equation below:
$fPbestPbestPbestbest$
If $(X_{p^{k+1}}) < (_{p^k})$ then $_{p^{k+1}} = X_{p^{k+1}}$ else $_{p^{k+1}} = P\ _{p^k}$
If $(X_{p^{k+1}}) < f\ (Gbest^k)$ then $Gbest^k = X_{p^{k+1}}$ else $Gbest^{k+1} = Gbest^k$
With Gbest been the optimum parameters of PSO optimized SVM, when applied to equation 16, that is replace older parameter with new parameters
$yi(Wnew.xinew + bnew) - 1 \geq 0 \quad for\ yi = +1, -1$ --------Equation 22 (PSO SVM Model)
Where $Wnew, xinew\ and\ bnew$ are the new optimized parameters.

4. 0 RESULTS AND DISCUSSION

4.1 Results

This chapter presents the results of the experiment carried out over the course of this research work, it present the findings and the outcome in which the quality of the research can be based on, it will show whether the research objective has been achieved or not.

International Journal of Biometrics, Image Processing and Computing Research
ISSN: 2997-688X |
Volume 11 Issue 1, January-March, 2023
Journal Homepage: https://ethanpub.online/Journals/index.php/E23
Official Journal of Ethan Publication

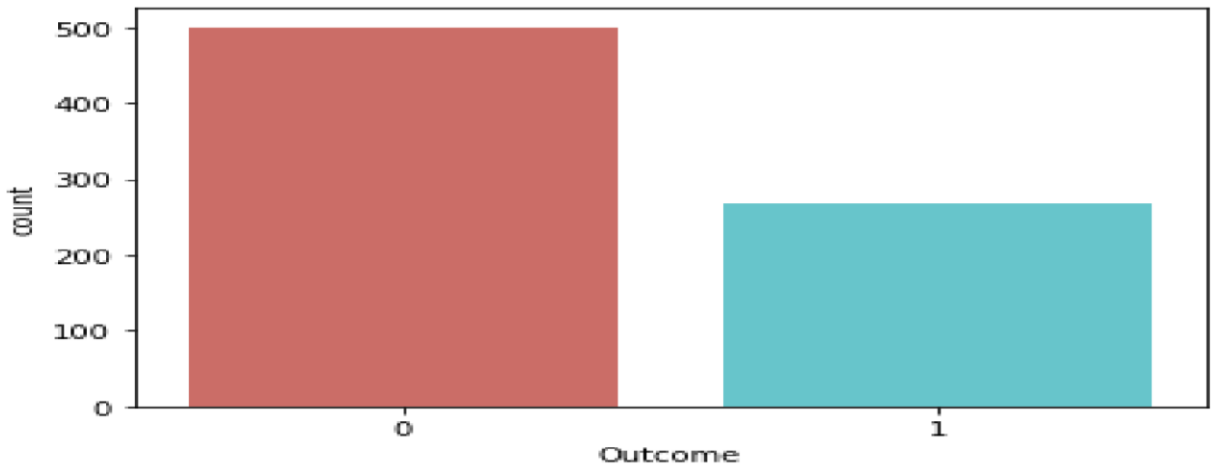


Fig. 4.1: A Barchat showing the real diabetes patient and non-diabetic from our dataset where 0=Non diabetic and 1=Diabetic

4.2 Logistic Regression results

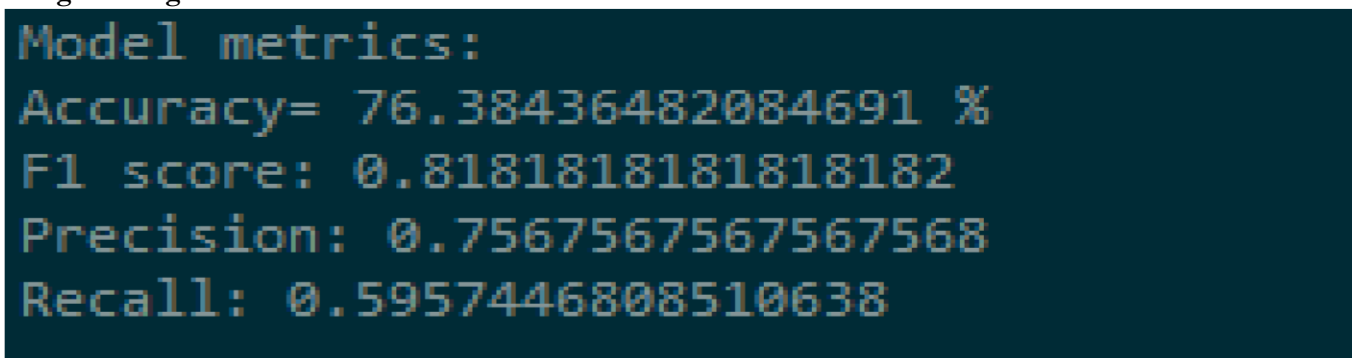


Fig. 4.3: Performance of ordinary Logistic Regression model when applied to the dataset

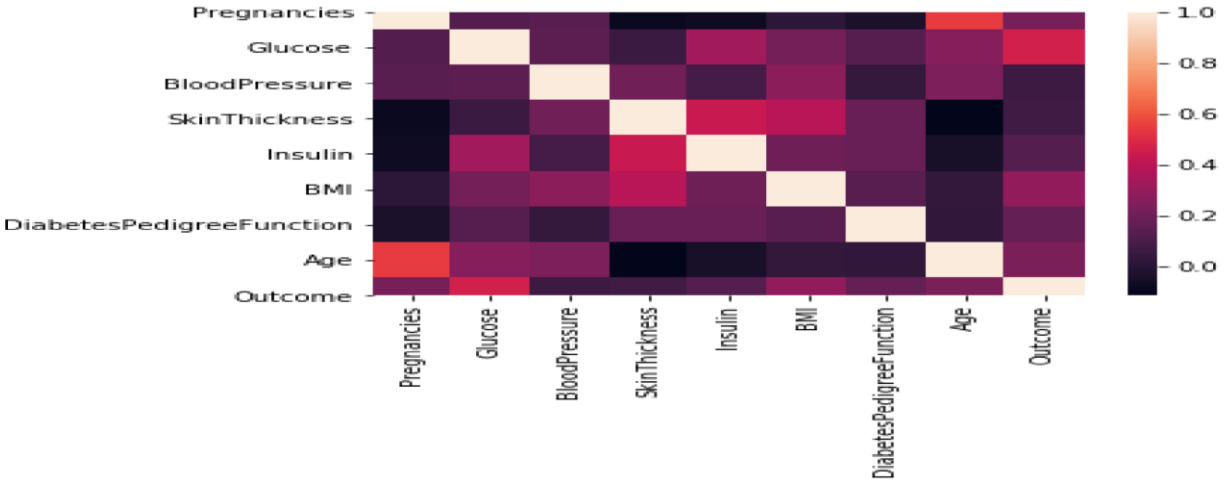


Fig. 4.2: Feature correlation showing the significance of each feature in the prediction

International Journal of Biometrics, Image Processing and Computing Research
ISSN: 2997-688X |
Volume 11 Issue 1, January-March, 2023
Journal Homepage: https://ethanpub.online/Journals/index.php/E23
Official Journal of Ethan Publication

From Fig 4.3 we can see the metrics of traditional (non-optimized) logistic model after been implemented and tested on the dataset prior to optimization. We can see the accuracy is not that commendable, it surely needs to be boost for better predictions.

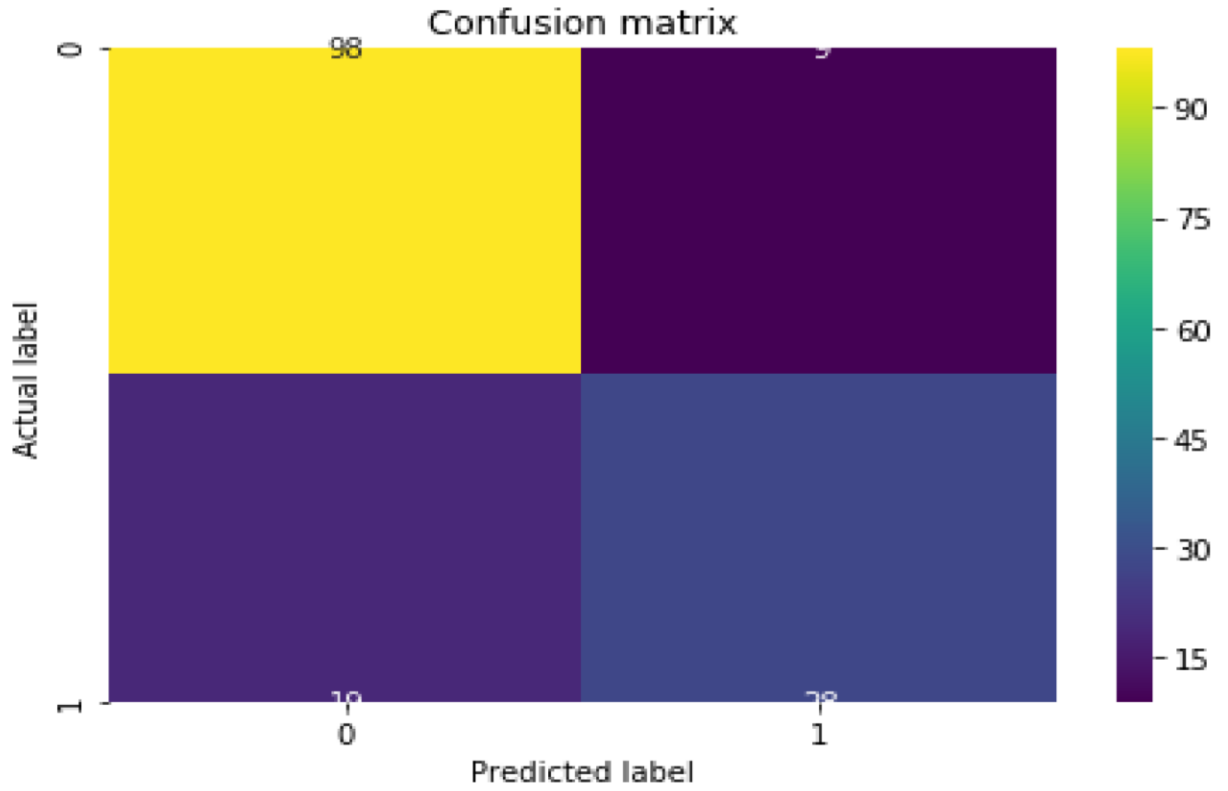


Fig.4.4: Confusion Matrix of the Logistic Regression

Confusion matrix is a metric measure that is used to describe the performance of classification model on datasets where the true values are known, in which it gives out the figures of true positive, true negative, false negative and false positive. Note that:

True positive: The instance where the model predicted yes and the actual value is also yes

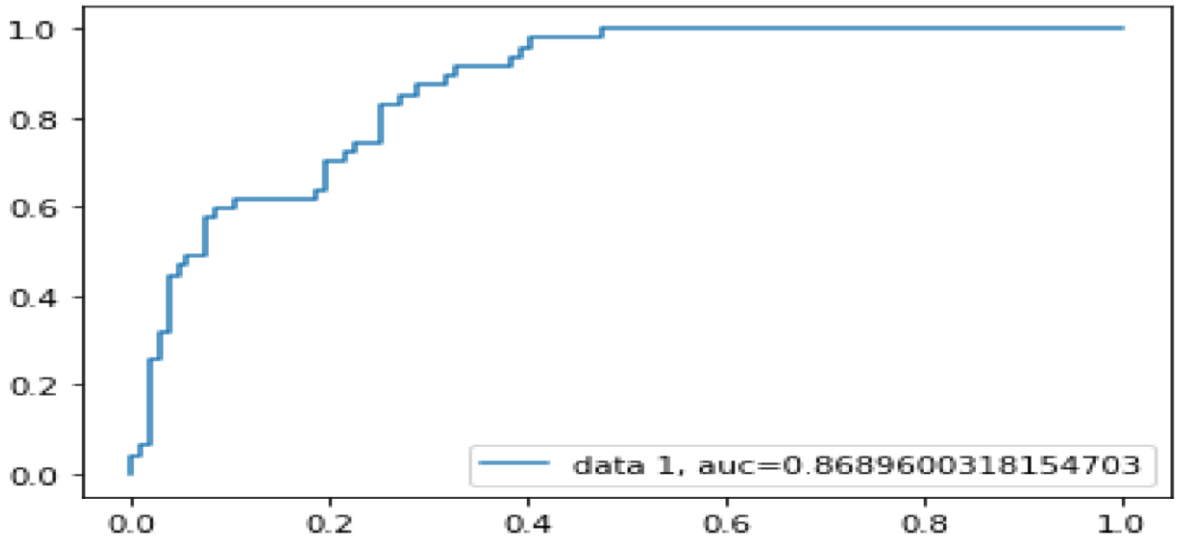
True Negative: The model predicted no, and it was actual no

False positive: The model predicted yes, but it was actual no

False negative: The model predicted no, but it was actual yes

Receiver Operating Characteristic (ROC) curve is a graphical plot of the true positive rate (TPR) (Recall) against the false positive rate (FPR). It shows a kind of tradeoff between sensitivity and specificity.AUC Score (Area Under Curve) is computed when a ROC was plotted the value of this area range between 0-1, a model with an AUC score of 1 is considered as perfect model while model with an AUC score of 0.5 and below are considered as worthless model.

International Journal of Biometrics, Image Processing and Computing Research
ISSN: 2997-688X |
Volume 11 Issue 1, January-March, 2023
Journal Homepage: https://ethanpub.online/Journals/index.php/E23
Official Journal of Ethan Publication



4

Fig.4.5: ROC Curve of Ordinary Logistic Regression

Accuracy: score of a model is the percentage of correct predictions the model performed during testing phase.

The precision: is the ratio tp / (tp + fp) where tp is the number of true positives and fp the number of false positives.

The precision is the ability of the classifier to not label a sample as positive if it is negative.

The recall: is the ratio tp / (tp + fn) where tp is the number of true positives and fn the number of false negatives. The recall is intuitively the ability of the classifier to find all the positive samples

F-Score:The F-score can be interpreted as a weighted harmonic mean of the precision and recall, F-score has a highest values of 1 and worst score at 0.

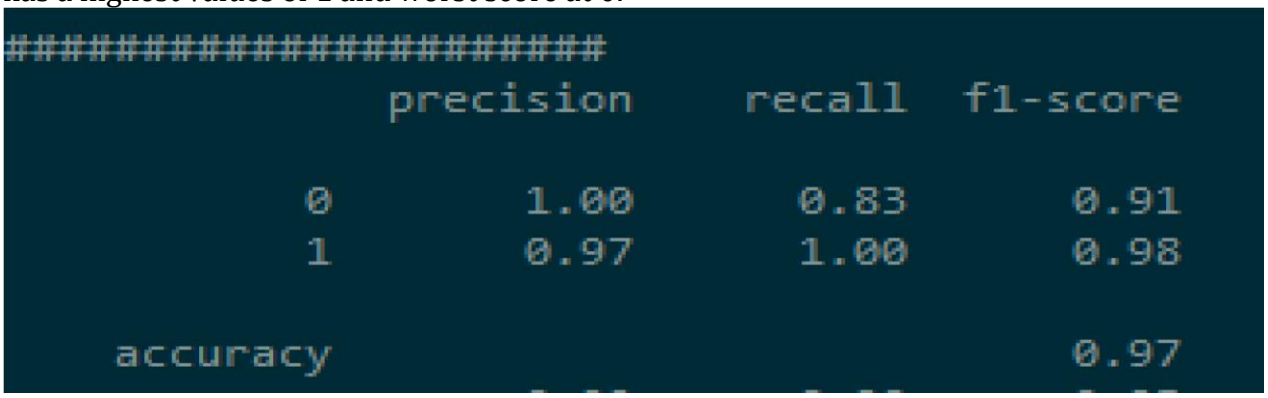


Fig.4.6: Performance of the PSO-Optimized Logistic Regression

Learning curve is a graphical illustration of performances between the training and testing phases of the model, it distinguished the performance a model achieved during the training as well as during the test.

International Journal of Biometrics, Image Processing and Computing Research
ISSN: 2997-688X |
Volume 11 Issue 1, January-March, 2023
Journal Homepage: https://ethanpub.online/Journals/index.php/E23
Official Journal of Ethan Publication

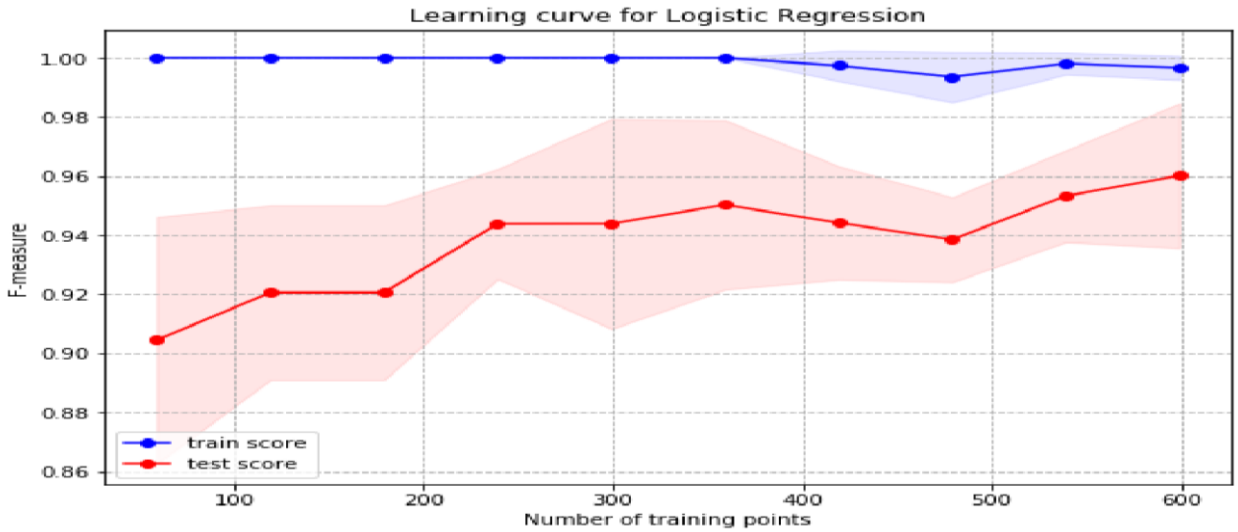


Fig. 4.7: Learning Curve of the Optimized Logistic Regression

Metrics	**Ordinary Logistic Regression (%)**	**PSO Optimized Logistic Regression (%)**
Accuracy	76.38	97
F1-Score	81.82	91
Precision	75.66	100
Recall	59.57	83

Table 4.1: Showing metrics comparison between the two models

International Journal of Biometrics, Image Processing and Computing Research
ISSN: 2997-688X |
Volume 11 Issue 1, January-March, 2023
Journal Homepage: https://ethanpub.online/Journals/index.php/E23
Official Journal of Ethan Publication

4.3 SVM Results

```
Accuracy_Score :  0.8203125
Accuracy: 0.8203125
Precision:0.89457
Recall: 0.764329
```

Fig: 4.8 Accuracy of Ordinary SVM Model

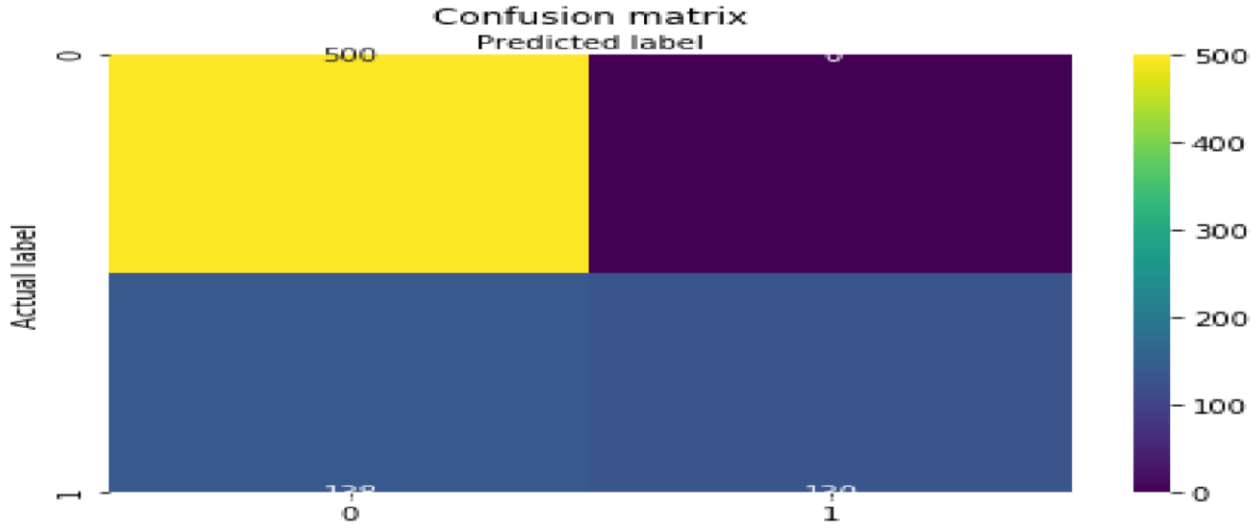


Fig 4.9: Confusion Matrix of Ordinary SVM

The receiver operating characteristic (ROC) curve is another metric tool used to explain the performance of a model it works well with binary classifiers. In the ROC curve below of SVM model the dotted line presumably represents the ROC curve of a purely random classifier; a good classifier stays as far away from that line as possible

(toward the top-left corner).

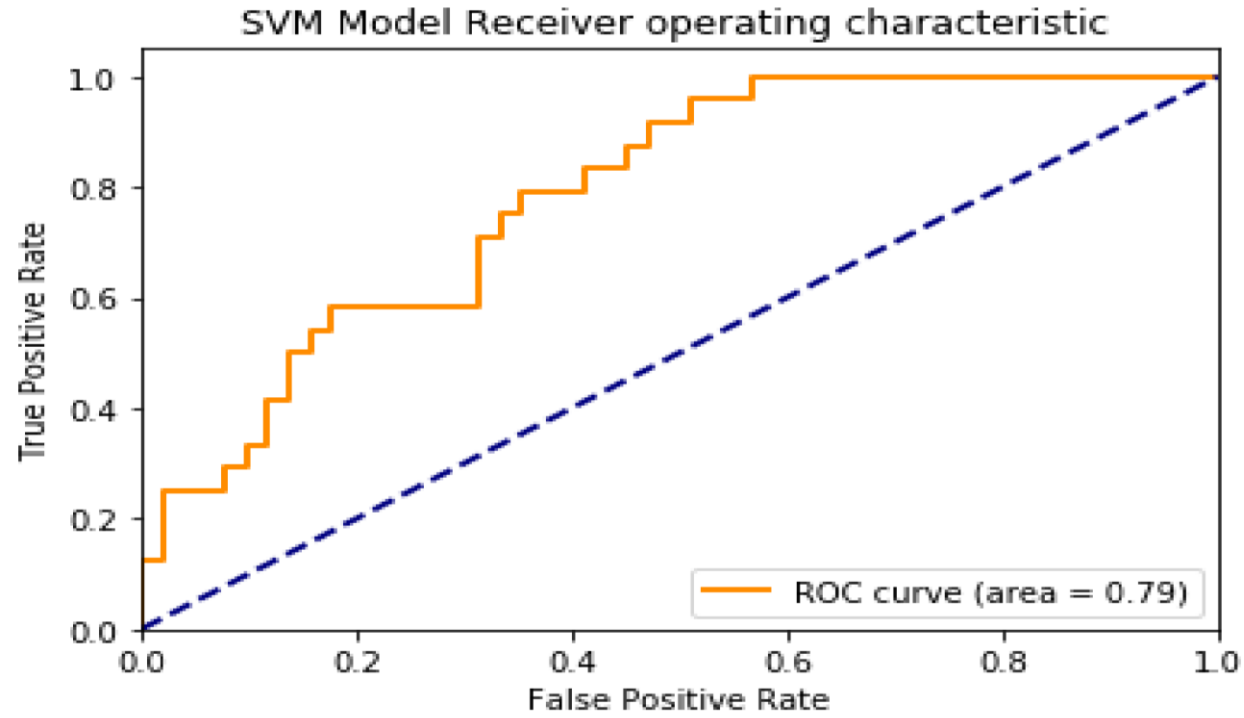


Fig: 4.10 ROC curve of Ordinary SVM Model

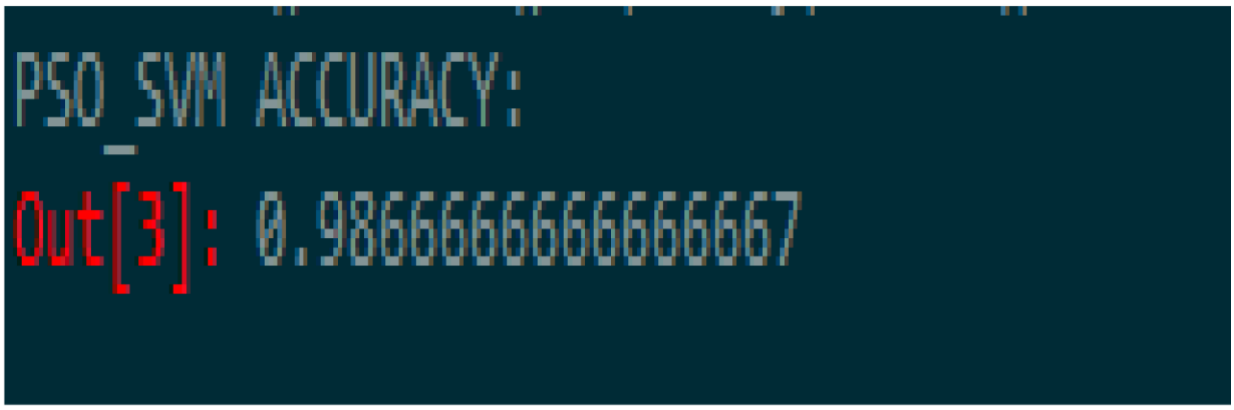


Fig: 4.11 Accuracy of Optimized PSO-SVM Model

International Journal of Biometrics, Image Processing and Computing Research
ISSN: 2997-688X |
Volume 11 Issue 1, January-March, 2023
Journal Homepage: https://ethanpub.online/Journals/index.php/E23
Official Journal of Ethan Publication

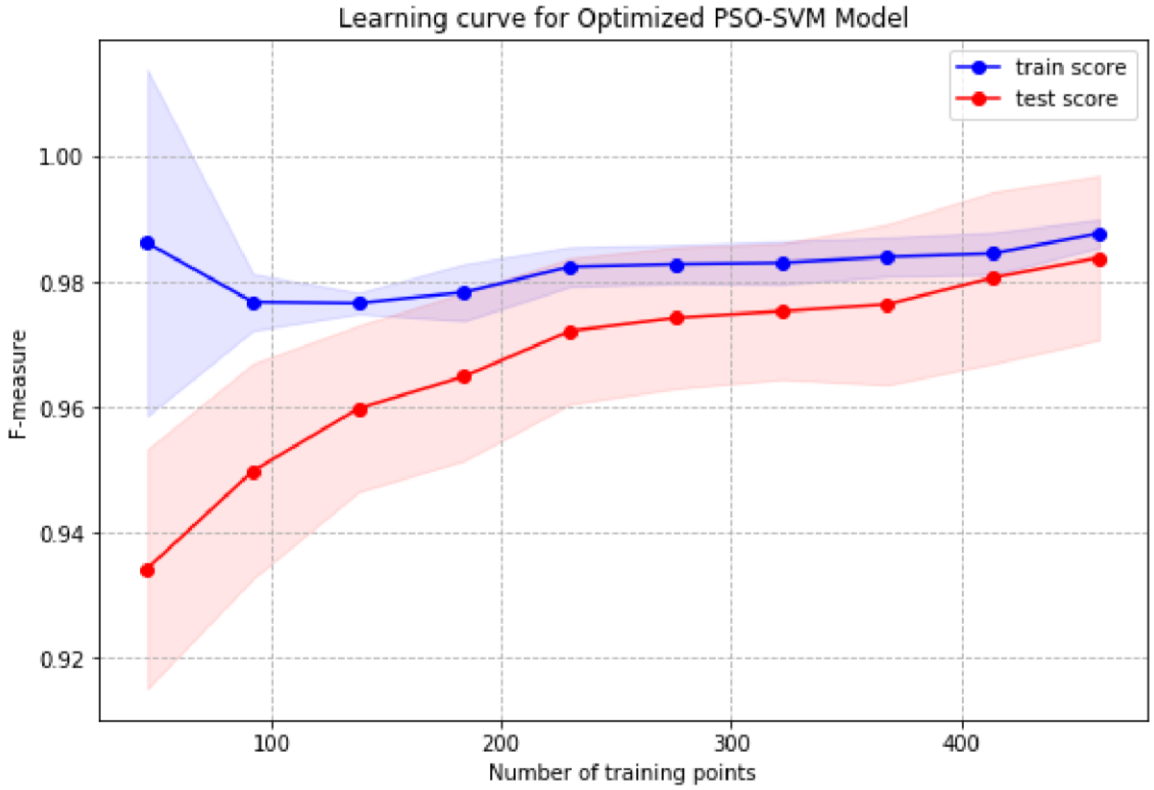


Fig: 4.12: Learning Curve of Optimized PSO-SVM Model

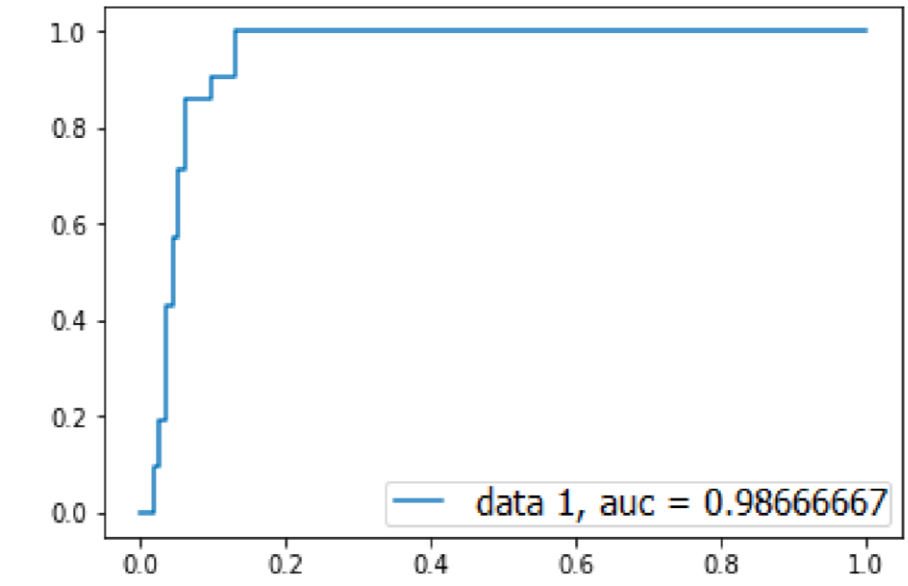


Fig: 4.13 Learning Curve of Optimized PSO-SVM Model

SVM VS PSO-SVM

Metrics	**Ordinary SVM (%)**	**PSO Optimized SVM (%)**
Accuracy	82.03	98.67
F1-Score	81	91
Precision	89.46	97
Recall	76.43	93

Table 4.2: Comparison between Ordinary and optimized SVM Model

4.4 General Models Performance Comparison

Metrics	**Ordinary Logistic (%)**	**PSO-Logistic (%)**	**Ordinary SVM (%)**	**PSO- SVM (%)**
Accuracy	76.38	97	82.03	98.67%
F1-Score	81.82	91	81	91
Precision	75.66	100	89.46	97
Recall	59.57	83	76.43	93

Table 4.3: Performance comparison of all the models both optimized and the ordinary

4.**5 Accuracy Comparison with other Experiments.**

S/N	Papers	Algorithms/Model	Accuracy (%)
1	Tarun J., et al. (2014)	PCA+SVM	93.66
2	Zou Q et al (2018)	PCA+RF on two datasets	76.04 & 80.84
3	Aishwarya, Jakka (2019)	KNN, DT, NB,RF, LR & SVM	LR with 77.6
4	ChangSHeng Zhu (2019)	Logistic regression+PCA+K-means	97.4
5	Han Wu et al (2018)	Logistic regression+K-means	95.45
6	Our Proposed Model	PSO_SVM and PSO_ LR	98.67 & 97

Table 4.5: comparison accuracy of the model with other works

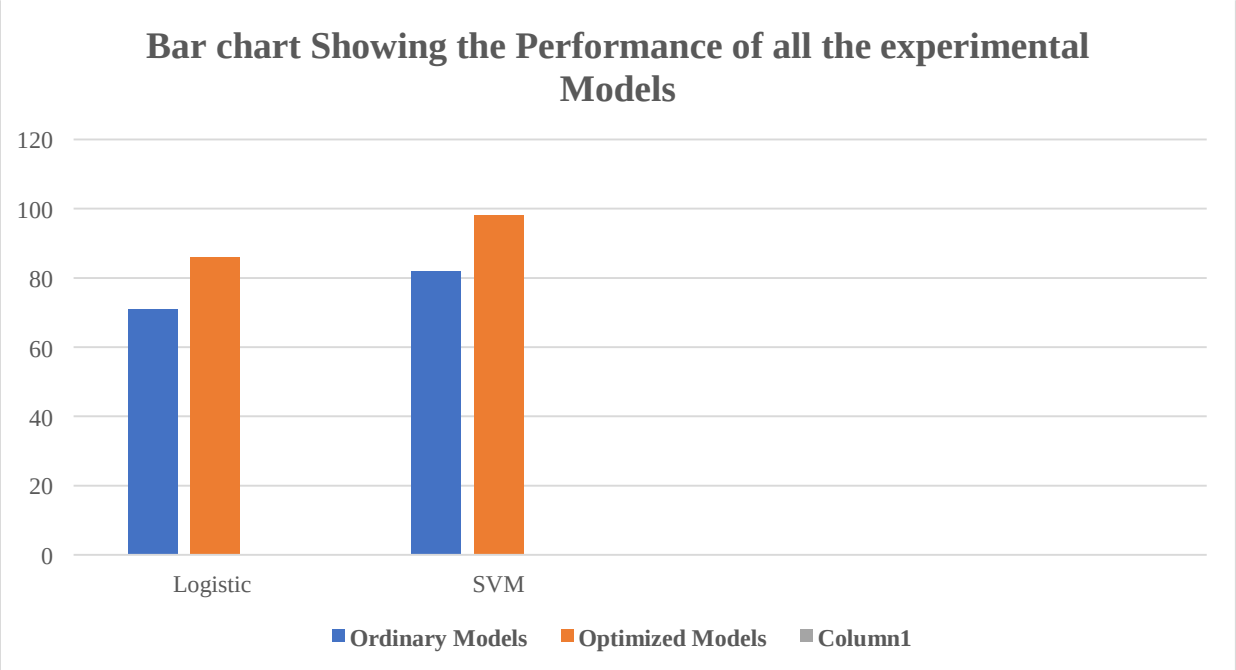

Bar chart Showing the Performance of all the experimental Models
120
100
80
60
40
20
0
Logistic
SVM
Ordinary Models
Optimized Models
Column1

Fig.4.14: A bar chart showing performances of the models

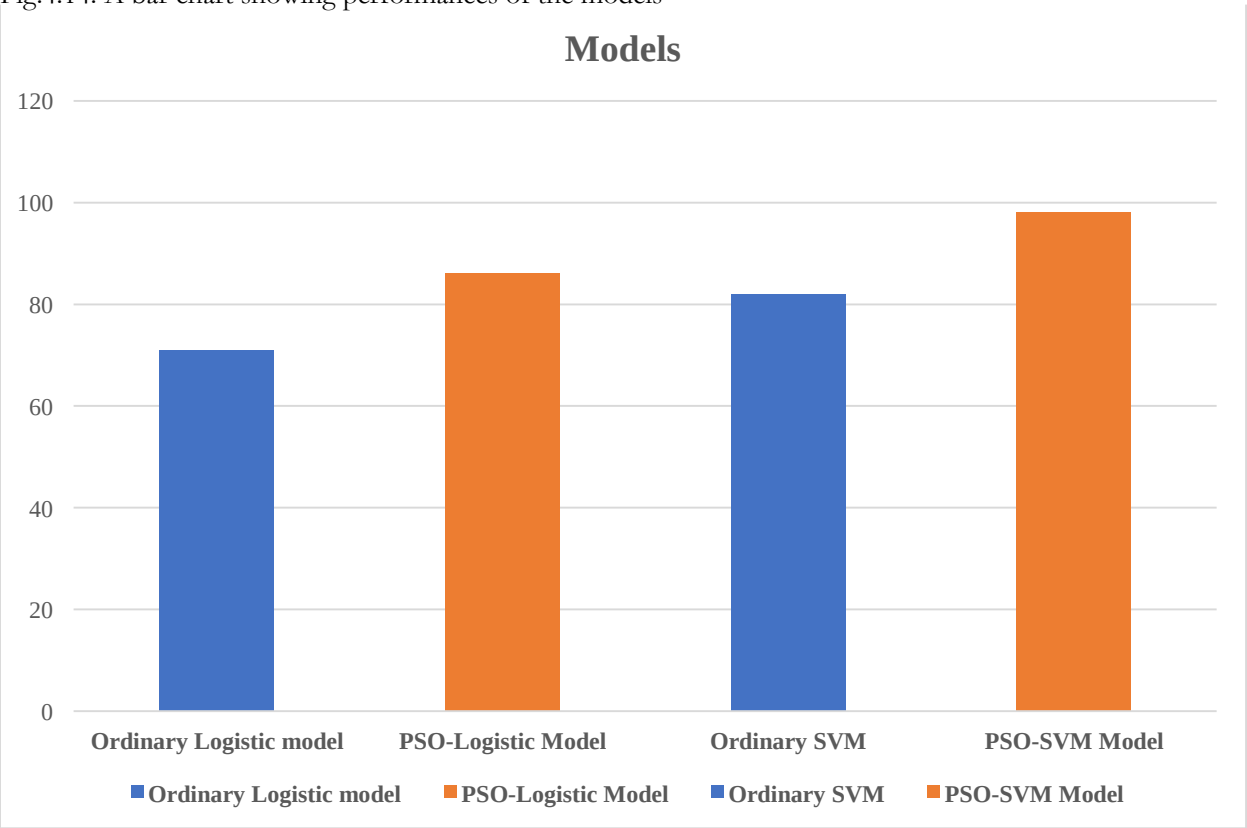


Fig 4.15: Bar chart showing all model the performances based on accuracy separately

International Journal of Biometrics, Image Processing and Computing Research
ISSN: 2997-688X |
Volume 11 Issue 1, January-March, 2023
Journal Homepage: https://ethanpub.online/Journals/index.php/E23
Official Journal of Ethan Publication

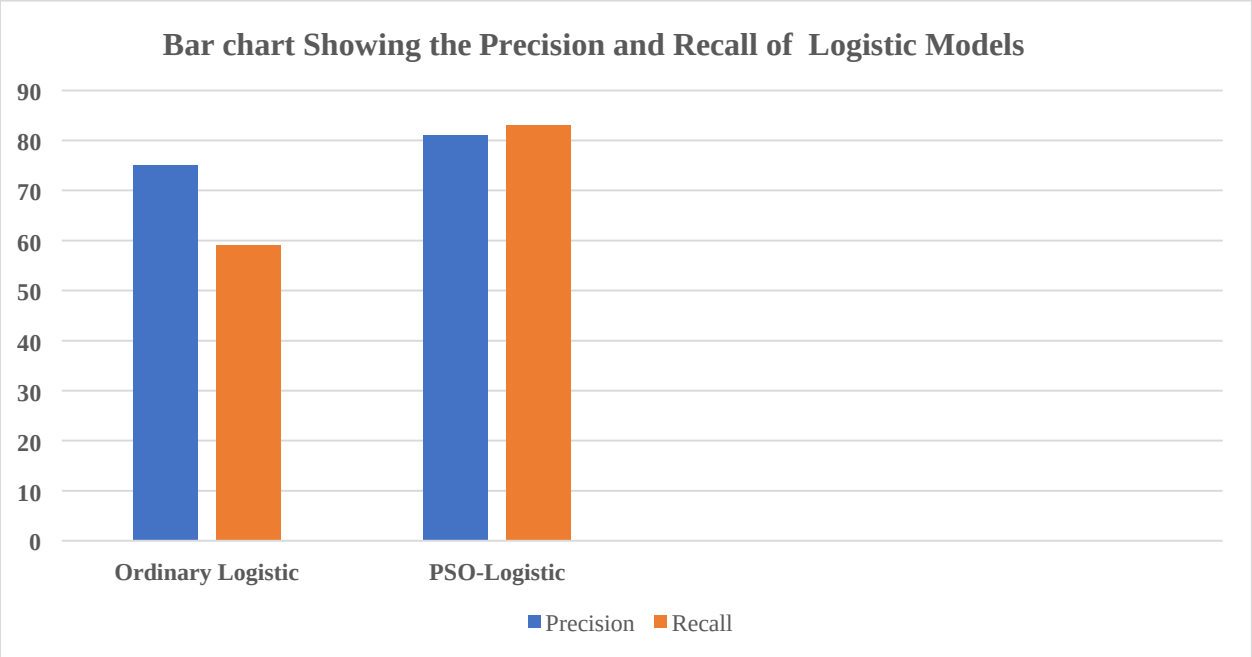
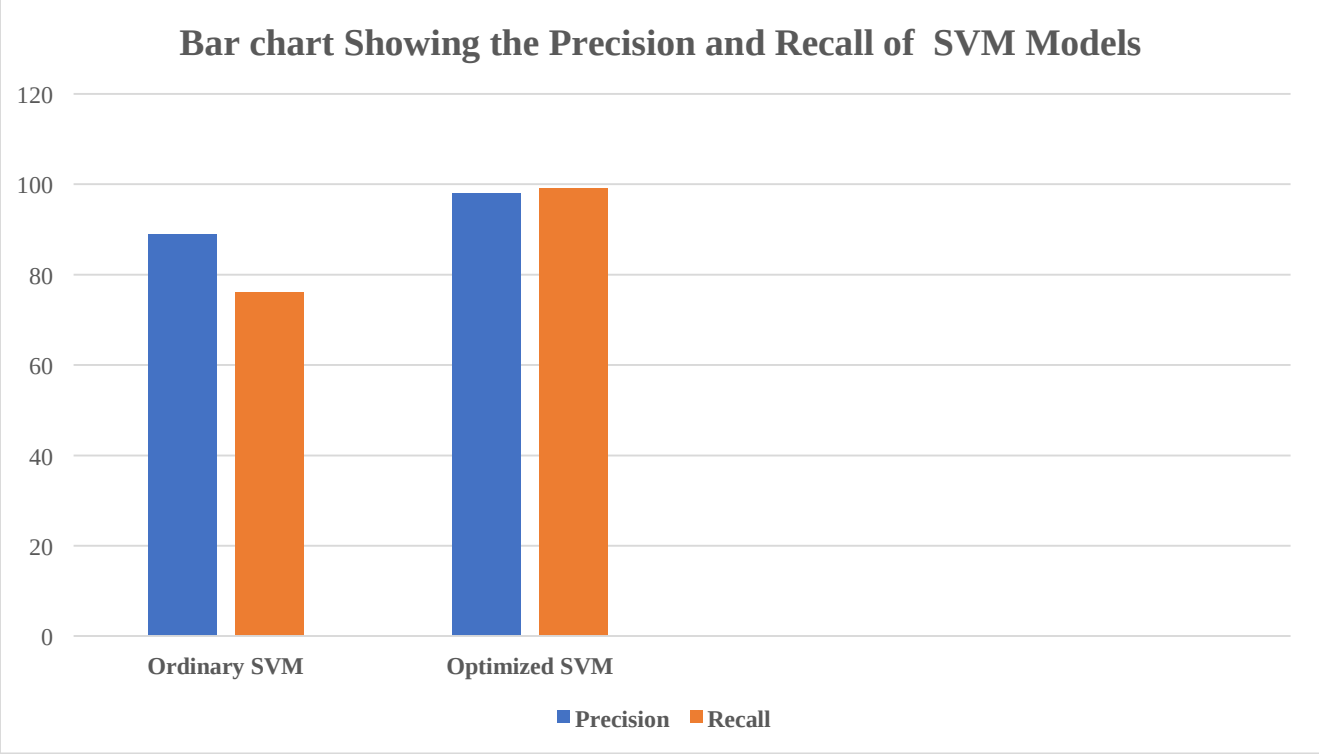


Fig. 4.16:Bar chart Showing the Precision and Recall of Logistic Models



International Journal of Biometrics, Image Processing and Computing Research
ISSN: 2997-688X |
Volume 11 Issue 1, January-March, 2023
Journal Homepage: https://ethanpub.online/Journals/index.php/E23
Official Journal of Ethan Publication

Fig.4.17:Bar chart Showing the Precision and Recall of SVM Models

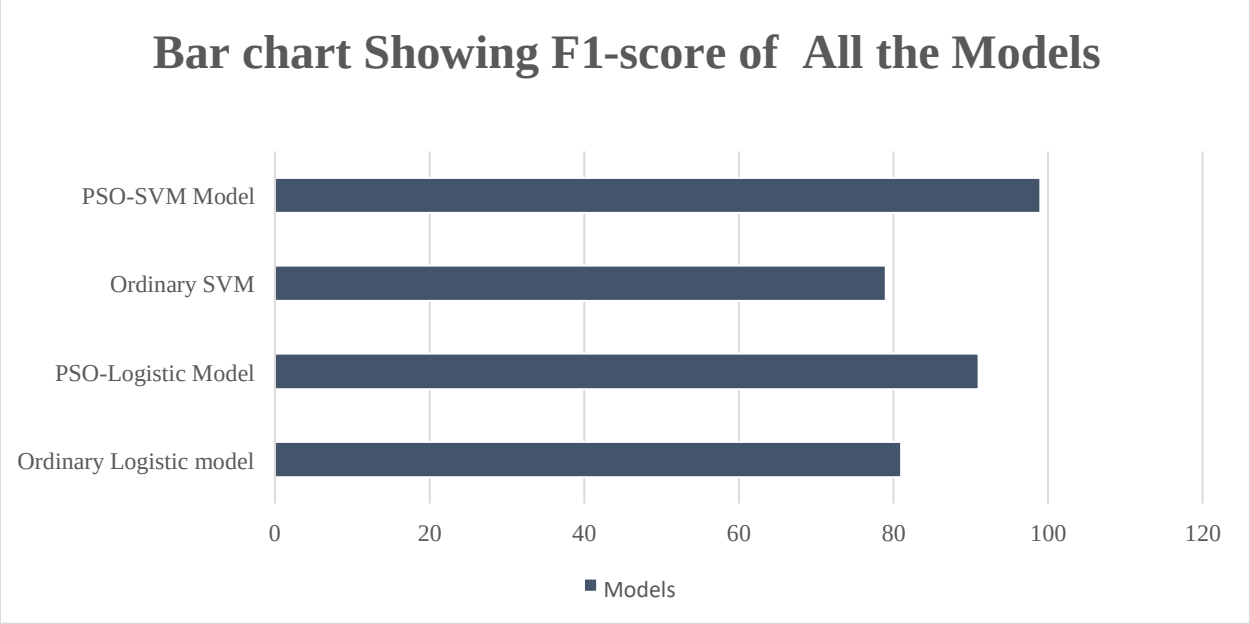


Fig.4.18:Bar chart Showing F1-score of all the Models

4.5 Discussion

From chapter 3 the traditional or ordinary models were built and implemented using python programing language alongside the PSO optimized version of them. In chapter 4, we evaluated the performance of each model evaluating its pros and cons in an effort to identify the optimal model or the model that will perform the best.

It was shown from the charts and tables above the performances of all the models based on accuracy, F1-score, precision and recall. At first ordinary Logistic was implemented and achieved an accuracy score of 76.38%, later on ordinary SVM was also implemented on the same dataset and machine setup and was able to achieve an accuracy score of 82.03% which shows a slight improvement on the logistic model.

The models were later Optimized with Particle Swarm Optimization method and re-implemented on the same dataset and machine in an effort to boost the performance of both and beat the benchmark accuracy, they were re-evaluated again using the same metric measures as was before but this time with different scores showing the effect of the optimization. Optimized Logistic model been the first to be optimized and implemented came up with an accuracy of 97% beating the previous accuracy of non-optimized model of 76.38%. The model has also beat the accuracy of non-optimized SVM model of 82.03% to 98.67% and both the optimized LRand optimized SVM model performance was commendable and promising. SVM was also optimized using PSO technique and was implemented on the diabetes dataset for classification, the new PSO-SVM model performance was also evaluated and from the results of the evaluation the model achieved an accuracy score of 98.67% approximately 99% in testing despite the fact that all the models achieved 100% accuracy on training. PSO-SVM outperformed the traditional SVM, outperformed the traditional and optimized logistic model significantly with a huge gap, it did not just outperforming the traditional models but also beat the accuracy of many related works.

The models were also compared with other models from different experiment and obtained an accuracy far above that which other researchers have obtained in similar studies such as the work of Tarun, J, and Pawan, K.,M,.(2014) and many other related works. The ability of our model to obtain accuracy higher than that of many similar works shows that the models accuracy have been improved and the entire study was a success in the sense that the intended objectives have been achieved.

It can be concluded that the best algorithm or model from our set of experiments was PSO_SVM in terms of all the metrics as it outperforms better all other models, it was very robust due to its high precision score and very accurate due to higher accuracy score, from its confusion matrix it was able to classify all the instances into their appropriate category with very few wrong classified instances, PSO-SVM is the best optimized model and a better one compared to Logistic regression on binary classification task on both higher and low performance machines.

The research work achieved its intended objectives of optimizing SVM and LR algorithms in diabetes prediction, the algorithms were implemented and tested using python programming language and Pima Indian diabetes datasets. It also provides answers to the entire research question mentioned in chapter 1, such as: the model shows that it can be optimized; it shows that the performances of the model do increased when optimized, and also serve the purpose of the research of achieving the state of the art accuracy and beating the benchmark paper accuracy.

5. CONCLUSIONS AND FUTURE SCOPE

In conclusion this paper tries to show the application of machine learning algorithms in chronic medical diagnosis specifically in diabetes prediction, it also extends the performance of the models used in this prediction by optimizing the algorithms with particle swarm optimization technique to achieve an optimal state of the art performance that is higher in terms of accuracy with fewer or no prediction errors. The research indicate that the algorithms can be optimized and when optimized the performance really improved significantly, and finally the research contributes to the paradigm of knowledge by building and implementing an optimized machine learning model of diabetes prediction system that is better than the current available systems by achieving accuracy of 98.67% above the benchmark accuracy of 93.66% in diabetes prediction system.At the end it was discovered both algorithms can be optimized using PSO to improve performance and shows that PSO-SVM (98.67%) was able to achieve an accuracy that was higher than that of PSO-LR (97%).

It is recommended that the future research should extend the scope of this research by experimenting the optimized models in different scenarios such as resources constrained machine, embedded system and so on, to check if it will affect the performance of the model.

The metrics which were used in this research were only able to provide information about the performance of the model with putting into consideration the complexities of the two model, it is recommended that the future research should investigate the time and space complexities of the model, to identify which model consume lesser resource and is fit for a certain situation.

It is also recommended that in future the optimized models should be tested on different datasets containing large and small number of instances and re-evaluated as it was discovered in other researches because some model tends to be good on small datasets, while others performed excellent on large datasets and poorly when the data is inadequate.

It is also recommended that in future work the ordinary model should be optimized with other optimization techniques such as Genetic Algorithm, Artificial Neural Network to investigate which

optimization technique should optimize the models best.The scope of this research covers only optimization of the Support Vector Machine (SVM) parameters and Logistic regression (LR) algorithms using Particle Swamp Optimization (PSO) in diabetes prediction. The scope covers only one dataset of diabetes diagnosis obtained from PIMA Indian diabetes datasets for data science. The research is limited to the above mentioned scope as any other algorithm or dataset beyond the ones identified above is simply beyond the scope of this research and can be used to conduct another research that will extend the limit of this one.

REFERENCES

https://www.healthline.com/health/top-10-deadliest-diseases Updated on Sep 9, 2022 Kumar.Pima-Indians-Diabetes.csv.Kaggle.2018.Availableonline: https://www.kaggle.com/kumargh/pimaindiansdiabetescsv (accessed on 18 June 2021).

Chaves, L.; Marques, G. Data Mining Techniques for Early Diagnosis of Diabetes: A Comparative Study. Appl. Sci. **2021**, 11, 2218.

Aljulifi, M.Z. Prevalence and reasons of increased type 2 diabetes in Gulf Cooperation Council Countries.Saudi Med. J. **2021**, 42,481–490.

Performance for Diabetes with Linear Discriminant Analysis and Genetic Algorithm. Available online: https://ieeexplore.ieee.org/stamp/stamp.jsp?tp=&arnumber=9637039 (accessed on 25 January 2022).

https://www.idf.org/news/240:diabetes-now-affects-one-in-10-adults-worldwide.html
https://www.cbsnews.com/amp/news/coronavirus-vaccine-san-diego-lab-inovio-pharmaceuticalsdiscovered-drug-testing/

Jakka, A., Vakula R. J., (2019) *Performance Evaluation of Machine Learning Models for Diabetes Prediction*, International Journal of Innovative Technology and Exploring Engineering (IJITEE) ISSN: 2278-3075, Vol. 8 (11), 19761980.

Akanksha_R and Pratyaksha S.,(2018) Machine Learning Definition and classification available online at https://www.geeksforgeeks.org/supervised-unsupervised-learning/ last accessed 14/12/2019.

Gabriell. O., (2018) *Making Predictions Using Logistic Regression — Machine Learning,* published at medium.com, available at: https://medium.com/analytics-vidhya/predict-population-growth-using-linear-regression-machinelearning-d555b1ff8f38.

Iyer A., Jeyalatha, S., Sumbaly, R.,(2015) *Diagnosis of diabetes using classification mining techniques*. Int J Data Min KnowlManag Process (IJDKP) 2015; 5(1).

Mahsa A.R.., Seyed, T., A., Seyed, A., N., (2019) *A bi-objective hybrid optimization algorithm to reduce noise and data dimension in diabetes diagnosis using support vector machines* Elsevier: Expert Systems with Applications 127 (2019) 47–57

Ming, Y. C. and Thi T. H., (2017) *Feature Selection and Parameters Optimization of SVM Using Particle Swarm Optimization for Fault Classification in Power Distribution Systems,* Hindawi Computational Intelligence and Neuroscience Volume 2017.

Quan, Zhou *et al,* (2018*) Predicting Diabetes Mellitus with Machine learning Techniques,* Frontiers in Genetics, Vol. 9 (515).

Suyash, S. ,Lokesh S., V. , and . Ajai, K., (2019). *Prediction of Diabetes using Artificial Neural Network approach*, Center for Development of Advanced Computing (C-DAC) Pune, India.

Swapna, S. Kip, R. Vinayakumar, G., (2018) *Automated detection of diabetes using CNN and CNN-LSTM network and heart rate signals*, ProcediaComput. Sci. 132 1253–1262.

Shen, Y. Wei, L.C. Zeng, A. and Chen, J. "Particle swarm optimization with double learning patterns," *Computational Intelligence and Neuroscience*, vol. 2016, 19 pages, 2016.

Tarun J. and Pawan K.(2014) Analysis and prediction of diabetes mellitus using PCA, REP and SVM, *International Journal of Engineering and Technical Research* (IJETR) ISSN: 2321-0869, Volume-2, Issue-8

Zhu C., *et al* (2019), *improved logistic regression model for diabetes prediction by integrating PCA and K-means techniques* Elsevier, April 2019.

Akanksha_Rand Pratyaksha S.,(2018) Machine Learning Definition and classification available online at https://www.geeksforgeeks.org/supervised-unsupervised-learning/ last accessed 14/12/2019..

Humar and Novruz M.,(2017), *Design of a hybrid system for the diabetes and heart diseases*, Department of Electronic and Computer Education, Selcuk University, Konya, Turkey.

Jakka, A., Vakula R. J., (2019) *Performance Evaluation of Machine Learning Models for Diabetes Prediction*, International Journal of Innovative Technology and Exploring Engineering (IJITEE) ISSN: 2278-3075, Vol. 8 (11), 19761980.

Ming, Y. C. and Thi T. H., (2017) *Feature Selection and Parameters Optimization of SVM Using Particle Swarm Optimization for Fault Classification in Power Distribution Systems,* Hindawi Computational Intelligence and Neuroscience Volume 2017.

Swapna, S. Kip, R. Vinayakumar, G., (2018) *Automated detection of diabetes using CNN and CNN-LSTM network and heart rate signals*, ProcediaComput. Sci. 132 1253–1262.

Shen, Y. Wei, L.C. Zeng, A. and Chen, J. "Particle swarm optimization with double learning patterns," *Computational Intelligenceand Neuroscience*, vol. 2016, 19 pages, 2016. retrievedhttp://www.who.int/news-room/fact-sheets/detail/the-top-10-causes-ofdeath, Accessed date: 27 July 2018.

http://www.who.int/news-room/fact-sheets/detail/diabetes retrieved 27/07/2018. https://www.diabetesdaily.com/learn-about-diabetes/what-is-diabetes/how-manypeople-have-diabetes/.

Tarun Jhaldiyal, Pawan Kumar Mishra Analysis and prediction of diabetes mellitus using PCA, REP and SVM 2014 Int J Eng Tech Res (IJETR) ISSN: 2321-0869,Volume-2, Issue-8.

Khandegar Anjali. KhushbuPawar diagnosis of diabetes mellitus using PCA, neural Network and cultural algorithm.Int J Digital ApplContemp Res 2017;5(6).

Wu Han, Yang Shengqi, Huang Zhangqin, He Jian. Xiaoyi Wang Type 2 diabetes mellitus prediction model based on data mining.Inf Med. 2018;10:100–7.Unlocked.

Orabi," Early Predictive System for Diabetes Mellitus Disease," , in Industrial Conference on DataMining, 2016, Springer.pp.420–427.

Perveen S., "Performance Analysis of Data Mining Classification Techniques to Predict Diabetes," in Proceedings Computer Science, 2016, vol. 82, pp 115–121.

PradeepKhanhasamy J., "Performance Analysis of classifier models to predict diabetes Mellitus," in Elsevier journal, 2016, vol. 82, pp 115 – 121.

DeeptiSisodiaa, Dilip Singh Sisodiab, 2018, Prediction of Diabetes using Classification Algorithms, International Conference on Computational Intelligence and Data Science - ICCIDS 2018 ,Science Direct Procedia Computer Science 132 (2018) 1578–1585.

Pima Indians Diabetes dataset (PIDD) from the UCI repository

Zhe Wei, Guangjian Ye and Nengcai Wang.Analysis for risk factors of type 2 diabetes mellitus based on FP-growth algorithm.China Medical Equipment, 2016. 13(5):45-48.

YiruiGuo. Application of artificial neural network to predict individual risk of type 2 diabetes mellitus. Journal of Zhengzhou University, 2014.49(3):180-183.

Shuaishuai Li, Enke Zhang, Min Li and Wei Pan, Research on the Effectiveness of Application of Diabetes Management APP, China Medical Devices, 2015. Vol 30.No.08.

OmprakashChandrakar, Dr. Jatinderkumar R. Saini. Development of Indian Weighted Diabetic Risk Score (IWDRS) using Machine Learning Techniques for Type-2 Diabetes, ACM COMPUTE '16, October 21-23, 2016, Gandhinagar, India.

Huan Li, Qi Zhang and Kejie Lu, Integrating Mobile Sensing and Social Network For Personalized Health-Care Application, Health care information systems(2016).

Yan Luo, Charles Ling, Jody Schuurman and Robert Petrella, GlucoGuide: An Intelligent Type-2 Diabetes Solution Using Data Mining and Mobile Computing, 2014 IEEE International Conference on Data Mining Workshop.

Jalan, S.; Tayade, A.A. Review paper on Diagnosis of Diabetic Retinopathy using KNN and SVM Algorithms.*Int. J. Adv. Res.Comput. Sci. Manag. Stud.* **2015**, *3*, 128–131.

World Health Organization.Noncommunicable Diseases (NCD) Country Profiles. 2018. Available online: https://www.who.int/nmh/countries/omn_en.pdf (accessed on 15 November 2021).

Wei, S.; Zhao, X.; Miao, C.A comprehensive exploration to the machine learning techniques for diabetes identification. In Proceedings of the 2018 IEEE 4th World Forum on Internet of Things (WF-IoT), Singapore, 5–8 February 2018.

Anwar, F.; Qurat-Ul-Ain, F.A.; Ejaz, M.Y.; Mosavi, A.A comparative analysis on diagnosis of diabetes mellitus using different approaches—A survey. Inform. Med. Unlocked **2020**, 21, 100482.

Swapna, G.; Vinayakumar, R.; Soman, K.P. Diabetes detection using deep learning algorithms. ICT Express **2018**, 4, 243–246.

Gra¸dalski, T.; Hołon´ , A. Diabetes mellitus in the last weeks of life—Case study and current literature review. Med. Paliatywna**2019**, 11, 67–72.

Perveen, S.; Shahbaz, M.; Guergachi, A.; Keshavjee, K. Performance analysis of data mining classification techniques to predict diabetes.ProcediaComput. Sci. **2016**, 82, 115–121.

Khan, N.S.; Muaz, M.H.; Kabir, A.; Islam, M.N.A machine learning-based intelligent system for predicting diabetes.Int. J. Big Data Anal.Healthc.**2019**, 4, 20

Barik, S.; Mohanty, S.; Mohanty, S.; Singh, D. Analysis of Prediction Accuracy of Diabetes Using Classifier and Hybrid Machine Learning Techniques. In Intelligent and Cloud Computing. Smart Innovation, Systems and Technologies; Mishra, D., Buyya, R.,Mohapatra, P., Patnaik, S., Eds.; Springer: Singapore, 2021; Volume 153, pp. 399–409.

Zheng, T.; Xie,W.; Xu, L.L.; He, X.Y.; Zhang, Y.; You, M.R.; Yang, G.; Chen, Y. A machine learning-based framework to identify type 2 diabetes through electronic health records. Int. J. Med. Inform. **2017**, 97, 120– 127

Lekha, S.; Suchetha, M. Real-Time Non-Invasive Detection and Classification of Diabetes Using Modified Convolution Neural Network. IEEE J. Biomed. Health Inform. **2018**, 22, 1630–1636.

Yuvaraj, N.; Sri Preethaa, K.R. Diabetes prediction in healthcare systems using machine learning algorithms on Hadoopcluster.Clust.Comput.**2019**, 22, 1–9.

Sisodia, D.; Sisodia, D.S. Prediction of Diabetes using Classification Algorithms.ProcediaComput. Sci. **2018**, 132, 1578–1585.

Mercaldo, F.; Nardone, V.; Santone, A. Diabetes Mellitus Affected Patients Classification and Diagnosis through Machine Learning Techniques. ProcediaComput. Sci. **2017**, 112, 2519–2528.

Maniruzzaman, M.; Rahman, M.J.; Ahammed, B.; Abedin, M.M. Classification and prediction of diabetes disease using machine learning paradigm.Health Inf. Sci. Syst. **2020**, 8, 7.

Find Missing Values—MATLAB. 2022. Available online: https://www.mathworks.com/help/matlab/ref/ismissing.html?s_tid=doc_ta. (accessed on 29 July 2021).

Detect and Replace Outliers in Data—MATLAB. Available online: https://www.mathworks.com/help/matlab/ref/filloutliers.html?s_tid=doc_ta (accessed on 30 August 2022).

Jain, D.; Singh, V. Feature selection and classification systems for chronic disease prediction: A review.Egypt. Inform. J. **2018**, 19, 179–189.

Mishra, S.; Mishra, B.K.; Sahoo, S.; Panda, B. Impact of swarm intelligence techniques in diabetes disease risk prediction. Int. J. Knowl. Discov.Bioinform. (IJKDB) **2016**, 6, 29–43.

Keerthi, P.; HemaLatha, N.; HariGokul, K.; Prasad, V.; Arun Kumar, T. Wrapper Based Feature Selection for Disease Diagnosis using Optimization Algorithms. Int. J. Eng. Res. Technol. **2018**, 6, 1–10.