iJRASET
International Journal For Research in Applied Science and Engineering Technology

INTERNATIONAL JOURNAL FOR RESEARCH

IN APPLIED SCIENCE & ENGINEERING TECHNOLOGY

Volume: 10 **Issue:** V **Month of publication:** May 2022

DOI: https://doi.org/10.22214/ijraset.2022.41968

www.ijraset.com

Call: 08813907089 | E-mail ID: ijraset@gmail.com

International Journal for Research in Applied Science & Engineering Technology (IJRASET)
ISSN: 2321-9653; IC Value: 45.98; SJ Impact Factor: 7.538
Volume 10 Issue V May 2022- Available at www.ijraset.com

Prediction of Diabetes Using Logistics Regression Algorithms with Flask

Sriram M[1], Nithish J[2], Harshavardhan B[3], Mrs. Steffina Muthukumar[4]
[1, 2, 3, 4]*Computer Science Department, SRM University*

Abstract: ***The diabetes is one of deadly diseases on the planet. There are many kinds of diseases which has different assortments of issues for instance: coronary disappointment, visual deficiency, urinary organ infections and so forth. In such case the patient is expected to visit an analytic focus, to get their reports after meeting. Because of each time they need to contribute their time and cash. Yet, with the development of Machine Learning strategies we have the adaptability to look out a solution to the recent concern the main necessity of Artificial intelligence is data The collected dataset is used to build the machine learning model The necessary pre- processing techniques are applied like univariate analysis and bivariate analysis are implemented. The data is visualized for better understanding of the features and based on that a classification model is built by utilizing AI calculation The point of this investigation is to foster a framework which could anticipate the diabetic gamble level of a patient with a superior precision***

I. INTRODUCTION

Diabetes is various sicknesses that include issues with the chemical insulin. Typically, the pancreas (an organ behind the stomach) discharges insulin to help your body store and utilize the sugar and fat from the food you eat Diabetes is a general diseases that is there all over the world. The fact that they have it makes generally 18.2 million Americans have the infection and close to 33% (or around 5.2 million) uninformed. An extra 41 million individuals have pre-diabetes. At this point, there is no fix. Individuals with diabetes need to deal with their illness to remain beneficial to comprehend the reason why insulin is significant in diabetes, it assists with find out about how the body involves nourishment for energy. Your body is comprised of millions of cells. To make energy, these cells need food in an extremely straightforward structure. Whenever you eat or drink, quite a bit of your food is separated into a basic sugar called "glucose." Then, glucose is shipped through the circulatory system to the phones of your body where it very well may be utilized to give a portion of the energy your body needs for day to day exercises This dataset is initially from the National Institute of Diabetes and Digestive and Kidney Diseases. The goal is to anticipate in view of indicative estimations whether a patient has diabetes. Several requirements were put on the choice of these occurrences from a bigger data set. Specifically, all patients here are females no less than 21 years of age of Pima Indian heritage. Pregnancies: Number of times pregnant, Glucose: Plasma glucose fixation a 2 hours in an oral glucose resilience test, Blood Pressure: Diastolic pulse (mm Hg), Skin Thickness: Triceps skin overlay thickness (mm),Insulin: 2-Hour serum insulin (mu U/ml),BMI: Body mass file (weight in kg/(level in m)^2), Diabetes Pedigree Function: Diabetes family function, Age: Age (years),Outcome: Class variable (0 or 1),Original proprietors: National Institute of Diabetes and Digestive and Kidney Diseases.

II. REVIEW OF LITERATURE SURVEY

1) A predict a predictive model for diabetes using machine learning techniques (a case study of some selected hospitals in kaduna metropolis), abdulkadir,2021,diabetes mellitus (dm) which refers to a metabolic disorder that occurs when the level of blood sugar in the body is considered high, which could be a resulting effect of inadequate availability of insulin in the body. It is a chronic disease which may lead to myriads of complications in the body system. Statistics by the World Health Organization (WHO) in 2013, indicated that DM was the cause of death of over 1.5 million people around the world and in 2016, 8.5% of adults within age seventeen (17) and above were reported to be diabetic and diabetic patients have continued to increase in recent years. It is therefore very glaring that these alarming figures calls for very urgent and effective attention. The research was based on the prevalence of diabetes amongst the masses of Kaduna metropolis using some selected hospitals as a case study after which a predictive model was designed for diabetes, using some selected supervised learning algorithms like Decision tree algorithm, K- Nearest Neighbour algorithm and Artificial Neural Networks on a dataset gotten from 44 Army Reference Hospital and Yusuf Danstoho Memorial Hospital Kaduna which constitutes of nine (9) attributes that was considered. The results indicated that ANN produced the highest accuracy with 97.40% followed by decision tree algorithm with 96.10% accuracy then K-NN algorithm with 88.31%

2) Detection and Prediction of Diabetes Using Machine Learning Techniques Priyanka Indoria· Yogesh Kumar Rathore, March- 2018 Diabetes is a one of the leading cause of blindness, kidney failure, amputations, heart failure and stroke. When we eat, ourbody turns food into sugars, or glucose. At that point, our pancreas is supposed to release insulin.

3) Insulin serves as a "key" to open our cells, to allow the glucose to enter -- and allow us to use the glucose for energy. But withdiabetes, this system does not work. Several major things can go wrong – causing the onset of diabetes. Type 1 and type 2 diabetes are the most common forms of the disease, but there are also other kinds, such as gestational diabetes, which occurs during pregnancy, as well as other forms. This paper focuses on recent developments in machine learning which have made significant impacts in the detection and diagnosis of diabetes

4) Prediction of Diabetics using Machine Learning G. Geetha, K.Mohana Prasad Year : May 2019 Around 50.9 Million People in India suffer from diabetics and Tamil Nadu stands second in the list of Indian states. The main objective of this paper is to develop prediction modeling of the given medical data of patients with and without diabetics. Through this paper, we aim to create hybrid models that can be easily used by doctors to treat patients with diabetics. Naïve Bayes and Random forest algorithms are used to predict whether a person having diabetics or not, by keeping his health conditions in mind. Thus this process enables doctors to easily group, classify and categorize the disease type accordingly treatment can be given tothem.. The Random Forest algorithm is used here in order to perform feature selection. It takes n inputs from the dataset and builds numerous uncorrelated decision trees during the time of training. It then displays the class that is the mode of all of the class outputs by individual Trees.

5) Effective Prediction of Diabetes Mellitus using Nine different Machine Learning Techniques and their Performances Sarvesh Vishwakarma·Anupam Agrawal, May 2017,Diabetes is a disease where the predominant finding is high blood sugar. The high blood sugar may either be because of deficient insulin production (Type 1) or insulin resistance in peripheral tissue cells (Type 2). Many problems occur if diabetes remains untreated and unidentified. It is additional inventor of various varieties of disorders for example: coronary failure, blindness, urinary organ diseases etc. Nine different machine learning techniques are used in this research work for prediction of diabetes. A dataset of diabetic patient's is taken and nine different machine learning techniques are applied on the dataset. Positive likelihood ratio, Negative likelihood ratio, Positive predictive value, Negative predictive value, Disease prevalence, Specificity, Precision, Recall, F1-Score ,True positive rate, False positive rate of the applied algorithms is discussed and compared. Diabetes is growing at an increasing in the world and it requires continuous monitoring. To check this we use Logical regression, Random forest, Logical regression CV, Support Vector Machine, Artificial Neural Network (ANN), Decision Tree, k-nearest neighbors (KNN), XGB classifier.

6) Literature Survey On Different Techniques Used For Predicting Diabetes Mellitus,Shashank Joshi, Vijayendra Gaikwad, Sairam Rathod,June 2020,Diabetes is a disease where the predominant finding is high blood sugar. The high blood sugar may either be because of deficient insulin production (Type 1) or insulin resistance in peripheral tissue cells (Type 2). Many problems occur if diabetes remains untreated and unidentified. It is additional inventor of various varieties of disorders for example: coronary failure, blindness, urinary organ diseases etc. Nine different machine learning techniques are used in this research work for prediction of diabetes. Positive likelihood ratio, Negative likelihood ratio, Positive predictive value, Negative predictive value, Disease prevalence, Specificity, Precision, Recall, F1-Score ,True positive rate, False positive rate of the applied algorithms is discussed and compared. Diabetes is growing at an increasing in the world and it requires continuous monitoring. To check this we use Logical regression, Random forest, Logical regression CV, Support Vector Machine, Artificial Neural Network (ANN), Decision Tree, k-nearest neighbors (KNN), XGB classifier.

Number of Features	Features	Descriptions and Features values
1	Number of times a person was pregnant	Numeric value
2	Glucose Concentration	Numeric value
3	Blood Pressure	Numeric value (in mm Hg)
4	Skin Thickness	Numeric value (in mm)
5	Insulin	Numeric value
6	Body Mass Index (BMI)	Numeric value (weight in $kg/(height\ in\ m)^2$)
7	Diabetes Pedigree Function	Numeric value
8	Age	Numeric value
9	Value of Diabetes Diseases	Yes = True No = False

Fig:1 Features of kaggle dataset for Diagnosing Diabetics Disease

III. METHODOLOGY

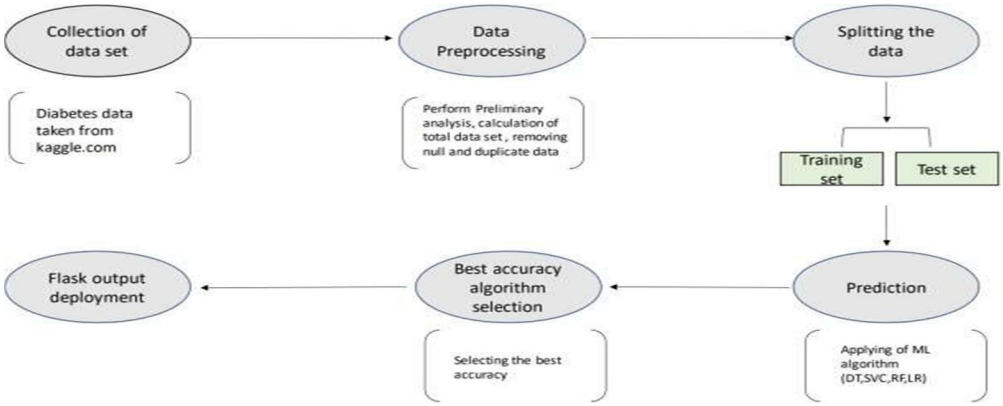


Fig:2 Architecture diagram

Number of Attributes	Attributes Name	Mean	Standard Deviation
1	Number of times a person was pregnant	3.8	3.4
2	Glucose Concentration	120.9	32.0
3	*Blood Pressure*	69.1	19.4
4	*Skin Thickness*	20.5	16.0
5	*Insulin*	79.8	115.2
6	*Body Mass Index (BMI)*	32.0	7.9
7	*Diabetes Pedigree Function*	0.5	0.3
8	*Age*	33.2	11.8

Fig:3 Dataset variables and their values

A. Procuring the Dataset

The dataset utilized here is the PIMA Indian Dataset. It is the information gotten from the National Institute for Diabetics. It comprises of a few clinical indicator factors and one objective variable. The different clinical factors are BMI, Glucose levels, Blood Pressure and so on. It contains 768 lines and 9 sections. The dataset document is in a .csv(Comma Separated Values) design.Utilizing the assistance of Python's inbuilt library Pandas, which is an information outline library, we import the document into our Python climate. Different libraries that are brought into the climate are: Numpy-a library that is utilized to work with huge layered exhibits and lattices, giving undeniable level numerical functionalities to primarily chip away at information. Matplotlib-the library that gives Python the usefulness of plotting diagrams and plots. It works pair with NumPy. Pandas have a capacity named read_csv(), which basically peruses a record of the configuration (.csv). Once the dataset is stacked into the climate, we can really look at the components of the dataset by the capacity .shape() which returns the quantity of lines and segments. The essential query of the information is done, by utilizing the inbuilt orders .head() and .tail() which print the quantity of columns from the beginning of the dataset and the bottom of the Dataset correspondingly.

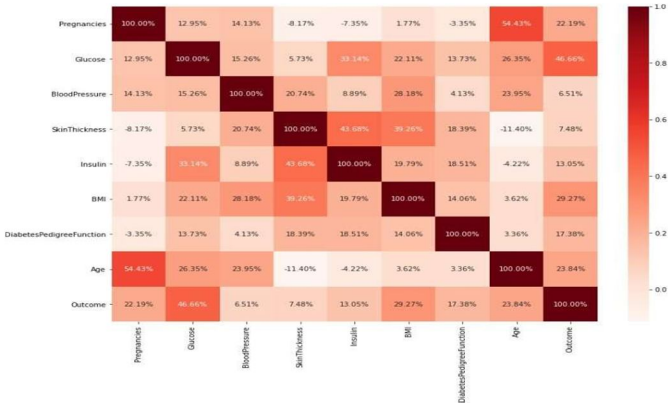


Fig:4 Correlation heatmap

B. Preparation of the Dataset

In the wake of getting the dataset, we check whether we can roll out any improvements to the dataset. Tasks, for example, instatement of the factors, purifying the information, making proper names for the information happens. For our situation, the dataset contains a boundary skin thickness, this segment has a powerless relationship to the commitment of an individual being diabetic. Thus, we eliminate the section for our examination. In this stage, we can compute the numeric parts of the information, for example, the normal of a specific section, number of instances of the segment in light of conditions, etc,The dataset containsthe qualities for individuals having diabetics and individuals who don't. Thus, we determined the count for each case People with Diabetics: 268 People without Diabetics: 500 In the given information, around 35% individuals have been determined to have diabetics.

C. Splitting the Data

Partitioning the dataset into preparing and test information is one of the significant stages in investigation. This interaction is fundamentally done to guarantee test information is not the same as the preparation information since we really want to test the model followed by the preparation cycle. In the first place, the preparation information goes through advancing and afterward, the information which is prepared is summed up on different information, in light of which the forecast is made. The dataset for our situation is parted into various variations and forecast is performed likewise. The dataset has different sections that are clinical indicators and one objective segment, that of the diabetcs result. The clinical indicators are given as contributions to a variable and the objective variable is given as contribution to another variable. Utilizing the inbuilt capacity, train_test_split, the dataset is parted into clusters and is planned to preparing and test subsets. For our situation, we are performing parts of 80/20,70/30,75/25,60/40 and the precision of each is recorded. It was seen that the dataset contain a few invalid qualities, to smooth out the examination and the expectation, the invalid qualities were loaded up with the mean upsides of the separate sections.

D. Logistic Regression

It is a factual technique for examining an informational index where there are at least one autonomous factors that decide a result. The result is estimated with a dichotomous variable (where there are just two potential results). The objective of calculated relapse is to track down the best fitting model to depict the connection between the dichotomous quality of interest (subordinate variable = reaction or result variable) and a bunch of free (indicator or illustrative) factors. Calculated relapse is a Machine Learning order calculation that is utilized to foresee the likelihood of a clear cut subordinate variable. In strategic relapse, the reliant variable is a paired variable that contains information coded as 1 (indeed, achievement, and so forth) or 0 (no, disappointment, and so on.). All in all, the calculated relapse model predicts P(Y=1) as a component of X. Calculated regression. Logistic relapse predicts the result of a clear cut subordinate variable. Hence the result should be a straight out or discrete worth. It very well may be either Yes or No, 0 or 1, valid or False, and so forth however rather than giving the specific worth as 0 and 1, it gives the probabilistic qualities which lie somewhere in the range of 0 and 1. Logistic Regression is much like the Linear Regression with the exception of that how they are utilized. Direct Regression is utilized for tackling Regression issues, though Logistic relapse is utilized for settling the arrangement problems. In Logistic relapse, rather than fitting a relapseline, we fit an "S" molded calculated work, which predicts two most extreme qualities (0 or 1).The bend from the strategic capacity shows the probability of something, for example, regardless of whether the cells are destructive, a mouse is stout or not in view of its weight, etc. Logistic Regression is a huge AI calculation since it has the capacity to give probabilities and arrangenew information utilizing ceaseless and discrete datasets Logistic Regression can be utilized to group the perceptions utilizing various kinds of information and can without much of a stretch decide the best factors utilized for the order. The underneath picture is showing the strategic capacity.

E. Decision Tree

As a rule, Decision tree investigation is a prescient displaying apparatus that can be applied across numerous areas. Choice trees can be developed by an algorithmic methodology that can part the dataset in various ways in view of various circumstances. Choices trees are the most impressive calculations that falls under the classification of managed algorithms They can be utilizedfor both grouping and relapse assignments. The two primary elements of a tree are choice hubs, where the information is partedand leaves, where we came by result.

The case of a parallel tree for foreseeing whether an individual is fit or unsuitable giving different data like age, dietary patterns and exercise habits two sorts of choice trees Classification choice trees − In this sort of choice trees, the choice variable is all out. The above choice tree is an illustration of arrangement choice tree Regression choicetrees − In this sort of choice trees, the choice variable is ceaseless.

F. Random forest Algorithm:

Random Forest is a famous AI calculation that has a place with the administered learning strategy. It very well may be utilized for both Classification and Regression issues in ML. It depends on the idea of gathering realizing, which is a course of joining various classifiers to tackle a complicated issue and work on the exhibition of the model.

Implementation in Scikit-learn

For each decision tree, Scikit-learn calculates a nodes importance using Gini Importance, assuming only two child nodes (binary tree):

$$ni_j = w_j C_j - w_{left(j)} C_{left(j)} - w_{right(j)} C_{right(j)}$$

The importance for each feature on a decision tree is then calculated as:

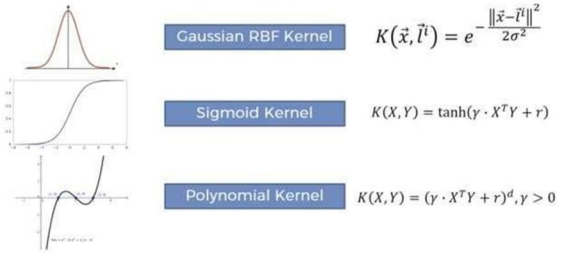
$$fi_i = \frac{\sum_{j:node\ j\ splits\ on\ feature\ i} ni_j}{\sum_{k \in all\ nodes} ni_k}$$

These can then be normalized to a value between 0 and 1 by dividing by the sum of all feature importance values:

G. Support Vector Machines

$$normfi_i = \frac{fi_i}{\sum_{j \in all\ features} fi_j}$$

Given a bunch of preparing models, each set apart as having a place with either of two classifications, a SVM preparing calculation constructs a model that allots new guides to one class or the other, making it a non-probabilistic twofold direct classifier The goal of applying SVMs is to find the best line in two aspects or the best hyperplane in multiple aspects to assist us with isolating our space into classes. The hyperplane (line) is gotten through the greatest time, i.e., the most extreme distance between data of interest of the two classes The vector focuses nearest to the hyperplane are known as the help vector focuses onthe grounds that main these two focuses are adding to the aftereffect of the calculation, and different focuses are not. On the offchance that an information point isn't a help vector, eliminating it significantly affects the model. Then again, erasing the help vectors will then, at that point, change the place of the hyperplane The element of the hyperplane relies on the quantity of highlights. In the event that the quantity of information highlights is 2, the hyperplane is only a line. On the off chance that the quantity of info highlights is 3, the hyperplane turns into a two-layered plane. It becomes challenging to envision when the quantity of highlights surpasses 3 In the SVM calculation, we are hoping to amplify the edge between the pieces of informationand the hyperplane. The misfortune work that amplifies the edge is pivot misfortune.



$$\max \frac{2}{\|\mathbf{w}\|} \rightarrow \max \frac{1}{\|\mathbf{w}\|} \rightarrow \min \|\mathbf{w}\| \rightarrow \min \frac{1}{2} \|\mathbf{w}\|^2$$

$$L(w) = \sum_{i=1} \underbrace{max(0, 1 - y_i[w^T x_i + b])}_{\text{Loss function}} \underbrace{+ \lambda ||w||_2^2}_{\text{regularization}}$$

H. Finding the Accuracy

In the first place, the exactness of the preparation information are checked by taking care of the contentions for the preparation information split. From that point onward, precision of the testing information is finished by the same way with the testing information as the boundaries. By contrasting these two, we can develop a disarray framework. The fundamental goal of disarray grid is to assess the precision of the characterization. By definition a disarray grid C is that, Ci,j addresses the quantity of perceptions known to be in bunch I however anticipated to be in bunch j. Subsequently in twofold characterization, the countof genuine negatives are C0,0, bogus negatives are C1,0, genuine up-sides are C1,1 and misleading up-sides are C0,1.

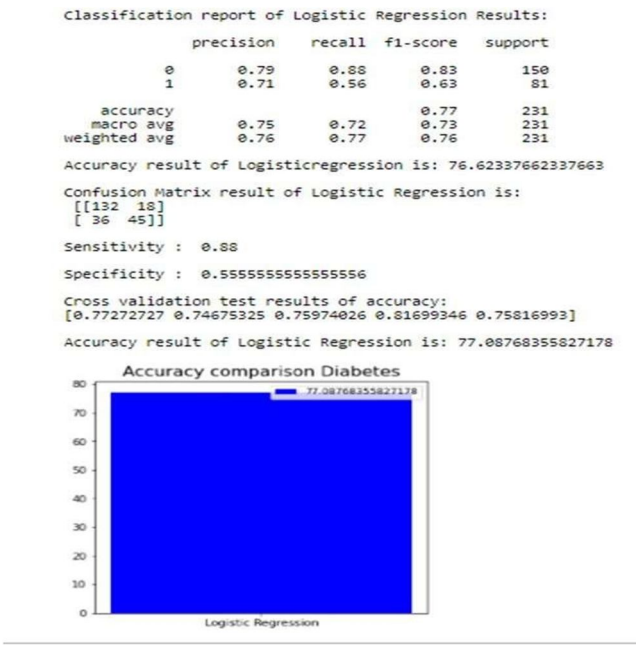


Fig:5 Classification Report for logisticsregression

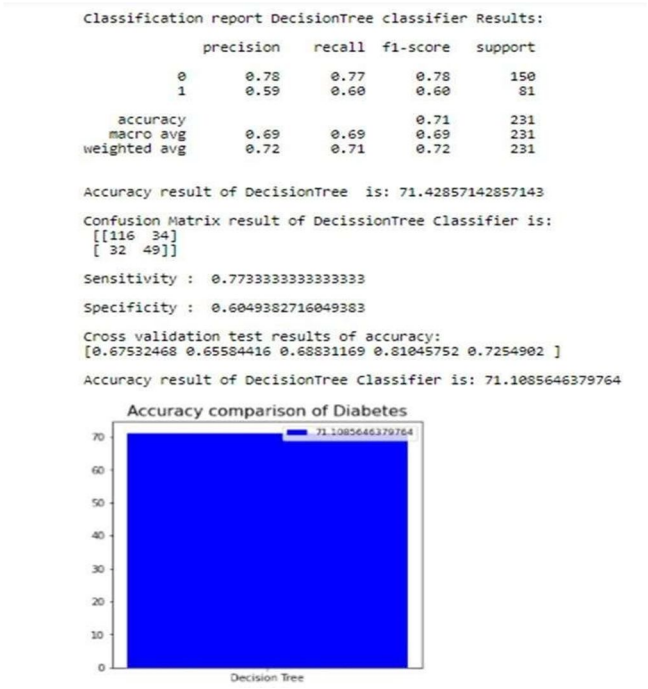


Fig :6 Classification report for Decision tree

```
Classification report of Random Forest Results:

              precision    recall  f1-score   support

           0       0.78      0.85      0.81       150
           1       0.67      0.54      0.60        81

    accuracy                           0.74       231
   macro avg       0.72      0.70      0.71       231
weighted avg       0.74      0.74      0.74       231

Accuracy result of Random Forest is: 74.45887445887446

Confusion Matrix result of Random Forest is:
 [[128  22]
 [ 37  44]]

Sensitivity :  0.8533333333333334

Specificity :  0.5432098765432098

Cross validation test results of accuracy:
[0.74675325 0.73376623 0.77272727 0.83006536 0.77777778]

Accuracy result of Random Forest is: 77.2217978100331
```

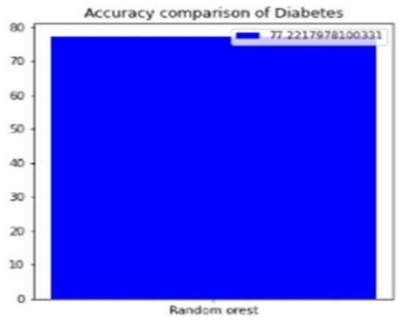


Fig:7 Classification report for random forest

IV. DEPLOYING

A. Flask (Web Framework)

Flask is a miniature web framework written in Python. It is named a miniature web-framework since it doesn't need specific devices or libraries. It has no data set reflection layer, structure approval, or whatever other parts where previous outsider libraries give normal functions. However, Flask upholds augmentations that can add application highlights as though they were executed in Flask itself .Extensions exist for object-social mappers, structure approval, transfer dealing with, different open verification advances and a few normal system related apparatuses.

Fig:8 Flask Framework of diabetics prediction

Recognizing diabetics or predicting the forthcoming of a diabetic life can be moved by utilizing different AI strategies like choice tree, logistic regression, support vector machine, Random Forest, and so on. In this paper, we can infer that the best technique for expectation of diabetics is strategic relapse and we have Deployed in the neighborhood site utilizing Flask Framework

V. CONCLUSION

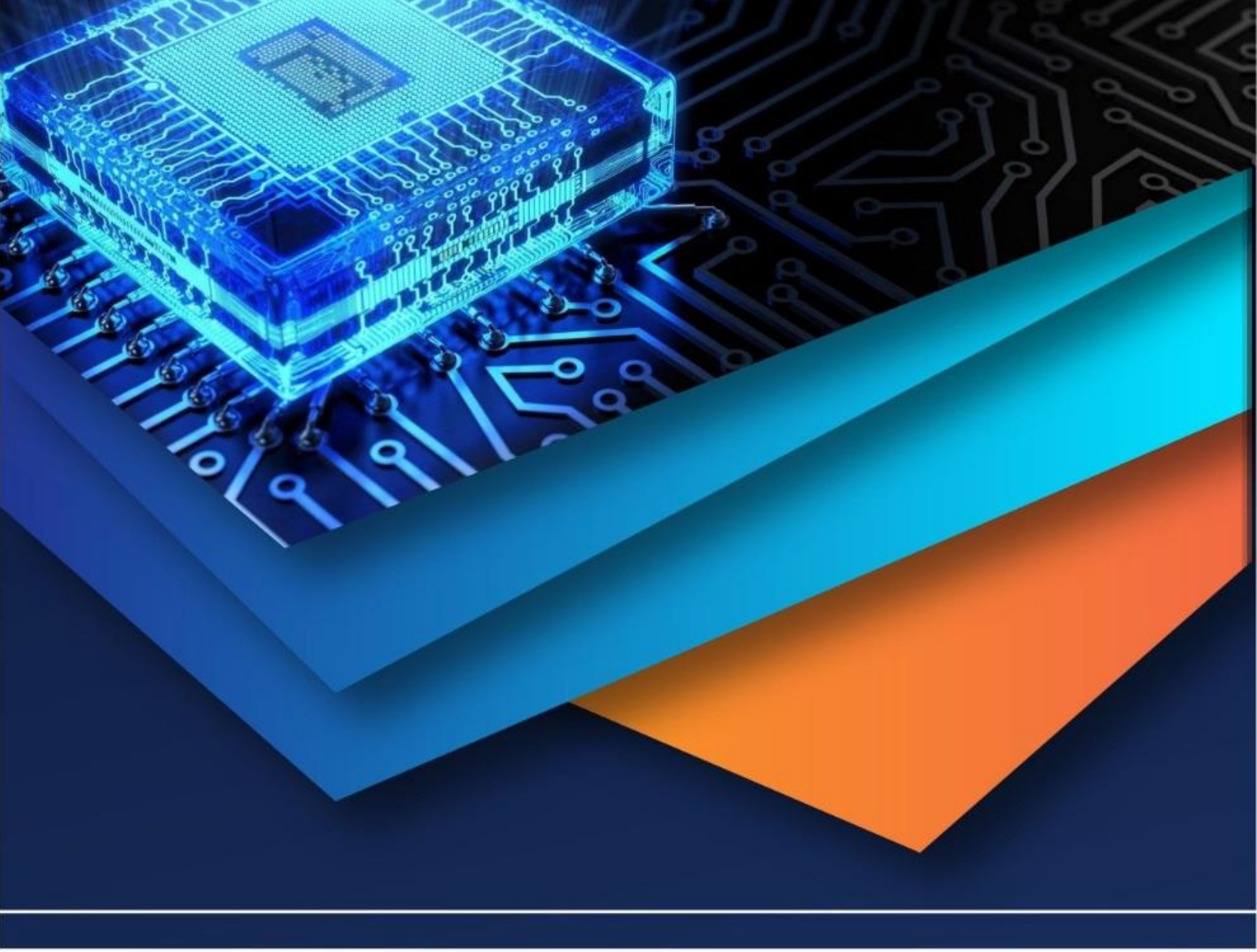
The analytical cycle began from data cleaning and handling, missing values, exploratory analysis lastly model structure buildingand evaluation. The best accuracy on open test set is higher precision score will be found out. This application can assist with tracking down the human diabetes issues.

VI. FUTURE WORK

Human diabetes problems connect with AI model. To computerize this cycle by show the prediction results about web application or work area application. To streamline the work and to carry out in Artificial Intelligence environment.

REFERENCES

[1] a e. evwiekpaefe, nafisat Abdulkadir . “a predictive model for diabetes using machine learning techniques (a case study of some selected hospitals in kaduna metropolis)” ,Jan 2021.
[2] Priyanka Indoria, Yogesh Kumar Rathore. A survey: Detection and Prediction of diabetics using machine learning techniques.IJERT, 2019.
[3] ZhengT, XieW, Xu L, He X, Zhang Y, You M, Yang G, Chen Y. A Machine Learning-Based Framework to identify Type 2 Diabetics through Electronic Health Records, International Journal of medical informatics (IJMI),2017, Vol9, pages120-127.
[4] R. Williams, S. Karuranga, B. Malanda et al., “Global and regional estimates and projections of diabetes-related health expenditure: results from the international diabetes federation diabetes atlas,” Diabetes Research and Clinical Practice, vol. 162,Article ID 108072, 2020.
[5] R. Ahuja, S. C. Sharma, and M. Ali, “A diabetic disease prediction model based on classification algorithms,” Annals of Emerging Technologies in Computing, vol. 3, no. 3, pp. 44–52, 2019.
[6] I. R. Rodríguez, M. Á. Z. Izquierdo, and J. V. Rodríguez, “Towards an ict- based platform for type 1 diabetes mellitus management,” Applied Sciences, vol. 8, no. 4, 2018.
[7] G. Cappon, G. Acciaroli, M. Vettoretti, A. Facchinetti, and G. Sparacino, “Wearable continuous glucose monitoring sensors: a revolution in diabetes treatment,” Electronics, vol. 6, no. 3, 2017.
[8] Rahul Joshi, Minyechil Alehegen . Analysis and Prediction of Diabetics Disease using Machine Learning Algorithm:Ensemble Approach, International Research Journal of Engineering and Technology(IRJET),2017Volume 04Issue10, e- ISSN:2395-0056.
[9] Mekruksavanich, S. Medical Expert System Based Ontology for Diabetics Disease Diagnosis. In Software Engineering and Service Science (ICSESS), 7th IEEE International Conference Pages 383-389,IEEE,2016.
[10] Francesco Mercaldo, Vittoria Nardone, Antonella Santone (2017). Diabetics Mellitus Affected Patients Classification and Diagnosis through Machine Learning Techniques, ProcediaComputerScience112, 2017,2519-2528
[11] Ioannis Kavakiotis, Olga Tsave, Athanasios Salifoglou, “Machine Learning and Data Mining Methods in Diabetics Research”, Jan 8, 2017.



10.22214/IJRASET

45.98

IMPACT FACTOR: 7.129

IMPACT FACTOR: 7.429

INTERNATIONAL JOURNAL FOR RESEARCH

IN APPLIED SCIENCE & ENGINEERING TECHNOLOGY