IJARCCE ISSN (O) 2278-1021, ISSN (P) 2319-5940

International Journal of Advanced Research in Computer and Communication Engineering

Impact Factor 8.102 ⵣ Peer-reviewed & Refereed journal ⵣ Vol. 13, Issue 3, March 2024

DOI: 10.17148/IJARCCE.2024.13327

Diabetes Analysis using Machine Learning with KNN

Dr. P.V.R.D. Prasada Rao[1],

Asritha Musunuru[2], **Subhash Alapati**[3], **Abhinava Reddy KAMIREDDY**[4], **Venkatesh Jajula**[5]

Professor, CSE, Koneru Lakshmaiah Education Foundation, Guntur, AP, India[1]

Student, CSE, Koneru Lakshmaiah Education Foundation, Guntur, AP, India[2-5]

Abstract: Diabetes mellitus, a prevalent global health issue, demands early detection and effective management. For deeper analysis and diabetes prediction, this study uses ML methodologies. A large dataset including clinical, sociological, and biological features is meticulously processed. A wide range of ML methods are used to initiate predictive models. This study enhances the science of diabetes prediction by giving effective tools for early risk assessment, personalized medications, and optimal healthcare management. These breakthroughs have the potential to improve public health outcomes and help combat the diabetes epidemic.

Keywords: Diabetes, prediction, analysis, e-Health, data processing, machine learning

I. INTRODUCTION

A greater number of individuals than ever are affected by diabetes mellitus, a chronic metabolic disease marked by increased blood glucose levels. The International Diabetes Federation (IDF) estimates that 463 million cases of diabetes were diagnosed globally in 2019. If proper preventive measures are not implemented, this figure is expected to rise to 700 million by 2045. Diabetes is an important issue for the community because of the consequences it might cause, such as cardiovascular disease, renal failure, and blindness.[10]

Diabetes risk prediction and early identification are critical for preventing these complications and improving the standards of life for those that are at risk. Machine learning, a subset of artificial intelligence, has the potential to improve diabetes research and prediction. Machine learning algorithms can examine massive datasets, uncover hidden patterns, and anticipate outcomes based on complicated data relationships. The analysis and prediction of diabetes using machine learning approaches is the main work. The main objective is to create reliable predictive algorithms that can pinpoint those whose lives are at risk of acquiring diabetes. Clinical data (such as age, gender, and family history), demographic data (such as BMI, waist circumference), and biochemical indicators (such as glucose levels, lipid profiles) are few of the input variables used by these models.

The healthcare industry can employ machine learning, a type of artificial intelligence, for risk assessment and predictive modelling. Because of its ability to examine massive and complex datasets, discover microscopic correlations, and generate data-driven projections, it has created new options for the treatment of diseases such as diabetes. Using machine learning technologies, we can change how we identify those that are at risk of developing diabetes.

In conclusion, the intent of the investigation is to improve diabetes analysis and prediction by leveraging the potential of machine learning. We may be able to transform the diagnosis and treatment of diabetes by using cutting-edge computational methods, which would ultimately result in better health outcomes and lower healthcare costs. The methodology, findings, and ramifications of the above study will be covered more fully in the parts that follow, providing insight into the exciting future of ML in diabetes care.

II. OBJECTIVE

These objectives all work in concert to enhance machine learning's analysis and prediction of diabetes, which is going to enhance healthcare outcomes of individuals who are at risk of developing this chronic condition.

1)Create Accurate Predictive Models: The major goal is to create theories of ML that can accurately predict a person's risk of acquiring diabetes. These models ought to incorporate a variety of input factors, such as clinical, demographic, and biochemical data, to provide reliable risk assessments.

DOI: 10.17148/IJARCCE.2024.13327

2)List the Important Predictive Features: Determine the most critical diabetes risk variables using feature selection and engineering approaches. Healthcare practitioners may focus on the crucial factors mentioned by this objective in order to identify risk and intervene.

3)Compare multiple machine learning algorithms [2]: Analyse and assess the performance of several machine learning computational methods, such as LR, SVM's, Random Forests, choice trees, and deep learning models. This comparison aids in the selection of the best diabetes prediction algorithm based on dataset properties and performance measures.

4)Access model generalization: Make sure the models you construct generalize effectively to brand-new, unexplored data. This entails evaluating their robustness and propensity to produce accurate predictions in situations outside of the training dataset.

5)Evaluate model interpretability: Make machine learning models easier to understand so that they can be used by healthcare practitioners. Understanding the variables influencing diabetes risk estimates requires the use of methodologies such as feature importance analysis and visualization.

6)Quantify prediction performance: To analyse the models' success in making predictions, use typical assessment metrics such as precision, recall, efficacy, F1-score, and AUC-ROC. These measures provide a comprehensive picture of the models' performance.

7)Contribute to public health [1]: Diabetes prevention and early intervention are made possible by identifying those who are at a high risk of developing the disease. By combining medicine and lifestyle changes, the objective is to enhance healthcare outcomes as soon as possible.

8)Disseminate findings and guidelines: Inform the healthcare industry, decision-makers, and the public about study techniques, findings, and recommendations. Encourage the application of machine learning-based diabetes prediction as a useful tool for enhancing healthcare.

III. LITERATURE REVIEW

a)The first study provides an overview of big data structures and machine learning algorithms used in healthcare, including diabetes detection, given in the first publication by authors Alsulaiman and Alshurideh. It provides details on the various strategies and how they apply to diabetes research.[1]

b)Rajalakshmi et al.'s third study investigates numerous ML algorithms used for diabetes risk assessment, management, and diagnosis. It explains the advantages and disadvantages of various tactics.[3]

c)A deep learning method for assessing retinal pictures and identifying diabetic retinopathy was reported by Gulshan et al. in their fifth paper. The model based on deep learning beat qualified ophthalmologists in the study's demonstration of the value of machine learning in determining the cause of diabetic eye illness by correctly identifying diabetic retinopathy and macular edema.[5]

d)In the sixth paper, Alharthi et al. address using machine learning to create diabetes prediction models based on food trends. In predicting diabetes risk, it underlines the significance of taking lifestyle and nutritional factors into account.[6]

e)The sixth study by Koli and Patil uses machine learning techniques to forecast how type 2 diabetes would develop in patients. The scientists created a model that is capable of determining those who are at risk to problems and offer early care.[7]

IV. PERFORMANCE EVALUATION METHODS AND METRICS

To determine their efficacy and make wise judgments about their implementation, ML models for diabetes detection and analysis must perform well. You can use the following metrics and techniques for performance evaluation in this situation:

a)F1 score: Harmonic mean of recall and precision is known as the F1-score. When you want to consider both false positives and false negatives, it offers equality between these measures and is helpful.

b)ROC-AUC: The model's capacity to distinguish between the beneficial and the contrary examples across various probability thresholds is measured by the ROC-AUC. It is appropriate for issues involving binary classification and unbalanced datasets.

DOI: 10.17148/IJARCCE.2024.13327

c)Confusion matrix: A confusion matrix, which provides a thorough split of genuine positives, genuine negatives, false positives, and erroneous negatives, can be applied to determine measurements such as precision and recall.

d)Mean absoluter error (MAE) and Mean Squared error (MSE): In regression issues, where you're attempting to predict a continuous target variable (such as blood sugar level), these metrics are frequently used. MSE represents the average squared difference, whereas MAE measures the average absolute distinction between the forecast and predictions and actual values.

e)Accuracy: Accuracy is a popular indicator for determining how accurate forecasts are overall. It is calculated as the proportion of cases correctly predicted to all instances. However, accuracy may be deceptive when dealing with biased data sets.

f)Precision: Precision is defined as the fraction of correct positive forecasts among all positive predictions; it helps to assess the model's ability to avoid false positives. Precision is essential since erroneous positives can have disastrous implications.

g)Recall (Sensitivity or True positive rate): Recall quantifies the percentage of correct beneficial predictions all of them instances of actual positive outcomes. When false negatives are expensive, it is essential because it assesses the model's capacity to catch all pertinent instances.

h)ROC curve: You can see the trade-off between sensitivity and specificity at various probability thresholds by plotting the ROC curve. For model comparison, the region under this curve (ROC- AUC) might be used.

i)Medel Interpretability: For healthcare applications like diabetes detection, model interpretability is crucial. Use tools like SHapley Additive exPlanations(SHAP) or Local Interpretable Model- Agnostic Explanations(LIME) to comprehend how the model generates predictions.

V. IMPLEMENTING ANY STRATEGY

Regarding your neutral working model to be successfully deployed, you must take several important actions while implementing a machine learning strategy. The fundamental actions you should think about are as follows:

1)Select the right algorithm: Pick the machine learning algorithm or algorithms that will be most beneficial to you solve your challenge. Consider factors including the task's kind (classification, regression, or clustering), the volume of data, and the available computer resources.

2)Data collection and preparation: Assemble pertinent information from numerous sources. Make sure the information is accurate, organized, and pertinent to the issue at hand. To develop useful features for modeling, perform data preprocessing, which may include Insufficient value handling, outlier detection, and feature engineering are all possible. For developing and assessing the model, divide the data into training, validation, and test sets.

3)Model Development: Utilize the training data to create, train, and improve your machine learning model(s). Use methods like grid search or randomized search to fine-tune hyperparameters. Utilize the right measurements and validation methods to assess the model's performance.

4)Model Evaluation: Utilize metrics pertinent to your issue to evaluate the model's performance on the validation dataset (examples: accuracy, F1-score, RMSE). Consider potential problems like overfitting and underfitting and make the required model revisions.

5)Monitoring and Maintenance: Keep an eye on the model's functionality in the real-world setting. Create notifications for abnormalities or performance degradation. To keep the model current, retrain it frequently with fresh data.

VI. METHODOLOGY

Implementing a machine learning strategy for diabetes analysis and prediction requires unique actions catered to the goals of project and the healthcare industry. Here are the essential actions to take:

1)Data collection and Preprocessing: The patient's demographics, medical history, lab results (such as blood glucose levels), medication information, and lifestyle choices are all important pieces of medical information to gather.

Ensure data privacy is upheld and abide by applicable rules, such as HIPAA in USA for healthcare. Perform data pretreatment to cope with missing values, outliers, and noise. normalize or standardize numerical properties.

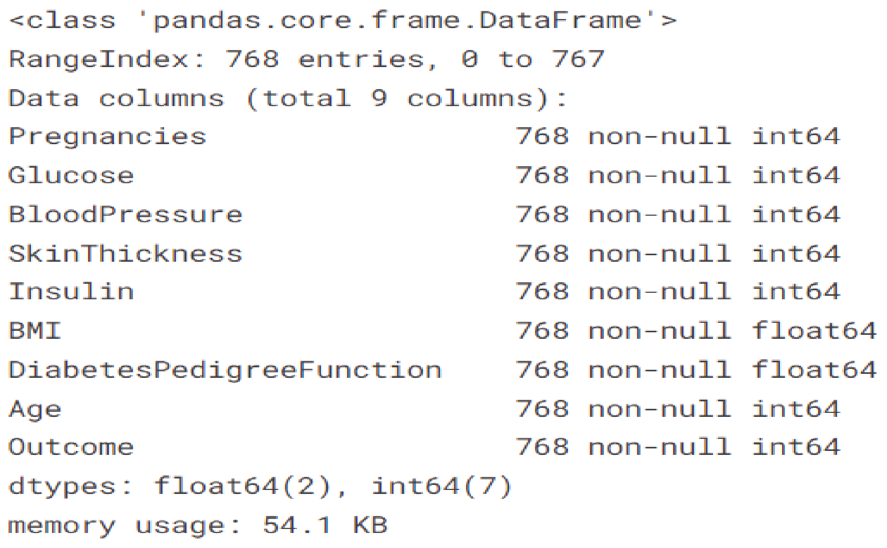
2)Feature Engineering: From raw data, produce useful characteristics. Calculating the body mass index (BMI), insulin sensitivity indices, or diabetes risk scores may be required. Include domain-specific attributes that experts in medicine believe are crucial for predicting diabetes. Make sure that the data is balanced and representative, divide the dataset into training, validation, and test sets.

3)Model Selection: Select AL algorithms that can accurately forecast diabetes. Common possibilities include SVM˙s, random forests, decision trees, LR, and deep learning models such as neural networks.

4)Model Development: Utilizing the training dataset, create and train your chosen machine learning model(s). To improve model performance, take into account strategies such hyperparameter tuning.

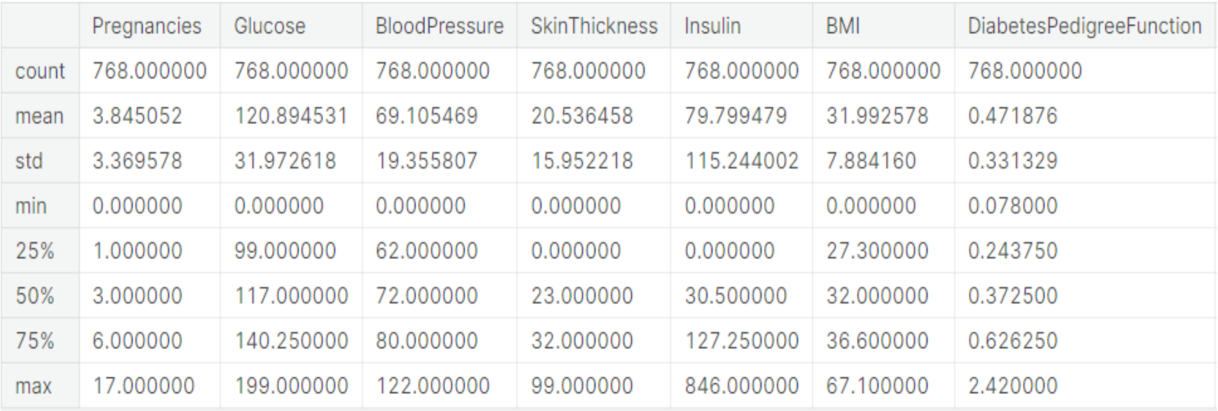
5)Model Evaluation: The F1-s, area hinter the ROC curve (AUC-ROC), precision, recall, and other pertinent assessment metrics can be used to evaluate the model's performance. Use approaches such as oversampling, under sampling, or cost-sensitive learning to resolve any difficulties, including class imbalance.

VII. RESULTS



```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 768 entries, 0 to 767
Data columns (total 9 columns):
Pregnancies                 768 non-null int64
Glucose                     768 non-null int64
BloodPressure               768 non-null int64
SkinThickness               768 non-null int64
Insulin                     768 non-null int64
BMI                         768 non-null float64
DiabetesPedigreeFunction    768 non-null float64
Age                         768 non-null int64
Outcome                     768 non-null int64
dtypes: float64(2), int64(7)
memory usage: 54.1 KB
```

Fig.1. Basic EDA and sttistical Analysis

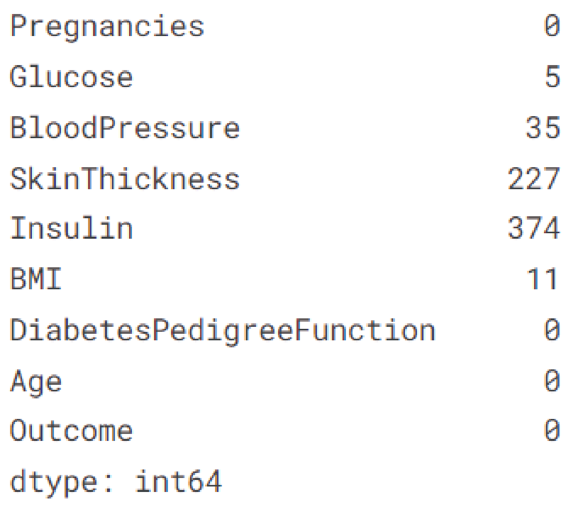


	Pregnancies	Glucose	BloodPressure	SkinThickness	Insulin	BMI	DiabetesPedigreeFunction
count	768.000000	768.000000	768.000000	768.000000	768.000000	768.000000	768.000000
mean	3.845052	120.894531	69.105469	20.536458	79.799479	31.992578	0.471876
std	3.369578	31.972618	19.355807	15.952218	115.244002	7.884160	0.331329
min	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000	0.078000
25%	1.000000	99.000000	62.000000	0.000000	0.000000	27.300000	0.243750
50%	3.000000	117.000000	72.000000	23.000000	30.500000	32.000000	0.372500
75%	6.000000	140.250000	80.000000	32.000000	127.250000	36.600000	0.626250
max	17.000000	199.000000	122.000000	99.000000	846.000000	67.100000	2.420000

Fig. 2 Describing data set

DOI: 10.17148/IJARCCE.2024.13327

diabetes_data_copy = diabetes_data.copy(deep = True)
diabetes_data_copy[['Glucose','BloodPressure','SkinThickness','Insulin','BMI']] = diabetes_data_copy[['Glucose','BloodPressure','SkinThickness','Insulin','BMI']].replace(0,np.NaN)print(diabetes_data_copy.isnull().sum())



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

Fig. 3 Replacing the 0 or missing values with NaN

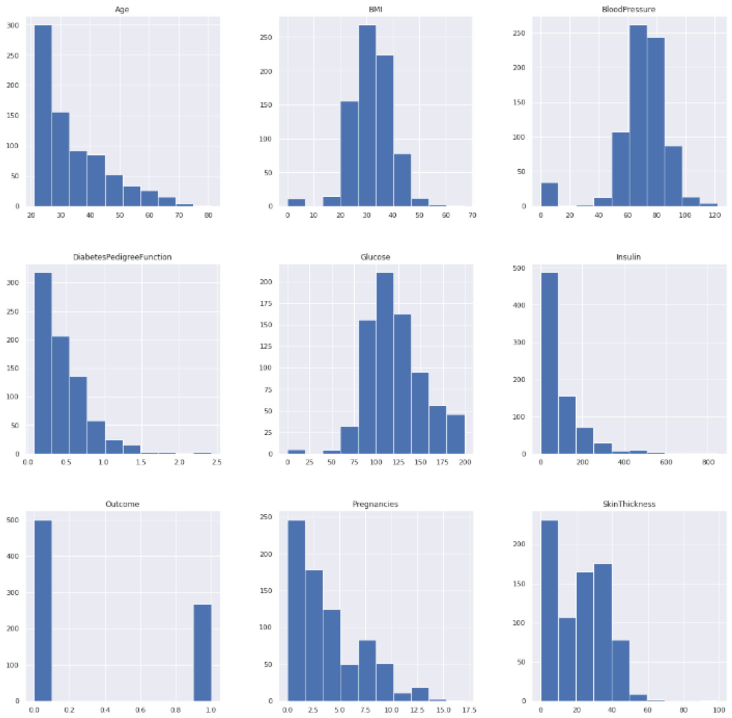


Fig.3.1. To fill those NaN values and plotting

diabetes_data_copy['Glucose'].fillna(diabetes_data_copy['Glucose'].mean(), inplace = True)
diabetes_data_copy['BloodPressure'].fillna(diabetes_data_copy['BloodPressure'].mean(), inplace = True)
diabetes_data_copy['SkinThickness'].fillna(diabetes_data_copy['SkinThickness'].median(), inplace = True)
diabetes_data_copy['Insulin'].fillna(diabetes_data_copy['Insulin'].median(), inplace = True)
diabetes_data_copy['BMI'].fillna(diabetes_data_copy['BMI'].median(), inplace = True)

ISSN (O) 2278-1021, ISSN (P) 2319-5940

International Journal of Advanced Research in Computer and Communication Engineering

Impact Factor 8.102 ⁜ Peer-reviewed & Refereed journal ⁜ Vol. 13, Issue 3, March 2024

DOI: 10.17148/IJARCCE.2024.13327

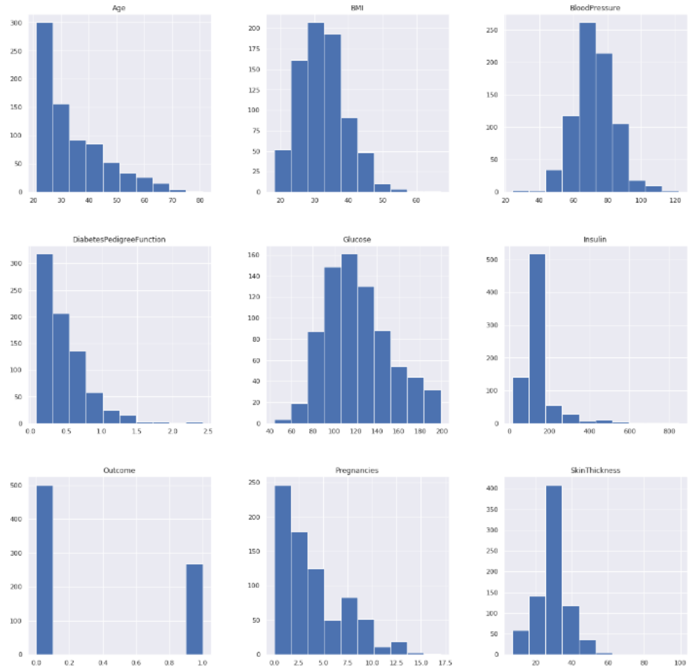


Fig.3.2. Inputing the values of Nan for the columns based on distribution

p = diabetes_data_copy.hist(figsize = (20,20))

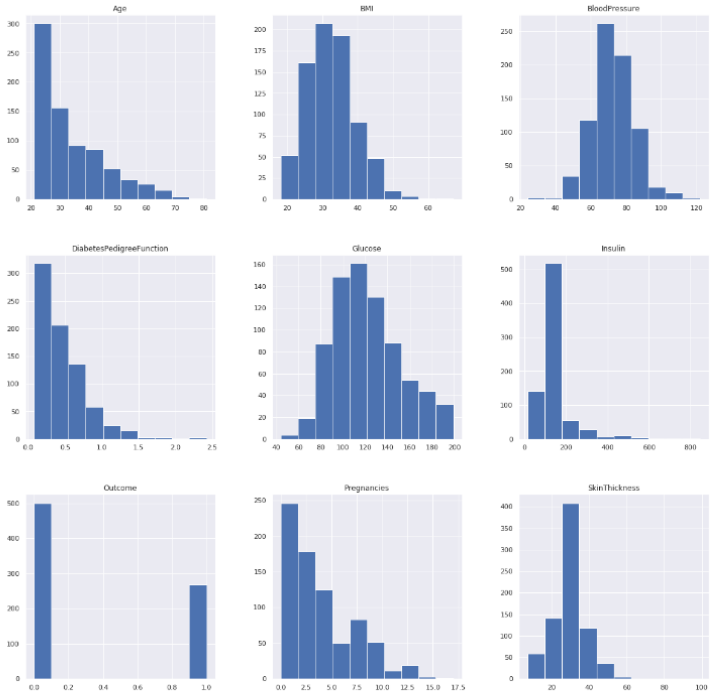


Fig.3. 3 Plotting after NaN removel

A) Skewness:

A lengthy left tail characterizes a left-skewing distribution. Left-skewed distributions are sometimes used to describe negatively skewed distributions. This is a consequence of the number line's big negative tail. In addition, the peak is to the left of the mean.

DOI: 10.17148/IJARCCE.2024.13327

The right tail of a right-skewed distribution is lengthy. Positive-skew distributions are another name for right-skewed distributions. That is because of a lengthy positive tail on the number line. The meaning is also to the peak's right.

- If we shape the data set now using " diabetis_data.shpe" it will show you value like (768,9)

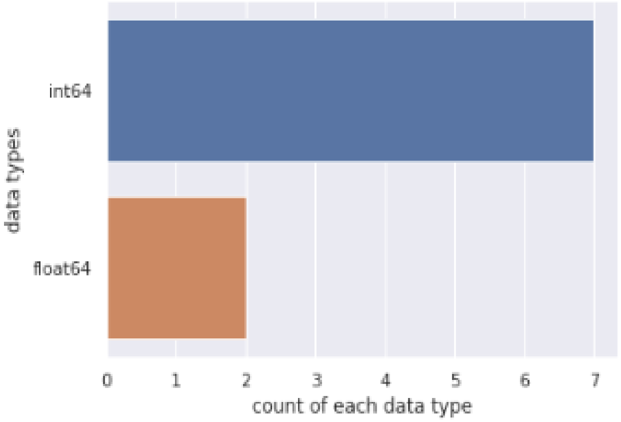


Fig.4 Plotting data types

B) Null count analysis:

import missingno as msno
p=msno.bar(diabetes_data)

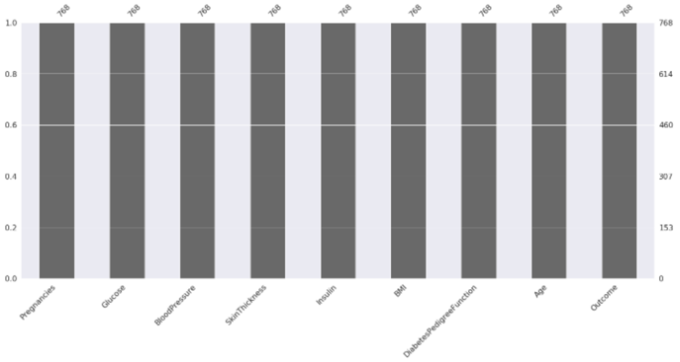
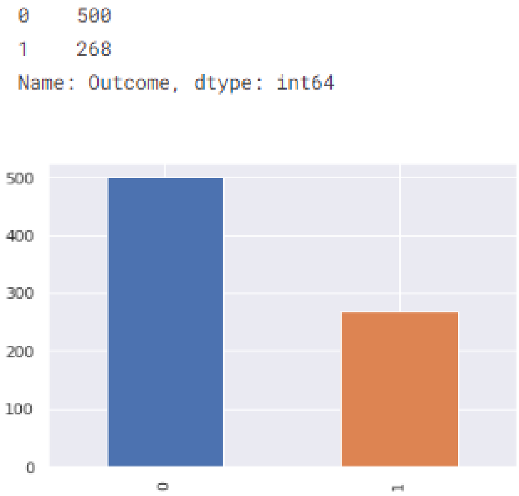


Fig.5 Null count analysis



```
0    500
1    268
Name: Outcome, dtype: int64
```

Fig.6. Checking the balance between data types

C) Scatter Matrix:

```
from pandas.tools.plotting import scatter_matrix
p=scatter_matrix(diabetes_data,figsize=(25, 25))
```

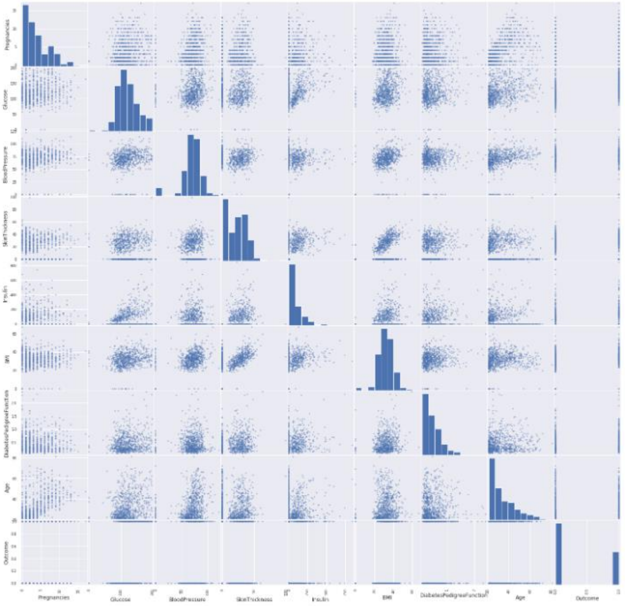


Fig. 7 Scatter matric for uncleaned data

D) Pait plot:

```
p=sns.pairplot(diabetes_data_copy, hue = 'Outcome')
```

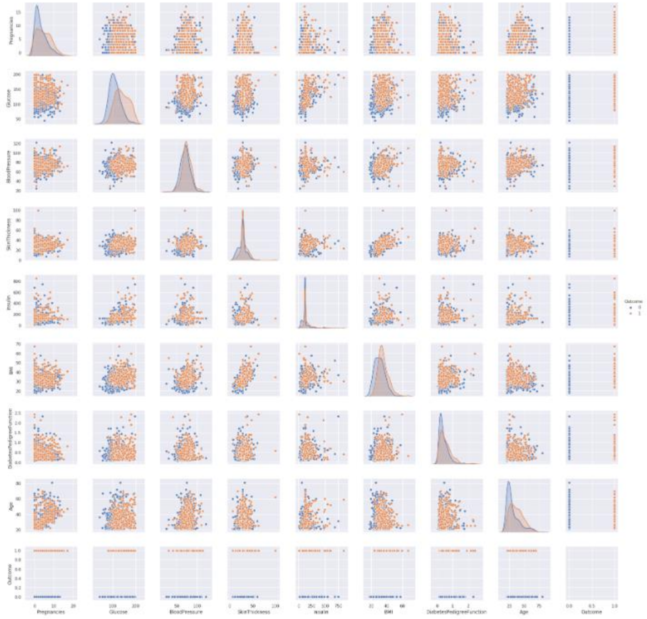


Fig.8 Pair plot for clean data

Finding the correlation between two quantities is made easier with the aid of Pearson's correlation coefficient. It gives you a measure of the degree to which both factors are related. The Pearson's Correlation Coefficient's value can range from -1 to +1. 1 denotes a strong link between them, while 0 denotes no association

A heat map is a two-dimensional representation of information with the help of colors. Heat maps can help the user visualize simple or complex information.

DOI: 10.17148/IJARCCE.2024.13327

E) Heat Map:

```
plt.figure(figsize=(12,10))
p=sns.heatmap(diabetes_data.corr(),annot=True,cmap ='RdYlGn')
```

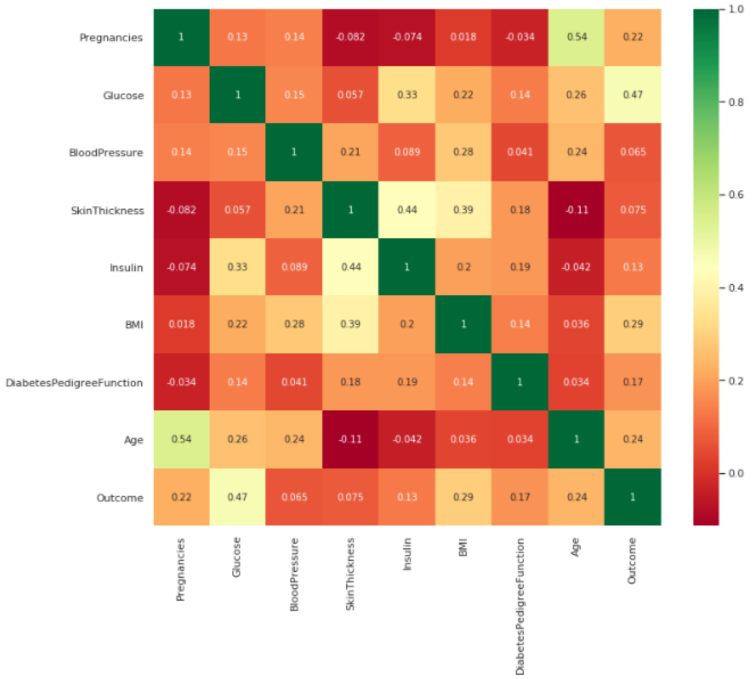


Fig.9.1 Unclean data

```
plt.figure(figsize=(12,10))
p=sns.heatmap(diabetes_data_copy.corr(), annot=True,cmap ='RdYlGn')
```

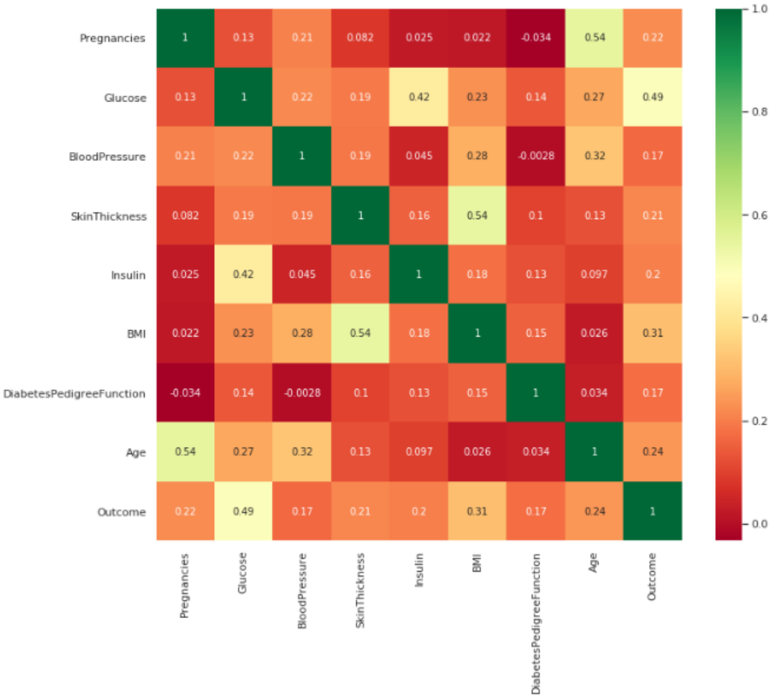


Fig.9.2 Clean data

IJARCCE ISSN (O) 2278-1021, ISSN (P) 2319-5940

International Journal of Advanced Research in Computer and Communication Engineering

Impact Factor 8.102 ⁜ Peer-reviewed & Refereed journal ⁜ Vol. 13, Issue 3, March 2024

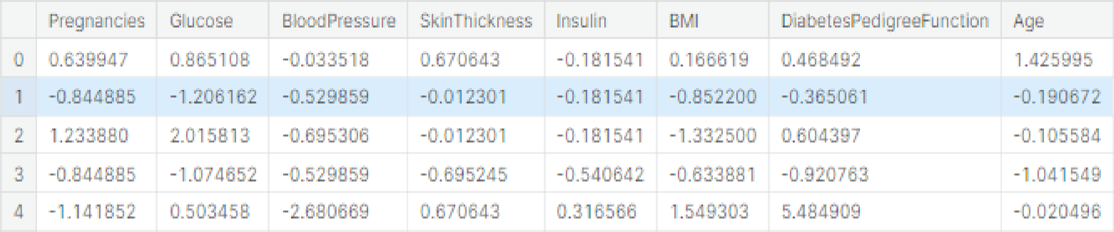
DOI: 10.17148/IJARCCE.2024.13327

F) Scaling the data:

Data Z is rescaled such that μ = 0 and **σ** = 1, and is done through this formula,

$$z = \frac{x_i - \mu}{\sigma}$$

```
from sklearn.preprocessing import StandardScaler
sc_X = StandardScaler()
 X=pd.DataFrame(sc_X.fit_transform(diabetes_data_copy.drop(["Outcome"],axis = 1),)columns=['Pregnancies', 'Glucose', 'BloodPressure', 'SkinThickness', 'Insulin', 'BMI', 'DiabetesPedigreeFunction', 'Age'])X.head()
Y=diabetis_data_copy.Outcome
```



	Pregnancies	Glucose	BloodPressure	SkinThickness	Insulin	BMI	DiabetesPedigreeFunction	Age
0	0.639947	0.865108	-0.033518	0.670643	-0.181541	0.166619	0.468492	1.425995
1	-0.844885	-1.206162	-0.529859	-0.012301	-0.181541	-0.852200	-0.365061	-0.190672
2	1.233880	2.015813	-0.695306	-0.012301	-0.181541	-1.332500	0.604397	-0.105584
3	-0.844885	-1.074652	-0.529859	-0.695245	-0.540642	-0.633881	-0.920763	-1.041549
4	-1.141852	0.503458	-2.680669	0.670643	0.316566	1.549303	5.484909	-0.020496

Fig.10 Scaling

G) Test Train Split and Cross Validation Methods:

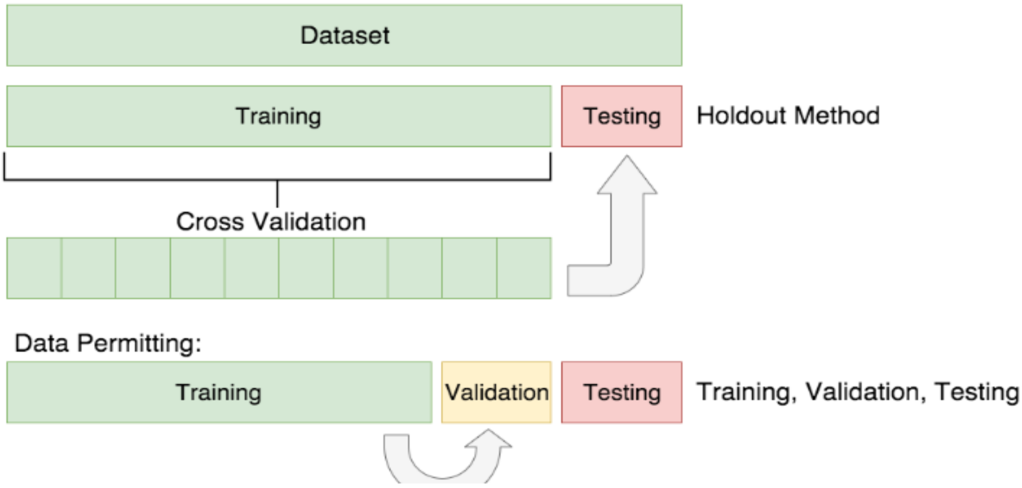


Fig.11 Flow Chart

```
from sklearn.model_selection import train_test_split
X_train,X_test,y_train,y_test = train_test_split(X,y,test_size=1/3,random_state=42, stratify=y)
from sklearn.neighbors import KNeighborsClassifier
test_scores = []
train_scores = []
for i in range(1,15):
  knn = KNeighborsClassifier(i)
  knn.fit(X_train,y_train)
   train_scores.append(knn.score(X_train,y_train))
   test_scores.append(knn.score(X_test,y_test))
max_train_score = max(train_scores)
train_scores_ind = [i for i, v in enumerate(train_scores) if v == max_train_score]
print('Max train score {} % and k = {}'.format(max_train_score*100,list(map(lambda x: x+1, train_scores_ind))))
```

DOI: 10.17148/IJARCCE.2024.13327

Max train score 100.0 % and k = [1]

```
max_test_score = max(test_scores)
test_scores_ind = [i for i, v in enumerate(test_scores) if v == max_test_score]
print('Max test score {} % and k = {}'.format(max_test_score*100,list(map(lambda x: x+1, test_scores_ind))))
```

Max test score 76.5625 % and k = [11]

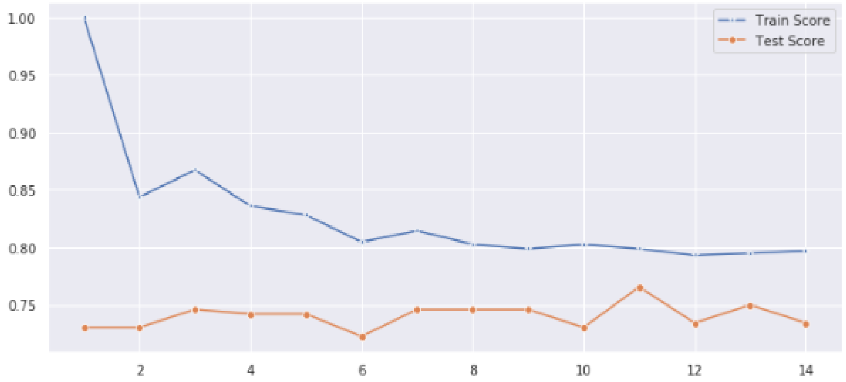


Fig. 11. 1 Visualization

The best result is captured at k = 11 hence 11 is used for the final model

```
knn = KNeighborsClassifier(11)
knn.fit(X_train,y_train)
knn.score(X_test,y_test)
```

- 0.765625

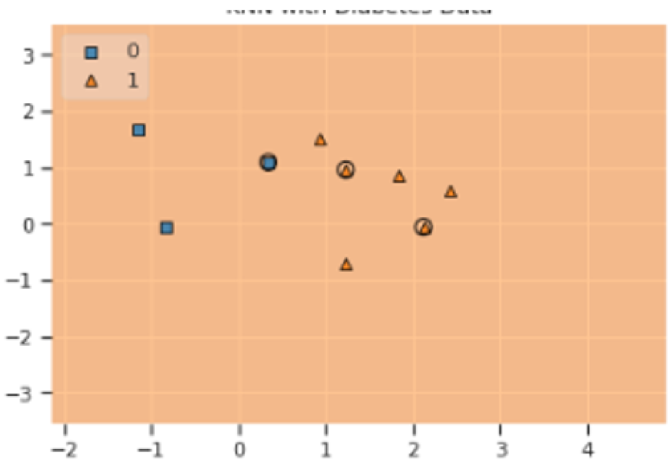
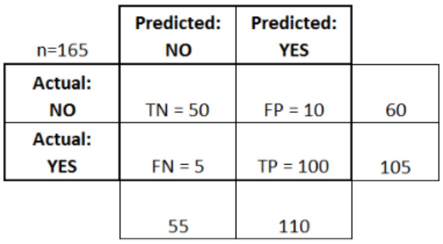


Fig. 12 KNN with Diabetes Data

H) Confusion Matrix:



n=165	Predicted: NO	Predicted: YES	
Actual: NO	TN = 50	FP = 10	60
Actual: YES	FN = 5	TP = 100	105
	55	110	

Fig.13.1 Data

DOI: 10.17148/IJARCCE.2024.13327

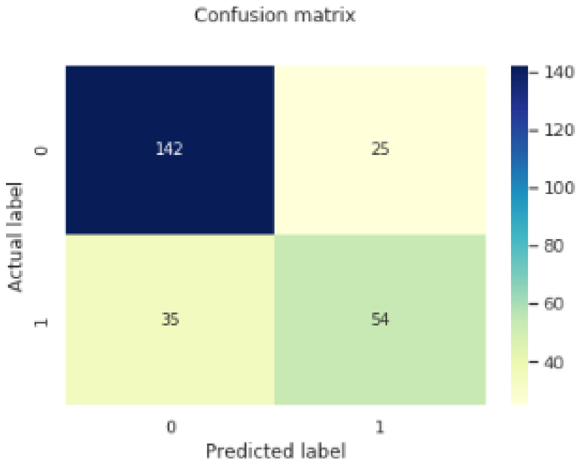


Fig.13.2 Confusion Matrix

VIII. CONCLUSION

To summarize, the development of early diabetes detection by machine learning is critical to tackling the global diabetes epidemic. These technologies allow for rapid reactions, which can enhance the results for patients while reducing the financial and personal costs associated with diabetes-related disorders. These technologies deliver precise risk assessments by utilizing a variety of datasets and cutting-edge algorithms. The idea of early diabetes identification offers promise for transforming the landscape of preventive healthcare as technology develops. In the not-too-distant future, the best diabetes analysis will be developed with the goal of forecasting the disease and promoting diabetes awareness through the use of neural network technology.

REFERENCES

[1]. Alsulaiman and Alshurideh " A Survey of Big Data Architectures and Machine Learning Algorithms in Healthcare" , Journal of King Saud University - Computer and Information Sciences, 2018

[2]. M. K. Hasan, M. A. Alam, D. Das, E. Hossain, and M. Hasan, ''Diabetes prediction using ensembling of different machine learning classifiers,'' IEEE, 2020.

[3]. Rajalakshmi et al, " A Review on Machine Learning Techniques for the Diagnosis of Diabetes", Diabetes, Metabolic Syndrome and Obesity: Targets and Therapy, 2018

[4]. P. Dua, F. J. Doyle, and E. N. Pistikopoulos, "Model-based blood glucose management for type 1 diabetes using parametric programming" IEEE, Aug. 2006.

[5]. Gulshan et al, "Automated Detection of Diabetic Retinopathy Using Deep Learning", Ophthalmology, 2016

[6]. "Diabetes Mellitus: Machine Learning-Based Predictive Modelling Using Dietary Patterns," Alharthi et al., Nutrients,2020

[7]. Koli and Patil, "Machine Learning method to forecast how patients' type 2 diabetes mellitus", SN Computer Science, 2021

[8]. Yasin, S. A., & Prasad Rao, P. V. R. D. (2018). Analysis of single and hybrid data mining techniques for prediction of heart disease using real time dataset, International Journal of Engineering and Technology (UAE), 7(2), 97-99. doi:10.14419/ijet. v 7i2.32.13536

[9]. Asif Hssan Syed, "Machine Learning-Based Aplication for Predecting Risk of Type 2 Diabetes Mellitus in Suadi Arabia: A retrospectie Cross- Sectional Study", IEEE Access, 2020

[10]. M. A. Al Mansour, "The prevalence and risk factors of type-2 diabetes mellitus in a semi-urban saudi population", Public Health, 2019

[11]. Balaraju, J., Prasada Rao, P.V.R.D. "Designing authentication for hadoop cluster using DNA algorithm" International Journal of Recent Technology and Engineering, 2019

[12]. U. Ahmed et al, "Prediction of diabetes Empowered with fused mahine learning", IEEE Journel 2022

[13]. S. Kumari, D. Kumar, and M. Mittal, ''An ensemble approach for classification and prediction of diabetes mellitus using soft voting classifier,'' Int. J. Cogn. Comput. Eng 2021