International Journal of Scientific Research in Engineering and Management (IJSREM)
Volume: 08 Issue: 11 | Nov - 2024 SJIF Rating: 8.448 ISSN: 2582-3930

GlucoPredict: Leveraging Logistic Regression to Predict Diabetes Risk

K.Rupa Sravanthi[1], Ch.Bhavannarayana[2]
1.Aditya College of Engineering and Technology, Surampalem.
2.Kakinada institute of Engineering and technology-II, Korangi.

ABSTRACT

GlucoPredict is a machine learning-based project designed to predict the risk of diabetes using logistic regression, a powerful statistical method. The goal of this project is to develop a model that can accurately classify individuals as either at risk of diabetes or not, based on a set of medical and lifestyle-related features. The dataset used for training and evaluation consists of key factors such as age, BMI, blood pressure, glucose levels, and family history of diabetes. By applying logistic regression, the model can identify complex patterns and relationships within these variables, providing a clear, probabilistic prediction of diabetes risk. This approach offers a transparent and interpretable way to understand how individual factors contribute to the likelihood of developing diabetes.

The project aims to address the growing global concern of diabetes by providing an accessible, data-driven tool for early detection. Early diagnosis is crucial in managing and preventing complications associated with diabetes, and this model offers healthcare professionals a powerful resource for decision-making. By leveraging logistic regression, **GlucoPredict** not only delivers high accuracy but also ensures that the predictions are interpretable, enabling users to understand the underlying factors influencing the outcomes. This project demonstrates the potential of machine learning in improving healthcare outcomes

***Key words*: *Machine learning, GlucoPredict, Logistic regression, Diabetes*.**

Introduction:

Machine Learning is a way of Manipulating and extraction of implicit, previously unknown/known and potential useful information about data. Machine learning incorporates
various classifiers of Supervised, Unsupervised and Ensemble Learning which are used to predict and Find the Accuracy of the given dataset. In this we are using an effective Machine Learning algorithm I.e. Logistic Regression Model.

Logistic Regression is a fundamental machine learning algorithm used for binary classification, where the goal is to predict one of two possible outcomes (e.g., "yes/no," "true/false," or "1/0"). It works by modeling the relationship between the input features and the target variable using a linear equation, similar to linear regression. However, instead of directly predicting a continuous value, logistic regression applies a **logistic (sigmoid)** function to the output of the linear equation, transforming it into a probability between 0 and 1. This probability indicates the likelihood of the instance belonging to a particular class, with a threshold (usually 0.5) used to make the final classification decision.

Diabetes is a chronic metabolic condition where the body fails to regulate blood sugar (glucose) levels effectively. This can occur due to insufficient production of insulin, a hormone that manages glucose levels, or because the body's cells fail to respond properly to insulin.

Diabetes prediction focuses on identifying individuals at high risk of developing the condition, enabling early intervention and prevention. By analyzing demographic, lifestyle, and medical data such as age, BMI, blood pressure, and glucose levels, predictive models can forecast diabetes onset. Advances in machine learning and data analytics have made it possible to develop accurate and efficient systems for this purpose. These models, including logistic regression, decision trees, and neural networks, aid healthcare providers in targeting at-risk populations with tailored preventive measures, improving outcomes, and reducing the burden of diabetes on healthcare systems.

Aim :

The aim of the project GlucoPredict: Leveraging Logistic Regression to Predict Diabetes Risk is to develop a reliable and data-driven model for predicting an individual's likelihood of developing diabetes. By utilizing logistic regression, the project seeks to analyze key risk factors such as age, BMI, glucose levels, and lifestyle habits to generate accurate predictions. This model aims to assist healthcare providers in identifying at-risk individuals early, enabling timely interventions, such as lifestyle changes or medical treatments, to prevent the onset of diabetes and reduce its long-term complications.

Objective :

The objective of the project GlucoPredict: Leveraging Logistic Regression to Predict Diabetes Risk is to design and implement a predictive tool that utilizes logistic regression to assess diabetes risk based on key health and demographic variables. The project aims to provide a user-friendly and interpretable solution for healthcare professionals and individuals to estimate the probability of developing diabetes, thus promoting early detection. By focusing on accuracy, accessibility, and practicality, the tool seeks to enhance preventive healthcare strategies and reduce the overall impact of diabetes in at-risk populations.

Literature review:

Yasodha et al. conducted a study on diabetes prediction by classifying a dataset of 200 instances with 9 attributes, focusing on blood and urine tests. Using WEKA, the authors applied different classification algorithms (Naïve Bayes, J48, REP Tree, and Random Tree) with 10-fold cross-validation. The results showed that J48 performed the best with an accuracy of 60.2%, outperforming the other models in predicting diabetes status from the dataset. This study highlights the effectiveness of decision tree-based models for small datasets.

Aiswarya et al. aimed to find efficient methods for diabetes detection using classification algorithms, particularly Decision Tree and Naïve Bayes. They used the PIMA dataset and cross-validation, concluding that the J48 algorithm achieved an accuracy of 74.8%, while Naïve Bayes performed slightly better, reaching 79.5% accuracy with a 70:30 data split. The study emphasizes improving the speed and accuracy of diabetes detection to enable timely medical intervention.

Gupta et al. evaluated the accuracy, sensitivity, and specificity of various classification algorithms, comparing their performance across WEKA, RapidMiner, and Matlab. The study used JRIP, Jgraft, and BayesNet algorithms and found that Jgraft achieved the highest accuracy of 81.3%, with a sensitivity of 59.7% and specificity of 81.4%. The results also showed that WEKA outperformed both Matlab and RapidMiner in terms of classification performance, making it the most effective tool for diabetes prediction in this study.

Lee et al. applied the CART (Classification and Regression Trees) algorithm to a diabetes dataset while addressing the issue of class imbalance. They emphasized the importance of detecting class imbalance early in the data preprocessing stage, as it often occurs in datasets with dichotomous values (two possible outcomes). By applying a resample filter, they aimed to correct the imbalance, which significantly improved the accuracy of the predictive model, showing that handling such issues is critical for enhancing the performance of classification algorithms.

Methodology:

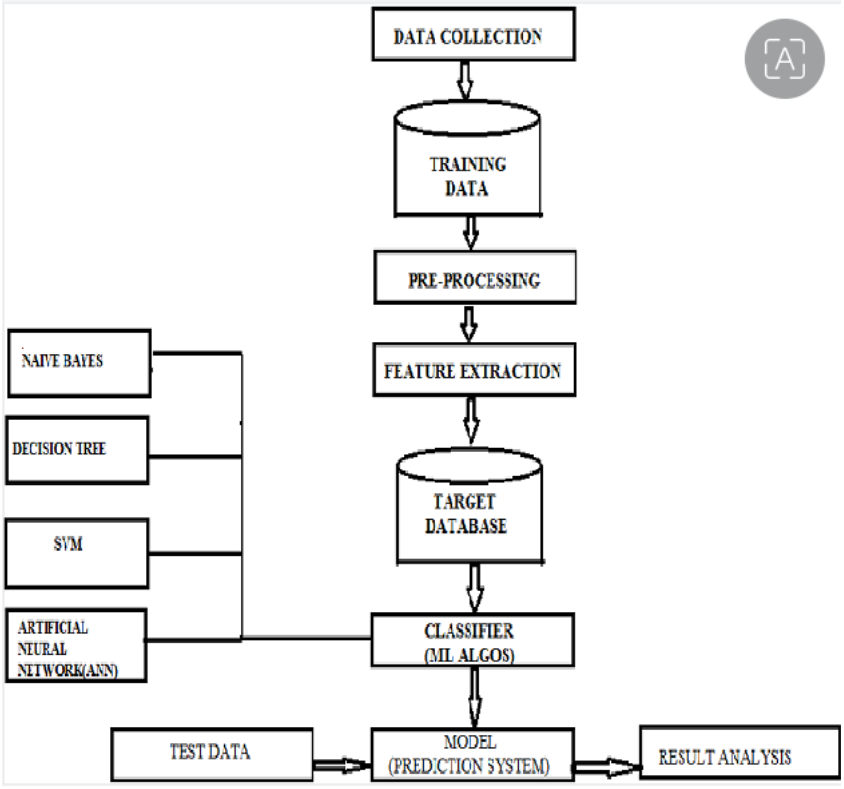
- **Existing system:**

Existing diabetes prediction systems using machine learning rely on models trained on medical datasets to identify individuals at risk based on health and demographic features. Popular datasets like the PIMA Indian Diabetes Dataset (PIDD) are widely used, containing attributes such as glucose levels, BMI, blood pressure, and age. These systems often utilize algorithms like Logistic Regression, Decision Trees, Random Forests, and Gradient Boosting for their simplicity and accuracy in structured datasets. Advanced models, such as Support Vector Machines (SVM) and Neural Networks, have shown high performance, particularly when datasets are large and include complex patterns. These systems are frequently implemented in clinical settings and mobile health applications for preliminary risk assessment.

Real-world applications of machine learning-based diabetes prediction include mobile apps, wearable device integration's, and hospital decision-support systems. Apps like MySugr or Fitbit employ basic predictive models to provide users with personalized diabetes risk insights, while hospital systems integrate predictive analytics into electronic health records (EHR) for early detection. However, existing systems often face challenges such as dataset bias, class imbalance, and lack of interpretability in complex models, which limit their reliability across diverse populations. Ongoing improvements in feature engineering, dataset quality, and explainable AI aim to make these systems more effective and widely applicable.

- **Existing Architecture :**

Supervised Machine Learning algorithms that are precisely used for diabetes prediction. The results indicate that the Decision Tree classification model predicted the cardiovascular diseases better than Naive Bayes, Logistic Regression, Random Forest, S VM and KNN based approaches.



The Decision Tree bequeathed the best result with the accuracy of 73%. This approach could be helpful for doctors to predict the occurrence of Diabetes in advance and provide appropriate treatment.

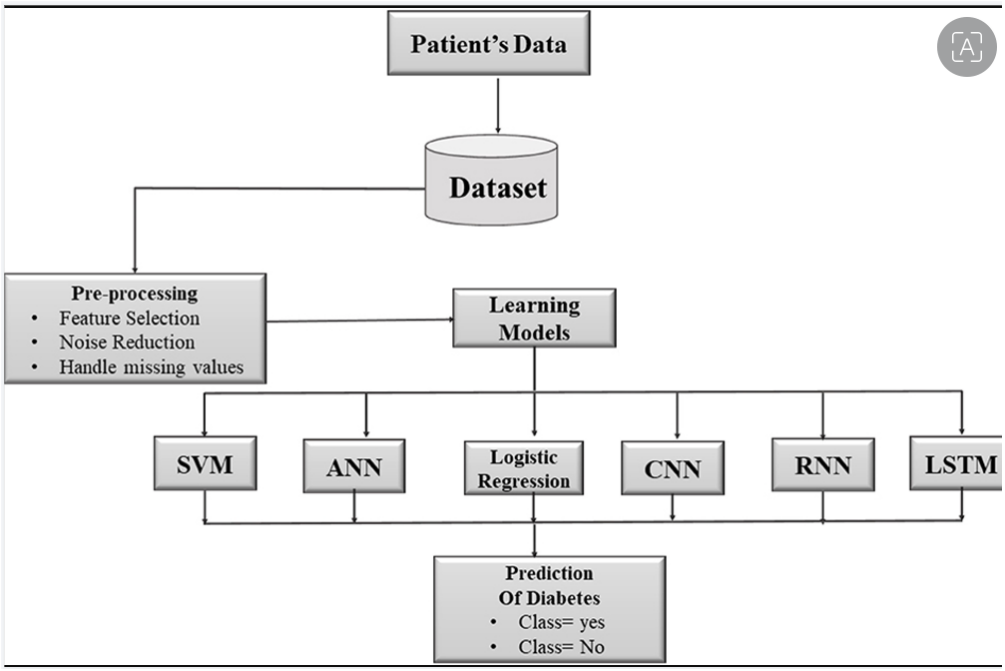
Disadvantages of existing system:

1. Risk of Overfitting it may this means they may perform very well on the training data but fail to generalize to new, unseen data,In the context of diabetes prediction, over-fitting can lead to inaccurate predictions.

2. In the existing system, practical use of various collected data is time consuming.

➢ **Proposed system:**

The proposed system for "GlucoPredict: Leveraging Logistic Regression to Predict Diabetes Risk" is designed to help identify people who are at risk of diabetes using a simple and effective method. By using logistic regression, the system will analyze health data such as glucose levels, age, BMI, blood pressure, and family history to predict whether someone is at risk of developing diabetes. The data will first be cleaned and prepared to ensure accuracy, with missing or incorrect values handled carefully. Features that are most important for predicting diabetes will be selected to make the model efficient and reliable. The system will also address any imbalance in the data, ensuring it works well for both diabetic and non-diabetic cases.

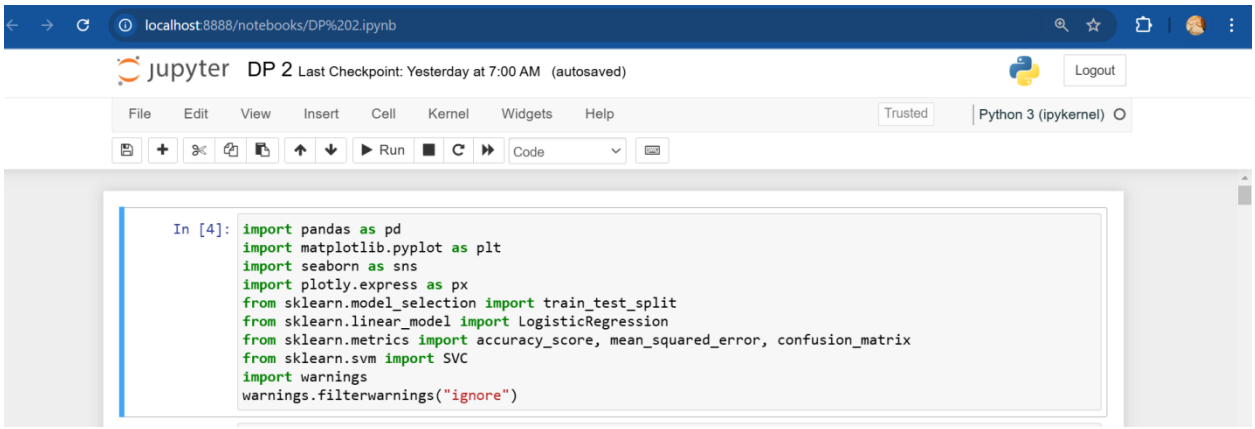


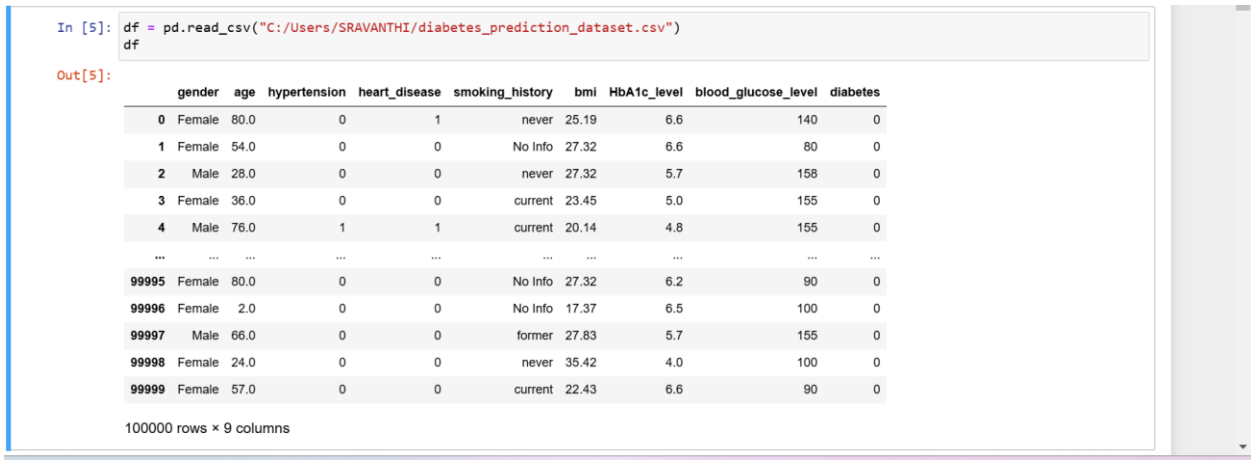
The system's predictions will not only show whether someone is at risk but also highlight the factors contributing to that risk, helping individuals and doctors take early action. It will be tested for accuracy, precision, and reliability to ensure it performs well. Since logistic regression is simple and easy to understand, healthcare providers can trust the predictions and use the system effectively. Overall, "GlucoPredict" aims to be a user-friendly tool that supports early detection and prevention of diabetes

The effectiveness and accuracy of the machine learning method can be evaluated using performance indicators. In this Logistic Regression model is used and obtained most accurate result I.e **96.235% accuracy**.

Results :

Step 1: This step imports necessary libraries such as pandas, numpy, matplotlib, seaborn, logistic regression, sklearn and loads the dataset into a pandas dataframe for further analysis.





In [5]:
```
df = pd.read_csv("C:/Users/SRAVANTHI/diabetes_prediction_dataset.csv")
df
```

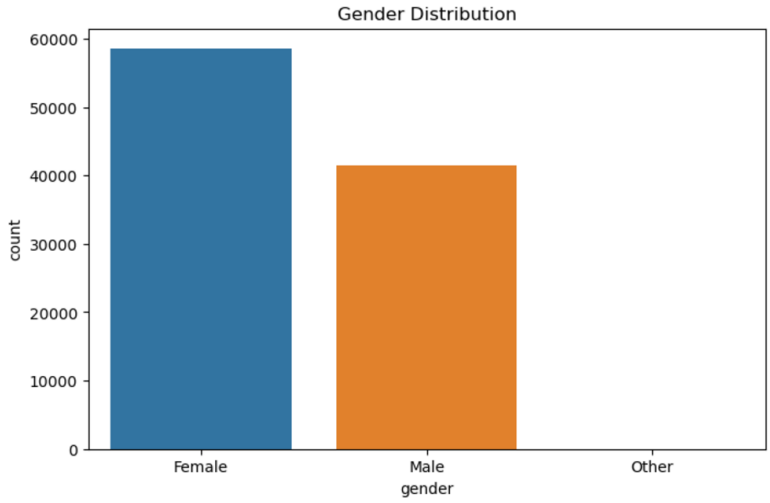
Out[5]:

	gender	age	hypertension	heart_disease	smoking_history	bmi	HbA1c_level	blood_glucose_level	diabetes
0	Female	80.0	0	1	never	25.19	6.6	140	0
1	Female	54.0	0	0	No Info	27.32	6.6	80	0
2	Male	28.0	0	0	never	27.32	5.7	158	0
3	Female	36.0	0	0	current	23.45	5.0	155	0
4	Male	76.0	1	1	current	20.14	4.8	155	0
...	...	...	...	...	...	...	...	...	...
99995	Female	80.0	0	0	No Info	27.32	6.2	90	0
99996	Female	2.0	0	0	No Info	17.37	6.5	100	0
99997	Male	66.0	0	0	former	27.83	5.7	155	0
99998	Female	24.0	0	0	never	35.42	4.0	100	0
99999	Female	57.0	0	0	current	22.43	6.6	90	0

100000 rows × 9 columns

Step 2 : In This step it will divide the data according to Gender wise and plot the graph.

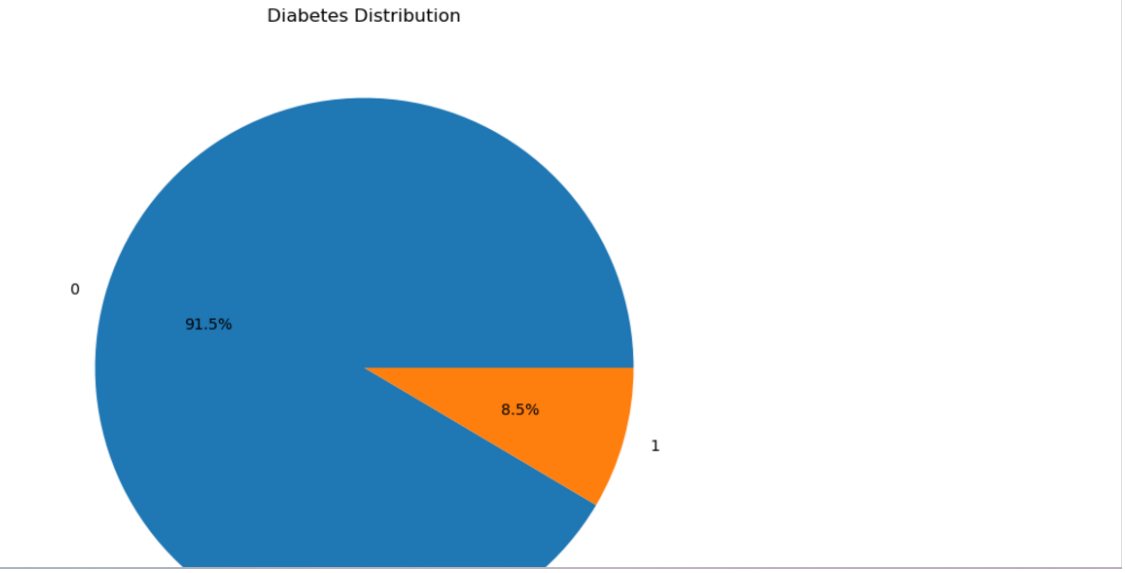
```
In [8]: plt.figure(figsize=(8, 5))
        sns.countplot(data=df, x='gender')
        plt.title("Gender Distribution")
        plt.show()
```



Step 3 : In This step it creates a pie chart of diabetes count between genders in respect to logistic regression of classifying the data into 0 or 1.

```
In [10]: diabetes_counts = df['diabetes'].value_counts()

         # Create a pie chart
         plt.figure(figsize=(8, 8))
         plt.pie(diabetes_counts, labels=diabetes_counts.index, autopct='%1.1f%%')
         plt.title("Diabetes Distribution")
         plt.show()
```



Step 4 : In This step it evaluates the Lasso regression model by computing accuracy score, cross validation score, classification report and confusion matrix using the testing data and obtained **96.235%** accuracy as a result.

```
In [21]: X = df.drop(columns = ["diabetes"])
         Y = df["diabetes"].values.reshape(-1, 1)

In [22]: Y.shape
Out[22]: (100000, 1)

In [23]: X_train, X_test, Y_train, Y_test = train_test_split(X, Y, test_size=0.2, shuffle=True, random_state= 55)

In [24]: model = LogisticRegression()
         model.fit(X_train, Y_train)
Out[24]: LogisticRegression()

In [25]: model.score(X_train, Y_train)
Out[25]: 0.959725

In [26]: y_pred = model.predict(X_test)

In [27]: print(accuracy_score(y_pred, Y_test))
         0.96235

In [28]: print(mean_squared_error(y_pred, Y_test))
         0.03765
```

Conclusion :

The "GlucoPredict" project concludes that logistic regression is an effective method for predicting diabetes risk, as it balances accuracy with simplicity. By using key health data like glucose levels, BMI, age, and family history, the system provides reliable predictions while remaining easy to interpret. This makes it particularly useful in healthcare, where transparency and trust in a tool are essential. The project also emphasizes the importance of preprocessing the data and addressing challenges like missing values and class imbalance to improve the system's performance and fairness.

Ultimately, implementing this proposed system has the potential to lead to earlier detection and more personalized treatment plans for patients. As a result, it could significantly improve patient outcomes while reducing the burden on healthcare resources. By making the diagnostic process more efficient and understandable, Logistic regression can play a vital role in transforming how diabetes is diagnosed and managed in clinical settings and obtained with highest accuracy **96.325%**.

References:

[1]. Aljumah, A.A., Ahamad, M.G., Siddiqui, M.K., 2013. Application of data mining: Diabetes health care in young and old patients. Journal of King Saud University - Computer and Information Sciences 25, 127–136. doi:10.1016/j.jksuci.2012.10.003.

[2]. Arora, R., Suman, 2012. Comparative Analysis of Classification Algorithms on Different Datasets using WEKA. International Journal of Computer Applications 54, 21–25. doi:10.5120/8626-2492.

[3]. Bamnote, M.P., G.R., 2014. Design of Classifier for Detection of Diabetes Mellitus Using Genetic Programming. Advances in Intelligent Systems and Computing 1, 763–770. doi:10.1007/978-3- 319-11933-5.

[4]. Choubey, D.K., Paul, S., Kumar, S., Kumar, S., 2017. Classification of Pima indian diabetes dataset using naive bayes with genetic algorithm as an attribute selection, in: Communication

and Computing Systems: Proceedings of the International Conference on Communication and Computing System (ICCCS 2016), pp. 451– 455.

[5]. Dhomse Kanchan B., M.K.M., 2016. Study of Machine Learning Algorithms for Special Disease Prediction using Principal of Component Analysis, in: 2016 International Conference on Global Trends in Signal Processing, Information Computing and Communication, IEEE. pp. 510.

[6]. Sharief, A.A., Sheta, A., 2014. Developing a Mathematical Model to Detect Diabetes Using Multigene Genetic Programming. International Journal of Advanced Research in Artificial Intelligence (IJARAI) 3, 54–59. doi:doi:10.14569/IJARAI.2014.031007.

[7]. Sisodia, D., Shrivastava, S.K., Jain, R.C., 2010. ISVM for face recognition. Proceedings - 2010 International Conference on Computational Intelligence and Communication Networks, CICN 2010 , 554–559doi:10.1109/CICN.2010.109.

[8]. Sisodia, D., Singh, L., Sisodia, S., 2014. Fast and Accurate Face Recognition Using SVM and DCT, in: Proceedings of the Second International Conference on Soft Computing for Problem Solving (SocProS 2012), December 28-30, 2012, Springer. pp. 1027–1038.

[9]. https://www.kaggle.com/johndasilva/diabetes

[10].Rani, A. S., & Jyothi, S. (2016, March). Performance analysis of classification algorithms under different datasets. In Computing for Sustainable Global Development (INDIACom), 2016 3rd International Conference on (pp. 1584- 1589). IEEE.

ACKNOWLEDGEMENT

We would like to thank Mr.P.V.Viswam Sri mahavishnu, Founder & Chairman of KIET Group of institutions, P.Aiswarya, Dr.D.Revathi, Mr.Y.Ramakrishna, Dr.A.Ravi Kumar and D.Ramkiran for their support and guidance in completing our project. It was a great learning experience. The completion of the project would not have been possible with out their help and insights.