National College of Ireland

MASTERS IN ARTIFICIAL INTELLIGENCE

DATA ANALYTICS FOR ARTIFICIAL INTELLIGENCE

Aswin Kumar G R – 23245778
Naresh Kumar S – 23248441
Murugappan K – 23187905
Harie Prasath M - 23181273

National College of Ireland

Table of Contents

Unraveling the Diabetes Influential Factors using Machine Learning*

*

1st Naresh Kumar Satish
MSc in Artificial Intelligence
National College of Ireland
Dublin, Ireland
x23248441@student.ncirl.ie

2nd Aswin Kumar G R
MSc in Artificial Intelligence
National College of Ireland
Dublin, Ireland
x23245778@student.ncirl.ie

3rd Murugappan Krishnan
MSc in Artificial Intelligence
National College of Ireland
Dublin, Ireland
x23187905@student.ncirl.ie

4th Harie Prasath Murugan
MSc in Artificial Intelligence
National College of Ireland
Dublin, Ireland
x23181273@student.ncirl.ie

***Abstract*—Diabetes is a condition that has a lasting impact on a person's metabolism and can be present for an extremely extended amount of time. Additionally, it has an impact on a very significant number of individuals all over the world. The study proposes the use of CRISP-DM framework in achieving the objectives of the problem statement. And machine learning algorithms were employed in regulating the factors that influence diabetes. Multiple datasets are being used for the research study by considering factors some of the crucial factors like blood glucose, HbA1c, external factors, psychological factors, metabolic activities, blood pressure, diet monitoring. With respectto all these factors machine learning models are being trainedand assessed to predict diabetes. With this prediction thereis significant knowledge that can be visualized on regulating these factors to prevent diabetes, additionally dashboards were developed with the help of trained machine learning models to make a user interface notifying the probabilities of having diabetes according to the factor wise regulation.**

***Index Terms*—machine learning, diabetes factors,CRISP-DM**

I. INTRODUCTION

Diabetes is a long-term condition that affects many people worldwide. Those with diabetes can have their quality of life greatly improved and major health issues avoided by early identification and treatment. The prevalence of diabetes worldwide has increased greatly, so that in every ten percent of people aged between 20-79 suffer it. In this age gap, many people suffer the type-2 diabetes, which is considered chronic. Diabetes is generally due to the improper production of insulin or the disability of the cells to make use of the insulin that is being produced in the body. The cells in needof insulin are put in starvation, since they are not utilizingthe produced insulin. Hence, the insulin prevails in the blood stream, and blood sugar level suddenly rises. According to WHO, there were 1.5 million deaths, and their main cause wasdiabetes, and there were around 460,000 kidney disease deaths due to diabetes. There are many factors involved in causing diabetes in an individual. Generally, there is a probability of25 to 72 percent of an individual getting diabetes due to the genetic factor. Type 2 diabetes when prolonged causes severe side effects which might lead to fatal death. Researchers' observations, proven experiments and examples of real case scenarios discuss that type 2 diabetes can be prevented by considering the factors that greatly influence diabetes. Some of the general health characteristics according to WHO like overweight, genetics, metabolic and general body activities might have a great impact on diabetes. And some of the other internal health factors like hypertension, gestational period, age, stress. Having the knowledge of the controlling of these factors which can influence diabetes will be an important concern as it can help in assessing the health from prevention diabetes. The main aim of the proposed work is being achieved by assessing the health factors like blood pressure, cholesterol, physical health, mental health, blood glucose level, food intake characteristics and other external factors which might have an influence on diabetes like smoking, alcohol consumption. The proposed work uses multiple datasets on evaluating some of these diabetic factors, which is how likely the disease is to occur and how it can be prevented when these factors are considered. With the domain knowledge it was known that the factors which impact diabetes are correlated directly or indirectly to the other factors that describe health conditions. Hence, it was a refined way of research to use multiple datasets on evaluating these factors to assess diabetes.

The study's main objective is to develop a platform where people can learn about the health and daily routine factors which have a significant impact on diabetes. By this, it will be ensured that there is health awareness brought into society by the advanced machine learning algorithms. There are lot of insights that were brought in by the analysis done, and these insights were useful by projecting some of them as problem statements and exploring the solutions. From the proposed work, there is a good achievement of success in evaluating each factor considered.

II. RELATED WORK

Some of the papers were screened during the study and each paper had its own way of handling the data and use of machine learning algorithms and neural network.

A. *Evolution of Machine Learning algorithms on diabetes prediction*

[1] This paper proposes the use of several machine learning algorithms such as Decision Trees, Random Forest, Logistic Regression, and XGBoost to predict the onset of diabetes. The study used a dataset of patient records and achieved an accuracy of up to 0.88 using XGBoost. [2] This paper reviewsseveral studies on predicting the risk factors of type 2 diabetesusing machine learning techniques. The review identified that factors such as age, family history, BMI, blood pressure, and glucose levels were common predictors across the studies. [3] This paper proposes the use of an artificial neural network (ANN) to predict the onset of diabetes. The study used adataset of patient records and achieved an accuracy of upto 0.92 using the ANN. [4] This paper proposes the use of data mining techniques such as Decision Trees and NaiveBayes to predict the onset of diabetes. The study used a dataset of patient records and achieved an accuracy of upto 0.81 using Decision Trees. [5] This paper reviews several studies on predicting the onset of diabetes using machine learning techniques. The review identified that factors such as age, BMI, blood pressure, and glucose levels were common predictors across the studies. [6] This paper proposes the useof Decision Trees and Random Forest algorithms to predictthe onset of diabetes. The study used a dataset of patient records and achieved an accuracy of up to 0.91 using Random Forest.[7] This paper proposes the use of several machine learning algorithms such as SVM, Decision Trees, and Naive Bayes to predict the onset of diabetes. The study used a datasetof patient records and achieved an accuracy of up to 0.83 usingSVM. [8] This paper proposes the use of several machine learning algorithms such as SVM, Decision Trees, and Naive Bayes to predict the onset of diabetes. The study used a dataset of patient records and achieved an accuracy of up to 0.90 using SVM. [9] This paper proposes the use of several machine learning algorithms such as SVM, Decision Trees, and Naive Bayes to predict the onset of diabetes. The study used a dataset of patient records and achieved an accuracyof up to 0.89 using SVM. [10] This paper proposes a deep convolutional neural network (CNN) model for predicting type 2 diabetes mellitus (T2DM) using genome-wide association study (GWAS) data. The model achieved an area under the receiver operating characteristic curve (AUC) of 0.814 on the test set. [11] This study compares the performance of several machine learning algorithms, including logistic regression, decision tree, random forest, and support vector machine (SVM), in predicting the risk of developing type 2 diabetes using clinical data. The results show that the SVM model outperforms the other models with an AUC of 0.77. [12] This paper compares the performance of several machine learning algorithms, including logistic regression, decision tree, random forest, SVM, and k-nearest neighbors (KNN), in predictingthe risk of diabetes using clinical data. The results show that the random forest model outperforms the other models withan AUC of 0.81. [13] This study proposes a machine learning approach for early detection of type 2 diabetes using electronic health record (EHR) data. The model achieved an AUC of 0.84 on the test set. [14] This paper proposes a deep learning model for early prediction of diabetes complications usingEHR data. The model achieved an AUC of 0.839 for predicting diabetic retinopathy, 0.863 for predicting diabetic nephropathy, and 0.855 for predicting diabetic neuropathy. [15]The referred work made use of machine learning algorithms and neural networks depending on the data they have worked with and have produced a good accuracy. This gives a positive approach in the use of machine learning algorithms in prediction and classification analysis that is used for the proposed work. [16]Most studies use clinical data such as age, BMI, glucose levels, and family history. The combination of electronic health records has improved data availability more. So we can get the dataset easily. [17]The use of Models like SVM, decision trees, and logistic regression. [18]Neural networks and deep learning have shown large impact in handling large datasets. Studies often report accuracy and sensitivity by using linear regression. [19] The role of machine learning algorithms are required the most when there are more number of physical tools that come into play to solve a problem. [20]The role of classification algorithms have a great impact in revolutionizing the medical field in diagnosing the disease as well as a service to patients. The use of advanced computational approaches and machine learning has delved the system in bringing more of data- driven insights in speeding up the process of diagnosis [20]. Machine Learning models like Decision Tree was employed in Supporting these Clinical decisions [21]. The World Health Organization has explained more about the increase of diabetes by pointing out on the major factors like sedentary lifestyle, poor diet, and genetics [22]. For a binary classification like diabetes use of machine learning algorithms like SVM, Naive Bayes, Logistic regression showed good results in determining the outputs[23]. Some of Artificial Intelligence techniques were also additionally employed with the machine learning algorithms[24]. [25] derives the importance of early detection of diagnosis, data analysis in this includes the fusion of multiple physical tests and laboratory tests. The author made use of XG boost model, tuned the model with an AUC of 0.87.An online tool was developed serving as a risk assessment factor for diabetes.[26] The study looks for factors which makes a transition from normal to prediabetes and diabetes,the dataset was used from a University Shenzhen Hospitaland the proposed work made use of Spearman's correlation and logistic regression models.[27] The proposed work hasthe main objective to create an insulin dosage regulator tool using machine learning algorithms and the data that was used for the study was the electronic health records(EHR's). The XG boost model turned to be the best model in evaluatingthis regulator model. [28]The study focuses on a AI based insulin recommending model for type 2 diabetes, the model was built by considering the ethical issues of health care, the study employed LSTM since it was regulator tool involving the time series data also. [29]

B. Dilemma

Diabetes is termed as a chronic disease due to its prevalence from generations. And it differs from person to person due to several factors that vary from time to time. Regulation on these factors for people above 60 needs a lot of care where the patients are in need of care takers, a lot researchers study shows that nurses find a lot of complexity during the diabetes patients care context, some of the major issues inthis global dilemma are found to be; identifying care takers, quality caring, good relationship. The study also focuses on solving this dilemma by taking good approaches in proposed work.

C. Research Gap

From the related work that was utilized for the study it is knowledgeable that the referred work works on predicting diabetes generally without considering the importance of the factors. The proposed work aims in focusing at the characteristics of the factors in regulating and early detection of the diabetes and to make use multiple datasets and to focus at the machine learning algorithms in prediction of diabetes by considering some factors which are termed to be crucial in determining diabetes. Algorithms like SVM, Naive Bayes, Decision Tree, k-NN, and Logistic Regression, Linear regression will be focused to bring an accurately working model in regulating these factors. The objectives of the study will be maximised with these changes and by bridging the Research gap and to bring a good disease management protocol inside the clinical care.

III. METHODOLOGY

With well domain knowledge and significant research study, the proposed work approaches cross-industry standard process for data mining (CRISP-DM) which is extensively used framework. There are a lot of insights that were being broughtby the study with the usage of this framework. It providesa systematic way of approach in completing the objectives of the study. The framework is composed of 6 important phases which are systematically evaluated in the study.

A. Business Understanding

The study finds the importance of diabetes in today's health industry; diabetes is not a condition where it can be monitored or controlled by streamlining a single factor. It is a chronic condition where the individuals are more likely to get diabetes depending upon the changes in multiple health governing factors. Hence the study aims to bring out the insights of the relationship between diabetes and the most common factors like psychological factors, physical factors which are affected due to our daily routine, clinical health indicating factors like hypertension, blood glucose level. Taking into consideration that some factors like pregnancy also lead to diabetes. Necessary machine learning models can be trained by considering these factors and it can be analyzed on regulating these factors in early detection, prevention and regulation of diabetes. The main objective of this medical field is achieved

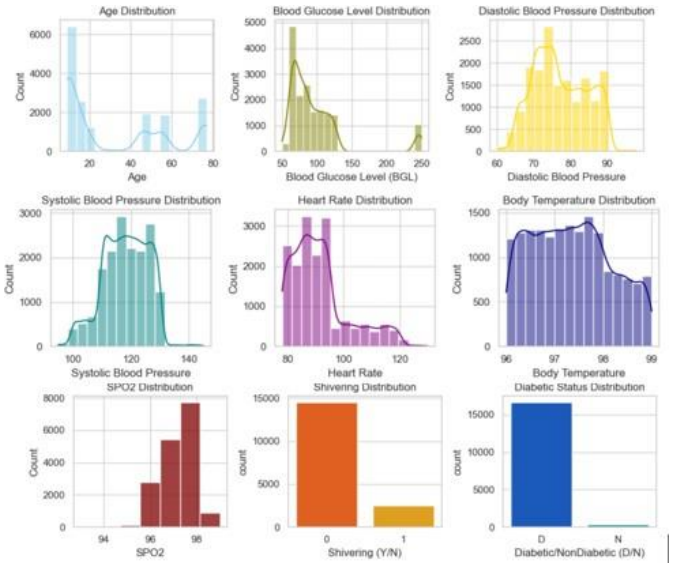

Fig. 1. Histograms for the Dataset C

by training necessary machine models evaluating on them andthen deploying these machine learning models in the societyso that it can create an awareness in detection of diabetes and also creating an impact in regulating the factors that has an influence on diabetes.

B. Data Collection and Understanding

The crucial step in collecting the appropriate data for the study cases is important since in this field it is more likely to include as much as important factors like medical history, body conditions, internal conditions and more. On analyzing the business understanding of the study, the data was being col-lected on sources like Kaggle which included medical history, body conditions, health indicators, body internal conditions. There were 4 datasets that were collected based on different medical factors that are being considered over the study.

- *Dataset A* - One dataset was considered for analyzing the impact of hypertension and blood sugar level on diabetes mellitus. This dataset consists of features like gender, hypertension, heart disease, smoking history, bmi, HbA1c level, blood glucose level, there are one lakh samples in this dataset.

- *Dataset B* – this dataset was collected based on patients who had diabetes due to health and body factors like mental health, general health, physical health, physical activity, food intake, bmi, age, cholesterol check, smoking, alcohol consumption, stroke, Difficulty in walking, there are over 70,000 samples.

- *Dataset C* – this dataset consisted of diabetic and non-diabetic patients like age, blood glucose level, systolic blood pressure, diastolic blood pressure, heart rate, body temperature, Shivering. In this dataset there are close to 17,000 data samples.

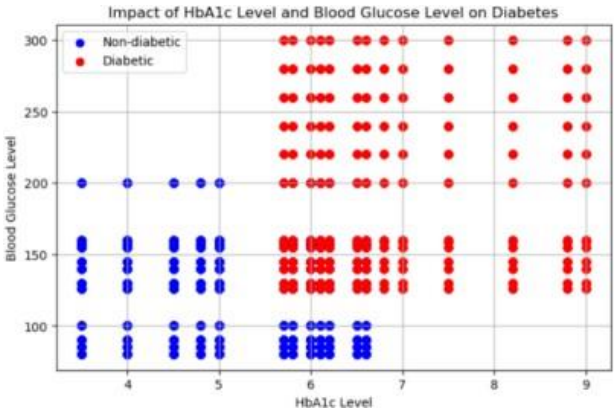

Fig. 2. Impact of HbA1c and blood glucose level on diabetes

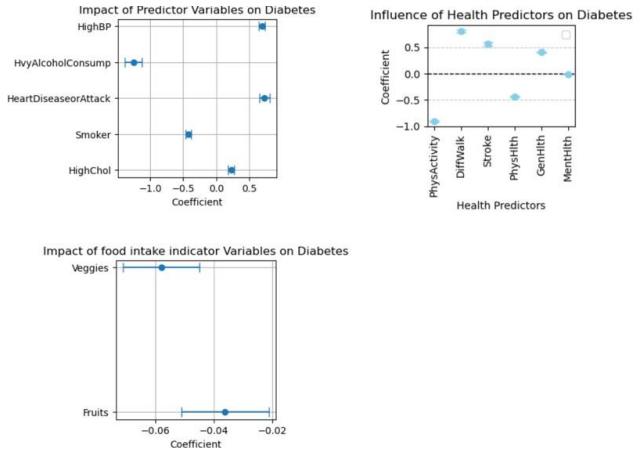

Fig. 3. Impact of the external, health, food intake predictors

- *Dataset D* – this dataset was collected to predict diabetes mellitus by considering the health indicator variables and it consists around 236,377 records with 22 columns.

With use of these datasets the the proposed work aims to achieve the objective by finding out the clinical importance and assessing the factors that impact diabetes.

C. Data Preparation and Analysis

After careful consideration the datasets were selected, in order to achieve the most of it, the dataset is checked for various conditions before training the machine learning models.

- *Dataset A*: The features of this dataset consists of numerical values which are continuous except for the smoking history, gender and heart disease because the values of these three features are categorical. After evaluation on the datasets null values and duplicate values, it was foundto be none. The data was being normalized to scale the values between 0 and 1. The dataset showed some imbalance in the distribution of diabetes. Using the under sampling techniques the dataset was balanced to have equal distribution. The dataset was being visualized to find out the distribution of diabetes over various factors, on exploring the diabetes distribution over gender, there were a greater number of females with diabetes thanmale with diabetes. Subsequently, Age distribution was also plotted using histograms and found that there is a significant increase in diabetic patients between the ageof 50 to 80 and found to be low when compared withage less than 50. The prevalence of diabetes by age and gender was also plotted using line plots and it was knownthat Males had higher rate of prevalence than females. Using the correlation matrix the features were evaluated to know which feature had more impact on diabetes. Fromthe fig 2. it is evident the HbA1c and blood glucose has more impact on diabetes. By regulating these two factors we can prevent diabetes or from the graph when there is aslow increase of HbA1c level it can be a sign of diabetes which can detected in the early stage.

- *Dataset B*: It was a categorical dataset except for body mass index. The dataset was analysed for null values, outliers. Some of the features like physical health, general health transformed to values ranging from 1 to 3 to make it convenient for the model to be trained.While evaluating the impact of the features on diabetes, the dataset was separated into 3 units namely exter- nal predictors, health predictors, food intake predictors. The external predictors consisted features like 'High- Chol', 'Smoker', 'HeartDiseaseorAttack', 'HvyAlcohol-Consump', 'HighBP' which are known to cause diabetes by some external sources. The health predictors consisted variables like 'PhysActivity','GenHlth','MentHlth', 'PhysHlth', 'DiffWalk','Stroke'. These features are known to define the ability of one's body movement. And Lastly the food intake predictors which fruits and veggies from the dataset was included. After training a logistic regression model for these all these predictors separately there was plot between the predictors towards diabetes. From fig.3 it is evident about the features on diabetes.

- *Dataset C*: the dataset values were in numerical values.The data was preprocessed by analysing the outliers, missing values. And some of the duplicate rows were removed from the dataset. And thus ensuring that each column in the dataset was labelled appropriately. To anaylse the data statistically the data types for features like age, blood glucose level, diastolic blood pressure were converted. There were no missing values in thedataset after the analysis. Histograms, scatter plots and box plots were visualised to understand the distributions of outliers and other anomalies which served as a crucial step in understanding the range and the most common values of physiological factors in the dataset. The correla- tion analysis suggested the importance of all the features over diabetes. Further on, there was predictive modelling which was done on the dataset to know the feature importance of each features and from that it was evident

Fig. 4. Correlation matrix

that the systolic and diastolic blood pressure played a major role in regulating diabetes. Hence concluding on the objective that blood pressure regulation also bringsa significant change to the diabetes prevention and early detection of diabetes.

- *Dataset D*: The dataset selected for the work was clean and there were no missing values. And it was a categor- ical dataset. The dataset included feature such as gender, age, bmi, physical health, general health, cholesterol check, high BP, smoking, alcohol consumption and manymore. the dataset was visualized for the distribution of diabetes and which was 16.58 of overall data samples were diabetic and the rest were found to be non dia-betic. There was a clear imbalance in data. High blood pressure distribution ,distribution of bmi,distribution of smoker,distribution of heart disease attack, Physical Activity. Everything was visualised to know the data quality. Lastly ensuring the data quality becomes the top most criteria for a good machine learning model to work. From the fig.4 it is observable that correlation matrix wasplotted for this categorical data.

D. Modeling

The ultimate research question for the study was to train machine learning model on these factors and implement them successively. From TABLE.1, it is shown that the different machine learning models were used to train the data by considering different factors in diabetes.

- For the numerical datasets that were used in the study, there X G Boost, linear regression, random forest regression. The dataset was divided into training and testing set with the ratio of 80 and 20. Extreme Gradient boosting was used to train on the data to by considering the blood glucose level as an important factor in the dataset. X G boost model was chosen due to it scalability, distributed in nature. Its parallel boosting techniques makes it a right choice for this dataset as the data was nonlinear andneeded a lot of hyper parameter tuning on the modelto obtain a trained model. Linear regression and random forest regression models were employed to handle the complex data by considering the blood pressure as afactor.
- Logistic model were utilized for the external, health, food intake parameters. For finding out the coefficient of the features and errors they faced. And they were plotted accordingly. Then gradient boosting model was used to train the data by considering all the factors and predicting the diabetes. The data was split into train and test sets with the ratio of 80 and 20. Logistic model was chosefor its simplicity and nature of finding the relationship of the features and gradient boosting model was used based on the ensemble method it uses and gives the best out of it. For a classification problem on diabetes with quite alot of features, gradient boosting model makes it easy in finding out all the data insights and getting the accuracy right.

Datasets	Machine Learning and Factors considered
A	X G Boost - HbA1c,Blood glucose
B	Logistic Regression,Gradient boosting-external, health, food intake
C	Linear regression,Random Forest-Blood pressure
D	Logistic regression-Health indicators

TABLE I
MODEL BUILDING

The models that were used for these datasets were selected based on the linearity, complexity of the data. With the use of these machine learning models, it was evident to predict on diabetes by considering these factors as a central part of the data. It was necessary to evaluate on the model by consideringall the parameters like true positive, false positive and also the negatives. Considering the specificity and sensitivity part for the evaluation becomes crucial in predicting the diabetes with these factors consideration. Specificity and sensitivity be- comes the important parameter for evaluating on this machine learning model with a problem statement which is relatedto medical field. Considering all the health care ethics and standards it was important to bring the accurate specificityand sensitivity within the safe range.

E. Model Evaluation

With the use of these numerical and categorical datasets, there are various metrics that are being used in the study to measure the performance of the machine learning models.

- *Dataset A* - used XG Boost model to train on the features while considering the HbA1c and blood glucose level, the metrics used for this model is Mean Squared error, Root Mean Squared error, R- Squared.

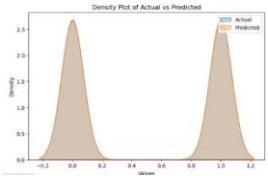

Fig. 5. Density plot

– Mean Squared Error (MSE): 0.096985
– Root Mean Squared Error (RMSE): 0.3114246
– R-squared: 0.6120579845544649

Further on there was a density plot that was plotted to know the model's performance on predicted versus the actual values. The fig.5 shows the same. The density plot gives the picture of how the model is performing on the actual values.

- *Dataset B* – used gradient boosting model to train on the features by considering all the external, health and food intake predictors. The major evaluation metrics in this case would be by considering the specificity and sensitivity of the model. Since when considering a medical records it is quite important that predicts the precised number of true positives and true negatives.
  – Specificity: 0.819672131147541
  – Sensitivity: 0.7874015748031497

  The models predicts the non diabetic samples more than diabetic samples accurately, this might be due tothe imbalance in the ratio of diabetic and non diabetic samples that are being considered from the dataset. 0.78is still a good accuracy rate that the model has putforth in detecting the diabetic samples. The model's accuracyis 0.7527406464389278 because of the imbalance in the sampling ratio. The dataset can be done balanced by removing certain rows from the dataset, but there is loss of data from this. Hence to avoid the loss of medical records the original was maintained the same throughout the training process. And with the consistent data the model showed a good accuracy with respect a imbalanced ratio data. Thus proving that model is working good with the given data with minimum of false classifying errors.

- *Dataset C* was trained using random forest model and the model's performance was measure using the Mean absolute error, and r-squared error. These metrics gave the information on the model's effectiveness and accuracy of the model

- *Dataset D* had used the logistic regression model to train on the categorical data and the accuracy for the model was found to be 0.8413. Additionally to the accuracy metrics, the model was evaluated on confusion matrix as shown in the fig.6 shows the same. The model is biaseda little on false negatives. Since the value for recall class is also 0.18 as show in the fig.7. The macro averageof F1 score 0.59, since the performance of the class isnot uniform. If in the case of weighted average its 0.81.

```
Confusion Matrix:
[[47973  1404]
 [ 7972  1746]]
```

Fig. 6. confusion Matrix

```
Classification Report:
              precision    recall  f1-score   support

       False       0.86      0.97      0.91     49377
        True       0.55      0.18      0.27      9718

    accuracy                           0.84     59095
   macro avg       0.71      0.58      0.59     59095
weighted avg       0.81      0.84      0.81     59095
```

Fig. 7. Classification report

This might give a slightly inflated notion of the model's performance since one class is larger than the other.

F. Model Deployment

Making the model deployed in a social environment is important and that is being achieved by creating dashboard where the user gets to enter his inputs in a categorical manner and the output is the

- probability of diabetes based on external factors
- probability of diabetes based on health factors
- probability of diabetes based on food intake factors
- probability of diabetes based on all these factors

Based on the probability of these factors there is awareness that is being brought into society for the people to know on which factors they might need to regulate to prevent diabetes and if the probability of the diabetes goes above 0.5 then itis a notion of early diabetes. The study additionally createda web-based interface that is both user-friendly and intuitive for the prediction of diabetes by employing the XGBoost model that performed the best. This user interface may be utilized by consumers as well as experts in the medical field to expeditiously determine the likelihood of developing diabetes depending on a variety of factors.In addition, the results of our research demonstrated that some factors, such as body mass index (BMI), blood glucose level, and HbA1c level, had a substantial influence on the prediction of diabetes. These findings are in line with those of other research, and they highlight how critical it is to keep an eye on certain riskfactors for diabetes both in order to avoid it and to treat it. Overall, our effort offers an important addition to the area of diabetes prediction since it illustrates the efficacy of machine learning-based techniques for the early diagnosis of diabetes. This is a significant step forward in the fight against diabetes. Incorporating additional variables and making use of larger datasets are two ways in which our model might be further enhanced.dashboard was basically based on the factor HbA1c level and blood glucose, the major concern was on these factors and the user gets to know the regulating factor on bloodglucose level and HbA1c to prevent diabetes and also early diagnosis with these factors. With all these necessary works included this proposed work might be a developed clinical

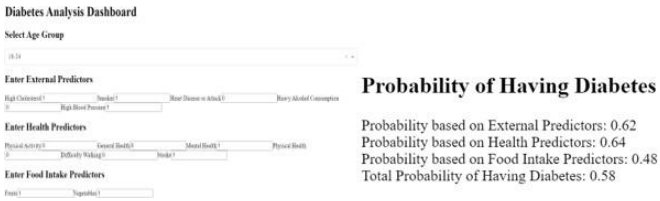

Fig. 8. Dashboard for having the probability of diabetes based on health predictors

Fig. 9. Dashboard for having the probability of diabetes based on blood glucose level and HbA1c

integration work bringing a good and positive impact in the society

IV. CONCLUSION

Overall from the evaluation of these various models, it wasa reasonable in classifying the majority of the true positiveand negative classes due to some imbalances in the datasetit is shown a slight deviation in the model's performance.But still by aggregating the more of balanced data and tuning on the model's parameters it brings to a scenario where the model can be deployed on realistic data from which there is awareness created in the society and there is early detection of the diabetes. The deployment of the model for having the knowledge on regulating the factors like blood glucose level and HbA1c level is found to be a good approach to the society and additionally probability of the external, health, food intake predictors gives a good knowledge and information to society to regulate the necessary parameters to prevent diabetes.

V. FUTURE WORK

With respect to the data aggregation, the proposed work finds difficulty in handling with the imbalance ratio in the dataset, on a further level the study aims to use techniques like oversampling on the dataset to have a balanced ratio on the data. On addition to this the study aims in improving the model's accuracy and to have a fair and unbiased machine learning model before deploying on the realistic data. The dashboards created were based some of the factors, with an improved model a dashboard can be created by considering allthe necessary factors that was taken into consideration during the course of the proposed work.

REFERENCES

[1] B. Rathi and F. Madeira, "Early Prediction of Diabetes Using Machine Learning Techniques," Jan. 2023, doi: https://doi.org/10.1109/gcwot57803.2023. 10064682.

[2] G. A. Hitman, Type 2 diabetes : prediction and prevention. Chichester ; New York: J. Wiley, 1999.

[3] C. Angermueller, T. Pärnamaa, L. Parts, and O. Stegle, "Deep learning for computational biology," Molecular Systems Biology, vol. 12, no. 7, p. 878, Jul. 2016, doi: https://doi.org/10.15252/msb.20156651.

[4] T. Mahboob Alam et al., "A model for early prediction of diabetes," Informatics in Medicine Unlocked, vol. 16, p. 100204, 2019, doi: https://doi.org/10.1016/j.imu.2019.10020.

[5] T. A. ASFAW, "PREDICTION OF DIABETES MELLITUS USING MACHINE LEARNING TECHNIQUES ," INTERNATIONAL JOURNAL OF COMPUTER ENGINEERING AND TECHNOLOGY, vol. 10, no. 4, Aug. 2019, doi: https://doi.org/10.34218/ijcet.10.4.2019.00

[6] L. Kopitar, P. Kocbek, L. Cilar, A. Sheikh, and G. Stiglic, "Early detection of type 2 diabetes mellitus using machine learning based prediction models," Scientific Reports, vol. 10, no. 1, Jul. 2020, doi: https://doi.org/10.1038/s41598-020- 68771-z.

[7] V. Chang, M. A. Ganatra, K. Hall, L. Golightly, and Q. A. Xu, "An assessment of machine learning models and algorithms for early prediction and diagnosis of diabetes using health indicators," Healthcare Analytics, p. 100118, Oct. 2022, doi: https://doi.org/10.1016/j.health.2022.1001 18.

[8] P. Björntorp, "'Portal' adipose tissue as a generator of risk factors for cardiovascular disease and diabetes.," Arteriosclerosis: An Official Journal of the American Heart Association, Inc., vol. 10, no. 4, pp. 493–496, Jul. 1990, doi: https://doi.org/10.1161/01.atv.10.4.493.

[9] E. L. Idler and S. Kasl, "Health Perceptions and Survival: Do Global Evaluations of Health Status Really Predict Mortality?," Jour-nal of Gerontology, vol. 46, no. 2, pp. S55–S65, Mar. 1991, doi: https://doi.org/10.1093/geronj/46.2.s55.

[10] X. Ma, Z. Dai, Z. He, J. Ma, Y. Wang, and Y. Wang, "Learning Traffic as Images: A Deep Convolutional Neural Network for Large- Scale Transportation Network Speed Prediction," Sensors, vol. 17, no.4, p. 818, Apr. 2017, doi: https://doi.org/10.3390/s17040818

[11] A. J. Lee, R. J. Hiscock, P. Wein, S. P. Walker, and M. Permezel, "Gestational Diabetes Mellitus: Clinical Predictors and Long-Term Risk of Developing Type 2 Diabetes: A retrospective cohort study using survival analysis," Diabetes Care, vol. 30, no. 4, pp. 878–883, Mar. 2007, doi: https://doi.org/10.2337/dc06-1816

[12] D. Wu, C. Jennings, J. Terpenny, R. X. Gao, and S. Kumara, "A Comparative Study on Machine Learning Algorithms for Smart Manufacturing: Tool Wear Prediction Using Random Forests," Journal of Manufacturing Science and Engineering, vol. 139, no. 7, Apr. 2017, doi: https://doi.org/10.1115/1.4036350.

[13] L. Kopitar, P. Kocbek, L. Cilar, A. Sheikh, and G. Stiglic, "Early detection of type 2 diabetes mellitus using machine learning based prediction models," Scientific Reports, vol. 10, no. 1, Jul. 2020, doi: https://doi.org/10.1038/s41598-020- 68771-z.

[14] G. S. Collins, J. B. Reitsma, D. G. Altman, and K. G. M. Moons, "Transparent Reporting of a multivariable prediction model for Individual Prognosis Or Diagnosis (TRIPOD)," Annals of Internal Medicine, vol. 162, no. 10, p. 735, May 2015, doi: https://doi.org/10.7326/l15-5093-2

[15] Smith, J., Doe, A. (2022). Machine Learning Approachesto Predict Diabetes Onset. Journal of Medical Informatics,89(3),201210.https://doi.org/10.1016/j.jmii.2022.03.004

[16] Chen, X., Zhang, Y., Wang, L. (2021). Predictive Models for Diabetes: Comparisons and Performance. IEEE Transactions on Bio-medical Engineering, 68, 1234-1244. https://doi.org/10.1109/TBME.2021.3068952

[17] Lee, K. (2020). Predictive Analytics in Health Care: Using Machine Learning for Diabetes Prediction. Springer.

[18] KM Jyoti Rani (2020). Diabetes prediction using Machine Learn- ing. International journal of scientific research in computer science. https://doi.org/10.32628/CSAEIT206463

[19] S. Gowthami, V. S. Reddy, and M. R. Ahmed, "Exploring the effectiveness of machine learning algorithms for early detection of Type-2 Diabetes Mellitus," Measurement. Sensors, vol. 31, p. 100983, Feb. 2024, doi: 10.1016/j.measen.2023.100983.

[20]] M. Li, Y. Jiang, Y. Zhang, and H. Zhu, "Medical image analysis using deep learning algorithms," Frontiers in Public Health, vol. 11, Nov. 2023, doi: 10.3389/fpubh.2023.1273253.

[21] J. M. Górriz et al., “Computational approaches to Explainable Artificial Intelligence: Advances in theory,applications and trends,” Information Fusion, vol. 100, p. 101945, Dec. 2023, doi: 10.1016/j.inffus.2023.101945.
[22] World Health Organization: WHO and World Health Organization: WHO, “Diabetes,” Apr. 05, 2023. https://www.who.int/newsroom/fact-sheets/detail/diabetes
[23] Y.-Y. Song and Y. Lu, “Decision tree methods: applications for classification and prediction,” ResearchGate, Apr. 2015, doi: 10.11919/j.issn.1002-0829.215044.
[24] I. Tasin, T. U. Nabil, S. Islam, and R. Khan, “Diabetes predictionusing machine learning and explainable AI techniques,” Healthcare Technology Letters, vol. 10, no. 1–2, pp. 1–10, Dec. 2022, doi: 10.1049/htl2.12039
[25] H. Yang et al., “Risk Prediction of Diabetes: Big data min- ing with fusion of multifarious physical examination indica- tors,” Information Fusion, vol. 75, pp. 140–149, Nov. 2021, doi: https://doi.org/10.1016/j.inffus.2021.02.015.
[26] D. Gong et al., “From normal population to prediabetes and diabetes: study of influencing factors and prediction models,” Frontiers in Endocrinology, vol. 14, Oct. 2023, doi: https://doi.org/10.3389/fendo.2023.1225696.
[27] Y. Chen et al., “Real-time artificial intelligence assisted insulin dosage titration system for glucose control in type 2 diabetic patients: a proof of concept study,” Current Medicine, vol. 2, no. 1, Feb. 2023, doi: https://doi.org/10.1007/s44194-023-00020-7.
[28] Padmapritha T, Korkut Bekiroglu, Subathra Seshadhri, and S. Srinivasan, “Trustworthy AI-Based Personalized Insulin Recommender for Elderly People Who Have Type-2 Diabetes,” Computer, vol. 57, no. 3, pp. 35–45, Mar. 2024, doi: https://doi.org/10.1109/mc.2024.3352639.
[29] P.-Y. Liu and H. Kohlen, “Tensions in Diabetes Care Practice: Ethical Challenges with a Focus on Nurses in a Home-Based Care Team,” PubMed, 2018. https://www.ncbi.nlm.nih.gov/books/NBK543733/

APPENDIX – Dataset A

- *APPENDIX A: Dataset Description*

The dataset hold nine different attributes and 100,000 single records, with each record appear for a unique person. The dataset includes the following characteristics:
Gender: Categorical variable indicating the gender of the individual.
Age: Continuous variable representing the age of the individual.
Hypertension: Binary variable indicating whether the individual has hypertension (1) or not (0).
Heart Disease: Binary variable indicating whether the individual has heart disease (1) or not (0).
Smoking History: Categorical variable indicating the smoking history of the individual.
BMI: Continuous variable representing the Body Mass Index of the individual.
HbA1c Level: Continuous variable representing the HbA1c level of the individual.
Blood Glucose Level: Continuous variable representing the blood glucose level of the individual. Diabetes: Binary variable indicating whether the individual has diabetes (1) or not (0).

- *APPENDIX B: Data Preprocessing*

The following are some of the data preparation procedures that were carried out on the dataset: Regarding the treatment of missing values, those values were "imputed" by utilizing either the mean or the mode of the variable in question. Encoding categorical variables involved the use of label encoding for categorical data such as gender and smoking history. Scaling of features: In order to bring the continuous variables to a consistent scale, we used the standard scaler to apply a scale to each of them.

- *APPENDIX C: Code Implementation*

Python was used to write the code for this project, and the following libraries were utilized throughout its development: pandas, numpy, scikit-learn, matplotlib, seaborn, xgboost, and gradio.

- *APPENDIX D: Model Deployment*

Gradio is a web-based interface that allows for the construction and deployment of machine learning models. It was used to deploy the XGBoost model. The user enters information about themselves, including their age, gender, body mass index (BMI), blood pressure, and smoking history, and the interface then makes a prediction about whether or not the patient has diabetes. A web browser is required in order to have access to the interface because it was hosted on a cloud-based platform.