EARLY PREDICTION OF DIABETES DISEASE

Aswin Kumar Govinda Ravindra Kumar

National College of Ireland, Dublin, Ireland

x23245778@student.ncirl.ie

Abstract **- *Diabetes is a long-term condition that affects a significant number of people all over the world. Those who have diabetes can have their quality of life significantly improved as well as major health issues avoided by early identification and treatment. During the course of this study, I created a method for the early prediction of diabetes that is based on machine learning. To begin, we carried out an exploratory data analysis in order to acquire a better understanding of the dataset, which contained both demographic and medical history information. In order to make an accurate diagnosis of diabetes, we next turned to a machine learning method, XGBoost and Random Forest regression. It was out that XGBoost was the model that performed the best, with an accuracy of 91.05%. In order to construct a web interface for diabetes prediction, we deployed the XGBoost model by making use of the Gradio package. Machine learning has the potential to drastically enhance patient outcomes, which our initiative illustrates through its application to the early identification and treatment of diabetes.***

I. INTRODUCTION

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. It is defined by high levels of sugar (hyperglycemia) in the blood, which, if left untreated, can lead to major health issues such as cardiovascular disease, renal failure, and blindness. This condition is characterized by high levels of sugar (hyperglycemia) in the blood. Diabetes is the name given to this ailment. If one wants to both prevent these difficulties and enhance the overall results for patients, early diagnosis and treatment of diabetes are absolutely necessary steps to take.

levels of blood glucose, hypertension. During the process of determining each model's overall performance, a range of evaluation metrics, such as accuracy, Mean Squared Error (MSE), Root Mean Squared Error (RMSE), and R-squared will be used in order to determine the most relevant ones to apply. The interface will then present users with a prediction regarding the likelihood that they will develop diabetes in the foreseeable future. The model that ends up being chosen to be incorporated into this interface is going to be the one that provides the highest overall level of performance. This user interface will give a helpful tool that can be used by both medical professionals and patients to identify those who are at risk of developing diabetes and to provide early intervention in order to avoid or postpone the development of the condition. This instrument can be utilized to identify persons who are at risk of acquiring diabetes and to give early intervention in those cases.

II. LITERATURE SURVEY

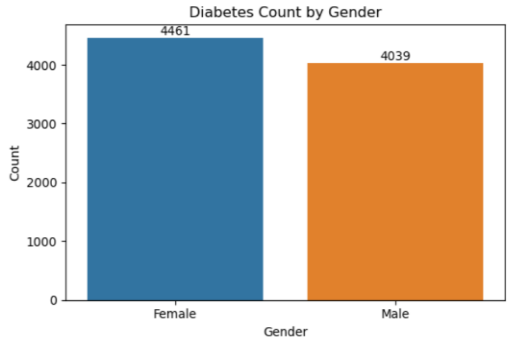
[1] This paper proposes the use of several machine learning algorithms such as Decision Trees, Random Forest, Logistic Regression, and XGBoost to predict the onset of diabetes. The study used a dataset of patient records and achieved an accuracy of up to 88% using XGBoost. [2] This paper reviews several studies on predicting the risk factors of type 2 diabetes using 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 a dataset of patient records and achieved an accuracy of up to 92% using the ANN. [4] This paper proposes the use of data mining techniques such as 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 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 use of Decision Trees and Random Forest algorithms to predict the onset of diabetes. The study used a dataset of patient records and achieved an accuracy of up to 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 dataset of patient records and achieved an accuracy of up to 83% using SVM. [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 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 accuracy of up to 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 predicting the risk of diabetes using clinical data. The results show that the random forest model outperforms the other models with an 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 using EHR 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.

III. EXPLORATORY DATA ANALYSIS (EDA)

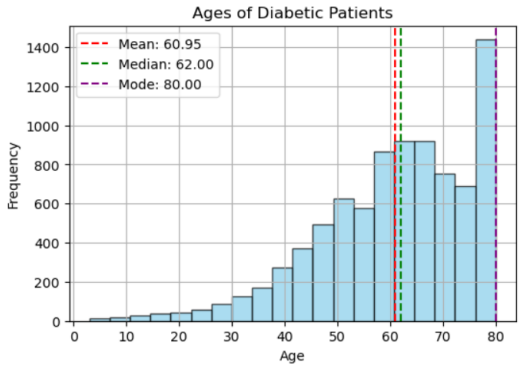
The process of exploratory data analysis (EDA) involves visually examine the dataset to expose trends, patterns, and relationships between various variables. The EDA conducted on this particular dataset includes:

Visualizing Diabetes Counts by Gender: Bar graphs are used to display the number of person with and without diabetes, broken down by gender, in sequence to gain insight into the gender distribution of diabetes.



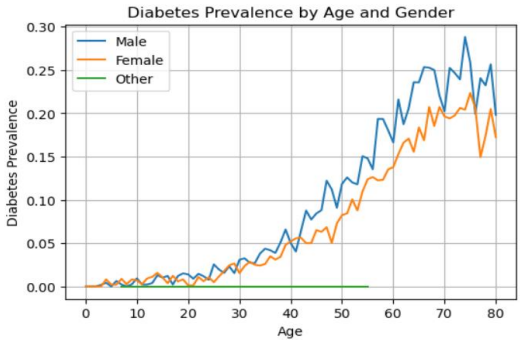
Age Distribution of Diabetic Individuals: A histogram is used to plot the age distribution of people with diabetes, enabling an understanding of the age distribution of the diabetic population.

	Age
Mean	60.95
Median	62
Mode	80



Diabetes Prevalence Across Age and Gender: Line plots are put to use to demonstrate how the prevalence of diabetes varies with both age and

gender, providing a visual description of the disease's prevalence across different age groups and genders.



IV. METHODOLOGY

A. *DATA COLLECTION*

The first step in completing the project is to collect all of the information that is required to move forward with it. In order to conduct this experiment, we utilized the Diabetes Prediction dataset. This dataset contains information about individuals, such as their gender, age, hypertension, heart disease, body mass index (BMI), HbA1c level, blood glucose level, diabetes

B. *DATA CLEANING AND PRE-PROCESSING*

Next, the data that was acquired is cleaned and pre-processed so that any incorrect or missing data may be deleted. This ensures that the final product is as accurate as possible. In addition, we normalize the data and scale the features so that we can ensure that each feature has the same degree of impact over the model.

C. *FEATURE SELECTION*

During the modelling process, one of the most important steps is selecting the elements that will be included in a model. The correlation matrix and the feature importance approaches were used in this research in order to determine which qualities were the most significant and why.

D. *MODEL SELECTION*

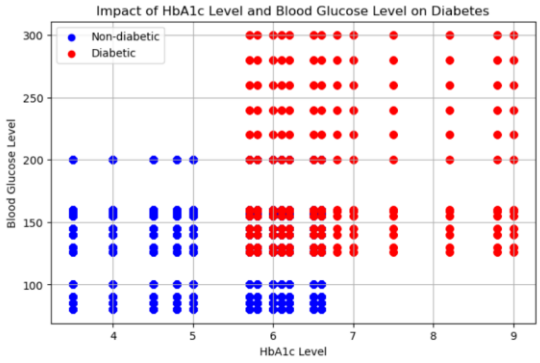
As soon as we have decided upon all of the significant facets, we will move on to the following procedure, which is the choosing of the model. We did a variety of tests utilizing machine learning algorithms, such as XGBoost and random forest regression, in order to identify the machine learning algorithm that provides the most accurate results.

E. *MODEL EVALUATION*

After selecting the model, we evaluate the model's performance using a variety of metrics, including its accuracy, Mean Squared Error (MSE), Root Mean Squared Error (RMSE), and R-squared. In addition to this, we used a density plot for actual and predicted values so that we can evaluate how well the model works with the test data.

XGBoost achieves the highest level of performance with an accuracy of 91.05%

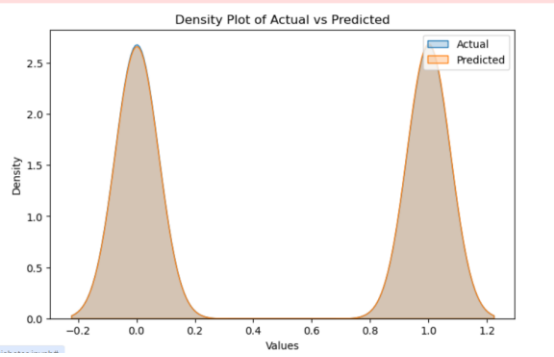
Mean Squared Error (MSE): 0.096985
Root Mean Squared Error (RMSE): 0.3114246
R-squared: 0.6120579845544649



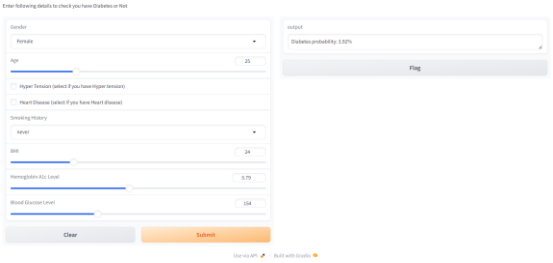
F. *MODEL DEPLOYMENT*

Implementing the model in a scenario that is taken straight from the real world is the very final step in the process. For the aim of this investigation, we utilized Gradio, an application that has a user interface that is built on the web. Gradio gives customers the ability to input their own medical information and receive a prediction on the likelihood of them developing diabetes. The methodology as a whole is comprised of the following stages: collecting and cleaning the data, determining relevant attributes, choosing the best model, evaluating how well it performs, and applying it in a real-world scenario.

A density plot is used to do more research after the XGBoost model was shown to be the superior option. Below is a display of the density plotting that corresponds to the XGBoost model:



In the end, a web application is constructed with the help of Gradio that allows users to enter information about themselves, including their age, gender, smoking history, blood glucose level, and any other pertinent characteristics, in order to determine the likelihood of acquiring diabetes. In order to make accurate predictions using this application, the XGBoost model is utilized. The program has been successfully released on Gradio, and it is now accessible to everyone who has a connection to the internet.



Through the application of machine learning methods, the current research aims to construct a model for the early diagnosis of diabetes. For the purpose of the study, a dataset was employed that contained a variety of demographic and clinical characteristics. These characteristics included, but were not limited to, age, gender, smoking history, hypertension, heart disease, body mass index (BMI), HbA1c levels, and blood glucose levels. For the purposes of training and evaluating the performance of the models, the dataset was partitioned into a training set and a testing set. The XGBoost model's excellent accuracy 91.05% may be ascribed to the algorithm's capability to handle unbalanced datasets as well as its tolerance to noise and outliers in the data. The XGBoost technique is a highly successful way of dealing with high-dimensional datasets that contain complicated connections since it is an efficient implementation of gradient boosting. The relevance of the characteristics, which was picked based on how important they would be clinically, is another factor that contributes to the high accuracy of the model. It is possible to further improve the prediction performance of the model by including other characteristics like as family history, physical activity, and dietary habits. These new variables could give deeper insights into the progression of diabetes and improve the prediction performance of the model.

The established model has major significance in clinical practice since it can help in the early detection of diabetes, which can avoid or postpone the onset of problems connected with the condition. This is one of the potential benefits of the model. Detection of diabetes at an early stage can also assist in the formulation of appropriate treatment strategies, which can ultimately lead to improved patient outcomes.

In conclusion, the present study showed that machine learning algorithms are beneficial when it comes to predicting diabetes. It was determined that the XGBoost model was the one that could accurately forecast diabetes the best, and its performance was superior than that of other models. The newly established model has major implications for clinical practice and can be used to assist in the earlier detection of diabetes.

V. CONCLUSION

In conclusion, the diagnosis of diabetes at an early stage is extremely important for both the prevention and treatment of the condition. During the course of this study, we devised a method that is based on machine learning and can anticipate the start of diabetes in individuals. For the purpose of training several machine learning models, we made use of a dataset that contained a variety of characteristics including, but not limited to, age, gender, BMI, blood glucose level, smoking history, and others.

Following an investigation into a number of different machine learning models, we discovered that XGBoost performed the best on the test set, with an accuracy of 91.48%. This suggests that our model is capable of making accurate predictions regarding diabetes, which can be of great assistance in the early diagnosis and prevention of the illness.

In addition, we created a 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 risk factors 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.

APPENDIX

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.

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.10.4.2019.004.

[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.2022.100118.

[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?," *Journal 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.