The Chang School
of Continuing
Education

Literature Review, Data Description, and Approach for the project

"Diabetes prediction using Machine Learning Techniques."

Course:CIND 820

Submitted by: Muhammad Maqsud Hussain
Ryerson ID: 500978290

Submitted to: Professor Dr. Ceni Babaoglo

Date: October 20, 2023

Toronto Metroploitan University

Abstract

Diabetes mellitus refers to a group of diseases that affect how the body uses blood glucose. Glucose is an important source of energy for the cells that make up the muscles and tissues. It's also the brain's main source of fuel. Diabetes can lead to excess sugar in the blood. Too much sugar in the blood can lead to serious health problems, such as heart disease, vision loss, and kidney diseases etc. As per the International Diabetes Federation, 537 million adults (20-79 years) are living with diabetes and there are more than 5.7 million Canadians living with diabetes. Machine learning is a branch of artificial intelligence (AI) and computer science which focuses on the use of data and algorithms to imitate the way that humans learn, gradually improving its accuracy. In this project, the aim is to develop a system to predict diabetes early for a patient using different machine learning techniques. The dataset is called "Diabetes prediction dataset" and it has 100,000 rows that consists of medical and demographic data collected from different patients along with their diabetes status (positive or negative). Besides, early detection of diabetes, analysis will be performed to find some of the related research questions which are:

- Which factors are highly correlated with positive diagnosis of diabetes?
- Does probability of being a diabetic vary with age and/or gender?
- How smoking history is related to the probability of becoming a diabetic patient in future?
- Is blood sugar level of an obese/overweight person always higher than a person with normal BMI?
- How accurately can diabetes be predicted, and which model works best?

At least 3-4 Machine Learning techniques will be adopted for this dataset to predict diabetes. Finally, results of the ML models will be compared to identify which one works better and why. Proposed algorithms to use for this project are logistic regression, k-nearest neighbor classifier, decision tree classifier, random forest classifier, support vector model classifier etc. Python with its different software library such as pandas will be used predominantly for performing the analysis. If the scope of the project permits, an interactive Tableau dashboard may be developed for visualization of the results and findings.

Literature Review

Several scholars used the machine learning (ML) method to predict diabetes using different datasets over the time. Due to the increasing trend of diabetes patients all over the world, researchers and data scientists from different part of the world have taken different approach to find a method for early prediction of this disease.

Soni & Varma, 2020 used six machine learning techniques in their project which are Which are K-Nearest Neighbor (KNN), Logistic Regression (LR), Decision Tree (DT), Support Vector Machine (SVM), Gradient Boosting (GB) and Random Forest (RF). They used a dataset with eight attributes and 768 instances. The dataset is called Pima Indian Diabetes (PID) and it contains attributes about female only. After applying the machine learning techniques, they measured the accuracy of the results of each method and Random Forest classifier achieved highest accuracy (77%). Also, importance of each feature has been measured and Glucose is found to the most important among the eight features considered in their project.

However, the paper titled *"A comparison of machine learning algorithms for diabetes prediction."* (Khanam & Foo, 2021) took a more detailed approach with the same dataset. They additionally used Neural Network (NN) methods in their research besides seven ML techniques to predict diabetes. They adopted feature reduction method and finally kept five input features (Glucose, BMI, Insulin, Pregnancy, and Age) from the PIMA dataset. All of their ML models provided an accuracy greater than 70%. However LR and SVM provided highest accuracy for both train/test split and K-fold cross-validation method. They used NN models with 1, 2, 3 hidden layers varying the epochs 200, 400, 800. Hidden layer 2 with 400 epochs provided highest (88.6%) accuracy. Finally, they recommended the Neural Network with two hidden layers as it has an accuracy rate of approximately 86% for all varying epochs (200, 400, 800).

Similar approach was taken with more detailed approach such as measuring feature importance for each of the adopted machine learning methods but using a different dataset of 2000 instances (Rani, 2020). In this paper, five different ML techniques were used, and an accuracy comparison table was presented. In this case, Decision tree showed highest accuracy for both the training (98%) and testing datasets (99%). Also, similar to the results found by Soni & Varma, 2020, Glucose has been identified as the most important feature for this dataset too.

Zou et al., 2018 adopted similar methods (decision tree, random forest and neural network) to predict diabetes mellitus. Their dataset originated from physical examination of patients in a hospital in a in

Luzhou, China. It contains 14 attributes. They used five-fold cross validation to examine the models. To neutralize the impact of data imbalance, the researchers randomly extracted 5 times data and adopted the final result as the average of the five experiments. They also used principal component analysis (PCA) and minimum redundancy maximum relevance (mRMR) to reduce the dimensionality. Finally, their results showed that the prediction with random forest achieved highest accuracy (0.8084) when all the attributes were used.

Project performed by Tigga & Garg, 2020 was focused on predicting Diabetes type-2 in people. They have applied six machine learning methods and their analysis resulted Random Forest (RF) as the most accurate method after comparison. The dataset originated from data collected through online and offline questionnaires consisting of 18 questions relevant to diabetes. Also, after observing significance of all the variables, 'Age', 'Family diabetes', 'Physically active', 'Regular Medicine' and 'Pdiabetes' or gestation diabetes has highest significance among all variables and have greater impact on predicting diabetes than the rest.

Two main algorithms were applied on a dataset that has 800 instances with 10 attributes instances by Mujumdar & V, 2020 in their project. Algorithm 1 is diabetes prediction using various machine learning algorithms and Algorithms 2 is diabetes prediction using pipeline. For algorithm 1, the results show that logistic regression has the highest accuracy of 96%. However, they also compared the result of their diabetes dataset with the PIMA Diabetes dataset (which was the dataset used in the project by Soni & Varma, 2020 and Khanam & Foo, 2021). The comparison shows that their models give significantly better accuracy for the diabetes dataset than the PIMA dataset. For algorithm 2 (Pipelining), they also got highest accuracy for logistic regression (97.2%).

The project titled" Machine learning in precision diabetes care and cardiovascular risk prediction" by Oikonomou & Khera, 2023 is one of the latest related scholarly works. Their research is different from others as it is focused on how to use AI and ML to find new ways to optimize the management of diabetes and care and predicting cardiovascular risk to reduce related to diabetes. The paper is a comprehensive review of the various ML methods in the context of ML in the diagnosis, prognostication, phenotyping, and treatment of diabetes and its cardiovascular complications. Furthermore, the authors tried to identify key issues on equity and bias mitigation in healthcare and ways in which the regulatory framework can ensure the efficacy and safety of ML and AI in transforming cardiovascular care and outcomes in diabetes (Oikonomou & Khera, 2023).

Data Description & Descriptive Statistics

For the project, the **Diabetes prediction dataset** is chosen which is a collection of medical and demographic data from patients, along with their diabetes status (positive or negative). The data includes features such as *age, gender, body mass index (BMI), hypertension, heart disease, smoking history, HbA1c level*, and *blood glucose level*. There are 100,000 instances/rows in this dataset. The dataset can be retrieved from the following link:

https://www.kaggle.com/datasets/iammustafatz/diabetes-prediction-dataset

Dataset Description:

The dataset has total eight features or independent variables and one target feature/dependent variable. Brief description of each feature is summarized in the following table:

Sl.	Feature	Description	Data Type
1.	Gender	Gender refers to the biological sex of the individual, which can have an impact on their susceptibility to diabetes. There are three categories for this feature which are: *male, female and other.*	Categorical
2.	Age	Age is an important factor as diabetes is more commonly diagnosed in older adults. Age ranges from 0-80 in this dataset.	Numerical
3.	Hypertension	Hypertension is a medical condition in which the blood pressure in the arteries is persistently elevated. It has values either a 0 or 1 where 0 indicates they don't have hypertension and for 1 it means they have hypertension.	Numerical
4	Heart disease	Heart disease is another medical condition that is associated with an increased risk of developing diabetes. It has values either a 0 or 1 where 0 indicates they don't have heart disease and for 1 it means they have heart disease.	Numerical
5	Smoking History	Smoking history is also considered a risk factor for diabetes and can exacerbate the complications associated with diabetes.	Categorical

		In this dataset, we have 5 categories which are not current, former, No Info, current, never and ever.	
6	BMI (Body Mass Index)	BMI (Body Mass Index) is a measure of body fat based on weight and height. Higher BMI values are linked to a higher risk of diabetes. The range of BMI in the dataset is from 10.16 to 71.55. BMI less than 18.5 is underweight, 18.5-24.9 is normal, 25-29.9 is overweight, and 30 or more is obese.	Numerical
7	HbA1c (Hemoglobin A1c)	HbA1c (Hemoglobin A1c) level is a measure of a person's average blood sugar level over the past 2-3 months. Higher levels indicate a greater risk of developing diabetes. Mostly more than 6.5% of HbA1c Level indicates diabetes.	Numerical
8	Blood glucose	Blood glucose level refers to the amount of glucose in the bloodstream at a given time. High blood glucose levels are a key indicator of diabetes.	Numerical
9	**Diabetes (Target Variable)**	Diabetes is the target variable being predicted, with values of 1 indicating the presence of diabetes and 0 indicating the absence of it.	Numerical

Descriptive Statistics:

There are no missing values in the dataset. However mean, mode, standard deviation for the numerical variables are in the following table:

	age	hypertension	heart_disease	bmi	HbA1c_level	blood_glucose_level	diabetes
count	100000.000000	100000.00000	100000.000000	100000.000000	100000.000000	100000.000000	100000.000000
mean	41.885856	0.07485	0.039420	27.320767	5.527507	138.058060	0.085000
std	22.516840	0.26315	0.194593	6.636783	1.070672	40.708136	0.278883
min	0.080000	0.00000	0.000000	10.010000	3.500000	80.000000	0.000000
25%	24.000000	0.00000	0.000000	23.630000	4.800000	100.000000	0.000000
50%	43.000000	0.00000	0.000000	27.320000	5.800000	140.000000	0.000000
75%	60.000000	0.00000	0.000000	29.580000	6.200000	159.000000	0.000000
max	80.000000	1.00000	1.000000	95.690000	9.000000	300.000000	1.000000

Fig.1: Statistical measures for numerical features.

- **<u>Univariate Analysis:</u>**

1) Gender|:

There are 58,552 instances of female and 41,430 instances of male in the gender column. There are only 18 instances of "other" gender which is very low and can be removed as it will not affect the final result significantly.

2) Age:

The age column is distributed from new born to 80 years details distribution is given in Fig.2.

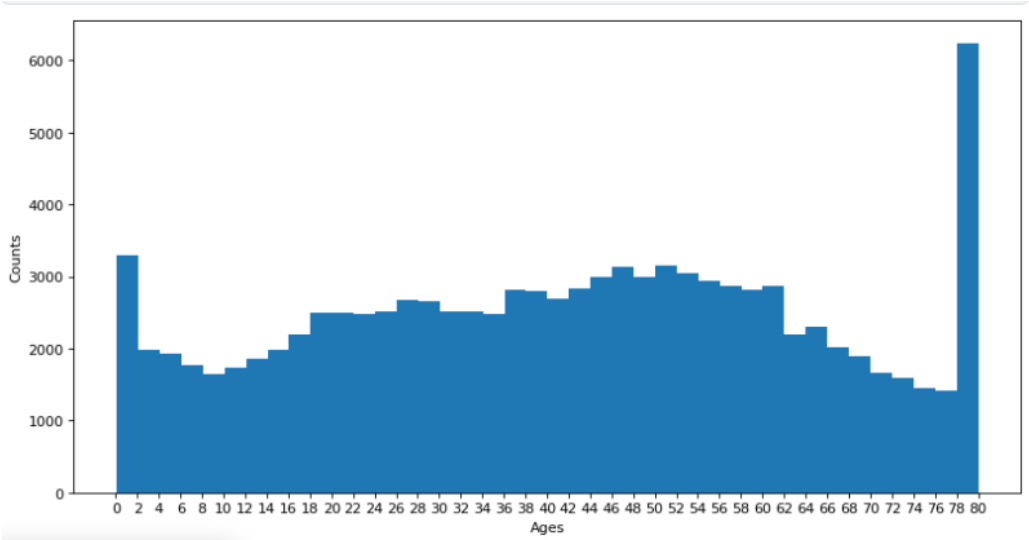

Fig.2: Distribution of "age" in the dataset

3) Hypertension:

Research indicates that hypertension or high blood pressure may worsen complications in diabetes patients. Percentage of patients with hypertension is 7.49% in this the dataset.

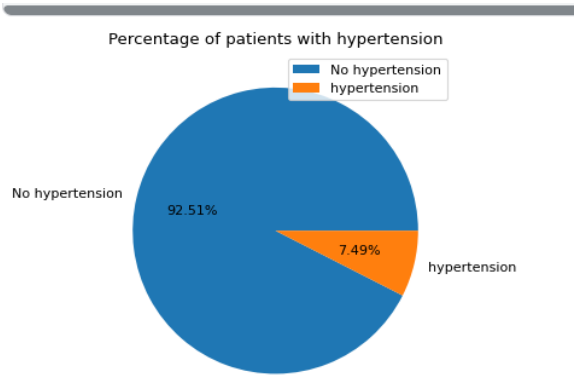

Fig.3: "Hypertension" in the dataset

4) Heart Disease:

Patient with heart disease is 3942 counts in the dataset which 3.94%.

5) Smoking History:

35.81% of instances in the smoking history feature has No info about smoking which is high and may impact the result significantly. This column may be removed from the dataset before performing further analysis.

6) Body Mas Index (BMI):

It is clear from both the boxplot and violin plot, there are outliers in the BMI features. Removing those outliers will help in getting better results.

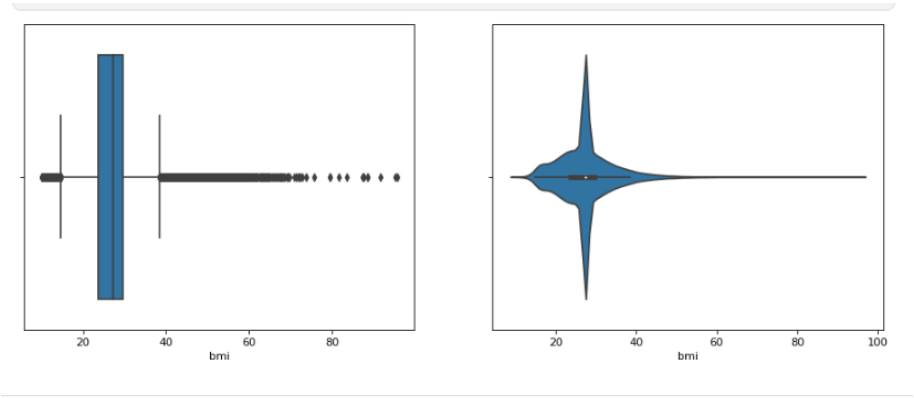

Fig.4: Boxplot and Violin plot for BMI in the dataset

7) Hba1c level:

Hba1c level for the patients in the dataset ranges from 3.5% to 9%. Any patient having 5.7% or below can be classified as non-diabetic while 5.7%-6.4% is prediabetic and 6.5% or above is a diabetic patient.

8) Blood glucose level:

Blood glucose level column ranges from 80 mg/dl to 300 mg/dl for the patients of this dataset.

9) Target Feature (Diabetes):

The percentage of patients having diabetes in this dataset is 7.56% which may indicate an imbalance. However, the global diabetes prevalence is estimated to be 9.30% (Saeedi et al., 2019). As there is no geographic location of the source of the dataset, the percentage seems to be acceptable.

➢ **Bivariate Analysis:**

Initially, some bivariate analysis have been performed between the target variable and other features

(i) HbA1c_level vs. diabetes:

The analysis of initial and actual diagnosis based on HbA1c level is summarized in the following table and pie charts:

HbA1c level	initial diagnosis	actual diagnosis
< 5.7	Normal	100% no diabetes
5.7 – 6.4	Prediabetes	7.47% have diabetes
>= 6.5	Diabetes	23.64% have diabetes

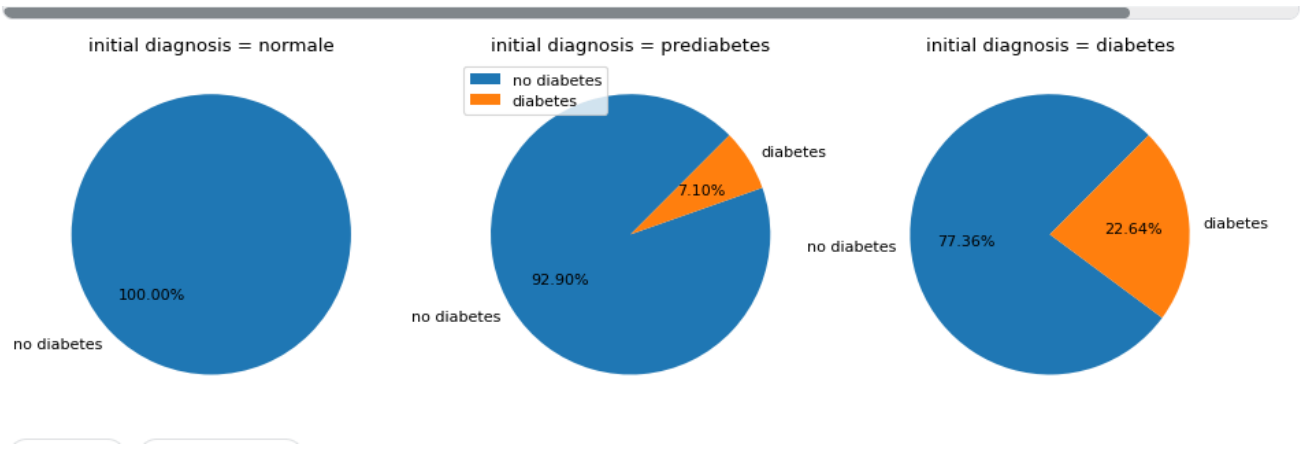

Fig.5: Pie Chart for HbA1C vs Diabetes

According to ordinal category [underweight, normal, overweight, obesity], as weight category increases, percentage of patients with diabetes increases.

(ii) BMI vs diabetes:

BMI	Category	Prediction
=< 18.5	Underweight	0.7% have diabetes
18.5 – 24.9	Normal	3.85% have diabetes
25 – 29.9	Overweight	7.88% have diabetes
>= 30	Obesity	15.94% have diabetes

(iii) Blood Glucose Level vs diabetes:

Blood Glucose Level	Category
=< 99	normal
100 – 125	Prediabetes
>= 126	Diabetes

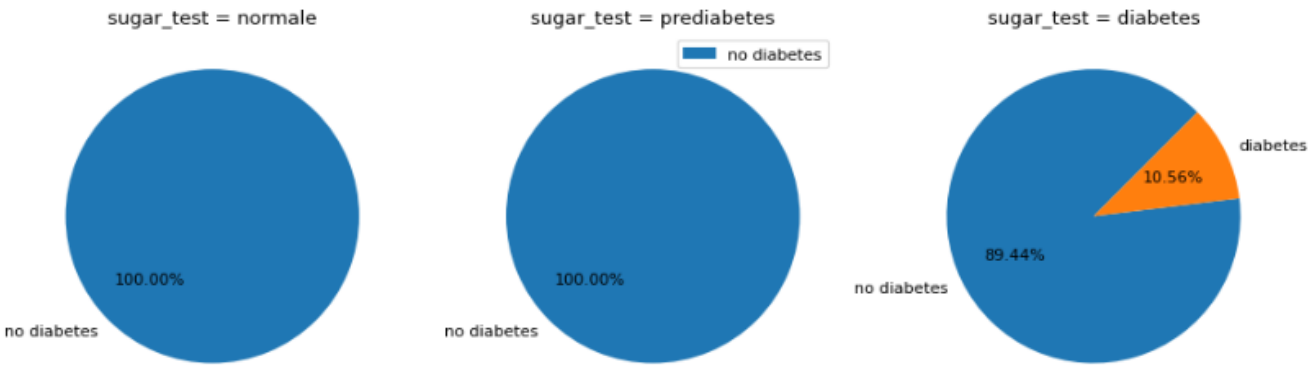

Fig.6: Pie Chart for Glucose Level vs Diabetes

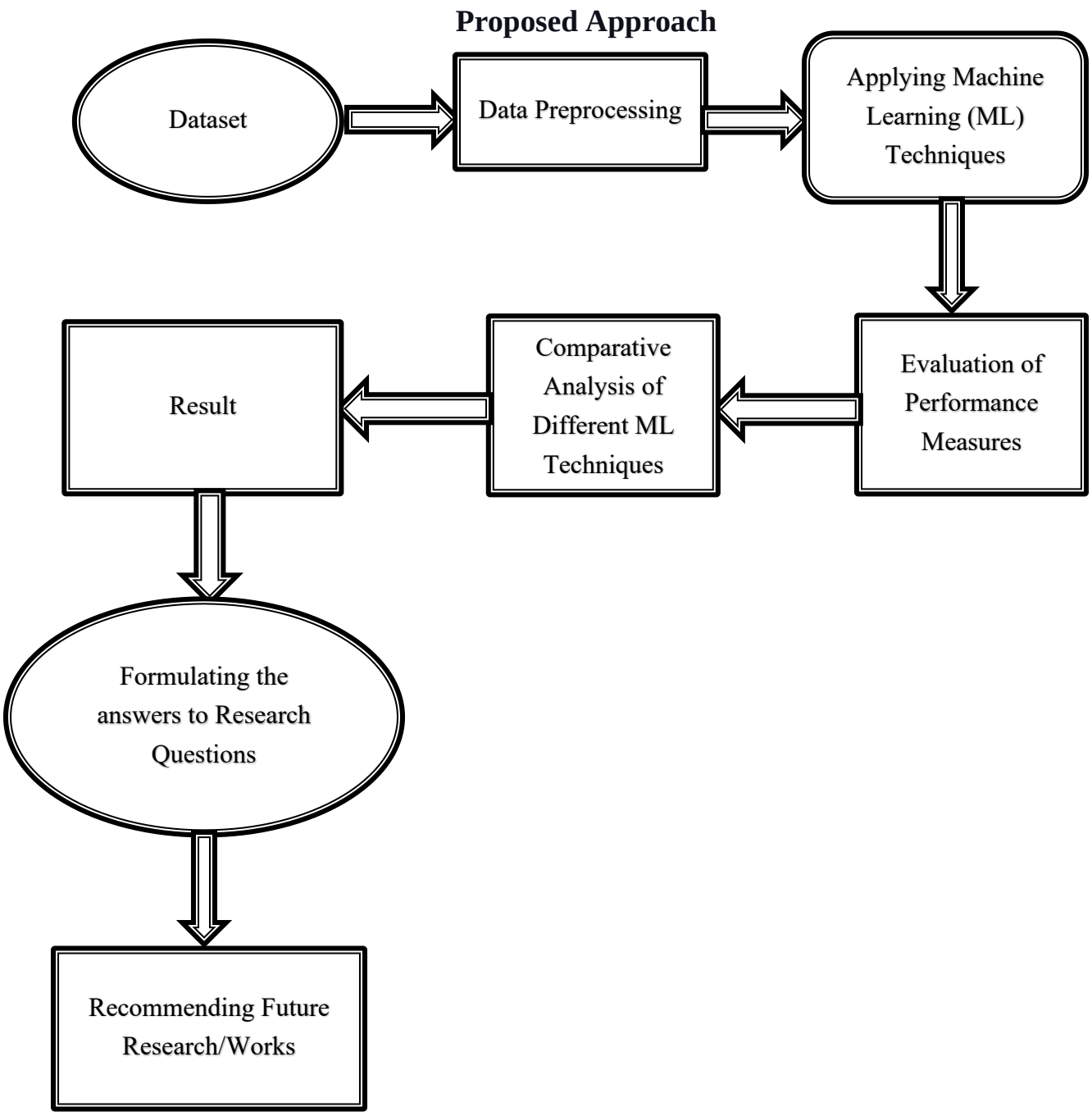

Fig.7: Proposed Approach/Process Diagram

Codes for Data Preprocessing

Codes for initial preprocessing is uploaded in the GitHub repository link of which is given below:
https://github.com/HussainM19/CIND-820-Predicting-Diabetes-using-MAchine-Learning

References

Soni, M., & Varma, Dr. S. (2020, October 4). *Diabetes prediction using Machine Learning Techniques*. International Journal of Engineering Research & Technology. https://www.ijert.org/diabetes-prediction-using-machine-learning-techniques

Rani, K. J. (2020, August 30). *Diabetes prediction using machine learning*. International Journal of Scientific Research in Computer Science, Engineering and Information Technology. https://ijsrcseit.com/CSEIT206463

Khanam, J. J., & Foo, S. Y. (2021, February 20). *A comparison of machine learning algorithms for diabetes prediction*. ICT Express. https://www.sciencedirect.com/science/article/pii/S240595952100020

Oikonomou, E. K., & Khera, R. (2023, September 25). *Machine learning in precision diabetes care and cardiovascular risk prediction - cardiovascular diabetology*. SpringerLink. https://link.springer.com/article/10.1186/s12933-023-01985-3

Zou, Q., Qu, K., Luo, Y., Yin, D., Ju, Y., & Tang, H. (2018, October 12). *Predicting diabetes mellitus with machine learning techniques*. Frontiers. https://www.frontiersin.org/articles/10.3389/fgene.2018.00515/full

Tigga, N. P., & Garg, S. (2020, April 16). *Prediction of type 2 diabetes using machine learning classification methods*. Procedia Computer Science. https://www.sciencedirect.com/science/article/pii/S1877050920308024

Mujumdar, A., & V, Dr. V. (2020, February 27). *Diabetes prediction using machine learning algorithms*. Procedia Computer Science. https://www.sciencedirect.com/science/article/pii/S1877050920300557

Saeedi P, Petersohn I, Salpea P, Malanda B, Karuranga S, Unwin N, Colagiuri S, Guariguata L, Motala AA, Ogurtsova K, Shaw JE, Bright D, Williams R; IDF Diabetes Atlas Committee. *Global and regional diabetes prevalence estimates for 2019 and projections for 2030 and 2045: Results from the International Diabetes Federation Diabetes Atlas*, 9th edition. Diabetes Res Clin Pract. 2019 Nov;157:107843. doi: 10.1016/j.diabres.2019.107843. Epub 2019 Sep 10. PMID: 31518657.