Identifying Diabetic Patients using Machine Learning Algorithms.

Presented by:

Gionian Kyros, v21gioky@du.se

Maha Vajeeshwaran, h21mahna@du.se

Abstract - The application of predictive analytics by using machine learning, in the medical field is a tough undertaking, but it ultimately has the potential to assist practitioners in making timely judgments. This research work makes use of six different machine learning algorithms, that tackles predictive analytics in the identification of diabetic, conducted on a dataset including medical records of patients. The effectiveness and precision of the various algorithms that were utilized were analyzed and contrasted. The many approaches to machine learning that were investigated for this study led to the identification of Random Forest and XGBoost algorithm as the ones that performed the most accurately when predicting diabetes.

Keywords: Machine Learning, Logistic regression, LDA, Decision Tree, XGB, Random Forest.

I. INTRODUCTION

Diabetes is a serious, long-lasting health illness that manifests itself either when the pancreas ceases to secrete insulin or when the human body is unable to make effective use of the insulin that the pancreas generates. According to a report by the International Diabetes Federation (IDF), diabetes has already affected more than 537 million people all over the world, the majority of whom are women. Furthermore, it is estimated that the upcoming rates will increase to 19.7% by 2030 and 29.9% by 2045 compared with the results of 2021. Indeed, more than half a billion individuals around the world have been identified as having diabetes, and it is anticipated that this figure will rise by 245 million by the year 2045 [1].

Diabetes cases and symptoms have been meticulously documented in recent years thanks to the development of information technology and its ongoing proliferation in the medical and healthcare industries. Throughout the past years different methods such as machine learning and data mining are being utilized for the aim of discovering knowledge that can be used for better prediction purposes [2].

Several studies have been carried out to investigate the diagnosis of diabetes utilizing a variety of classification algorithms and machine learning methodologies. [3]

Since there are so many algorithms, determining which one has the best performance is difficult. In most circumstances, depending on the data input, some machine learning algorithms perform better than others in some cases and perform worse in others.

The purpose of this study is to utilize various machine learning algorithms, which will be

evaluated for their level of efficiency and compared to one another on medical records of patients in order to identify if a one has a particular set of characteristics that can result in a positive or negative diagnosis of diabetes. [4] [5].

Our research focus on the answering the below question:

Q1. How different algorithm perform into classifying diabetes?

The remaining parts of the paper are structured as described below: The work of several different machine learning models for predicting diabetics is summed up in the second section of this article. The methodology that will be used for the proposed model is detailed in Section 3. In Section 4, you'll find a presentation of the findings obtained using the proposed model. In the fifth section, we offer both our findings and conclusion.

II. RELATED WORK

Researchers have utilized a variety of data mining strategies and machine learning algorithms to build a variety of prediction approaches, which have been proposed and developed across the scientific literature. Over the course of the past decade, one of the primaries focuses of research has been on the development of more accurate predictive models for the progression of diabetes.

Patients' risk of developing diabetes can be estimated with the use of a predictive model that was developed by Hang Lei et al. [6] and is based on demographic information and laboratory findings. The model makes use of Gradient Boosting Machine and Logistic Regression techniques (Diabetes Mellitus). For the purpose of analyzing and contrasting the outcomes of different approaches, machine learning strategies such as Rpart and Random Forest were utilized.

Karim et al. [7] presented a diabetes prediction system with the purpose of forecasting a candidate at a given age. The system would be based on a technique called a decision tree. The findings were encouraging; the method accurately forecasts the occurrence of diabetes depending on the patient's age.

Decision Tree, Artificial Neural Networks, Logistic Regression, and Naive Bayes were the four categorization models that Nongyao and Rungruttikarn looked into [8]. The bagging and boosting optimization strategies were utilized to make the model more robust. Based on the findings of the studies, it was determined that out of all of the algorithms that were utilized, the random forest algorithm delivers the greatest outcomes.

In his research, Alajlan [9] focuses on analyzing the dataset through classification analysis by employing decision trees, adaptive boosting, and K-nearest neighbor algorithms. Alajlan does this in order to determine how accurate the dataset is. As a result, a new, more efficient model for predicting diabetes has been developed, and the objective was to create the most effective model possible that can reach a conclusion on the early diagnosis of diabetes that has not yet been diagnosed.

Ohad Houriet al. [10] using machine learning skills, his objective was to determine whether or not a woman who had previously been diagnosed with gestational diabetes (GDM) would go on to develop type 2 diabetes after giving birth. He employs a decision tree-based fitting technique called

XGBoost, which was applied to the training data. They were able to rank the elements according to the impact that they had on the prediction that were made. In order to eliminate any possible biases, they gave a weighting to each category, either positive or negative.

These papers explore a variety of methods for estimating the risk of developing diabetes based on factors such as age, gender, or demographic information. However, despite the fact that each of these parts is significant, the primary focus of our research study is that on the accuracy of our models to make predictions based on the constellation of symptoms that a patient would have.

III. MATERIAL AND METHOD

3.1 Data Set

The data that was used in this research was obtained from the website of the UCI [11]. The information is organized into 520 rows and 17 columns total. The following are characteristics of the patient that have been identified: *age*, *gender*, and whether or not they exhibit any of the following symptoms if they are *male* or *female*: In addition to *polyuria, polydipsia, sudden weight loss, weakness, polyphagia, genital thrush, visual blurring, itching, irritability, delayed healing, partial paresis, muscle stiffness, alopecia, and obesity*, the last piece of information in the dataset is the *class* to which the individual belongs, which indicates whether or not they have diabetes.

3.2 Exploratory Data Analysis (EDA):

An EDA procedure on our dataset to better understand the information we were working with and the trends and details we required. Before analysis, all non-numerical categorical features are converted to numerical features using LabelEncoder(). We looked over the data and discovered that there were no records of missing values. Using the boxplot(), we also discovered that the dataset contained few outliers and that all of the features except age are categorical. We conducted some analysis to have a better understanding of our dataset.

3.3 Data Preparation

We were ready to run the ML algorithm after completing the preceding stages and ensuring that our data was clean. The feature selection was done with the help of SelectKBest and chi2, and we chose the top 10 features based on their scores. The top ten features were determined to be polydipsia, polyuria, abrupt weight loss, partial paresis, gender, irritability, age, polyphagia, baldness, and visual blurring.

However, we discovered that some predictors from the selected 10 features are highly connected using the correlation plot Figure 1, therefore we excluded polyuria, sudden

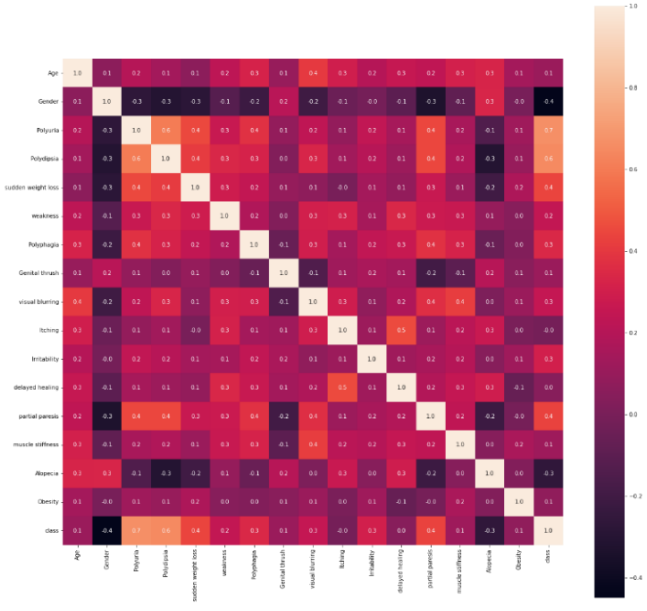

Figure 1*. Correlation matrix of the features in the dataset*

weight loss, and partial paresis to avoid overfitting. The data was then separated into features (X) as predictors (7 variables) and Y as a response (Y) (class). To accommodate highly changing values, we divided 80% of thhe data for training and 20% of the data for testing, and feature scaling was accomplished using StandardScaler(). Because after that step the data were now unbalanced, we used RandomUnderSampler() to even it out. To avoid overfitting, it randomly selects examples from the majority class and removes them from the training dataset.

3.4 Training and Testing with ML models

The scaled data is now trained using different machine learning model, and the train and test accuracy are calculated.

Logistic Regression with parameters (*penalty='l2', tol=0.0001, C=1.0, fit_intercept=True, intercept_scaling=1, random_state=46,solver='lbfgs',max_iter=100,multi_class='auto', verbose=0*),

LDA with parameters (*solver='svd', shrinkage=None, priors=None, tol=0.0001),*

KNeighbors Classifier *with parameters (n_neighbors=5, weights='uniform', algorithm='auto', leaf_size=30, p=2, metric='minkowski'*).

Decision Tree Classifier with parameters (*criterion='gini',splitter='best',min_samples_split=2,min_samples_leaf=1,min_weight_fraction_leaf=0.0,min_impurity_decrease=0.0).*

Xgboost with parameters (*'reg_lambda': 0.1, 'n_estimators': 500, 'max_depth': 4, 'learning_rate': 0.05).*

Random Forest classifier *with parameters (n_estimators = 200, min_samples_split = 2, min_samples_leaf = 1, max_features = 'auto', max_depth = 16, criterion = 'entropy', random_state = 40*) were used.

Then K fold cross validation was utilized, the key advantage of this method is that each sample is only used once for training and validation (as part of a test fold). Cross-validation is commonly used because it allows models to train on several train-test splits and provides a better indicator of how well the model will perform on unknown data. The variance estimate of the model performance is smaller than with the holdout technique.

IV. RESULT

After the EDA process of our data, from Figure 2, we were able to determine that despite the fact that female patients make up approximately one-third of our dataset, they have the largest number of patients that test positive for diabetes.

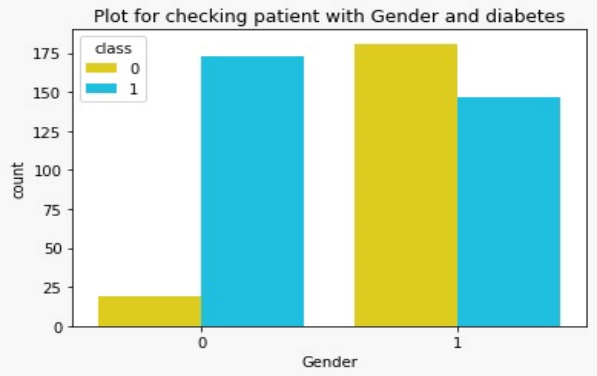

Figure 2. *Count of Positive (1) & Negative (2) of Man (1) & Woman (0)*

From Figure 3 we can see that between the ages of 46 and 55, the age range with the highest number of positive cases was that of 46–55 year-olds. This pattern is consistent with the one that was identified by IDF in the very first paragraph of our study.

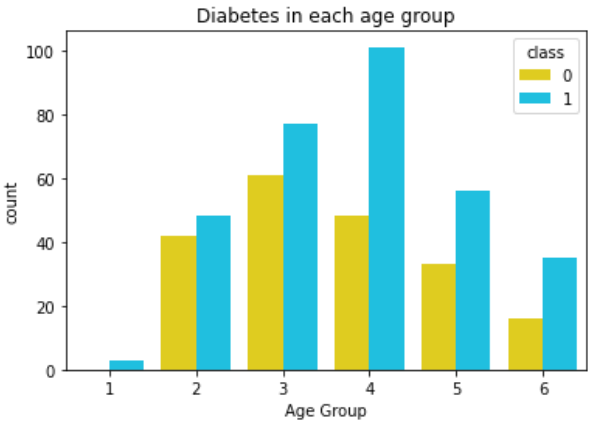

Figure 3.*.Count of Positive (1) & Negative (0) patients per age group, (1)15–25, (2) 26-35, (3)36–45, (4)46–(5)5.56–65, (6). above 65*

Having selected the 7 features which were chosen from the feature importance we predicted the train and test accuracy by using different machine learning models. As mentioned earlier in section **3.4** we trained the model using training data set and tested the model with the help of test datasets. The Table 1 shows the result of our train and test set and their accuracy.

We performed the hyperparameter tuning in:

XGboost with parameters *('reg_lambda': 0.1, 'n_estimators': 500, 'max_depth': 5, 'learning_rate': 0.05, random_state = 42).*

Random Forest with parameters *(n_estimators = 200, min_samples_split = 2, min_samples_leaf = 1, max_features = 'auto', max_depth = 15, criterion = 'entropy', random_state = 42).*

We found the rise in accuracy with train and test data for XGB with 99% (Train) and 93%(Test) and for RF 99% & 94% accordingly.

We checked with the K fold cross-validation approach with 5-fold to avoid higher variance and it gives the model the opportunity to train on multiple train-test splits.

Table 1. *Machine Learning Results*

	Accuracy		Precision		Recall		F1	
	Train	Test	Train	Test	Train	Test	Train	Test
Logistic Regression	86.2	84.6	85.4	86.0	86.6	85.9	85.6	85.3
LDA	87.5	87.2	100.0	96.0	99.4	80.0	83.6	87.3
KNN Classifier	98.2	90.4	97.4	94.7	95.5	90.0	96.4	92.3
Decision Tree	99.0	95.2	100.0	96.6	99.4	93.3	99.7	94.9
XGBoost	99.0	93.0	100.0	96.6	98.7	95.0	99.4	94.9
Random Forest	99.0	94.2	100.0	96.6	99.4	93.3	99.7	94.9

Table 2 shows the cross-validation scores for each algorithm of the full data set.

Table 2. *CV score of ML models*

Model	CV Score (5-fold)
Logistic Regression	0.855
LDA	0.836
KNN Classifier	0.846
Decision tree	0.936
XGBoost	0.948
Random Forest	0.95

Here we identified that Decision tree, XGBoost and Random Forest with hyperparameter tuning perform better than the other models. As they are non-parametric models it is more flexible with low bias and high variance, no assumptions are made about underlying functions which results in higher performance of model prediction when compared to parametric models (logistic regression, LDA). Therefore, we tried to explore further by checking the cross-validation score for XGBoost and Random Forest with hyperparameter tuning of the model with 10-fold. We noticed that there is

not much difference between the 5-fold to the 10-fold as we can see from the table below.

Table 3.CV score of XGBoost and Random Forest

Model	CV Score (10-fold)
XGBoost	0.9519
Random Forest	0.9596

By looking at the accuracy score from Table 3 and into the confusion matrix Random Forest classifier performed better than XGBoost.

V. CONCLUSION

One key area for the healthcare sector and experts is the ability to detect health disorders and diseases sooner and more correctly, allowing more individuals to receive treatment quickly and effectively. Diabetes is one of these instances. Our research focuses on comparing different machine learning algorithms to see which one performs best. The data model was used to train the algorithms after we processed our dataset and identified the most key features that would bring us the best results, such as seven input features (Polydipsia, Gender, Irritability, Age, Polyphagia, Alopecia, and visual blurring) and one output feature (class) in the diabetes dataset using feature selection techniques. The machine learning algorithms we employed to predict diabetes were LR, KNN, RF, XGB, LDA, and Decision Tree, and all of them had an accuracy of more than 80% based on their performance. As noticed, RF had the best results compared with the other ML models, if it was 5-fold cross-validation or 10-fold validation. These findings were encouraging, demonstrating that incorporating technology into our daily lives might improve certain aspects of it, particularly when it comes to human life and health. As a result, there will be much more to discover in the future, as more complex and advanced analyses will be possible.

REFERENSES

[1] *IDF Diabetes Atlas 2021 | IDF Diabetes Atlas*. (n.d.). Retrieved May 21, 2022, from https://diabetesatlas.org/atlas/tenth-edition/

[2] Kavakiotis, I., Tsave, O., Salifoglou, A., Maglaveras, N., Vlahavas, I., & Chouvarda, I. (2017). Machine Learning and Data Mining Methods in Diabetes Research. *Computational and Structural Biotechnology Journal*, *15*, 104–116. https://doi.org/10.1016/J.CSBJ.2016.12.005

[3] 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. 5– 10.

[4] Faniqul Islam, M. M., et al. "Likelihood Prediction of Diabetes at Early Stage Using Data Mining Techniques | SpringerLink." *Likelihood Prediction of Diabetes at Early Stage Using Data Mining Techniques | SpringerLink*, link.springer.com, 29 Aug. 2019, https://link.springer.com/chapter/10.1007/978-981-13-8798-2_12.

[5] "Diabetes Prediction Using Machine Learning Algorithms - ScienceDirect." *Diabetes Prediction Using Machine Learning Algorithms - ScienceDirect*, www.sciencedirect.com, 27 Feb. 2020, https://www.sciencedirect.com/science/article/pii/S1877050920300557.

[6] Lai, H., Huang, H., Keshavjee, K. et al. Predictive models for diabetes mellitus using machine learning techniques. BMC Endocr Disord 19, 101 (2019).

[7] Orabi K.M., Kamal Y.M., Rabah T.M. (2016) Early Predictive System for Diabetes Mellitus Disease. In: Perner P. (eds) Advances in Data Mining. Applications and Theoretical Aspects. ICDM 2016. Lecture Notes in Computer Science, vol 9728. Springer, Cham.

[8] Nongyao Nai-arun, Rungruttikarn Moungmai, Comparison of Classifiers for the Risk of Diabetes Prediction, Procedia Computer Science, Vol 69, 2015, pp. 132-142.

[9] lajlan, A. M. (2021). A Model-Based Approach for an Early Diabetes Prediction Using Machine Learning Algorithms. In *Turkish Journal of Computer and Mathematics Education* (Vol. 12, Issue 3)

[10] Ohad houri, Gil, Y., Berezowsky, A., Wiznitzer, A., Hadar, E., & Chen, R. (2020). 339: Future Type-2 diabetes prediction following pregnancy - using a novel machine learning algorithm. *American Journal of Obstetrics and Gynecology*, *222*(1), S228. https://doi.org/10.1016/J.AJOG.2019.11.355

[11] Index of /ml/machine-learning-databases/00529. (n.d.). Retrieved June 10, 2022, from https://archive.ics.uci.edu/ml/machine-learning-databases/00529/, https://archive.ics.uci.edu/ml/index.php