ISSN No:-2456-2165

A Transformer on Tabular Data Comparative Analysis with Linear and Tree Base Machine Learning Algorithm on Diabetic Dataset

Kamin Gorettie Precody [1,]
Komiwe Faith Phiri [2]
Dr. Ashish Kumar Chakraverti[3]
[1,2,3] Department of Computer Science & Engineering, School of Engineering and Technology, Sharda University, Greater Noida.

Abstract:- **Lifestyle diseases have a rating of 80% as one of the top causes of death. About over 41 million lives are claimed just by lifestyle diseases, which are over 70% of all deaths around the world. In this same percentage about roughly 15 million deaths happen to people of the age range 30 to about 69 years. Lifestyle diseases are primarily originated due to the day-to-day habits of an individual. These habits that detract from activities and push people towards a sedentary routine can cause numerous health issues that may lead to harmful diseases that are nearly life-threatening. Furthermore, there are two common complex diseases that are heart disease and diabetes, researchers have discovered diabetes to be a silent but deadly disease, and many researchers use machine learning methods to help medical professionals for the diagnosing of lifestyle diseases. This paper reviewed the literature on predictions and diagnoses of lifestyle diseases with the use of transformers and machine learning techniques it is presented and used on Diabetics data of patients. Our research paper will highlight the importance of transformers and machine learning in analyzing huge datasets of patients to predict the whole kinds of diabetes and how they can be treated and how they can be prevented. Further, we have utilized Transformers on tabular data (Tabpfn), Random Forest, Decision Tree, Support Vector Machine K-Nearest Neighbors, Gradient Boosting, Histogram Gradient Boosting, and Adaptive Boosting for predicting how likely a person will have a bank account. The stratified holdout cross-validation method has been used to split the training dataset randomly into 90% train and 10% test sets. The result was collected and further compared with some existing approaches, which indicates that using transformers on tabular data (Tabpfn) outperforms the existing state-of-the-art approach. The Tabpfn transformer on tabular data was optimal among adapted models based on F1-score, which are 98.46 %, 98.0694%, 91.736%, and 91.541% respectively.**

Keywords:- *Transformer, Lifestyle Diseases, Machine Learning Techniques, Prediction.*

I. INTRODUCTION

Diabetes is a condition caused as a result of high glucose levels in the human body. Diabetes ought not be overlooked on the off chance that it is untreated, Diabetes might cause a few significant issues in an individual like heart related issues, kidney issue, pulse, eye harm and it can likewise influence different organs of human body. Diabetes can be controlled on the off chance that it is anticipated before. To accomplish this objective this venture work we will do early expectation of Diabetes in a human body or a patient for a higher precision through applying, Different AI Methods. AI methods Give improved results to forecast by developing models from datasets gathered from patients.

As per (WHO) World Wellbeing Association around 422 million individuals experience the ill effects of diabetes especially from low-or inactive pay nations. Furthermore, this could be expanded to 490 billion up to the extended time of 2030. Nonetheless, predominance of diabetes is found among different Nations like Canada, China, and India and so on.

AI methods can assume a fundamental part in foreseeing diabetes at a beginning phase. These strategies use calculations and measurable models to dissect enormous datasets and anticipate results in view of the information. By applying AI calculations to information gathered from patients, it is feasible to recognize designs and foresee the probability of creating diabetes. This can be accomplished by dissecting different variables, for example, age, family ancestry, way of life propensities, and clinical history. AI procedures can give higher exactness in anticipating diabetes, which can assist patients with going to early preventive lengths to deal with their condition.

The utilization of AI methods for foreseeing diabetes is certainly not another idea. Nonetheless, with headways in innovation, it has become simpler to gather and examine a lot of information from patients. AI models can be prepared to distinguish examples and patterns in the information, which can assist with foreseeing the probability of creating diabetes. These models can similarly be invigorated regularly to chip away at their precision as new data opens up.

All in all, diabetes is a serious medical problem that can prompt significant confusions in the event that did not oversee as expected. Early identification is critical to forestalling complexities and working on the personal satisfaction for those living with diabetes. AI strategies can give a more exact expectation of diabetes, which can assist patients with going to early preventive lengths. With additional innovative work, AI could assume a huge part in further developing diabetes the board and lessening the weight of this constant sickness on people and medical care frameworks.

II. RELATED WORK

Investigating other written works, the number of researchers used various datasets from patients Investigating records and other sources like Kaggle. These where all investigated, it was also visible that the highest accuracies were obtained using Artificial Neural Networks or SVM.Many of these diseases use prediction algorithms and approaches which can be applied by machine learning, ensemble learning approaches and association rules for achieving the perfect classification precision.There is actually a close connection linking data mining and machine learning such that machine learning techniques are also known as data mining techniques. In this paper we came into possession of some diabetes datasets from Kaggle, medical records of both for genders and of all ages. Countless of researchers have used ML and DM based Algorithm Few of them are listed and explained below:

K.VijiyaKumar et al. [1] proposed random Forest algorithm for the Prediction of diabetes develop a system which can perform early prediction of diabetes for a patient with a higher accuracy by using Random Forest algorithm in ma- chine learning technique. The proposed model gives the best results for diabetic prediction and the result showed that the prediction system is capable of predicting the diabetes disease effectively, efficiently and most importantly, instantly.

Nonso Nnamoko et al. [2] presented predicting diabetes onset: an ensemble supervised learning approach they used five widely used classifiers are employed for the ensembles and a meta-classifier is used to aggregate their outputs. The results are presented and compared with similar studies that used the same dataset within the literature. It is shown that by using the proposed method, diabetes onset prediction can be done with higher accuracy.

Tejas N. Joshi et al. [3] presented Diabetes Prediction Using Machine Learning Techniques aims to predict diabetes via three different supervised machine learning methods including: SVM, Logistic regression, ANN. This project pro- poses an effective technique for earlier detection of the diabetes disease.

Deeraj Shetty et al. [4] proposed diabetes disease prediction using data mining assemble Intelligent Diabetes Disease Prediction System that gives analysis of diabetes malady utilizing diabetes patient's database. In this system, they propose the use of algorithms like Bayesian and KNN (K-Nearest Neighbor) to apply on diabetes patient's database and analyze them by taking various attributes of diabetes for prediction of diabetes disease.

Muhammad Azeem Sarwar et al. [5] proposed study on prediction of diabetes using machine learning algorithms in healthcare they applied six different machine learning algorithms Performance and accuracy of the applied algorithms are discussed and compared.

A comparison of the different machine learning techniques used in this study reveals which algorithm is best suited for the prediction of diabetes. Diabetes Prediction is becoming an area of interest for researchers in order to train the program to identify the patient are diabetic or not by applying proper classifiers to the dataset.

Based on previous research work, it has been observed that the classification process is not much improved. Hence a system is required as Diabetes Prediction is an important area in computers, to handle the issues identified based on previous research.

III. MACHINE LEARNING ALGORITHM

Arthur Samuel in 1959 coined the term ML—a branch of computer science (CS) that helps computers to learn independently without explicit programming. In ML, an algorithm manipulates a dataset enabling it to make predictions by learning patterns from previous data.

ML can be categorized as supervised, unsupervised and reinforcement learning [12]. Supervised learning contains classified data having labels. When such data is supplied to an algorithm, it can predict a test case's outcome. Classification and regression are the main methods of supervised learning. Supervised ML can be achieved using various algorithms such as Naive Bayes classification, decision tree, and SVMs.

In this paper, we have utilized Transformers on tabular data (Tabpfn), Random Forest, Decision Tree, Support Vector Machine K-Nearest Neighbors, Gradient Boosting, Histogram Gradient Boosting, and Adaptive Boosting for predicting how likely a person will have diabetes. The stratified holdout cross-validation method has been used to split the training dataset randomly into 90% train and 10% test sets. The result was collected and further compared with some existing approaches, which indicates that using transformers on tabular data (Tabpfn) outperforms the existing state-of-the-art approach. The Tabpfn transformer on tabular data was optimal among adapted models based on F1-score, which are 98.46 %, 98.0694%, 91.736%, and 91.541% respectively.

IV. PROPOSED SYSTEM

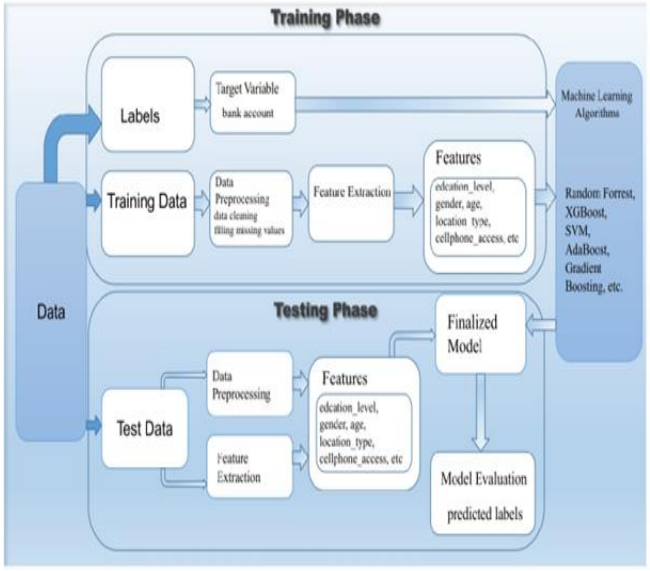

Fig 1 Training Phase and Testing Phase

We are going to build a system that will be able to efficiently predict if a patient is a diabetic or not. The system is utilizing the new techniques known as transformers which is going to use the new technique which we call Active Learning. Active Learning is a new technique with the aim of

➢ *Data Collection*

The dataset used for this project was obtained from Kaggle, a popular platform for sharing datasets and conducting data-driven research. The dataset is an updated version of the Pima Indians Diabetes Database, which includes demographic, diagnostic, and historical medical data of patients. The updated dataset consists of 768 instances with eight features, which are age, number of pregnancies, glucose concentration, blood pressure, skin thickness, insulin, BMI, and diabetes pedigree function.

To explore the dataset and prepare it for analysis, relational views were created for the features. The age feature represents the age of the patient, which is a continuous variable. The number of pregnancies is a discrete variable that represents the number of times a patient has been pregnant. The glucose concentration, blood pressure, skin thickness, and insulin features are continuous variables that represent different diagnostic measurements. The BMI feature is a continuous variable that represents the body mass index of the patient, and the diabetes pedigree function is a continuous variable that represents the genetic predisposition to diabetes.

➢ *Data Processing*

The data was preprocessed before being used for the prediction task. The preprocessing steps included handling missing values, normalizing the features, and encoding categorical variables. The dataset was split into training and testing sets using a stratified holdout cross-validation method. This ensured that the distribution of classes in the training and testing sets was similar, which is important for building an accurate prediction model.

Overall, the data used for this project was obtained from a reliable source and was preprocessed to ensure its quality and suitability for the prediction task. The relational views created for the features helped in understanding the dataset and preparing it for analysis.

➢ Training Data

Training data needs to be collected alongside the testing data further preprocessing is needed to know better the predictors. Training data helps us to prepare a budget request at some point and it's a proper document for building a business case and justifying budget requests.

➢ Predictive Features

To understand what drives the target outcome, there should be some research or an investigation to get ideas on the data points. Once the quality of understanding of what will fit well, the target outcome is achieved, further process of data requests can help build a business case. The main predictive features that are taken into feasibility criteria are: Age.

- *Gender*
- *Polyuria*
- *Polydipsia Sudden*
- *Weight Loss*
- *Weakness*
- *Polyphagia*
- *Genital Thrush*
- *Visual Blurring*
- *Itching*
- *Irritability*
- *Delayed Healing*
- *Partial Paresis*
- *Muscle Stiffness*
- *Alopecia*
- *Obesity*
- *Class*

➢ *Working of the Model*

The first task of the project would be to gather and clean the dataset. This would involve finding a reliable source of data and performing data cleaning and preprocessing to ensure that the data is ready for analysis. The duration of this task could be 2 weeks.

The next task would be to perform exploratory data analysis on the dataset to identify trends and patterns. This could take 3 weeks, and the output would be a report on the findings.

The third task would be to apply feature engineering techniques to the dataset to improve the accuracy of the models. This could take 2 weeks, and the output would be a dataset that is ready for modeling.

The fourth task would be to build and train machine learning models using both classical algorithms and transformers. This could take 6 weeks, and the output would be a set of models with their respective accuracy and performance metrics.

Finally, the last task would be to analyze and compare the performance of the models and draw conclusions on the suitability of transformers in early disease prediction. This could take 2 weeks, and the output would be a report summarizing the findings and conclusions.

A. Algorithms

- *Input:* Data set from Kaggle
- *Output*: a prediction model
- *Variables*:

✓ Required Accuracy--Minimum threshold accuracy of the model (%)
✓ Current Accuracy--Accuracy of the model after training (%)
✓ X train--Training data for the model: predictor
✓ Y train--Training data for the model: target
✓ X test--Testing data for the model: predictor
✓ Y test--Testing data for the model: target model--lifestyle disease prediction model

- *SVM Parameters--Kernel, C, Gamma, and Degree BEGIN*

✓ STEP 1: Determine the value of the required Accuracy
✓ STEP 2: Prepare the dataset from the questionnaire
✓ STEP 3: Note the predictor and target values
✓ STEP 4: Preprocess the dataset:
✓ STEP 4.1: Data integration
✓ STEP 4.2: Data transformation
✓ STEP 4.3: Data reduction
✓ STEP 4.4: Data cleaning
✓ STEP 5: X train, Y train--70% of data collected
✓ STEP 6: X test, Y test--30% of data collected
✓ STEP 7: current Accuracy--0
✓ STEP 8: while(current Accuracy < required Accuracy)
✓ STEP 9: Deployment
✓ STOP

B. Algorithm Part 2

- *Input: Predictor values of a web user*
- *Output: Yes, if a user suffers a lifestyle disease (with his/her name). No, if a user does not suffer from a lifestyle disease.*
- *Variables: user Input--a web user's values*

✓ *Model--Trained Model From Algorithm I*
✓ *Prediction--Output From The Model*

- BEGIN STEP 1: user Input Æ Store user input in an appropriate format
- STEP 2: prediction Æ predict if an individual suffers from any lifestyle disease using user Input and model
- STEP 3: Display the prediction to a user in an appropriate format.
- STOP

V. SIMULATION RESULTS

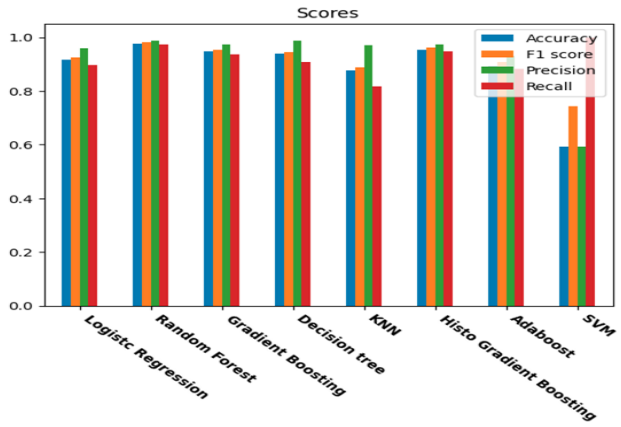

Fig 2 Scores

➢ *Required Accuracy = 90%*

Type of testing adapted: To test the accuracy or the performance of each algorithm that was adapted to this project, we have used a different classification algorithm since the data was a bit unbalanced. Besides that, we consider the fact that it's medical data so it's prone to bias. We have utilized the F1 score metric to evaluate how good or bad the model performs.

In terms of medical data having more False Positives means that the model is performing good. We care more about false positives than true negatives.

Test results of various stages: We have used accuracy, f1 score, recall, et precision to test the performance of each machine learning Algorithm that was adapted to this project.

We used transformers on tabular data and the performance of the algorithm was quite good and we collected the result, and we compared the result that was apply on the same sample data to each algorithm such as decision tree, random first, gradient boosting and logistic regression.

We notice that the performance of transformers on the tabular data was quite impressive and optimal as compared to many others.

VI. CONCLUSION

The result of the proposed model has shown us a very important step towards using the concept of transformer and self-attention to solve tabular data classification tasks for a small data set.

It shows us that we can in the future use a transformer to train a huge amount of data for different classification problems.

We got the following result after our first run for the comparative analysis on the diabetes dataset.

The stratified holdout cross-validation method has been used to split the training dataset randomly into 90% train and 10% test set. The result was collected and further compared with some existing approaches, which indicates that using transformers on tabular data (Tabpfn) outperforms the existing state-of-the-art approach. The Tabpfn transformer on tabular data was optimal among adapted models based on F1-score, which are 98.46 %, 98.0694%, 91.736%, and 91.541% respectively.

- *Scope of Improvement*

The scope of this project is to use machine learning techniques to predict, diagnose, and prevent lifestyle diseases, with a particular focus on diabetes. With lifestyle diseases responsible for over 80% of deaths worldwide, this project aims to develop a system that can accurately identify individuals at risk of developing diabetes at an early stage, enabling timely intervention and treatment.

The project involves a comprehensive review of the existing literature on the prediction and diagnosis of lifestyle diseases using machine learning techniques. The results of this review will inform the development of a system that is reliable, efficient, and easy to use for medical professionals of different levels of expertise.

The project also involves the collection and analysis of patient data to train the machine learning algorithms. The data will be used to develop models that can accurately predict the likelihood of an individual developing diabetes based on their lifestyle habits and other risk factors.

Another scope of the project is to compare different machine learning techniques to determine which algorithm is best suited for predicting diabetes. The project will evaluate various algorithms, including Transformers on tabular data (Tabpfn), Random Forest, Decision Tree, Support Vector Machine K-Nearest Neighbors, Gradient Boosting, Histogram Gradient Boosting, and Adaptive Boosting. The project will then identify the most effective algorithm for predicting diabetes.

Finally, the project aims to provide insights into the lifestyle habits of patients, which could be used to design targeted interventions that promote healthy habits and prevent the onset of lifestyle diseases. This will improve the overall health of the patient population and reduce the burden on healthcare systems, resulting in significant cost savings.

In conclusion, lifestyle diseases are a major health concern worldwide, and efforts must be made to prevent and treat them. Machine learning techniques have been used in this study to predict the likelihood of diabetes, a common lifestyle disease. The results of this study show that using transformers on tabular data (Tabpfn) outperformed other existing approaches, indicating that this technique could be used to improve the accuracy of diabetes prediction.

The use of machine learning in healthcare is rapidly growing, and this study is an example of how it can be used to improve diagnosis and treatment. The use of machine learning can help doctors and other healthcare professionals to make more accurate predictions about disease outcomes and can also help identify patients who are at high risk for developing lifestyle diseases. This information can then be used to develop targeted prevention and treatment strategies.

Furthermore, the findings of this study suggest that machine learning techniques can be used to improve the accuracy of diabetes prediction. This is important because early detection of diabetes can lead to better management and prevention of complications. Machine learning can help identify patients who are at high risk of developing diabetes, enabling healthcare professionals to develop targeted prevention strategies.

In conclusion, this study provides evidence that machine learning techniques can be used to improve the accuracy of diabetes prediction. This has significant implications for healthcare, as it can help healthcare professionals to identify patients who are at high risk of developing lifestyle diseases such as diabetes. This information can then be used to develop targeted prevention and treatment strategies, ultimately leading to improved health outcomes. However, it is important to note that further research is needed to fully understand the potential of machine learning in healthcare and to develop effective strategies for implementing these techniques in clinical setting

REFERENCES

[1]. International Diabetes federation. Diabetic Atlas Fifth Edition 2011.
[2]. American Diabetes Association. Diagnosis and classification of diabetes mellitus. Diabetes Care 2009;32(Suppl. 1): S62–
[3]. Krentz AJ, Bailey CJ. Oral antidiabetic agents: current role in type 2 diabetes mellitus. Drugs 2005;65(3):385–411.
[4]. Tsave O, Halevas E, Yavropoulou MP, Kosmidis Papadimitriou A, Yovos JG, Hatzidimitriou A, et al. Structure-specific adipogenic capacity of novel, welldefined ternary Zn(II)-Schiff base materials. Biomolecular correlations in zincinduced differentiation of 3T3-L1 pre-adipocytes to adipocytes. J Inorg Biochem Nov 2015; 152:123–37.
[5]. Halevas E, Tsave O, Yavropoulou MP, Hatzidimitriou A, Yovos JG, Psycharis V, et al. Design, synthesis and characterization of novel binary V(V)-Schiff base materials linked with insulinmimetic vanadium-induced differentiation of 3T3-L1 fibroblasts to adipocytes. Structure–function correlations at the molecular level. J Inorg Biochem Jun 2015; 147:99–115.

[6]. Tsave O, Yavropoulou MP, Kafantari M, Gabriel C, Yovos JG, Salifoglou A. The adipogenic potential of Cr(III). A molecular approach exemplifying metalinduced enhancement of insulin mimesis in diabetes mellitus II. J Inorg Biochem Oct 2016; 163:323–31.
[7]. Sakurai H, Kojima Y, Yoshikawa Y, Kawabe K, Yasui H. Antidiabetic vanadium(IV) and zinc(II) complexes review article coordination. Chem Rev March 2002; 226(1–2):187–98.
[8]. Nongyao Nai-arun, Rungruttikarn Moungmai(2015)Comparison of classifiers for the risk of diabetes ELSEVIER Procedia Computer Science 69 (2015) 132-142.
[9]. Pima Indian diabetes datasets from UCI Repository.
[10]. Çalisir D, Dogantekin E. An automatic diabetes diagnosis system based on LDA Wavelet Support Vector Machine Classifier. Expert Syst Appl 2011;38(7):8311–5
[11]. HianChyeKoh and Gerald Tan: Data Mining Applications in Healthcare. Journal of Healthcare Information Management, Vol 19, No 2.
[12]. P. Giudici: Applied Data Mining Statistical Methods for Business and Industry. Wiley & sons, 2003.
[13]. G.Piatetsky-shapiro, U.Fayyed and P.Smith: From data mining to Knowledge discovery: An overview. Advances in knowledge Discovery and Data Mining. pages 1-35, MIT Press, 1996.
[14]. S.Vijiyarani, S.Sudha: Disease Prediction In Data Mining Technique – A Survey. International Journal of Computer Applications & Information Technology Vol. II, Issue I, January 2013.
[15]. Huy Nguyen Anh Pham and Evangelos Triantaphyllou: Prediction of Diabetes by Employing a New Data Mining Approach Which Balances Fitting and generalization.
[16]. Og uz Karan, Canan Bayraktara, Haluk Gumus_kaya, Bekir Karlıkc: Diagnosing diabetes using neural networks on small mobile devices. Expert Systems with Applications 39 (2012) 54–60.