DIABETES PREDICTION USING MACHINE LEARNING: ENHANCING ACCURACY USING HOT ENCODING TECHNIQUE

Jose Akkarapatty, Samay Shetty Soham Pal, Sayan Panja, Sheetal M

Vidyalankar Institute of Technology, Wadala, Mumbai

ABSTRACT

Abstract: From conventional diagnostic methods to modern technology developments, the evolution of diabetes detection methods has undergone a revolutionary journey. This paper explores the development of diabetes detection techniques over time, exposing its flaws and laying the groundwork for machine learning (ML) application in healthcare. The identification of diabetes has been transformed by machine learning algorithms, which are renowned for their capacity to examine intricate patterns in sizable datasets. This study examines the critical function of machine learning in the prediction of diabetes, placing particular emphasis on the use of cutting-edge algorithms and data pretreatment methods like hot encoding. Accuracy is greatly improved by using hot encoding, a process for transforming category data into a binary matrix. This study exhibits the efficiency of ML models and highlights their excellent forecasting skills through a thorough investigation. The results highlight how ML has the potential to revolutionize healthcare by providing accurate, effective, and early diabetes diagnosis, enabling prompt interventions and individualized treatment plans.

Keywords: Machine learning, confusion matrix, diabetics, deep neural network, limited data

1.1 INTRODUCTION

Diabetes is a chronic metabolic condition characterized by elevated quantities of blood glucose. There are two types: Type 1 (in which the immune system targets insulin-producing cells) and Type 2 (which frequently corresponds to lifestyle variables). Insulin is a hormone that controls glucose absorption. In diabetes, insufficient insulin or poor insulin utilization results in hyper-glycemia, which causes symptoms such as excessive thirst, frequent urination, and exhaustion. Cardiovascular disease, kidney failure, and eye issues are examples of long-term consequences. Medication, lifestyle adjustments, and blood sugar monitoring are all part of the treatment plan. Early detection and effective management are critical for minimizing complications and preserving quality of life.

Diabetes is caused by a mix of genetic predisposition and lifestyle factors. Type 1 diabetes is mostly caused by a genetic predisposition, with the immune system mistakenly attacking insulin-producing cells. Type 2 diabetes, on the other hand, is closely connected to sedentary lifestyles, poor food habits, and obesity. Physical inactivity contributes to insulin resistance, a condition in which cells fail to respond to insulin adequately. Weight gain and metabolic dysfunction are exacerbated by unhealthy diets heavy in refined carbohydrates and saturated fats. Age and race are also important considerations. It is critical for diabetes prevention and control to address these causes through healthy lifestyle, frequent exercise, and balanced eating.

Diabetes has a pervasive effect on the body, resulting in a variety of devastating symptoms. Elevated blood sugar levels contribute to cardiovascular problems over time, increasing the risk of heart disease and stroke. Diabetic nephropathy, or kidney disease, is a common condition that can progress to renal failure. Peripheral neuropathy is characterized by nerve injury, which causes pain, tingling, and numbness, primarily in the limbs. Diabetes also endangers vision, resulting in diabetic retinopathy and an increased risk of blindness. Furthermore, weakened immune function increases vulnerability to

infections. Proper management, which includes medication, lifestyle changes, and regular monitoring, is critical in reducing these potentially severe outcomes.

Diabetes is classified into numerous categories, the most common of which are Type 1 and Type 2. Type 1 diabetes is caused by the immune system incorrectly attacking insulin-producing cells, necessitating lifelong insulin therapy. Type 2 diabetes, which is frequently associated with lifestyle factors, is characterized by insulin resistance and inadequate insulin production. Gestational diabetes affects blood sugar levels throughout pregnancy. Other less common types of diabetes include monogenic diabetes, which is caused by specific gene mutations, and secondary diabetes, which is caused by other medical diseases or drugs. Each kind necessitates customized management measures ranging from medicine and lifestyle changes to insulin therapy, stressing the importance of personalized care approaches.

1.2 Diabetes Detection Methods: A brief history

The physicians Charak and Sushrut, who lived between 400 and 500 BC, were most likely the first to notice the sweetness of diabetic urine. The diagnosis was made after noticing ants congregating around it. Charak and Sushrut discovered that the condition was more common in people who were sedentary, overweight, and gluttonous, and who ate sugary and fatty foods. Physical activity and plenty of vegetables were the mainstays of treatment for the obese, while thin persons, whose sickness was deemed more serious, were given a healthy diet. The critical fact that diabetic urine tasted sweet was also highlighted by medical books from the 9th to 11th centuries ad, most notably in Avicenna's (980-1037) medical encyclopaedia.

Diabetes was largely ignored in Europe until Thomas Willis (1621-1675) published Diabetes, or the Pissing Evil. [2] Matthew Dobson (1735-1784) of Liverpool presented the first description of hyperglycaemia in a paper published in 1776. He discovered that both the serum and urine of his patient Peter Dickonson (who passed 28 quarts of urine per day) tasted pleasant. Dobson came to the conclusion that sugar was expelled by the kidneys and that it was not "formed in the secretory organ but previously existed in the serum of the blood."[3] John Rollo (d. 1809), an Edinburgh-trained surgeon, was the first to use the adjective "mellitus" (from the Latin word meaning "honey"). He was also famous for his "animal diet," which remained the standard treatment for most of the nineteenth century. Rollo reasoned that sugar was generated in the stomach from vegetables and that the obvious remedy was to eat animal food. As a result, the regimen outlined in his 1797 book "An Account of Two Cases of Diabetes Mellitus"[4].

The French chemist Michel Chevreul (1786-1889) demonstrated in 1815 that the sugar in diabetic urine was glucose [5]. In the middle of the century, chemical tests for reducing agents such as glucose were established by Trommer in 1841, Moore in 1844, and, most notably, Fehling in 1848 to make the diagnosis. Blood glucose measurements could only be performed by trained chemists and required so much blood that they were rarely employed in clinical care or research. It was only after the Norwegian-born physician Ivar Christian Bang (1869-1918) introduced a micro-method in 1913 that it became realistic, and it was the capacity to measure glucose frequently that led to the invention of the glucose tolerance test between 1913 and 1915.

Bernard observed in between 1846 and 1848 that glucose was present in the blood of normal animals, even when they were malnourished. He also discovered larger glucose concentrations in the hepatic than in the liver, in the portal vein, as well as "enormous quantities" of a starch-like material in the liver that could be easily turned to sugar. He referred to this as "glycogen" (sugar-forming) and saw it as similar to starch in plants. His theory—the "glycogenic"-the sugar received from the intestine was

transformed, according to the notion, in the liver into glycogen, which is then continuously released into the blood during a fast, is lost. Bernard's other discovery made a big effect in a time when neurological control of body functioning was rare concept that is scientifically fashionable. He discovered that a lesion on the floor of the fourth ventricle caused hyper-glycemia (piqure diabetes) [7]. This discovery sparked a long period in which nervous influences were thought to be important causes of diabetes; indeed, J.J.R. Macleod cited "evidence" as late as 1914 that diabetes was more common among engine drivers than other railway workers due to the mental strain involved [8].

Murray's cure of myxedema in 1891 inspired hope that pancreatic extract would soon result in a diabetes cure, but after repeated failures over the next 30 years, even believers in an anti-diabetic internal secretion were pessimistic about the possibility of success isolating it, and shifted their focus to nutrition as a cure for the illness. Frederick Madison's fasting program was the most well-known. Allen (1876-1964), as described by Joslin in 1915 as the most significant advancement since Rollo's time. This strategy was an extreme version of one that had previously promoted intense exercise and "manger le moins possible." In a limited sense, starvation treatment did succeed, since some people were able to survive for months or even years. However, the quality of life was very low, and several died of starvation rather than diabetes.[9]

1.3 MACHINE LEARNING IN DIABETES DETECTION

Predictive modelling is performed by algorithms used in data mining, machine learning, or any branch of artificial intelligence. Predictive modelling is the use of data and statistics to forecast future outcomes based on historical data. Diabetes' most prevalent symptoms are improper metabolism, hyperglycaemia, and an increased risk of particular complications affecting the eyes, kidneys, and nervous system, all of which are vital organs. These symptoms are utilized to collect data, and then modelling is done based on age and gender groups. Machine learning is a method for a computing system to learn the characteristics of input data. Such approaches have been shown to be effective in the detection of diabetes. Many machine learning algorithms, including supervised, unsupervised, and reinforcement learning methods, have been created. This is feasible as machine learning methods are data-driven. Machine learning can save significant human labour when vast amounts of data are entered into the database. Models are trained on this data and output the most appropriate output based on the input data. The models can be trained on any parameters that are realistic and meet medical standards. Some may look at facial traits, while others may look at blood test results collected from patients. Because the disease has many symptoms, the parameters vary correspondingly. Researchers have examined many algorithms and altered several hyperparameters with many proposed ways to produce findings that appear most acceptable for real-life applications[11].

2 Method

The aim of the diabetes analysis project using Python is to develop a robust and accurate system for detecting and classifying different types of diabetes. The aim to deliver significant insights to healthcare professionals and enhance the diagnosis and treatment of diabetic patients by harnessing the capabilities of Python programming and data analysis packages. Diabetes analysis and detection entail a number of goals that encompass data analysis, feature engineering, modeling, and evaluation. The following are some major objectives to consider when working on a diabetes analysis and detection:

1. Data Exploration: Understanding the dataset's features is the goal. Explore the dataset's fundamental statistics, including the mean, median, and standard deviation.
2. Documentation: Create comprehensive documentation for the project, including code explanations, data descriptions, and model performance evaluations.
3. Feature Engineering: The goal is to improve the dataset by adding fresh, pertinent features.
4. Splitting data: Data preparation is the goal in order to train and assess models. Split the dataset into training and testing sets as part of your tasks. To keep the class distribution the same across both sets, think about stratified sampling. Create a validation set if you want to fine-tune your model.
5. Model Selection: Choosing the best algorithms for diabetes detection is the goal. Consider the following tasks while using conventional machine learning algorithms: logistic regression, decision trees, and random forests.
6. Model Education: Use the training dataset to run the chosen models. Use the relevant libraries (e.g., scikit-learn, TensorFlow, Keras) to implement the selected algorithms.
7. Model Evaluation: Evaluation of the trained models' performance is the goal. Perform model evaluations using metrics such area under the ROC curve, F1-score, recall, accuracy, and precision.
8. Analysis of Feature Importance: Recognize the significance of various features in prediction. Extract feature importance ratings from models (such as tree-based models) is the task at hand.
9. Interpretability of the model: Make the model predictions understandable is the goal. For interpretability, use model-independent methods (like SHAP values, for instance).
10. Fine-Tuning and Optimization: The goal is to improve the model's performance. To adapt the model architecture and hyperparameters based on the findings of the evaluation.
11. Reporting: The entire analysis process and findings should be documented.

By achieving these objectives, the analysis of diabetes project aims to make a positive impact on healthcare by providing a valuable tool for accurate, detection of diabetes and assisting medical professionals in providing better care for patients.

3 WORKING PRINCIPLE

1. Data Acquisition:

Appropriate dataset from the data repository-based website "Kaggle" as per the requirements is downloaded. The dataset is uploaded in the colab notebook.

2. Data Processing:

Libraries required like pandas, tensorflow, sklearn, matplotlib required for processing the data are imported. The dataset is displayed to extract relevant information.

3. Training and Testing of Data:

The data is now seperated under training and testing areas. This is done using sklearns which is a pre-built function "train_test_split" which serves the purpose of splitting the data randomly from the original dataset into train and test data.

A deep neural network model is trained on the train data. The network contains 3 layers with loss function binary cross entropy since the prediction outcome is in terms of yes or no. Optimizer used is

Adam with learning rate 0.01 and model fitting on train data with 100 epochs. Then the model is evaluated on the test data which gives an average accuracy of 73%.

4. Increasing Accuracy:

To increase the accuracy one hot encoding technique on the data was also implemented to observe the change in results. This data is then applied on another neural network with more layers and RMSprop optimizer. This gave an average accuracy of 95% on the training data. But since the test data was not one hot encoded it gave a validation accuracy of 75%.

Finally, a confusion matrix is created for the purpose of analysis

4 OBSERVATIONS

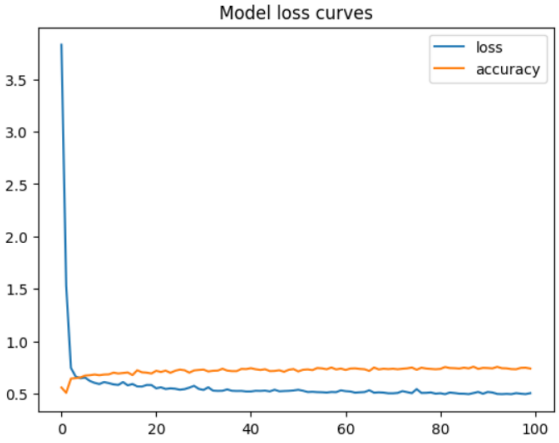

Fig 1: Model Loss Curves (Loss & Accuracy)

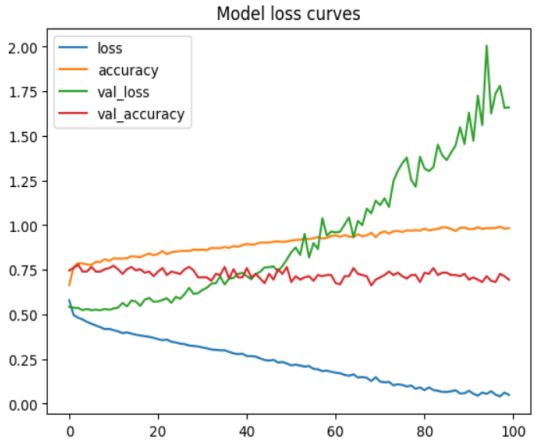

Fig 2: Model Loss Curves

Figure 1 shows the model loss curves which is observed without the use of hot encoding technique. It simultaneously displays “loss” & “accuracy” The accuracy achieved is 73%. An inverse variation is observed between the parameters. With each epoch, the loss decreases and accuracy increases.

Figure 2 shows the model loss curves which is observed with the use of hot encoding technique. It simultaneously displays “loss”,”accuracy”,”val_loss”&”val_accuracy”. The accuracy achieved is 97%. An inverse variation is observed between the parameters. With each epoch, the loss decreases and accuracy increases.

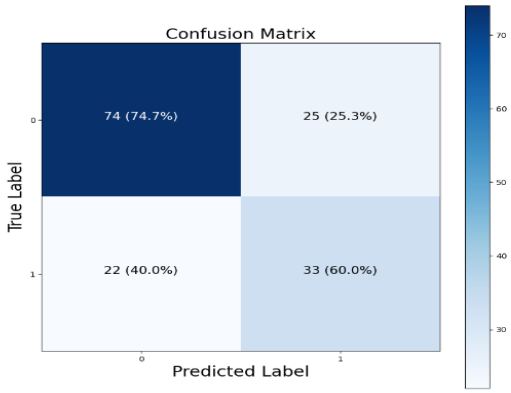

Fig 3: Confusion Matrix

The confusion matrix as observed in Figure 3 displays the accuracy of our prediction model.

4 CONCLUSION

The proposed model is a comparative study between the accuracy obtained with & without the use of hot encoding technique. After reviewing the epoch results, graphs & the confusion matrix, we observed that there is a significant increase in the accuracy percentage when the hot encoding technique is used. The future scope of the proposed model includes improving the efficiency percentage, integrating it into hardware model for practical & generic use & inclusion of improved sorting into Type 1 and Type 2 diabetes.

5 REFERENCES

1. Tattersall, R. B. (2016). The History of Diabetes Mellitus. In Textbook of Diabetes, Fifth Edition (5th ed., pp. 1–22). https://doi.org/10.1002/9781118924853.ch1
2. Willis T. Pharmaceutice Rationalis; sive, Diatriba de Medicamentorum Operationibus in Humano Corpore [2 parts in 1 vol]. Oxford: Sheldonian Theatre, 1674.
3. DOBSON, M., 1967. throughout his hospital stay. All seven patients did. JAMA, 201, pp.291-296.
4. Rollo, John, and William Cruickshank. An Account of Two Cases of the Diabetes Mellitus: With Remarks, as They Arose During the Progress of the Cure. To which are Added, a General View of the Nature of the Disease and Its Appropriate Treatment, Including Observations on Some Diseases Depending on Stomach Affection; and a Detail of the Communications Received on the Subject Since the Dispersion of the Notes on the First Case. By John Rollo, MD Surgeon-general, Royal Artillery. With the Results of the Trails of Various Acids and Other .... T. Gillet, 1797.
5. Chevreul ME. Ann Chim (Paris) 1815; 95:319–320.
6. Olmsted, J.M.D., 1953. Claude Bernard, 1813-1878: A pioneer in the study of carbohydrate metabolism. Diabetes, 2(2), pp.162-164.Bernard C. C R Seances Soc Biol (Paris) 1850; 1:60.
7. Tattersall, R.B., 2017. The history of diabetes mellitus. Textbook of diabetes, pp.1-22..
8. Tattersall, R.B., 1995. A force of magical activity: the introduction of insulin treatment in Britain 1922–1926. Diabetic medicine, 12(9), pp.739-755.Barth, S. (2023, September 26). Machine learning in healthcare - benefits & use cases. ForeSee Medical. https://www.foreseemed.com/blog/machine-learning-in-healthcare
9. Sharma, T., Shah, M. A comprehensive review of machine learning techniques on diabetes detection. Vis. Comput. Ind. Biomed. Art 4, 30 (2021). https://doi.org/10.1186/s42492-021-00097-7
10. Chevreul ME. Ann Chim (Paris) 1815; 95: 319–320.
11. Olmsted JMD. Diabetes 1953; 2: 162–164.
12. Bernard C. C R Seances Soc Biol (Paris) 1850; 1: 60
13. Macleod JJR. JAMA 1914; 113: 1226–1235.