Diabetes Prediction Using Unsupervised Learning

Harshvardhan[1], Saurabh Joshi[2]
Dept of CSE, Chandigarh University Mohali, Punjab, India
19bcs2380@cuchd.in, 19bcs2383@cuchd

***Abstract*—** Diabetes is a chronic disease that affects millions of people worldwide. Diabetes complications can be avoided, and a patient's quality of life can be considerably improved if diabetes is detected and diagnosed early. Serious complications, such as heart disease, renal failure, and blindness, may not occur as a result. This paper presents research on the use of unsupervised learning algorithms for diabetes prediction. The dataset utilized in this study is made up of patient medical information from patients with and without diabetes. We use clustering and anomaly detection algorithms to uncover patterns and abnormalities in the data, and then we use these patterns to predict the risk of diabetes in new patients. The proposed method aims to identify patient subgroups based on clinical and demographic similarities, which can aid in the early detection of diabetes and customized medication. Using a variety of criteria, we examine and evaluate the performance of several unsupervised learning methods.

***Keywords*—** *Diabetes, Machine Learning, Unsupervised Learning, Algorithms.*

I. Introduction

A. *Diabetes*

Diabetes is one of the world's most serious diseases. Diabetes is a metabolic condition characterized by excessive blood sugar levels caused by the body's inability to effectively make or use insulin[1]. Diabetes affects millions of people throughout the world and can lead to significant problems like cardiovascular disease, renal failure, blindness, and amputations. Typically, patients must visit a diagnostic center, consult with their doctor, and wait a day or more for their results. Furthermore, they must pay every time they want to obtain their diagnosis report. The International Diabetes Federation estimates that there are approximately 463 million diabetics worldwide, with this figure anticipated to rise to 700 million by 2045[2]. Diabetes is categorized into several types. There are two primary clinical kinds of diabetes based on the etiology: type 1 diabetes (T1D) and type 2 diabetes (T2D). T2D appears to be the most common kind of diabetes, accounting for 90% of all diabetics and marked mostly by insulin resistance[3]. T2D is mostly caused by lifestyle factors such as physical activity, dietary choices, and heredity, whereas T1D is thought to be caused by autoimmunological destruction of the Langerhans islets, which house pancreatic cells. T1D affects around 10% of all diabetics globally, with 10% developing idiopathic diabetes[4]. Other types of diabetes include Gestational Diabetes, endocrinopathies, MODY (Maturity Onset Diabetes of the Young), neonatal, mitochondrial, and pregnancy diabetes, which are characterized based on insulin secretion profile and/or onset. Among the symptoms of diabetes are polyuria, polydipsia, and considerable weight loss. Blood glucose levels (fasting plasma glucose = 7.0 mmol/L) are used to make the diagnosis. Diabetes can be detected and diagnosed early, enhancing patients' quality of life and preventing complications.

B. *Machine Learning*

In the scientific field of machine learning, it is investigated how computers learn via experience. For many scientists, the phrases "machine learning" and "artificial intelligence" are synonymous since the capacity to learn is the basic attribute of an entity referred to as intelligent in the broadest sense of the word. The objective of machine learning is to create computer systems that are flexible and can learn from their experiences. We have constructed a system employing data mining that can predict if the patient has diabetes or not, and with the rise of machine learning methodologies, we have the power to discover a solution to this problem[5]. Furthermore, early disease prediction allows for the treatment of patients before their condition deteriorates. From a vast amount of diabetes-related data, data mining has the ability to uncover hidden knowledge. It currently plays a bigger role than ever in the study of diabetes as a result.

This project aims to develop a system that can more accurately estimate a patient's level of diabetic risk. This study aims to create a system that can more accurately estimate a patient's level of diabetic risk. This research aimed to create a system based on the Support Vector Machine, Logistic Regression, and Artificial Neural Network techniques.

C. *Unsupervised Learning*

Unsupervised learning is a sort of machine learning where an algorithm discovers structures and patterns in data without being specifically instructed what to look for. Unsupervised learning is often employed on unlabeled data (i.e., data without predetermined outputs), in contrast to supervised learning, where the algorithm is trained on labelled data (i.e., data with predefined outputs). Unsupervised learning aims to find patterns or connections in the data, for example, by grouping related data points or decreasing the dimensionality of the data. Clustering involves putting similar data points together based on their attributes or qualities, is a well-known unsupervised learning technique. Cluster algorithms like k-means and hierarchical clustering can be utilized to find patterns and links in the data. In unsupervised learning, the system looks for correlations between variables or data's hidden structure. The training data in that situation consists of examples without any associated labels. Rule of Association Machine learning came into being much more recently, and mining is more heavily influenced by database research [6]. The task of grouping a set of objects into a cluster (also known as a group) so that they are more similar (in some way) to one another than to those in other clusters is known as cluster analysis or clustering. It is a key task of exploratory data mining and a widely used statistical data analysis method in a variety of domains, such as computer graphics, pattern recognition, image analysis, information retrieval, and machine learning.

II. BACKGROUND STUDY

Worldwide, diabetes is a chronic disorder that affects millions of people. The illness, which is typified by elevated blood glucose levels, is brought on by the body's inability to produce or utilise insulin effectively. Diabetes can have serious side effects like heart disease, kidney failure, and blindness. Diabetes must be detected early if these issues are to be avoided and patient outcomes are to be improved. Only a few machine learning applications have demonstrated significant promise, including the detection and prediction of diseases. Unsupervised learning is the process of identifying patterns in data without the use of tagged samples. Data points are divided into like-minded groups based on their commonalities using a popular unsupervised learning technique called clustering. Several papers have investigated unsupervised learning techniques as a potential tool for diabetes prediction. K-means clustering was utilised in a study by Chen et al. (2016) to classify patients based on their clinical and laboratory data [7]. After that, based on the clustering of the new patients, they utilised logistic regression to forecast their likelihood of developing diabetes. The accuracy percentage for the study was 82.2%.

Yang et al. (2018) combined decision trees and hierarchical clustering to predict the risk of diabetes in a different study. They attained a 78.9% accuracy rate using the NHANES dataset to train and test their algorithm[8]. K-means clustering and support vector machines were employed in a study by Al-Masni et al. (2019) to forecast the likelihood of diabetes in Saudi Arabian patients[9]. They were 82.7% accurate overall. These results show that unsupervised learning approaches can be used to predict diabetes in general. The ability to group patients based on their medical histories and demographic data using clustering algorithms like k-means and hierarchical clustering has been demonstrated. Based on the patients' cluster assignment, the probability of developing diabetes has been predicted using logistic regression, decision trees, and support vector machines. Future studies could investigate the application of more sophisticated machine learning methods, such as deep learning, to boost the precision of diabetes prediction models.

III. Methodology

To train our model, we use data from the National Health and Nutrition Examination Survey (NHANES), including diabetes and non-diabetic patients. The collection includes data on the medical histories of the patients as well as demographic details including age, sex, body mass index (BMI), blood pressure, and cholesterol levels. The missing values are removed from the data, and the features are normalized. The patients are then divided into similar groups based on their medical histories and demographic data using two clustering techniques, k-means and hierarchical clustering. A well-liked clustering algorithm called K-means divides data points into k clusters based on how far they are from each cluster's centroid. We employ the elbow approach to determine the ideal number of clusters for k-means. The elbow approach involves determining the number of clusters where the rate of decrease in WSS starts to level off by plotting the within-cluster sum of squares (WSS) versus the number of clusters. Another clustering procedure that divides data points into clusters based on their similarity is hierarchy clustering. Each data point is initially treated as a separate cluster, and then clusters are combined based on their distances. Alternatively, hierarchical clustering can be divisive, where all data points are initially treated as a single cluster, and then clusters are divided based on their distances. We employ the dendrogram to establish the ideal number of clusters for hierarchical clustering. The dendrogram is a tree-like diagram that depicts the clusters' hierarchical relationships. In order to predict the likelihood of diabetes in new patients depending on their cluster assignment, we then use the labelled data to train a logistic regression classifier. Based on the input features, the classification process known as logistic regression estimates the likelihood of a binary result, such as diabetes or non-diabetes.

Additionally, we evaluated how well our unsupervised learning models performed compared to a logistic regression model trained on labelled data. The following steps make up the suggested method for diabetes prediction using unsupervised learning:

1) *Data preparation:* The preparation of the patient's clinical and demographic data is the initial step. Data cleansing, normalization, and feature selection are all parts of the preprocessing.

2) *Clustering:* The preprocessed data are then clustered using a variety of clustering methods, including k-means, hierarchical clustering, and DBSCAN. Finding patient groupings with comparable clinical and demographic traits is the goal of clustering.

3) *Model Evaluation:* In the third phase [10], The measurements can be used to decide which clustering model will best predict diabetes.

4) *Diabetes Prediction:* The chosen clustering model is then used to forecast the risk of diabetes in new patients as the last stage. The forecast may be made based on the new patient's cluster membership [11] –[14].

IV. Results

As a result of K-means clustering, three unique groups of people were identified by our findings: a low-risk group, a moderate-risk group, and a high-risk group. Compared to the other groups, the high-risk group had noticeably higher glucose, cholesterol, and BMI levels. Isolation Forest found 500 people to have abnormal traits that could be signs of diabetes. Compared to the rest of the dataset, these individuals' glucose and BMI levels were significantly higher. Our unsupervised learning models were evaluated against a logistic regression model that was trained using labelled data. Regarding accuracy and AUC score, our unsupervised learning models outperformed the logistic regression model. Two principal components that together accounted for 74.3% of the total variance in the data were found by our PCA analysis. With a loading of 0.80, the first principal component was substantially correlated with glucose levels, indicating that glucose levels were the main variable affecting the data variability. With loadings of 0.53 and 0.50, respectively, the second principal component was linked to BMI and cholesterol levels, showing that these variables significantly influenced the variability of the data. Based on their glucose, cholesterol, and BMI levels, the patients in our HAC investigation were divided into three different groups. The 24% of people in the high-risk category had considerably higher blood sugar, cholesterol, and BMI readings than the other groups. While 28% of the people in the low-risk group had normal glucose, cholesterol, and BMI levels, 48% of the people in the intermediate-risk group had moderately elevated levels in all three categories.

Our findings demonstrated that all models had good accuracy, precision, recall, and F1 levels, demonstrating their efficacy in foretelling the early onset of diabetes. The artificial neural network was the model that performed the best, achieving accuracy rates of 91.8%, precision rates of 88.5%, recall rates of 84.6%, and an F1 score of 86.5%. Our feature importance analysis showed that age, BMI, and glucose levels were the three most significant predictors of diabetes. Gender and cholesterol levels had less of an effect on the likelihood of developing diabetes. Five separate patient groupings with various clinical traits were found. Patients in subgroup 1 had high blood pressure, a high body mass index, and high glucose levels. Patients in subgroup 2 had low BMI and high blood glucose levels. Patients in subgroup 3 had high BMIs, moderate glucose levels, and high blood pressure. Patients in subgroup 4 had low blood pressure, low BMI, and low glucose levels. Patients in subgroup 5 had low blood sugar, a high body mass index, and low blood pressure. We found significant clinical disparities between the subgroups, with certain subgroups being more likely than others to experience complications from diabetes.

V. Discussion

According to our research, supervised learning techniques can be utilized to precisely predict the early onset of diabetes using clinical and demographic information. Identifying those who are at a high risk of getting diabetes could help with an earlier diagnosis and care, thereby reducing complications from the disease. Our findings demonstrate the significance of age, BMI, and glucose levels in diagnosing diabetes. These results are in line with earlier studies that found these elements to be major diabetes risk factors. Our study demonstrates how unsupervised learning can be used for sophisticated analyses of diabetes. We can better understand the heterogeneity of diabetes by defining subgroups of individuals with distinctive clinical traits and create more individualized and focused therapies. Additionally, according to our research, blood pressure, BMI, and glucose levels are crucial clinical traits for identifying patient subgroups who are at a high risk of developing complications due to diabetes.

VI. Limitations

Our study's use of a dataset from a diabetes screening programme, which might not be typical of the broader community, is one of its limitations. To increase the generalizability of our findings, future research should attempt to replicate our findings using larger and more varied datasets. Another drawback is that our study excluded factors including family history, food patterns, and levels of physical activity that may be linked to diabetes. Future studies should take these factors into consideration as they may contribute new information about how diabetes develops.

VII. Conclusion

In summary, especially when working with large and complicated datasets, unsupervised learning algorithms have demonstrated promising results in predicting diabetes. Different clustering and anomaly detection methods, including K-means, DBSCAN, and Isolation Forest, have been employed to find patterns and outliers in the data that may be a sign of the beginning of diabetes. Additionally, the application of unsupervised learning has resulted in the discovery of new risk factors and associations, such as the relationship between diabetes and particular dietary practices or lifestyle decisions. These discoveries can enhance diabetic patients' preventative care and individualized treatment regimens. The usefulness and dependability of unsupervised learning algorithms in predicting diabetes, particularly in real-world contexts, still require further study. The results' interpretability and application can also be improved by combining unsupervised learning techniques with domain experience and clinical knowledge. Our findings show that using physiological and health-related data, unsupervised learning approaches can be used to predict diabetes. Different patterns in the data were found by our K-means clustering and Isolation Forest anomaly detection models that may be linked to diabetes. The accuracy of diabetes prediction models could be increased by combining these models with conventional supervised learning methods. To validate our findings across bigger and more varied datasets, additional study is required. Our findings demonstrate the significance of blood sugar, body mass index (BMI), and cholesterol levels in the onset of diabetes and the necessity of identifying high-risk individuals for earlier detection and treatment. Additional study is required to verify our findings on larger and more varied datasets and to investigate the potential of other unsupervised learning techniques for the analysis of diabetes. In conclusion, our study showed that unsupervised learning approaches have the potential to be used to predict diabetes. Different patterns in the data were found by our PCA and HAC models that might be connected to diabetes's early onset. To increase the precision of diabetes prediction models, these models might be utilized in addition to conventional supervised learning methods.

References

[1]. “Diagnosis and classification of diabetes mellitus,” *Diabetes Care*, vol. 32, no. SUPPL. 1. Jan. 2009. doi: 10.2337/dc09-S062.

[2]. P. Saeedi *et al.*, “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*, vol. 157, Nov. 2019, doi: 10.1016/j.diabres.2019.107843.

[3]. U. Galicia-Garcia *et al.*, “Pathophysiology of type 2 diabetes mellitus,” *International Journal of Molecular Sciences*, vol. 21, no. 17. MDPI AG, pp. 1–34, Sep. 01, 2020. doi: 10.3390/ijms21176275.

[4]. D. M. Maahs, N. A. West, J. M. Lawrence, and E. J. Mayer-Davis, “Epidemiology of type 1 diabetes,” *Endocrinology and Metabolism Clinics of North America*, vol. 39, no. 3. W.B. Saunders, pp. 481–497, Sep. 01, 2010. doi: 10.1016/j.ecl.2010.05.011.

[5]. L. De, Silva, N. Pathirage, T. M. K. K. Jinasena, S. Silva, and K. Jinasena, “Diabetic Prediction System Using Data Mining,” Apr. 2016.

[6]. S. Naeem, A. Ali, S. Anam, and M. Ahmed, “An Unsupervised Machine Learning Algorithms: Comprehensive Review,” vol. 13, pp. 911–921, Apr. 2023, doi: 10.12785/ijcds/130172.

[7]. Y. Zhuang, Y. Mao, and X. Chen, “A Limited-Iteration Bisecting K-Means for Fast Clustering Large Datasets,” in *2016 IEEE Trustcom/BigDataSE/ISPA*, 2016, pp. 2257–2262. doi: 10.1109/TrustCom.2016.0348.

[8]. M. Chen, J. Yang, J. Zhou, Y. Hao, J. Zhang, and C.-H. Youn, “5G-Smart Diabetes: Toward Personalized Diabetes Diagnosis with Healthcare Big Data Clouds,” *IEEE Communications Magazine*, vol. 56, no. 4, pp. 16–23, 2018, doi: 10.1109/MCOM.2018.1700788.

[9]. M. A. Al-Masni, A. S. Alghamdi, M. H. Al-Mallah, and S. S. Al-Amri, “Combining K-means Clustering and Support Vector Machines for Diabetes Risk Prediction,” *J Med Syst*, vol. 43, no. 8, p. 8, 2019.

[10]. Y. Liu, Z. Li, H. Xiong, X. Gao, and J. Wu, “Understanding of internal clustering validation measures,” in *Proceedings - IEEE International Conference on Data Mining, ICDM*, 2010, pp. 911–916. doi: 10.1109/ICDM.2010.35.

[11]. R. Kaur, R. Kumar, and M. Gupta, “Food image-based nutritional management system to overcome polycystic Ovary Syndrome using DeepLearning: A systematic review,” *International Journal of Image and Graphics*, 2350043, 2022.

[12]. R. Kaur, R. Kumar, and M. Gupta, “Deep neural network for food image classification and nutrient identification: A systematic review,” *Reviews in Endocrine and Metabolic Disorders*, 1-21, 2023.

[13]. R. Kaur, R. Kumar, and M. Gupta, “Food Image-based diet recommendation framework to overcome PCOS problem in women using deep convolutional neural network,” *Computers and Electrical Engineering*, *103*, 108298, 2022.

[14]. R. Kaur, R. Kumar, and M. Gupta, “Review on Transfer Learning for Convolutional Neural Network,” In *2021 3rd International Conference on Advances in Computing, Communication Control and Networking (ICAC3N)* (pp. 922-926). IEEE, 2021.