Analysing the risk level of diabetic neuropathy using data mining techniques

Priyadarshani, A.W.S.M.
Department of Information and Communication Technology
University of Ruhuna
sitharamp1996@gmail.com

Jayaneththi, J.K.D.B.G.
Department of Information and Communication Technology
University of Ruhuna
buddika@ictec.ruh.ac.lk

ABSTRACT

Diabetic neuropathy is one of the severe complications that damage diabetic patient's nerves throughout their bodies. This complication will cause pain and numbness in the legs and feet and affect their digestive system, urinary tract, blood vessels, and heart. This may affect as many as 50% of people; however, diabetic patients can prevent this complication or slow its progress with consistent blood sugar management and a healthy lifestyle. The main objective of this research is to analyse the risk level of getting diabetic neuropathy, and using these results; patients can perfectly manage their blood sugar level and get medical treatment before getting any serious medical issue. Mainly, data mining methodologies are used to build this model. In this model, a *k*-means algorithm was applied to analyse diabetes patient databases by taking various attributes of diabetes to predict diabetes disease. There are five main factors considered to build this model and considered factors are haemoglobin A1c (HbA1c), fasting blood sugar, cholesterol level, gender, and age. The finding from this study suggests that the data set is possibly divided into three main clusters with high accuracy. With ten iterations, the data set could be successfully clustered into three clusters, namely low-risk level (cluster 0), high-risk level (cluster 1), and intermediate-risk level (cluster 2) using the *k*-means algorithm. Furthermore, 6772 patients' data divide into several groups; cluster 0 with 5377, cluster 1 with 318, and cluster 2 with 1077.

Keywords: Clustering, Davies-Bouldin, Diseases, *k*-means.

INTRODUCTION

By considering globally and locally the situation in the world, diabetes is considered one of the diseases that cause more deaths than any other disease in the world. Diabetes is a severe, common, and costly yet manageable disease. Diabetes is one of the top 10 leading causes of death globally, and 463 million people suffer from this disease. According to the national diabetes website, the total adult population is 14,109,200, while the prevalence of diabetes in adults is 8.7%, Total cases of diabetes in adults:1232800. Diabetes is one of the vast medical issues in Sri Lanka (Cdc.gov., 2020). There were 1198100 cases reported from Sri Lanka in 2017. To prevent dangerous complications of diabetes, patients should control blood glucose. Most people are not aware of this disease, but diabetes will become the root cause of many other serious health conditions such as heart disease, stroke, kidney disease, eye problems, and nerve damage that may lead to amputation. Diabetes also increases the risk of microvascular damage and macrovascular complications. This microvascular damage and consequent cardiovascular disease ultimately lead to retinopathy, nephropathy, and neuropathy.

Diabetic neuropathy is defined as a diabetes-related complication that can develop when chronically elevated blood sugar levels damage the body's nerves. Neuropathies are the most common complication among diabetes patients, and this complication occurs because of uncontrolled diabetes or long-standing diabetes. The occurrence or the medical pathophysiology of diabetic peripheral neuropathy is a complex process. This complication can affect nerves in many different areas of the body, and it most commonly damages the nerves of the legs and feet. Up to 50% of diabetic patients may suffer from diabetic neuropathy. Diabetic neuropathy may also damage nerves of the digestive system, cardiovascular system, and genitourinary system and cause a myriad of system-specific symptoms. Diabetic neuropathy is classified into four types, and each type of diabetic neuropathy has specific characteristics. According to medical researchers, diabetic neuropathy is associated with chronically elevated blood sugar levels, damages the nerves, and reduces their transmission of nerve signals. Moreover, chronically elevated blood sugar levels can damage the small blood vessels that provide the body nerves with needed nutrients and oxygen.

Preeclampsia or eclampsia are the two types of blood pressure conditions that affect pregnant women. Moreover, diabetes raises the risk of developing high blood pressure and strain on the patient heart. This can contribute to fatty deposits in blood vessel walls when patients have high blood glucose levels. This reason can restrict blood flow and increase the risk of atherosclerosis or the hardening of the blood vessels. People with diabetes have decreased ability to filter waste products from the blood.

Moreover, this led to damage to the kidneys. Diabetic neuropathy is a type of nerve damage complication. This complication occurs because of high blood sugar (glucose), which will injure the patient's body nerves. Diabetic neuropathy is one of the severe complications, but patients can often prevent diabetic neuropathy or slow its progress with better blood sugar management and a healthy lifestyle. To prevent diabetes, several data analysis processes can be done, such as historical medical data analysis, future prediction, and identifying risk factors involved in diabetic neuropathy. Diabetic neuropathy may

Table 1: Normal, prediabetes, and diabetes levels of fasting blood glucose, HbA1c, and cholesterol

Level	Fasting blood glucose (mg/dL)	HbA1c (%)	Cholesterol (mg/dL)
Normal	<99	<5.7	<200
Prediabetes	100–125	5.7–6.4	200 239
Diabetes	126	6.5%	240

also damage nerves of the digestive system, cardiovascular system, and genitourinary system and cause a myriad of system-specific symptoms. Mainly this research is focused on future risk levels of diabetic neuropathy. Since patients are not aware of the stages of diabetic neuropathy, this research is helpful for the patients to identify which stage they belong to. For data analysis and prediction, there are risk factors involved in diabetic neuropathy, such as HbA1c test result, Fasting Blood glucose result, age, gender, and cholesterol level.

As shown in Table 1, these standard test levels are helpful for data analysis, pattern analysis, and future predictions. Most researchers have used these test levels for predictions. Moreover, other risk factors like age gender are involved in diabetic neuropathy complications in different manners. For example, according to this study, male patients with diabetes live 7.5 years less than other men who do not have diabetes. This number has increased to 8.2 years among women who have diabetes to those who do not have diabetes. In the early stages, there are no symptoms in diabetic peripheral neuropathy. However, even without symptoms, its presence raises the risk for foot ulcers by 5-7%. Suppose patients have diabetes in long. Identifying the pattern of the risk factors will be helpful to decrease the risk level of the complication because after getting future prediction results, patients can maintain the sugar level and follow good health tips according to predictions and patterns. The main reason for this kind of complication is inadequate knowledge about future risks. The majority of diabetic patients do not know which risk level they belong to. If a patient has some idea about his risk, he can prevent the complication by using the result of the predictions.

Most research is conducted to predict people getting diabetic, stages of diabetes, or chances of getting other complications such as heart attacks and kidney problems. In this study, systematic efforts were made in designing a system that predicts diseases like diabetes (Sisodia and Sisodia, 2018). During this work, three machine learning classification Anuradha Sharma (Sharma, 2016) developed data mining applications using biomedical records of the pathological attribute. Their main objective is to predict the presence of diabetes for an efficient allotment of minimum input variables biomedical signal using ANFIS and demonstrated that models developed using the ANFIS technique could be used for solving this critical issue. Diabetic neuropathy is one of the dangerous complications, and there is no model to predict or analyse the risk level of patients getting diabetic neuropathy using data mining methodologies. When diabetic patients with long-term diabetes will be at risk of getting diabetic neuropathy. Moreover, a patient's poor knowledge of diabetic neuropathy and specific factors which are the root cause for diabetic neuropathy will lead to severe health conditions such as heart disease, stroke, kidney disease, eye problems, and nerve damage. The main objective of this research is to predict the future risks of diabetic patients. Moreover, predict the risk level of diabetes patients getting diabetic neuropathy. To achieve this goal, the data set should be appropriately analysed with specific factors. Afterwards, the dataset should be arranged into groups according to their similarity. Since most diabetic patients and their medical experts are not aware of their historical medical records and patterns, this model will be helpful to keep track of their medical records and give knowledge about how critical their health condition.

METHODOLOGY

In this research, there were four main phases. Those four phases are proposed a theoretical design, collect data, data pre-processing, and data analysis. In the first phase of the methodology, a literature survey collected specific information after discussion with experts.

Analyse dataset: Initially, 8000 patient's 5 years of medical data were collected from survey and these personal data, safely stored and handled, there are 22 factors considered to build this model and considered factors are fbs_2020, hba1c_2020, cholesterol_2020, fbs_2019, hba1c_2019, cholesterol_2019, fbs_2018, hba1c_2018, cholesterol_2018, fbs_2018, hba1c_2017, cholesterol_2017, fbs_2016, hba1c_2016, cholesterol_2016, fbs_2015, hba1c_2015, cholesterol_2015, age and gender. After the data collecting process, analysed all the data according to 5 years. To conducted research gender attribute was represented as 1 and 2. Moreover, female patients represented using 1, and male patients were represented using 2. RapidMiner software is used to conduct this research. RapidMiner is a data science software platform used to do data preparation, machine learning, deep learning, text mining, and predictive analytics.

Data pre-processing: The dataset of size 8000 was decreased to 6772. There were missing values in each considered factor such as fbs_2020, hba1c_2020, cholesterol_2020, fbs_2019, hba1c_2019, cholesterol_2019, fbs_2018, hba1c_2018, cholesterol_2018, fbs_2018, hba1c_2017, cholesterol_2017, fbs_2016, hba1c_2016, cholesterol_2016, fbs_2015, hba1c_2015 and cholesterol_2015.Mainly, average values were used to fill the missing values in the data set. Replace Missing Value

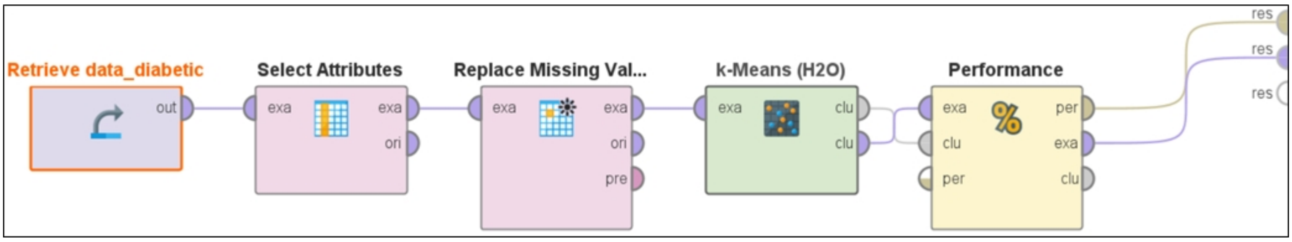


Figure 1: Deploying the model (Diabetic_Neuropathy_model)

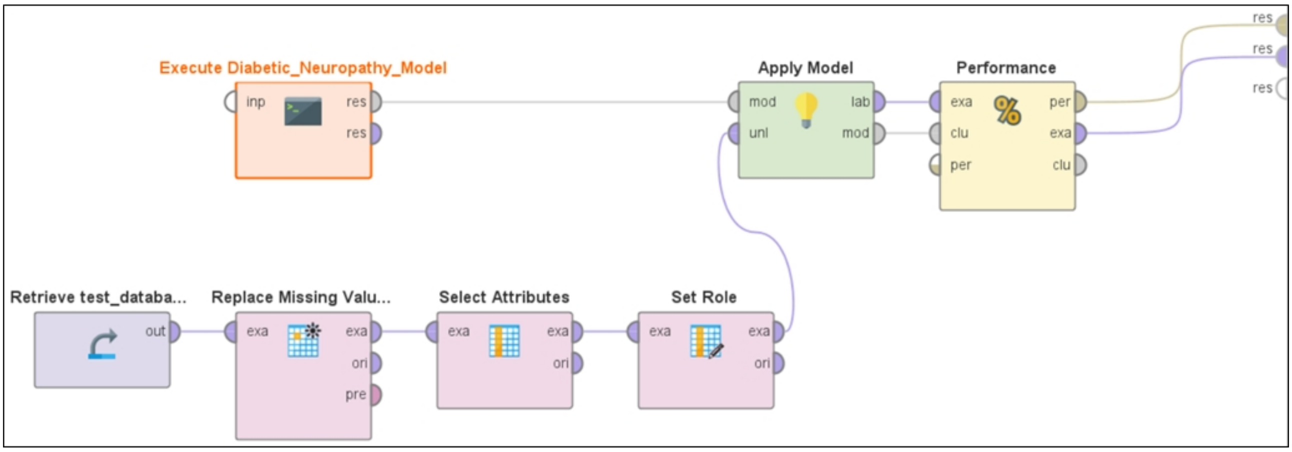


Figure 2: Testing and implementing the model

operator was applied for data pre-processing (GmbH, 2021).

Analysis: This research was conducted using the *k*-means (H2O) operator. *K*-means (H2O) operator is used to perform clustering on the provided data set, produces a model, and performs clustering using the H2O *k*-means algorithm. After using the *k*-means (H2O) operator, it is easy to visualise clusters using the cluster model visualiser operator. Data were clustered with different *k* values such as *k*=3, *k*=4, and *k*=5. As shown in Figure 1 cluster distance operator was used to validate and analyse the risk levels such as low-risk, high-risk, and intermediate-risk levels using average centroid distance. Finally, the Diabetic_Neuropathy_Model was built, as shown in Figure 1.

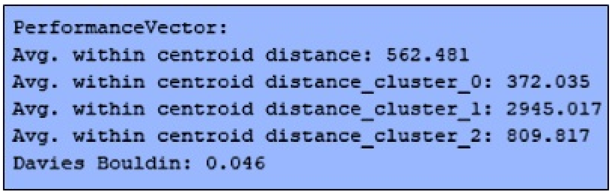
Testing and implementation: Validation is critical in this research because it validates mining models by understanding their quality and characteristics before deployment. As shown in Figure 2, Diabetic_neuropathy_Model was tested using the Cluster Distance Performance operator. By using Apply Model operator, the trained model can be applied to another data set. To test the model 134 labelled patient data were used.

Cluster	Number of items
cluster_0	5377
cluster_1	318
cluster_2	1077

Table 2: Item count of clusters (*k*=3)

RESULTS AND DISCUSSION

As shown in Table 2, the *k*-means clustering algorithm divides the data set into three main clusters. By using this model, patients can quickly identify their risk level of getting diabetic neuropathy. According to this prediction, this result can be divided into different levels: high-risk, intermediate, and low-risk. In this model, 318 patients are at high risk,1077 in intermediate-risk level, and 5377 in low risk. The main objective of this research was accomplished by dividing the data set into clusters by using specific factors and analysing their similarity.



```
PerformanceVector:
Avg. within centroid distance: 562.481
Avg. within centroid distance_cluster_0: 372.035
Avg. within centroid distance_cluster_1: 2945.017
Avg. within centroid distance_cluster_2: 809.817
Davies Bouldin: 0.046
```

Figure 3: Performance vector

K-means is an unsupervised clustering algorithm, and there are no predicted labels. Moreover, accuracy cannot be directly applied to *k*-means clustering for evaluation but comparing two examples of metrics it can be evaluated. Above mentioned result in Figure 3 was calculated by using the Davies-Bouldin method. Davis Bouldin method is an internal evaluation scheme that is used to validate clusters. Davies-Bouldin index is bounded to 0-1 (Tom, 2021).

Moreover, the lower score is the best value. When there are fewer Davies Bouldin values, comparing to other clusterings, there is high accuracy. The accuracy of this model is 0.046 with three clusters (*k*=3). In rapid miner used operator was Cluster Distance Performance operator to evaluate the model. According to this research, the accuracy level is high when there are three clusters (*k*=3), and there is a good average distance by comparing to others. Above mentioned result in Table 3 clearly shows that if there are three clusters(*k*=3), there is a high value of average centroid distance between each cluster.

Table 3: Clustered test data

Index	Nominal value	Absolute Count	Fraction
1	cluster_0	53	0.396
2	cluster_1	42	0.313
3	cluster_2	39	0.291

Finally, as shown in the Table 3, 134 of labelled patient's data were divided into three clusters. Medical experts or patients can easily identify their risk levels by entering their medical details into the test database, as shown in Table 3.

The finding from this study suggests that the data set can be divided into three main clusters with high accuracy. With ten iterations, the data set could be successfully clustered into three clusters, namely low-risk level (cluster_0), high-risk level (Cluster_1), and intermediate-risk level (cluster_2) using the *k*-means algorithm. These risk levels were identified using centroid distance as shown in table 1. Since *k*-means is an unsupervised data mining methodology, validation can be done using comparing different metrics. This model was validated using different *k* values such as three, four, and five. With three clusters model represent high accuracy. Moreover, amount of three clusters (*k*=3) showed unbalanced data distribution. Furthermore,6772 patient's data divide into cluster_0 with 5377, cluster_1 with 318, and cluster_2 with 1077 because medical data are difficult to analyse. When considering about average centroid distance between each cluster, with three clusters (*k*=3), there is a high value other than when there are four or five clusters. One of the important real-world medical problems is the analysing risk level of getting diabetic neuropathy. In this research, systematic efforts are made in designing a model which can analyse risk level of diabetic patients getting diabetic neuropathy. There is no model to analyse and predict the risk level of getting diabetic neuropathy, and this research solves this problem by introducing a new model to predict the risk level of getting diabetic neuropathy. This model was built using the *k*-means clustering algorithm and validated using the Davies Bouldin method. The most accurate result was taken, and analysed the clinical data set was. Medical experts and diabetic patients will get huge benefits from this model.

CONCLUSION

Data mining methodologies hold great potential for the healthcare industry to enable health systems to systematically use data and analytics to identify inefficiencies and best practices that improve care and reduce costs. Healthcare organisations produce vast amounts of data daily, and these data should be collected and stored in organised forms for making new knowledge. By analysing medical data, patients and medical experts can identify how critical the area of the patient's health is. In this research, 6772 patients' data were analysed to create new knowledge, and data mining methodologies such as the *k*-means algorithm were used to develop this model. The *k*-means clustering is one of the best clustering algorithms, and this algorithm was used to achieve the main objective of this research. Davis-Bouldin evaluation method was used to validate this model, and this model gives accurate results.

Furthermore, there were missing values in each considered factor, and we handled those missing values by replacing average values. This model gives patients and medical experts a good understanding of their risk level of diabetic neuropathy and the risk factors that are affected by this complication. This study solved this problem by using this developed model. One of the critical real-world medical problems is the analysing risk level of getting diabetic neuropathy. In this research, systematic efforts are made in designing a model which can analyse the risk level of diabetic patients getting diabetic neuropathy. There is no model to analyse and predict the risk level of getting diabetic neuropathy, and this research solves this problem by introducing a new model to predict the risk level of getting diabetic neuropathy. This model was built using the *k*-means clustering algorithm and validated using the Davies Bouldin method. The most accurate result was taken, and analysed the clinical data set. Medical experts and diabetic patients will get huge benefits from this model.

REFERENCES

Cdc.gov. (2020) National Diabetes Statistics Report, 2020 | CDC. [online] Available at: `https://www.cdc.gov/diabetes/data/statistics-report/` [Accessed 30 July 2021].

GmbH, R. (2021) Replace Missing Values - RapidMiner Documentation. [online] Docs.rapidminer.com. Available at: `https://docs.rapidminer.com/latest/studio/operators/cleansing/missing/replace_missing_values.html` [Accessed 30 July 2021].

Sharma, A. (2016) Data Mining Application in Diabetes Diagnosis using Biomedical Records of Pathological Attribute. *International Journal of Science and Research (IJSR)*, 5(6): 1077-1083.

Sisodia, D. and Sisodia, D.S. (2018) Prediction of Diabetes using Classification Algorithms. *Procedia Computer Science*, 132: 1578-1585.

Tom, V. (2021) Davies-Bouldin Index. [online] Tom Ron. Available at: `https://tomron.net/2016/11/30/davies-bouldin-index/` [Accessed 30 July 2021].