IJIRCCE

e-ISSN: 2320-9801 | p-ISSN: 2320-9798

INTERNATIONAL JOURNAL OF INNOVATIVE RESEARCH

IN COMPUTER & COMMUNICATION ENGINEERING

Volume 9, Issue 6, June 2021

INTERNATIONAL STANDARD SERIAL NUMBER INDIA

Impact Factor: 7.542

International Journal of Innovative Research in Computer and Communication Engineering

| e-ISSN: 2320-9801, p-ISSN: 2320-9798| www.ijircce.com | |Impact Factor: 7.542 |

|| Volume 9, Issue 6, June 2021 ||

| DOI: 10.15680/IJIRCCE.2021.0906257 |

Diabetes Prediction System Using Machine Learning Algorithms

S. Ayyappa[*1], S. Sai Krishna[*2], T. Sarath Chandra[*3], V. Sathvik[*4],

UG Student, Dept. of Electronics and communication Engineering, Vasireddy Venkatadri Institute of Technology, Namburu, Andhra Pradesh, India [*1,*2, *3, *4,]

ABSTRACT: The diabetes is one of lethal diseases in the world. Diabetes should not be ignored if it is untreated then Diabetes may cause some major issues in a person like: heart related problems, kidney problem, blood pressure, eye damage and it can also affect other organs of human body. But with the growth of Machine Learning methods we have got the flexibility to predict the Diabetes at early stage. The aim of this analysis is to develop a system which might predict the diabetic risk level of a patient with a better accuracy. Machine learning techniques Provide better result for prediction by constructing models from datasets collected from patients.

In this project models developed based on algorithms like K-Nearest Neighbour (KNN), Logistic Regression (LR), Support Vector Machine (SVM), Gradient boosting and Random Forest (RF) are compared by training them on PIMA India data base. The accuracy is different for every model when compared to other models. The Random forest algorithm gives the higher accuracy model which shows that the model is capable of predicting diabetes effectively.

I. INTRODUCTION

Diabetes is a dangerous disease that causes a great threat to human life. It affects people worldwide. The main reason behind diabetes is that it happens when the body is not capable of producing the insulin it needed. Insulin which is secreted by the pancreas is needed to maintain the glucose level in the human body. The symptoms of diabetes are frequent urination, increased thirst, and hunger. It can also be the reason for other diseases like heart disease, kidney failure, blindness, and many more. It can be controlled by insulin injection and regular exercise.

According to the research, about 422 million people are suffering from this disease in many countries. And this won't stop the cases will be increased more in the upcoming years. So to prevent this from happening we have to find the person who will be affected with diabetes or not at an early stage which helps to save more people's lives. To accomplish this we take various attributes like glucose level, blood pressure, skin thickness into consideration to predict diabetes. For this purpose, we use machine learning algorithms like classification and ensembling methods on the Pima Indian diabetes dataset to predict which model gives the best accuracy in predicting diabetes.

There are two types of diabetes Type-1 and Type-2.

TYPE 1 DIABETES

Type 1 diabetes develops when the cells of the pancreas stop producing insulin. Without insulin, glucose cannot enter the cells of the muscles for energy. Instead the glucose rises in the blood causing a person to become extremely unwell. Type 1 diabetes is life threatening if insulin is not replaced. People with type 1 diabetes need to inject insulin for the rest of their lives.

Type 1 diabetes often occurs in children and people under 30 years of age, but it can occur at any age. This condition is not caused by lifestyle factors. Its exact cause is not known but research shows that something in the environment can trigger it in a person that has a genetic risk.

The body's immune system attacks and destroys the beta cells of the pancreas after the person gets a virus because it sees the cells as foreign. Most people diagnosed with type 1 diabetes do not have family members with this condition.

TYPE 2 DIABETES

Type 2 diabetes develops when the pancreas does not make enough insulin and the insulin that is made does not work as well as it should (also known as insulin resistance). As a result, the glucose begins to rise above normal levels in the blood. Half the people with type 2 diabetes do not know they have the condition because they have no symptoms.

Type 2 diabetes (once known as adult-onset diabetes) affects 85 to 90 per cent of all people with diabetes. People who develop type 2 diabetes are very likely to also have someone in their family with the condition. It is considered a

| e-ISSN: 2320-9801, p-ISSN: 2320-9798| www.ijircce.com | |Impact Factor: 7.542 |

|| Volume 9, Issue 6, June 2021 ||

| DOI: 10.15680/IJIRCCE.2021.0906257 |

lifestyle condition because being overweight and not doing enough physical activity increases the risk of developing type 2 diabetes.

People from certain ethnic backgrounds, such as Aboriginal or Torres Strait Islander, Polynesian, Asian or Indian are more likely to develop type 2 diabetes.

When first diagnosed, many people with type 2 diabetes can manage their condition with healthy diet and increased physical activity.

Over time, most people with type 2 diabetes will need diabetes tablets to help keep their blood glucose levels in the target range. (Regular blood glucose monitoring may be necessary in order to keep track of the effectiveness of the treatment.) The starting time for diabetes tablets varies according to individual need. About 50 per cent of people with type 2 diabetes need insulin injections within 6 to 10 years of diagnosis.

II. RELATED WORK

The proposed methodology has been organized into different phases. Each phase is explained in detail.

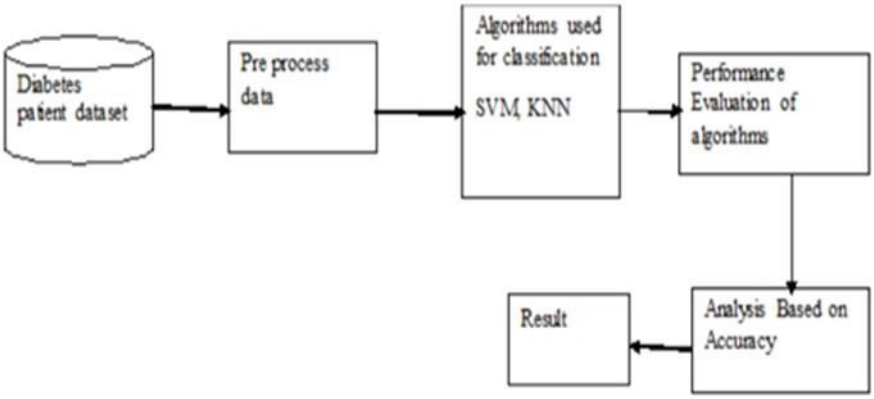


Fig 2.1 Block Diagram of diabetes prediction system using ML algorithms

ALGORITHMS

After the dataset is cleaned, we apply machine learning algorithms on the dataset. In this project we use classification and ensemble techniques to predict the diabetes. By this we can analyse the performance of the algorithms and find accuracy of them

2.1 SUPPORT VECTOR MACHINE

SVM is most popular supervised classification technique. SVM creates a hyper plane that separates the two classes. This hyper plane is used mainly for classification. To find the better hyper plane you have to calculate the distance between the planes and the data which is called Margin. If the distance between the classes is low then the chance of miss conception is high and vice versa. So, we need to select the class which has the high margin.

Algorithm-

- Select the hyper plane which divides the class bet- ter.
- To find the better hyper plane you have to calculate the distance between the planes and the data which is called Margin.
- If the distance between the classes is low then the chance of miss conception is high and vice versa. So we need to select the class which has high margin.
- Select the class which has the high margin. Margin = distance to positive point + Distance to negative point.

2.2 K-NEAREST NEIGHBOUR

KNN is the one of the best set of rules used in gadget getting to know. In this set of rules, the complete dataset is sorted. The data is divided into classes, if other data is wanted to classify then it finds the neighbours of that element based on the majority number of votes for the label. Initialize the data it calculates the distance between the classes and finding neighbours and voting for labels

To make a prediction for a new data point, the algorithm finds the closest data points in the training data set its nearest neighbours. Here K= Number of nearby neighbours, it's always a positive integer. Neighbour's value is chosen

| e-ISSN: 2320-9801, p-ISSN: 2320-9798| www.ijircce.com | |Impact Factor: 7.542 |

|| Volume 9, Issue 6, June 2021 ||

| DOI: 10.15680/IJIRCCE.2021.0906257 |

from set of class. Closeness is mainly defined in terms of Euclidean distance. The Euclidean distance between two points P and Q i.e. P (p1, p2, …. Pn) and Q (q1, q2...qn) is defined by the following equation

Algorithm-

- Take a sample dataset of columns and rows named as Pima Indian Diabetes data set.
- Take a test dataset of attributes and rows.
- Find the Euclidean distance by the help of formula

- Then, decide a random value of K. is the no. of nearest neighbours
- Then with the help of these minimum distance and Euclidean distance find out the nth column of each.
- Find out the same output values.

2.3 LOGISTIC REGRESSION

Logistic regression is a machine algorithm know as classifier. This set of rules is used to split the observations for discrete classes. The outputs given by using the logistic regression is based totally on the opportunity feature. It uses the fee function that's known as sigma characteristic. Sigma function is more complex than the normal linear function. Logistic regression limits the cost function value between 0 to1.

It classifies the data in binary form means only in 0 and 1 which refer case to classify patient that is positive or negative for diabetes.

Main aim of logistic regression is to best fit which is responsible for describing the relationship between target and predictor variable. Logistic regression is a based on Linear regression model. Logistic regression model uses sigmoid function to predict probability of positive and negative class. Sigmoid function P = 1/1+e - (a+bx) Here P = probability, a and b = parameter of Model.

Ensembling:

Ensembling is a machine learning technique Ensemble means using multiple learning algorithms together for some task. It provides better prediction than any other individual model that's why it is used. The main cause of error is noise bias and variance, ensemble methods help to reduce or minimize these errors. There are many popular ensemble methods such as – Bagging, Boosting, ada-boosting, Gradient boosting, voting, averaging etc. Here In these works we have used Bagging (Random forest) and Gradient boosting ensemble methods for predicting diabetes.

2.4 DECISION TREE

Decision tree set of rules is a supervised studying algorithm. It is used to remedy the type problems. In this algorithm whole facts are represented inside the shape of tree in which every leaf is corresponds to the class label and attribute are corresponds to inner node of the tree. The fundamental venture is to discover the foundation node in each node.

Algorithm-

- Construct tree with nodes as input feature.
- Select feature to predict the output from input feature whose information gain is highest.
- The highest information gain is calculated for each attribute in each node of tree.
- Repeat step 2 to form a subtree using the feature which is not used in above node.

2.5 RANDOM FOREST

Random forest is a supervised system getting to know set of rules. It's also used to remedy classification and regression additionally. In this algorithm it consists of the trees. The number of tree structures present in the data is directly proportional to the accuracy of the result. Each internal node within the tree corresponds to an attribute and every leaf node represents class label.

Algorithm-

- The first step is to select the "R" features from the total features' "m" where R<<M.
- Among the "R" features, the node using the best split point.
- Split the node into sub nodes using the best split.
- Repeat a to c steps until" l" number of nodes has been reached.
- Built forest by repeating steps a to d for "a" number of times to create "n" number of trees.

The random forest finds the best split using the Gin-Index Cost Function which is given by:

| e-ISSN: 2320-9801, p-ISSN: 2320-9798| www.ijircce.com | |Impact Factor: 7.542 |

|| Volume 9, Issue 6, June 2021 ||

| DOI: 10.15680/IJIRCCE.2021.0906257 |

The first step is to need the take a glance at choices and use the foundations of each indiscriminately created decision tree to predict the result and stores the anticipated outcome at intervals the target place. Secondly, calculate the votes for each predicted target and ultimately, admit the high voted predicted target as a result of the ultimate prediction from the random forest formula. Some of the options of Random Forest does correct predictions result for a spread of applications are offered.

III. PROPOSED ALGORITHM

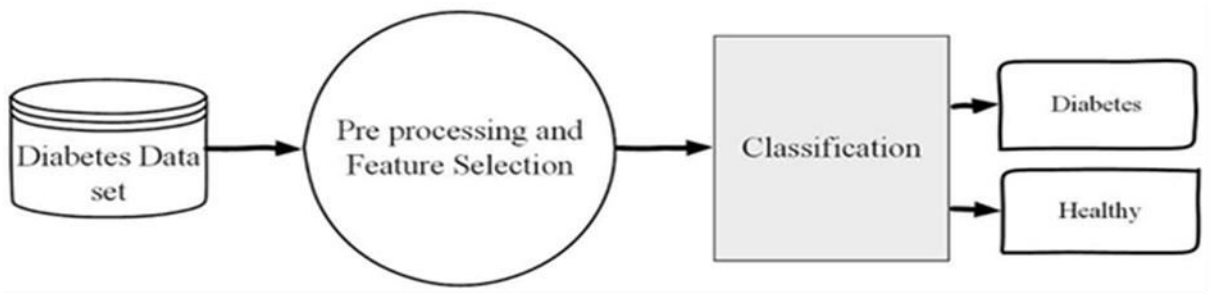


Fig 5.1 Data processing and Classification

3.1 DATA PREPROCESSING:

It involves transformation of raw data into a format that can be easily be further processed. In this project we are using healthcare data that may contains missing values that causes effectiveness of data. For that reason Data preprocessing is done. For Pima dataset we have to perform two steps for data preprocessing.

3.1.1 MISSING VALUES REMOVAL

We will remove all the instances with 0 in it. By removing the irrelevant features from the dataset we reduce dimensionality of data and works faster.

3.1.2 DATA SPLITTING

After cleaning the data we split the dataset into two segments namely training and testing sets. When the data is splitted we train the model using the training set which creates the model and we test the model using the test set which test the accuracy of the model.

3.2 CLASSIFICATION ALGORITHMS

ALGORITHMS USED

1.logistic regression

2.support vector machine

3.KNN

4.Random forest classifier

5.Naive Bayes

6.gradient boosting

3.3 DATA SET:

We are taking a CSV file of various data about huge number of diabetic patients. By applying MACHINE LEARNING algorithms and classification rules we are able to track and trace the chance of getting DIABETES.

| e-ISSN: 2320-9801, p-ISSN: 2320-9798| www.ijircce.com | |Impact Factor: 7.542 |

|| Volume 9, Issue 6, June 2021 ||

| DOI: 10.15680/IJIRCCE.2021.0906257 |

3.4 DATA FLOW CONTROL:

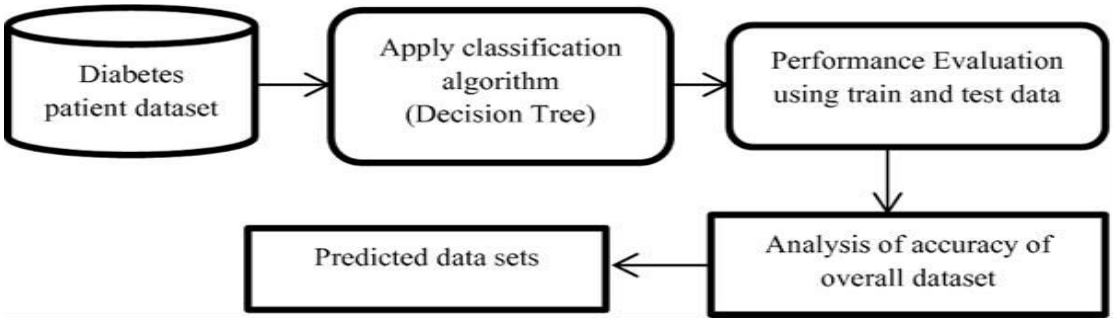
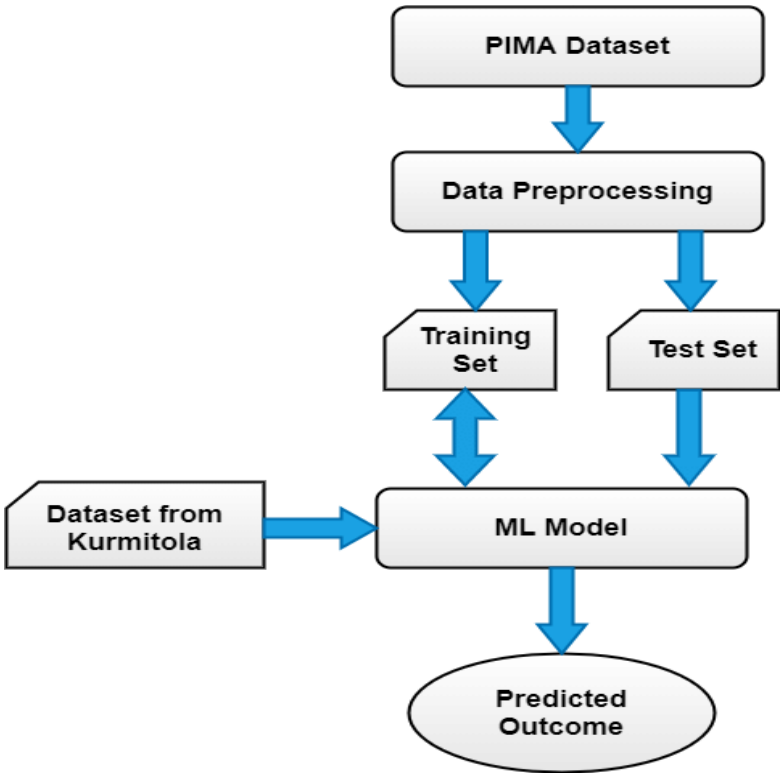


Fig 5.2 Data flow control

5.5 SIMPLE USE CASE DIAGRAM:

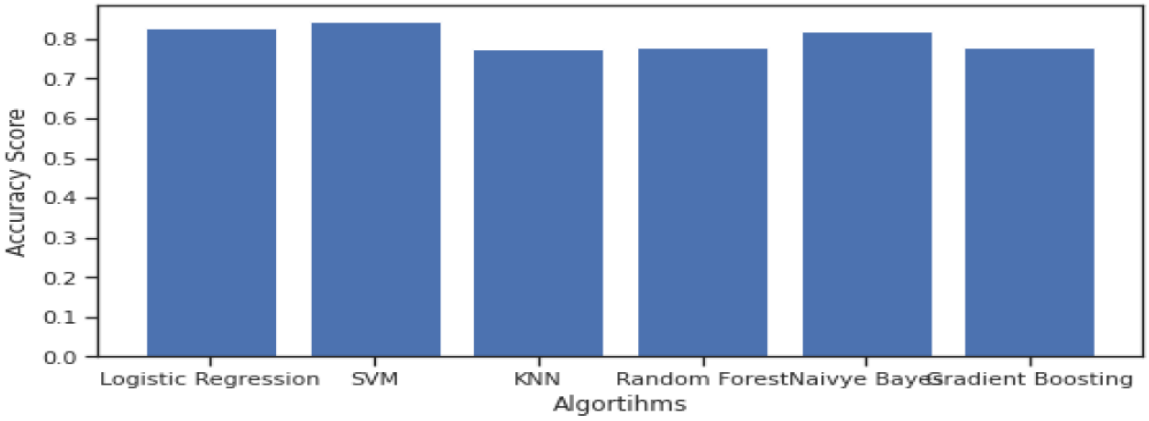


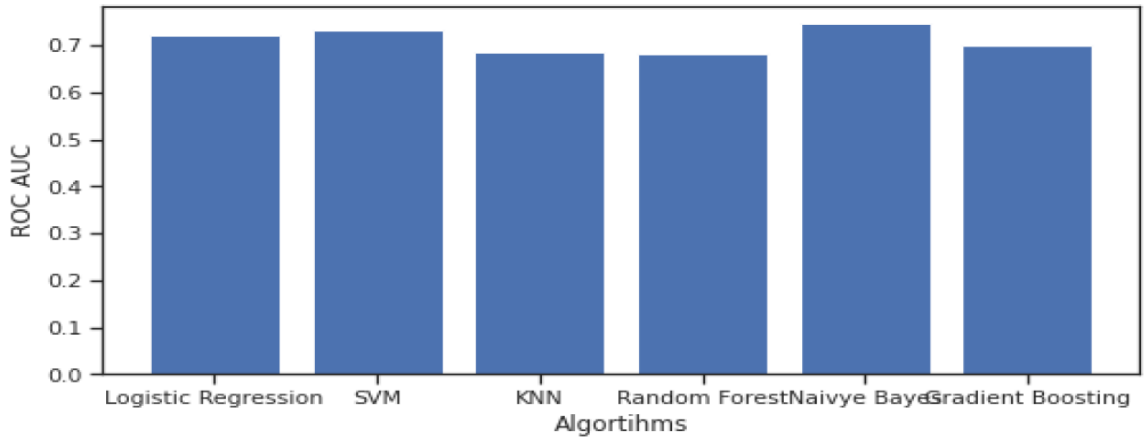
| e-ISSN: 2320-9801, p-ISSN: 2320-9798| www.ijircce.com | |Impact Factor: 7.542 |

|| Volume 9, Issue 6, June 2021 ||

| DOI: 10.15680/IJIRCCE.2021.0906257 |

IV. SIMULATION RESULTS





Algorithms	Accuracy	Misclassification
SVM	73.43	26.5
Decision Tree	72.91	27.08
Random Forest	74.4	25.5
KNN	71.3	28.64
Logistic Regression	72.39	27.60

| e-ISSN: 2320-9801, p-ISSN: 2320-9798| www.ijircce.com | |Impact Factor: 7.542 |

|| Volume 9, Issue 6, June 2021 ||

| DOI: 10.15680/IJIRCCE.2021.0906257 |

V. CONCLUSION AND FUTURE WORK

At last by using all these five machine learning algorithms we had measured different parameters within the dataset and we had come through better accuracy rate with RANDOM FOREST with nearly 75%.

This work can be extended by adding any other algorithm which can give better accuracy than RANDOM FOREST.

By measuring the accuracy of the different algorithms, we found that the most suitable algorithm for predicting the chance of getting diabetes based on the various data points from the data set is the RANDOM FOREST algorithm. The algorithm will be great asset for the medical students and healthcare industries since this algorithm is the most popular and effective compared to the others. The project demonstrates the machine learning model to predict the health issues is with morepromiment, accuracy than the previously used machine learning algorithms.

Future scope of this project will involve adding more parameters like Pregnancy, glucose, blood pressure, skin thickness, insulin, BMI, age. The more the parameters taken will be more helpful and will give more accuracy. The algorithms can also be applied for other health issues like cancer, tv, tumor. The use of traditional machine learning algorithms and data mining techniques can also help for predicting in healthcare industries.

Building the diabetes prediction system is useful for hospitals and doctors. This system predicts the disease at an early stage so doctors can treat their patients in a better way. As we use machine learning algorithms for disease prediction, we will get more accurate and efficient results

REFERENCES

[1] Deepti Sisodia, Dilip Singh Sisodia," Prediction of Diabetes Using Classification Algorithm", www.elsevier.com/locate/procedia, Procedia computer science 132(2018) 1578-1585.

[2] Xue-Hui Meng, Yi-Xiang Huang, Dong-PingRao, Qing Liu,2013," Comparison of Three Data Mining Models For Predicting Diabetes of Prediabetes By Rick Factors", Kaohsiung journal of medical science(2013) 29,93-99.

[3] V.AnujaKumari, R.Chithra. "Classification of Diabetes Disease Using Support Vector Machine". March-April 2013,pp.1797-1801.www.ijera.com.

[4] S.Selvakumar, K.Senthamarai Kannan and S.GothaiNachiyar, "Prediction of Diabetes Diagnosis Using Classification Based Data Mining Techniques", International Journal of Statistics and Systems, ISSN 0973- 2675 Volume12, Number2(2017), PP.183-188.http://www.ripublication.com.

[5] VeenaVijayan.V, Anjali. C," Decision Support Systems for Predicting Diabetes Mellitus –A Review", Proceedings of 2015 Global Conference on Communication Technologies(GCCT 2015), 978-1-4799-8553-1/15/$31.00 © s2015 IEEE.

[6] Rahul Joshi, MinyechilAlehegn, "Analysis and prediction of diabetes diseases using machine learning algorithm: Ensemble approach", International Research Journal of Engineering and Technology (IRJET), Volume: 04 Issue: 10 |Oct -2017,e-ISSN: 23950056,p-ISSN: 2395-0072. www.irjet.net.

[7] S M Hasan Mahmud, Md AltabHossin, Md. Razu Ahmed, Sheak Rashed Haider Noori, MdNazirul Islam Sarkar, "Machine Learning-Based Unified Framework for DiabetesPrediction", BDET 2018, August 25–27, 2018, Chengdu, China. © 2018 Association for Computing Machinery. ACM ISBN 978-1-4503-6582-6/18/08…$15.00.DOI:https://doi.org/10.1145/3297730.32 97737.

INNO SPACE
SJIF Scientific Journal Impact Factor
Impact Factor: 7.542

doi cross ref

ISSN INTERNATIONAL STANDARD SERIAL NUMBER INDIA

निस्केयर NISCAIR

INTERNATIONAL JOURNAL OF INNOVATIVE RESEARCH

IN COMPUTER & COMMUNICATION ENGINEERING

9940 572 462 6381 907 438 ijircce@gmail.com

www.ijircce.com

Scan to save the contact details