iJRASET
International Journal For Research in Applied Science and Engineering Technology

INTERNATIONAL JOURNAL FOR RESEARCH

IN APPLIED SCIENCE & ENGINEERING TECHNOLOGY

Volume: 8 **Issue:** VII **Month of publication:** July 2020

DOI: https://doi.org/10.22214/ijraset.2020.30632

www.ijraset.com

Call: 08813907089 | E-mail ID: ijraset@gmail.com

International Journal for Research in Applied Science & Engineering Technology (IJRASET)
ISSN: 2321-9653; IC Value: 45.98; SJ Impact Factor: 7.429
Volume 8 Issue VII July 2020- Available at www.ijraset.com

Diabetes Mellitus Prediction using Machine Learning Algorithms

A. R. Bindiya [1], K. Nikhil [2], M. S. Sindhu Rashmi [3], Shafinaz Banu [4]
[1]*Asst. Prof, Computer Science Department*
[2, 3, 4] *Students, JSS Science and Technology University, Mysore -57006*

Abstract: ***Diabetes mellitus is related to the high sugar level in the blood. According to the International Diabetes Federation (IDF), there are currently 422 million diabetic people worldwide or 7.7% of the world's population, and this number is expected to rise to 350 billion by 2030. Furthermore, 3.8 million deaths are attributable to diabetes complications every year with, an annual increase of 2.7% from 1990. In this paper, we have proposed the system to predict diabetes using a machine learning algorithm. Early detection of diabetes mellitus would lead to a decrease in the mortality rate. This paper presents an algorithm for naïve Bayes and KNN, which we have implemented using C#. KNN gave the highest accuracy (100%) compared to other algorithms. The other algorithms used are naïve Bayes, Decision tree, Logistic Regression, Random Forest, Support vector machine. A dataset that we have used to build this product contains 21 columns. This product helps in decreasing the mortality rate.***

Keywords: Data Science, Decrease mortality rate, Naïve Bayes Algorithm, KNN Algorithm, Machine learning

I. INTRODUCTION

Diabetes Mellitus is one of the major diseases that is associated with an increase in the level of blood glucose. Diabetes occurs mainly due to two reasons one reason is, Insulin cannot be used effectively by the body or when the pancreas cannot produce enough Insulin.

This high blood sugar affects different parts of our human body in particular blood veins and nerves, together with some symptoms like increased thirst, increased hunger, and weight loss. Patients of diabetes usually need constant treatment or, it may lead to many life-threatening complications. Diabetes is diagnosed, with the 2-hour post-load plasma glucose being at least 200mg/dL and the necessity of identifying diabetes timely calls in various studies about diabetes recognition. So the prediction and prevention of diabetes mellitus are increasing a lot of interest in medical sciences.

According to the International Diabetic Federation and World Health Organization, the number of diabetic patients is increasing rapidly for the past ten years, and it is one of the major chronic diseases in India.

The annual report of IDF says that 1 of every 12 had diabetes and more than 65.1 million people are suffering from diabetes in India and the annual report of world Health association adds that 422 million people are affecting from diabetes in the world that brings about 7.7 % of the world's growing population.

So the early detection of diabetes would be a great value and save human life. For this purpose, we collected the dataset, which has 21 attributes of 170 diabetic patients. So based on these attributes, we build a prediction model using various machine language techniques to predict diabetes.

Various machine learning techniques can predict diabetes mellitus, and many previous research studies have also been done about machine learning in Diabetes identification. The research work was mainly focusing on GDA (Generalized Discriminant Analysis) and SVM (Support Vector Machine). Comparing to the earlier work, we make a more comprehensive study containing several commonly used techniques used to diabetes identification, intending to compare their performance and find the best one among them. The disease diagnosis and decision makings mainly based on the Pima Indian diabetic database. We employed six popular machine learning algorithms, namely Support Vector Machine (SVM), Naive Bayes (NB), K-Nearest Neighbor (KNN), Decision tree (DT), and Random forest, Logistic regression to predict diabetic Mellitus.

The main aim of this study is to find the best optimal algorithm to predict diabetes faster and quicker. For this purpose, we used old patient data for diabetes identification of the new patient. The rest part of this paper is composed as follows Section II is about the related works, and we presented our methodology in Section III. In Section IV, we reported the experimental results. Finally, we close this paper in Section V.

ISSN: 2321-9653; IC Value: 45.98; SJ Impact Factor: 7.429
Volume 8 Issue VII July 2020- Available at www.ijraset.com

II. RELATED WORK

Predicting Diabetes using common risk factors was developed by comparing the performance of three data mining models like logistic regression, artificial neural networks (ANNs), and decision tree by using 12 risk factors. Here common risk factors of Diabetes were administered to all the participants to obtain information on demographic characteristics, family diabetes history, anthropometric measurements, and lifestyle risk factors. These models have measured in terms of accuracy, sensitivity, and specificity. The logistic regression model achieved a classification accuracy of 76.13% with a sensitivity of 79.59% and a specificity of 72.74%. The ANN model reached a classification accuracy of 73.23% with a sensitivity of 82.18% and a specificity of 64.49%. Decision tree (C5.0) achieved a classification accuracy of 77.87% with a sensitivity of 80.68% and specificity of 75.13%. Thus the results indicated that the C5.0decision tree model performed best on classification accuracy.

Performance Analysis of Classifier Models to Predict Diabetes Mellitus was developed mainly for comparing the performance of algorithms using data mining techniques for the prediction of Diabetes. The data samples have been downloaded from the UCI machine learning data repository. Here four prediction models have been employed using J48 Decision Tree, K-Nearest Neighbors, and Random Forest, Support Vector Machines for predicting Diabetes using eight important attributes under two different situations. The experiment results concluded that the decision tree J48 classifier achieved the highest accuracy of 73.82 % (before pre-processing the dataset). On the other hand, that is after pre-processing the dataset achieved accurate results for both KNN (k=1) and Random Forest and provided 100% accuracy. Thus removing the noisy data from the dataset, one can get a good result for the problems.

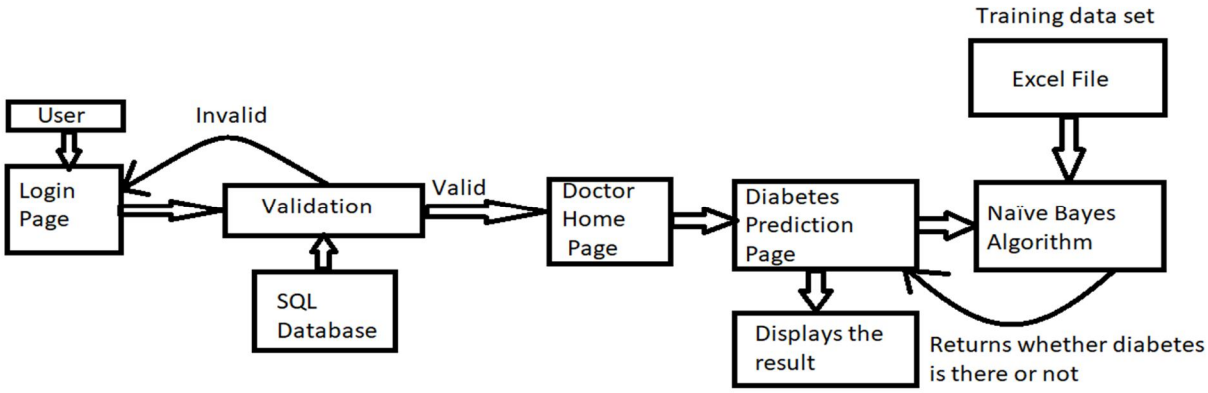
Application of Data Mining Methods and Techniques for Diabetes mellitus was developed mainly for mining the relationship in Diabetes data for efficient classification and predicting whether the likelihood of a patient being affected with Diabetes or not. So data mining methods and techniques were applied to identify the suitable methods and techniques for efficient classification of Diabetes datasets and in mining useful patterns. The training dataset used for data mining classification was the Pima Indians Diabetes Database of National Institute of Diabetes and Digestive and Kidney Diseases. This dataset contains about 768 record samples, each having eight attributes. Many classification algorithms have been employed on Diabetes datasets like C-RT, CS-RT, C 4.5, ID3, K-NN, LDA, NAÏVE BAYES, PLS-DA, SVM, RND TREE. The experimental results show that the performance of the C4.5 decision tree is significantly superior by achieving a classification rate of 91%.

Intelligible Support Vector Machines for Diagnosis of Diabetes Mellitus was developed mainly for the prediction of type 2 diabetes using a real-life data set. Support Vector Machines have been utilized for the diagnosis of Diabetes particularly, the use of an additional explanation module, which turns the “black box” model of an SVM into an intelligible representation of the SVM’s diagnostic (classification) decision. The data set used was from the National Survey of Diabetes data collected in the Sultanate of Oman. Results on a real-life diabetes data set show that intelligible SVMs provided a promising tool for the prediction of Diabetes. A comprehensible ruleset has been generated, with a prediction accuracy of 94%, sensitivity of 93% and specificity of 94%.

Analysis of various Data mining techniques to predict diabetes mellitus was developed mainly for an early prediction of Diabetes. The dataset has taken 768 instances from PIMA Indian Diabetes Dataset to determine the accuracy of the data mining techniques in prediction. Five predictive models have been employed like Naïve Bayes, Logistic Regression, C5.0, SVM, and ANN, using nine input variables and one output variable from the dataset. A comparative analysis was made between the models by making use of their Metric Measures say Accuracy, Precision, Sensitivity, Specificity, and the F1 Score. As a result, C5.0 and Logistic Regression were equally good based on their Accuracy measures, the Naïve Bayes algorithm has the Second highest accuracy, followed by ANN and the lowest accuracy is predicted, in the SVM algorithms.

III. METHODOLOGY

To get the advantage of Diabetes Mellitus prediction using a Machine learning algorithm in real-time, a friendly user interface and server to store the credentials are required. A friendly user interface is developed using Microsoft visual studio 2010. ASP.net is used for web development to produce dynamic web pages. Different machine learning algorithms are used, and the algorithm which gave high accuracy is used to predict the diabetes of the new patient. The data of the new patient is entered manually by the doctor. That data is passed as the string array to the algorithm. The algorithm will predict whether the new patient has diabetes or not. The only doctor has the privilege to use the diabetes prediction page.



A. Dataset Used

1) *Age*- Glucose tolerance decreases with age, as age increases glucose level increases but decreases the Insulin capacity to reduce glucose.
2) *Duration of Diabetes* - As the duration of diabetes increases, the blood sugar level also increases.
3) Last eye examination – Diabetes increases the risk of glaucoma and other eye problems. Early detection helps in early treatment and prevent complications.
4) *Diabetes Treatment* – Early and proper medication helps in controlling the rise of blood sugar and prevents complications.
5) *Smoke per Day* – It may lead to many serious complications and infections, heart, and kidney problems.
6) *Weight* – The chance of weight gain increases in diabetes, which may lead to an increase in the cholesterol level.
7) *B.P Systolic* - 1-4% increases in type 2 diabetes.
8) *Family History of Diabetes* - Most individuals with a family history have a risk of pre-diabetes.
9) *Past Smoked* – Smoking has ill effects on health.
10) *Drinking* – Mild moderate amount of alcohol may cause blood sugar level to rise, excess alcohol decreases the blood sugar level to dangerous levels.
11) *Abdominal Circumference* – Increased waist circumference is a strong prediction of diabetes in many cases.
12) *Gender* - Both genders are equally affected. The Ratio varies depending on lifestyle, hormones, Genetics.
13) *Diabetes Diagnosed* – Early diagnosis helps in early cure and prevents further complications.
14) *Symptom* – Symptoms help in the diagnosis of disease.
15) *Started Smoking* – As smoking is associated with many health risks, patients are advised to quit smoking.
16) *Height* – Tall structure is associated with a lower risk of developing type 2 diabetes.
17) *B.P Diastolic* – Usually systolic B.P increases four times the Diastolic B.P

Test Name	cut off for diabetes
SIC	>6.5%
Fasting plasma blood glucose	>126 mg/dl
Oral glucose tolerance test	> 2-hour blood glucose >200mg/dl
Random plasma glucose test	> 200mg/dl

B. Algorithms Considered for our Comparison Analysis for Predicting Diabetes.

1) Decision tree
2) KNN
3) Logistic Regression
4) Random Forest
5) Support vector machine
6) Naïve Byes

C. Naïve Bayes Algorithm

```
Naïve Bayes (String values)
        Step 1: Store the outcomes in the Array lists
        Step 2: Initialize _outcome array with ones.
        Step 3: Find the initial p
            In our case there only two classes therefore p = 0.5
        Step 4: Reading Training Dataset.
        Step 5:
            for i=0 to s.count
             for j=0 to features.Length
                for d = 0 to Number of rows in the training data set
                    if data in the training set == values[j]
                            ++n
                            If result of the same row == s[i]
                                    ++n_c
                              End if
                    End if
                End for
                Pi = (n_c + m*p) / (n+m)
                mul.add(Pi)
            End for

            for j=0 to j<mul.count
                    if mul[j] != 0
                        _outcome[i] = _outcome[i] * mul[j]
                    end if
             end for
             _outcome[i] = _outcome[i] * p
             end for
             if  _outcome[0] > _outcome[1]
                    return "0"
            else
                    return "1"
```

D. KNN Algorithm

```
KNN (string [] values)
```

1) Step 1: Store the outcomes in the Array Lists
2) Step 2: Set the value m=8 K-nearest neighbor
3) Step 3: Read the data from diabetes Dataset and store it in the data table dt
4) Step 4: Finding the distance between the objects

```
            for i=0 to dt.Rows.count
            for j=0 to j < values.length
             _val+=Math.Pow(double.Parse(dt.Rows[i][j+1].ToString())-double.Parse(values[j].ToString()),2);
            end for
            val = Math.Sqrt(_val);
            Append the value to the _Distance Arraylist
            Append the corresponding ID of the patient to the _Patient_d arraylist
            Copy the _Distance arraylist to temp Arraylist
            Sort the temp Arraylist
            //To get top m nearest neighbors
```

ISSN: 2321-9653; IC Value: 45.98; SJ Impact Factor: 7.429
Volume 8 Issue VII July 2020- Available at www.ijraset.com

```
  for y=0 to y<m
  d=0;
for z=0 to z<_Distance.count
     if(_Distance[z].Equals(temp[y]))
        if(d==0&&!arrayExists.contains(_PatientId[z]))
              append the corresponding patient ID to arrayPatients and arrayExists
            d++
          end if
        end if
      end for
 end for
Add cnt to arrayCnt Arraylist
Add s[i] to arrayOutcome Arraylist
Copy the arrayCnt Arraylist to temp1 Arraylist
              Sort the temp1 Arraylist
Reverse the temp1 Arraylist
For y=0 to y<arrayCnt.Count
       if(arrayCnt[y].Equals(temp1[0]))
                  _output = s[y].ToString()
                return _output
          end if
    end for
  end if
end for
```

E. Hardware Requirements

1) *RAM:* 4GB and Plus
2) *Processor:* Intel Quad Core and Higher versions
3) *Processor Speed:* 2.4ghz+
4) *Hard Disk:* 40GB and more

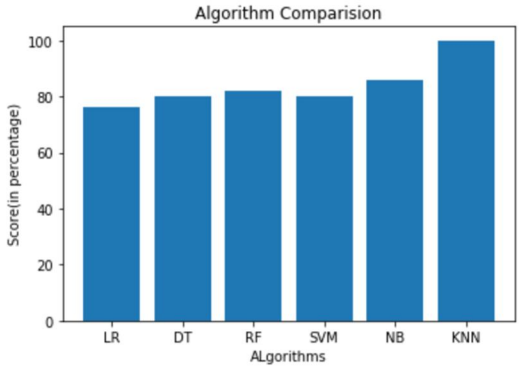
F. Software Requirements

1) *Frame Work:* DOTNET
2) *IDE:* Visual Studio 2010 or higher
3) *Front end:* ASP.NET 4.0
4) *Programming Language:* C#.NET
5) *Back End:* SQL Server
6) *OS:* XP, Win7, Win8, Win10
7) *Browsers:* IE, Firefox, Google chrome etc.

IV. RESULTS AND DISCUSSION

The naïve Bayes and KNN algorithms are written in C#. Logistic regression, Random forest classifier, decision tree classifier, Support vector machine algorithms are implemented using the modules in python language. A comparative analysis was done among the above algorithms. Accuracy of the K-nearest neighbor algorithm was highest with 100% accuracy for k = 8, and the second-highest accuracy was for the Naïve Bayes algorithm with 86% accuracy. Accuracy of the other algorithms are as bellow:

1) Logistic Regression – 76%
2) Random Forest Classifier – 82%
3) Decision Tree Classifier – 80%
4) Support vector Machine – 80%

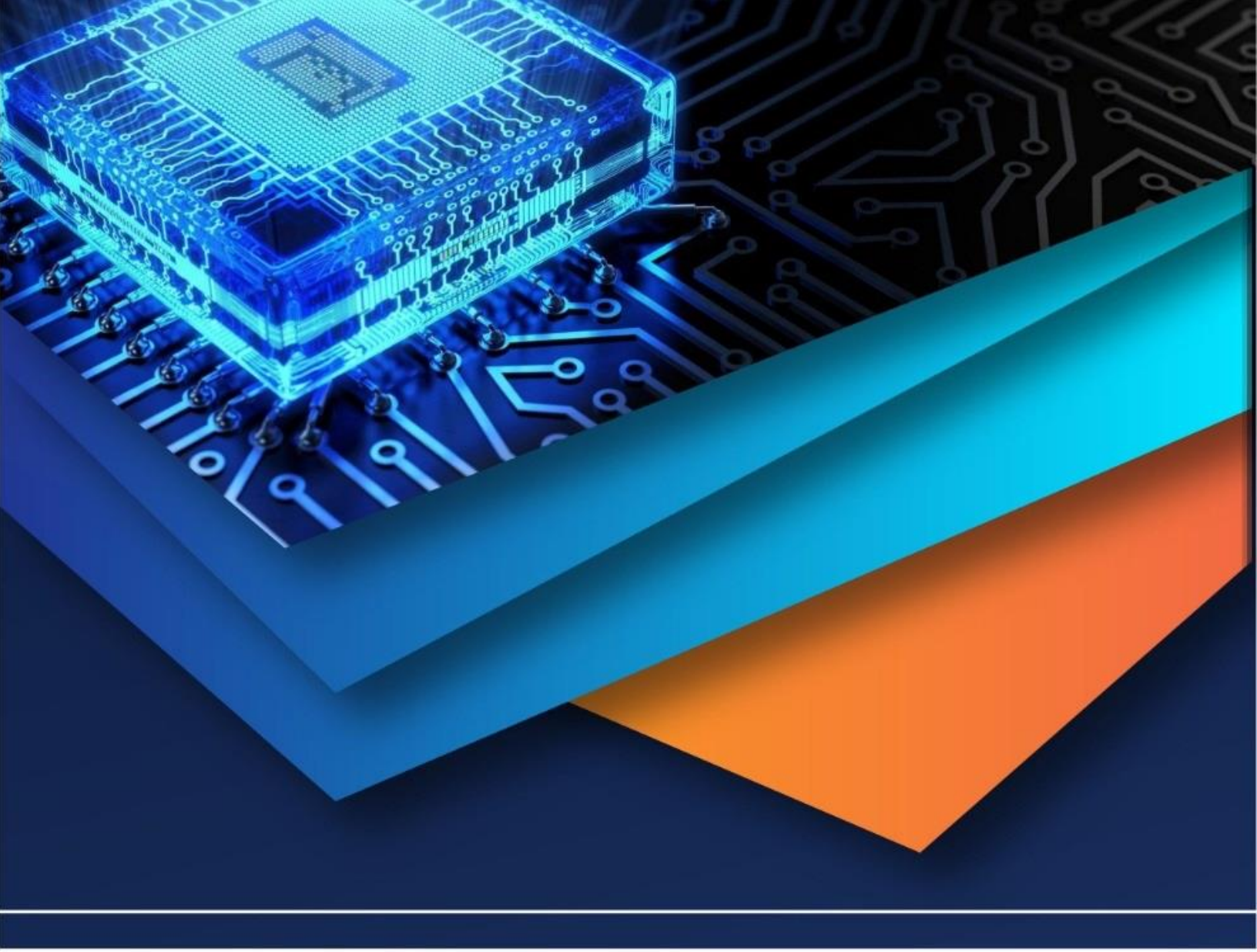


V. CONCLUSION

Today, machine learning algorithms have great importance. Using technology in the medical field will help in decreasing the mortality rate. Diabetes leads to many other problems. Maintaining a sugar level will help people to have good health. Using the KNN algorithm, we can help people to predict diabetes as it has the highest accuracy. The algorithm is written for a dynamic dataset. At regular intervals, we can add any information about new diabetes patients, which will help to predict diabetes accurately. Using algorithms mentioned in the methodology section can be used to customize for hospitals using the data collected by hospitals. Front-end can be developed by using more advanced technology like React JS.

REFERENCES

[1] World Health Organization, “Report of a study group: Diabetes Mellitus,” World Health Organization Technical Report Series, Geneva, 727, 1985.
[2] Kemal Polat, Salih Gunes, and Ahmet Arslan, “A cascade learning system for classification of diabetes disease: Generalized Discriminant Analysis and Least Square Support Vector Machine,” Expert Systems with Applications, vol. 34. 1, January. 2008, pp. 482-487.
[3] Kayaer K and Yildirim T, “Medical diagnosis on Pima Indian diabetes using general regression neural networks,” Proceedings of the international conference on artificial neural networks and neural information processing, 2003, pp. 181-184.
[4] Jack W. Smith, J. E. Everhart, W. C. Dickson, W. C. Knowler, and R. S. Johannes, “Using the ADAP learning algorithm to forecast the onset of diabetes mellitus,” Proc. Annu. Symp. Comput. Appl. Med. Care, November 9. 1988, pp. 261-265.
[5] Karegowda A. G., Manjunath A. S. and Jayaram M. A., “Application of genetic algorithm optimized neural network connection weights for medical diagnosis of pima Indians diabetes,” International Journal on Soft Computing, vol. 2. 2, 2011, pp. 15-23.
[6] Carpenter G. A. and Markuzon N., “ARTMAP-IC and medical diagnosis: Instance counting and inconsistent cases,” Neural Networks, vol. 11. 2, 1998, pp. 323-336.
[7] Wold S., Esbensen K. and Geladi P., “Principal component analysis,” Chemometrics and intelligent laboratory systems, vol. 2. 1-3, 1987, pp. 37-52.
[8] Balakrishnama S. and Ganapathiraju A., “Linear discriminant analysis-a brief tutorial,” Institute for Signal and information Processing, vol. 18, 1998.
[9] Deng L. and Yu D., “Deep learning: methods and applications,” Foundations and Trends in Signal Processing, vol. 7. 3-4, 2014, pp. 197-387.
[10] Lee H., “Tutorial on deep learning and applications,” NIPS 2010 Workshop on Deep Learning and Unsupervised Feature Learning, 2010.
[11] Safavian S. R. and Landgrebe D., “A survey of decision tree classifier methodology,” IEEE transactions on systems, man, and cybernetics, vol. 21. 3, 1991, pp. 660-674.
[12] Suykens J. A. K. and Vandewalle J., “Least squares support vector machine classifiers,” Neural processing letters, vol. 9. 3, 1999, pp. 293-300.
[13] Hosmer Jr. D. W., Lemeshow S. and Sturdivant R. X., “Applied logistic regression,” John Wiley & Sons, 2013.
[14] Lin Y., “Support vector machines and the Bayes rule in classification,” Data Mining and Knowledge Discovery, vol. 6. 3, 2002, pp. 259-275.



10.22214/IJRASET

45.98

IMPACT FACTOR: 7.129

IMPACT FACTOR: 7.429

INTERNATIONAL JOURNAL FOR RESEARCH

IN APPLIED SCIENCE & ENGINEERING TECHNOLOGY