Comparing Classification algorithms using PIMA database

ABSTRACT

Data mining has become one of the subfield subject of software engineering. It has become the technology for the process of filtering data based on the own business need from the large sets of data using various machine learning techniques. Classification plays tremendous role in data mining process, especially for huge amount of data and it is suitable for predict new knowledge and discover patterns. This process can work with different types of data whether it was nominal or continuous. In this paper different classification algorithms comparisons will be performed on (Pima-Indian-Diabetes PID) diseases database to compare giving the best accurate result and do the prediction. These Predictive analysis is a method that integrates various data mining techniques, machine learning algorithms and statistics that use current and past data sets to gain insight and predict future risks

INTRODUCTION

Healthcare information systems tend to capture data in databases for research and analysis to assist in making medical decisions. As a result, medical information systems in hospitals and medical institutions become larger and larger and the process of extracting useful information becomes more difficult. Traditional manual data analysis has become inefficient and methods for efficient computer-based analysis are essential. To this aim, many approaches to computerized data analysis have been considered and examined. Data mining represents a significant advance in the type of analytical tools currently available. It has been shown to be a valid, sensitive, and reliable method to discover patterns and relationships. It has been proven that the benefits of introducing data mining into medical analysis are to increase diagnostic accuracy, to reduce costs and to reduce human resources [1,2,3]

Classification is mainly used to grouping up or categorizing statistics or something on the values in their similarity and it provides meaningful group or category that we can use for another purpose. Classification is used in every field because classifying a data according to its value and similarity will help for resolving the risk assessment process. To study the Attribute feature selection and perform classification, dataset was collected from UCI Machine Learning repository to test whether pregnant women is infected or not infected with diabetes [4]

The steps involved in knowledge extraction are as follows:

1. Data Cleaning: The information obtained may contain some errors which is preprocessed in this stage.
2. Data Integration: Data available in various forms that are to be integrated.
3. Data Selection: The data which is suitable for user application.

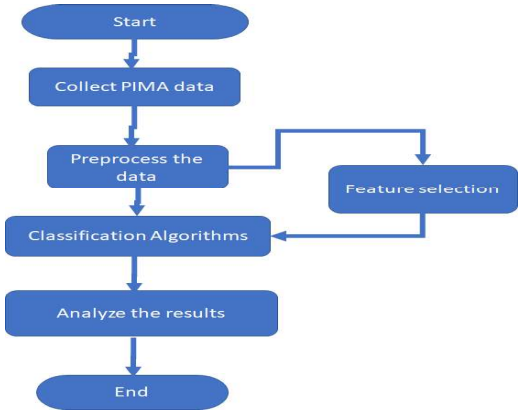
4. Data Reduction: Since they are large amount of data it occupies more space, so using this method we reduce the space but it achieves the same results.
5. Data Mining: A new methodology to extract the essential data.
6. Pattern Evaluation: It is the process in which a pattern is identified.
7. Knowledge Representation: This is the final stage in which the knowledge is represented using different visualization techniques.

RELATED WORK

D. Shetty, K. Rit, S. Shaikh and N. Patil found that prediction on small datasets does not give accurate result, when they predicted a recommendation of how to control diabetes, even the minor case of the diabetes gave the information. They used two algorithms Naïve Bayes and K-Nearest Neighbor to predict the results and compare the accuracy. In their proposed system they will be working on tackling with the diabetes attributes, improving the accuracy of the system. In I. S. Jasim, A. Deniz Duru, K. Shaker, B. M. Abed and H. M. Saleh used two classification algorithms ANN and KNN to compare the accuracy. Out of these two algorithms, firstly they trained the dataset and processed ANN and KNN algorithms out of which ANN gave the better results compared to KNN. [7]. For predicting diabetes disease, paper [8] has presented a comparison between Decision Tree (accuracy- 76.96%) and Naïve Bayes algorithm (accuracy- 79.56%) with the help of WEKA tool. Milan Kumari.et al [13] used decision tree to predict cardiovascular disease and it classify the patients who have swollen glands and diagnose the patients whether they are infected from fever, cold or throat pain. The dataset used here is Cleveland cardiovascular disease dataset from UCI repository.

METHODOLOGY

The study will perform different classification algorithms and give the comparison of the accuracy, by selection which algorithm works the best and do the prediction using that algorithm.



DATA COLLECTION

In this study Pima-Indian-Diabetes PID used as the dataset to performs classification, this dataset collected from pregnant women with diabetes by university of California, Irvine Repository (UCI) of Machine learning databases [9]. (Pima-Indian-Diabetes) dataset is real dataset consist of 768 instances each instance has 9 attributes including the class attribute (the last one), these attributes are:

Sr No	Attribute Name
1	Number of times pregnant.
2	Triceps skin fold thickness (mm).
3	2-hour serum insulin (mu U/ms).
4	Diastolic blood pressure (mmHg).
5	Body mass index (weight in kg/ (height in m)) 2.
6	Diabetes pedigrees function.
7	Age (years).
8	Class attribute (infected 1 not 0).

DATA PREPROCESSING

Most of the data sets used in data mining were not necessarily gathered with a specific goal in mind. Some of them may contain errors, outliers or missing values [9].

The predictions accuracy of any classification depends on both quality and quantity of the data. So, if the data has many instances but the data that is inconsistent, leads to the low accuracy prediction. To solve this, data processing technique plays a very important role to improve the quality of the data. There are many techniques to detect data anomalies and preprocess the data such as data cleaning, data integration, data transformation and data reduction. In the PIMA dataset, there are many instances having missing data and inaccurate values for which data cleaning and data reductions techniques will be used.

Cleaning: Firstly, duplication of data is checked on columns through which each record is identified. Duplicates were deleted as we have numerous rows. Also as a method to handle outliers without reducing granularity.

Noisy means incorrect or erroneous data. So, check columns, their format, e.g. is date is in correct format, validate true/false values attributes for other values etc.

Dimensionality Reduction

Out of the attributes we eliminated few which had repetitive data, and which presented unnecessary information. We used step-wise forward selection method to choose from attributes and place in an empty set, thereby building the selected attribute set.

Normalization

To fit a dataset to a model, we need to convert it into numerical form. After same has been done, they all need to be in a range. So, used various methods like: Decimal Scaling, StandardAero method from SciKit library to normalize the data.

Dealing with 0 or null values.

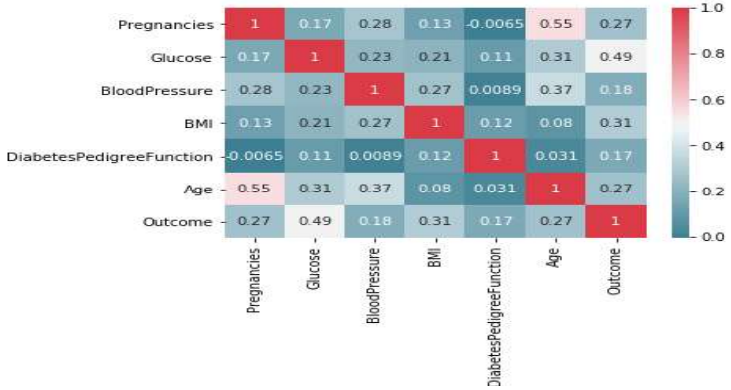
Pregnancies	110
Glucose	5
Blood Pressure	35
Skin Thickness	227
Insulin	374
BMI	11
Diabetes Pedigree Function	0
Age	0
Outcome	0

Seeing the above table, some of the data like Blood pressure, Skin thickness, Insulin, etc. cannot be 0. Hence, there are difference way to handle this data. So, the first and easy way to handle missing data is deleting the null values from the data. The second approach is to replace all missing values with the mean. The third technique is to replace all the missing values with zeros. However, this technique may lead to poor classification. The fourth technique K-nearest neighbor method replaces missing values in data with the corresponding value from the nearest-neighbor column. The nearest-neighbor column is the closest column in Euclidean distance. However, this technique also might bias the dataset.

So, in this study, data was handled in the following way

- As Glucose, Blood pressure and BMI has less missing values. It was replaced by the median value of their instance.
- For the zero values of the attribute Pregnant, the zero values are left assuming that they are the real values.
- The attributes Skin thickness and Insulin have very large number of missing values (227 zero values for skin thickness and 374 zero values Insulin), so the instances cannot be removed because it might miss the important data by deleting it, instead the attribute themselves were removed to keep as much accurate information as possible

Outcome	Infected	Not infected	Total
Before preprocessing	268	500	768
After preprocessing	230	427	657



Looking to the heatmap it is showing that age is playing an important role in pregnancy, also glucose level is next important factor to be considered for diabetic patient.

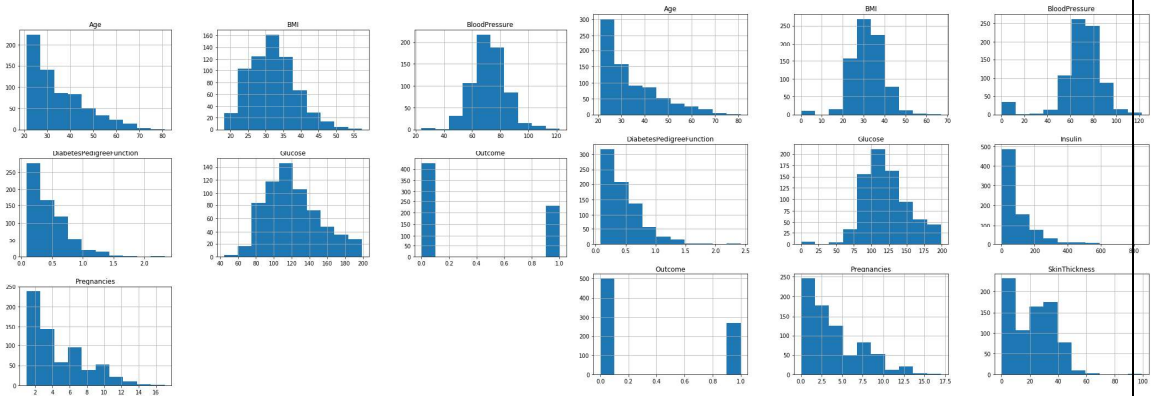


Fig:1 After Preprocessing

Fig 2 : Before preprocessing

FEATURE SELECTION AND FEATURE EXTRACTION [10]

Feature Selection is very dynamic and productive field and research area of machine learning and data mining. The main goal of feature selection is to choose a subset by eliminating nonpredictive data. Furthermore, it increases the predictive accuracy and reduces the complexity of learned results

Feature selection mechanism is done to eliminate the redundant features and extract the information. Feature should be selected in such a way that the:

1. The accuracy and performance should not be affected.
2. output must be the same.

A better feature selection mechanism helps in facilitating the data visualization, data understanding, reducing the storage requirements and utilization in time and in reducing the dimensionality.

Feature extraction creates new variables as a combination of others to reduce the dimensionality of the selected features. Feature extraction can be classified into two:

1. Linear and
2. Non-linear

Feature extraction has been one of the most important issues of pattern recognition. Most of the feature extraction literature has centered on finding linear transformations, which map the original high-dimensional sample space into a lower-dimensional space that hopefully contains all discriminatory information. The principal motivation behind dimensionality reduction by feature extraction is that it may reduce the worst effects of the curse of dimensionality. Also, linear feature extractions techniques are often used as pre-processors before more complex nonlinear classifiers.

All these attributes are numeric, to present the standard data attributes is needed. First, and then fit a numerical probability distribution for each node. Convert into standard format by transforming values.

CLASSIFICATION

Machine learning are categorized into supervised and unsupervised learning.

In Supervised learning, the input and its relevant output is already known because the data is already trained and creates the model which helps in prediction. Association, clustering and classification. Decision Tree, random forest etc. are the examples of supervised learning.

In Unsupervised learning, we have only input data and no corresponding output variable. The main job of unsupervised learning is to create class labels and the relationship between the data can be found using unsupervised learning algorithms to get whether the data can characterize to form a group. This group is known as clusters. K Means Clustering, KNN etc. are the examples of unsupervised learning

Classification is mainly used to grouping up or categorizing statistics or something on the values in their similarity and it provides meaningful group or category that we can use for another purpose. Classification is used in every field because classifying a data according to its value and similarity will help for resolving the risk assessment process.

A **decision tree** is widely used classification technique. The methodology used here is Divide and conquer. As there were huge amount of data, first we need to divide those data into sub data. The structure of the decision tree is organized in a manner that it contains the root the topmost node in the tree, Branches which are the internal nodes and leaf node is one which is not further classified. The internal nodes

represent a question and the branch which connects the node denotes the solution and the leaf node tries to predict the solution. It is widely used in decision making process. Say for example to predict the patients who have swollen glands and diagnose the type of infection.

SVC It is extensively used in pattern recognition, classification or regression challenges. It is particularly used in noisy and complex domains. The parameters are identified by solving a quadratic equation which involves equality and inequality constraints. The data item is plotted in n-dimensional space with the value of each feature being the value of a coordinate. The hyper plane must be created to differentiate the classes.

GaussianNB approach used in Naïve Bayes classifier is very simple. With the help of small amount of training data, it is possible to classify the given instances. For example, to predict the fruit as "apple", based on red, and its shape round it is classified as apple which shows it as an independent model. This method is also suitable for complex situations.

RandomForestClassifier

Random forests introduced by Breiman [12] are an integration of tree predictors so that every tree depend on the values of a random vector separately and through similar distribution for the whole trees in the forest. The tree classifier of a forest has a generalization error which relies on the strength correlation of each single tree in the forest.

KNeighborsClassifier

K-nearest neighborhood classification algorithm KNN is supervised classification algorithm depends on the distances between the test dataset and the training dataset and finds out which one is closest and take the majority class from K-list according to K samples chooses randomly

LOGISTIC REGRESSION

In statistical modeling, regression analysis is a statistical process for estimating the relationships among variables. It includes techniques for modeling and analyzing several variables, when the focus is on the relationship between a dependent variable and one or more independent variables

PREDICTION

For each algorithm applied so far, the performance must be evaluated. Only based on these evaluation criteria we will able to find out which algorithm suits for certain kinds of application. The tool used here is confusion matrix and receiver operating curve

confusion matrix is one of the means used to measure the performance of a classification. Confusion matrix is a table used to record the results of classification performance.

Confusion matrix		Prediction class		
		True	False	Total
Original Class	True	TP	FN	Total class
	False	FP	TN	Total class
	Total	Total prediction class	Total prediction class	Grand total

Based on the above table it will be known the amount of data from each class correctly classified (TP + TN) and incorrectly classified (FN + FP). Quantity of confusion matrix can be defined as accuracy

$$Accuracy = \frac{TP + TN}{TP + TN + FN + FP}$$

Precision (positive prediction value) is a metric used to measure system performance to obtain relevant data. While the recall (sensitivity) is a metric used to measure the performance of the system to obtain relevant data read in the search for information. Here is the formula for calculating precision and recall.

$$Precision = \frac{TP}{TP + FP}$$

$$Recall = \frac{TP}{TP + FN}$$

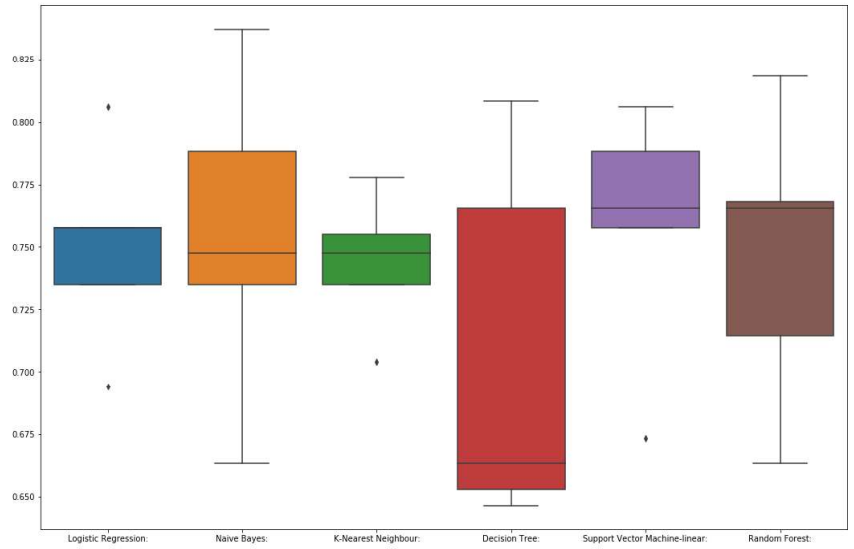
ANALYZING THE RESULT

From the given dataset, there were 768 instances out of which while doing the preprocessing of the data set the analysis of the result was done on 657 instances. After preprocessing the data was converted to the standard scalar format, where all classified data was transformed into the normalized form.

Different classification algorithms are applied to test the accuracy of the data and to find out which algorithms give the accurate results, two approaches were used

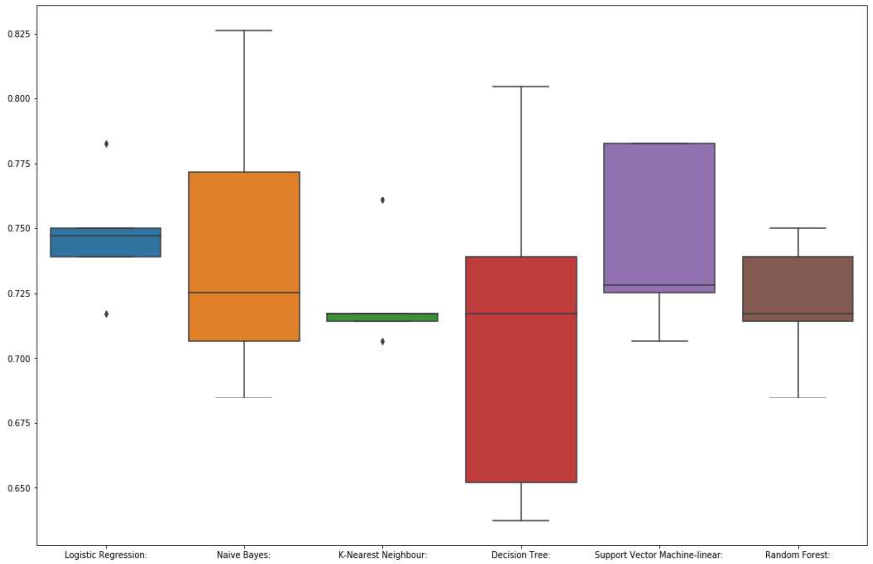
Without feature selection:

In this approach, the accuracy of the algorithms was checked after the Data preprocessing:



Algorithms	**Accuracy**
Logistic Regression	74.9969078540507
Naive Bayes	**75.40094825809112**
K-Nearest Neighbor	74.38260152545867
Decision Tree	70.93176664605237
Support Vector Machine-linear	**75.80705009276438**
Random Forest	70.93382807668522

With Feature selection: In this approach the accuracy of the classification algorithms was checked after doing the Feature selection. There are four types of feature selection. Univariate feature selection works by selecting the best features based on univariate statistical tests. It is a preprocessing step to an estimator. recursive feature elimination is to select features by recursively considering smaller and smaller sets of features. Select from model, features are considered unimportant and removed, if the corresponding coefficient or feature importance's values are below the provided threshold parameter. Out of the four methods we applied the Univariate feature selection.



Algorithm	Accuracy
Logistic Regression	**74.7276636407071**
Naive Bayes	74.28810320114668
K-Nearest Neighbor	72.32919254658385
Decision Tree	70.7955088389871
Support Vector Machine-linear	74.50549450549451
Random Forest	72.55613951266126

From the above tables, it shows that the Logistic Regression has the same accuracy before and after applying the feature selection.

	precision	recall	f1-score	support
0.0	0.78	0.88	0.83	132
1.0	0.67	0.50	0.57	66
avg / total	0.74	0.75	0.74	198

CONCLUSION

In this paper, we saw the basic idea of the data mining process, classification techniques and feature selection. Different algorithms Logistic Regression, Naive Bayes-Nearest Neighbour, Decision Tree, Support Vector Machine-linear, Random Forest were discussed. This algorithm performance was evaluated based on their accuracy levels. Also, the algorithms accuracy was calculated based on with and without feature selection. So, In the medical field accuracy in prediction of the diseases is the most important factor. To correctly know whether the patient is diabetic or not, also a system which will be developed to do the prediction for the diabetes patients. To do this, logistic classification algorithm gave the same accuracy result of 74.73% with and without the feature selection. The goal of Classification algorithms is to produce precise and accurate results.

REFERENCES

1. A. A. Al Jarullah, "Decision tree discovery for the diagnosis of type II diabetes," *2011 International Conference on Innovations in Information Technology*, Abu Dhabi, 2011, pp. 303-307.doi: 10.1109/INNOVATIONS.2011.5893838
2. Marjan Khajehei, Faried Etemady, "Data Mining and Medical Research Studies", *2010 Second International Conference on Computational Intelligence Modelling and Simulation*, pp. 119-122, 2010.
3. T. Jayalakshmi, A. Santhakumaran, "A Novel Classification Method for Diagnosis of Diabetes Mellitus Using Artificial Neural Networks", *Data Storage and Data Engineering (DSDE) 2010 International Conference*, pp. 159-163, 9–10 Feb. 2010.
4. http://archive.ics.uci.edu/ml/datasets/ Pima+Indians+Diabets, Irvine, CA: University of California, School of Information and Computer Science.
5. Yang Guo, Guohua Bai and Yan Hu, "Using Bayes Network for Prediction of Type-2 diabetes," *2012 International Conference for Internet Technology and Secured Transactions*, London, 2012, pp. 471-472.
6. D. Shetty, K. Rit, S. Shaikh and N. Patil, "Diabetes disease prediction using data mining," *2017 International Conference on Innovations in Information, Embedded and Communication Systems (ICIIECS)*, Coimbatore,2017, pp.1-5.
7. I. S. Jasim, A. Deniz Duru, K. Shaker, B. M. Abed and H. M. Saleh, "Evaluation and measuring classifiers of diabetes diseases," *2017 International Conference on Engineering and Technology (ICET)*, Antalya, Turkey,2017, pp.1-4.
8. A. Naik, L. Samant, Int. Con. on Computa. Mod. and Sec., ELSEVIER, vol. 85, pp. 662-668, 2016.
9. Larose, D. T. (2006) Data Mining Methods and Models, Hoboken: John Wiley & Sons, Inc.
10. Aparna U. R. and S. Paul, "Feature selection and extraction in data mining," *2016 Online International Conference on Green Engineering and Technologies (IC-GET)*, Coimbatore, 2016, pp. 1-3.
11. S. Umadevi and K. S. J. Marseline, "A survey on data mining classification algorithms," *2017 International Conference on Signal Processing and Communication (ICSPC)*, Coimbatore, 2017, pp. 264-268.
12. L. Breiman, Randomforests. Machine learning, vol. 45, no. 1, pp. 5-32, 2001.
13. R. Savundhariyalachmi, N. Pandimeena, P. Ramya," Study of Classification algorithm in Data mining", International Journal of Science and Research
14. http://scikit-learn.org/stable/modules/feature_selection.html#feature-selection
15. R. Kavitha and E. Kannan, "An efficient framework for heart disease classification using feature extraction and feature selection technique in data mining," *2016 International Conference on Emerging Trends in Engineering, Technology and Science (ICETETS)*, Pudukkottai, 2016, pp. 1-5.