DOI: 10.xx/ijdsml.2023.xx

MACHINE LEARNING TECHNIQUES FOR DIABETES CLASSIFICATION

P. Subham[1], Aryan Sinha[2], Adarsh Kumar[3], Shantilata Palei[4] and Puspanjali Mohapatra[5]
[1]Department of Computer Science and Engineering, National Institute of Technology, Rourkela, India
[2,3]Department of Computer Engineering, International Institute of Information Technology, Bhubaneswar, India
[4,5]Department of Computer Science and Engineering, International Institute of Information Technology, Bhubaneswar, India

Abstract

Driven by the explosion in the generation of Biomedical Data and their complexities, Machine learning approaches have been found to be extremely compelling for detection, diagnosis and necessary medical decision making of diseases. The objective of this paper is to investigate the efficiency of various Machine Learning based algorithms in the analysis of very common and fatal disease like Diabetes. These algorithms are not only classifying the diabetic patient into different categories but also, they are advising the diabetic patient suffering from other associated diseases (originated due to diabetes) for immediate medical attention or not. The two datasets used in the study are Pima Indians Diabetes Dataset and 130 US Hospitals Data for the Year 1999-2008. The various Machine Learning algorithms used in the study include Logistic Regression, K- Nearest Neighbors, XGBoost, Decision Tree, Random Forest, Support Vector Machines and Neural Network based MLP Classifiers. The efficacy of the models is tested on the basis of Classification Accuracy and F1 Score. The results are analyzed and compared. It demonstrates that Logistic Regression model outperforms other models in the study of Pima Indian Diabetes Data whereas Neural Network based MLP Classifier outperforms other models in the study of the Diabetes 130-Us Hospitals Data for Years 1999-2008.

Keywords:

Feature Selection, Feature Extraction, Exploratory Data Analysis (EDA), K-Fold Cross Validation, Classification

1. INTRODUCTION

Healthcare is a very important aspect of every human being. With the population explosion count of human beings suffering from diseases is ever increasing. It is creating a lot of complexities on the part of medical fraternity with a huge demand of medical resources for the detection and diagnosis of diseases. Many new technologies have been adapted by the medical practitioners but due to the level of complexities involved in the medical treatment procedures, time consumed to generate the reports and treatment expenses. There is a very high need to have an alternative and fast solution to deal with these diseases. Machine learning (ML) techniques can help doctors make almost perfect and fast diagnoses. It also helps in choosing the best medications and treatment process for their patients [1]. It also helps to predict re-admissions, identify patients at high risk. In general, it helps to have treatment with minimum costs. ML has shown a great result in prediction of many diseases like Cancer, Cardiac Health, Liver Disorder, Obesity, etc. Here, a Machine Learning approach to tackle Diabetes is presented.

In this paper the sections are organized as follows, Section 2 contains the Background, Section 3 discusses the datasets used in the study. In Section 4, the experimental results with analysis of Pima Indians dataset have been discussed. Similarly, Section 5 includes the experimental results with analysis of 130 US Hospitals dataset and the Section 4 contains the conclusion and future prospects.

2. BACKGROUND

2.1 A BRIEF INTRODUCTION TO DISEASES

A disease is a state of body which disables the proper functioning of the human body due to internal issues. Every disease is classified by their symptoms. The cause of a disease may be due to some external factors like pathogens, bacteria, infections, etc. Disease is often used to refer to a state that causes pain, failure in functioning, suffering or even survival risks to the person suffering from it, or similar issues for person who gets in contact with the person suffering from a communicable disease.

2.2 DIABETES – A SURVEY

The fatal diseases in humans include Cancer, Diabetes, Liver Disorder, Obesity, Heart Diseases, etc. The growth in population of diabetic patient is increasing steadily, as a survey by the International Diabetes Federation (IDC) [3]-[4]. The number of people suffering from diabetes was around 382 million in the year 2013 and count of patients is estimated to be around 595 million till the end of the year 2035. The cases of Type 2 diabetes in 2010 was around 285 million people out of the total population and is expected to be around 438 million by 2030, with a daily increase of 21,000 new cases. 90% of cases of diabetes are of Type 2 and the remaining 10% is basically due to Type 1 diabetes mellitus. Cause of Type 2 diabetes is mainly due to Obesity in humans who are suffering from it due to hereditary problems. Regular Exercise and proper diet can be helpful in treating Type 2 diabetes. Metformin or insulin are the drugs advised to be taken if the initial treatment fails. Patients are required to check blood sugar level regularly if they are taking insulin [5].

2.3 MACHINE LEARNING METHODOLOGIES

Machine Learning is a relatively new and fast-growing concept mainly focused on techniques and algorithms that allows the computers/machines to learn and gain knowledge based on the past recorded experiences [6]-[8]. Machine Learning is highly related to statistics. In Machine learning computers learn from data so that it can make predictions and take decisions. The first step of machine learning is acquiring the dataset from various data repositories. The next step is data preparation in which data related problems are solved by pre-processing and also to minimize the dimension of the dataset by dropping the useless data attributes from the dataset. Generally, the data is large so algorithms are developed using statistical and other computational methods for the machine to learn and take decisions experiences. Proceeding the steps testing of the model

to calculate the accuracy and performance of the machine was done. And the final step is improving the parameters of the models for the best efficiency, also known as optimization of the system. Classification, prediction and pattern recognition are the common use cases of Machine Learning [7]-[8].

ML models in clinical settings encounter challenges like poor data quality, interpretability issues, and limited generalizability. Clinical data is often incomplete or biased, affecting model accuracy. The "black box" nature of many ML models complicates decision-making, which is crucial in healthcare. Variability across patient populations further limits their broad application. Moreover, integrating ML into healthcare systems demands significant infrastructure, training, and regulatory compliance. Thus, while ML holds potential, its implementation requires careful, context-sensitive consideration.

Selecting an appropriate model involves balancing accuracy, complexity, and interpretability to address the specific needs of a clinical application. Justification for model choice should consider the type of data available, the problem's nature, and performance metrics. During model training and tuning, challenges include managing large and heterogeneous datasets, mitigating overfitting, and optimizing hyperparameters. Ensuring that the model generalizes well across diverse patient populations is critical, as is adapting to the evolving nature of clinical data. Additionally, computational constraints and the need for reproducibility can further complicate the training process, necessitating careful planning and validation.

2.4 METHODS OF DIMENTIONALITY REDUCTION

Machine Learning techniques are applied considering too many factors. Features or variables of the dataset are these factors. Analysis of training set becomes tougher as the number of attribute increases [9]. These features may be correlated in some cases; this means redundant data is available [10]. This requires some methods to be followed and the method which come into play here Dimensionality Reduction [11]. It has two types known till date:

2.4.1 Feature Selection:

The process of choosing the most useful features for the training of machine is known as Feature Selection. Inessential features decrease the efficiency of model and slows the learning speed due to problems like overfitting [11]. The methods of feature selection followed on both the datasets are as follows:

- In the set of features, the features with a high percentage of missing values or null values are operated on or removed whichever seems to be suitable.
- The features are tested for collinearity and extra features are dropped based on correlation.
- Attributes having 0 or negligible importance as found by a Tree based model are dropped.
- Attributes having less importance are generally removed from the dataset.

2.4.2 Feature Extraction:

As discussed earlier working with huge no. of dataset requires high computational power and memory. It also leads to overfitting and therefore reduces the efficiency of the model. In feature extraction the features are combined in such a way that each combination gives a useful insight of the data whilst increasing the training speed and efficiency of the model. It is believed that properly optimized feature extraction is the key to effective model building [11].

The different methods of feature extraction are as follows:

- PCA (Principal Component Analysis)
- LDA (Linear Discriminant Analysis)
- Kernel PCA

2.5 VARIOUS CLASSIFIERS USED IN THE STUDY

The following Classifiers were used to fit the data and predict the output:

2.5.1 Logistic Regression:

Logistic Regression is used for classification problem and not to be confused by its name as it is not used for regression problems [9]. It is used to predict binary values such as 0 or 1, yes or no, rain or no rain, etc. In this algorithm the probability of an event to be correct is found by fitting the data into a logit function. Therefore, it is also called as logit regression [12]. The output lies between [0,1] as it gives probability. Logit function is derived as follows:

Odds of an event (odds)= prob / (1-prob) = probability of event to be true / probability of event to be false (1)

$$\ln(odds) = \ln(prob/(1-prob)) \quad (2)$$

$$logit(prob) = \ln(prob/(1-prob)) \quad (3)$$

2.5.2 Decision Tree Classifier:

Decision trees are used for classification problems and is a type of supervised learning technique. Both categorical and continuous variables can be worked using this Classifier. Data is divided into two or more homogeneous sets in this algorithm. The classifier uses the most essential attributes and divides it into maximum different groups possible [13]. The techniques like Gini, Information Gain, Chi-square, entropy is used to split the data into different heterogeneous groups [14]-[15].

2.5.3 Random Forest Classifier

Random Forest is a hybrid version of decision trees. It is formed by the collection of Decision Trees, hence defining Forest [10]. Voting is done by the trees for classification of the target based on the features. The forest chooses the classification with most no. of votes from all the trees presents in the forest [13]-[16].

2.5.4 Support Vector Machines (SVM)

It is a classification algorithm in which the data units are plotted as a point in an n-dimensional space where, n = number of variables where each coordinate in the space corresponds to the value of the variable. The algorithm maximizes the distance between the hyper-plane and the data points. The loss function that is used maximize the margin is hinge loss [17]. The Kernel is an important factor in SVM algorithm which converts the form of inputs by using some mathematical functions [17]-[19]. The various Kernels of SVM are:

$$\text{Linear: } K(x, y) = x^T y K(x, y) = x^T y \quad (4)$$

$$\text{Polynomial: } K(x, y) = (x^T y+1)\, dK(x, y) = (x^T y+1)d \quad (5)$$

$$\text{Sigmoid: } K(x, y) = \tanh(ax^T y+b)\, K(x, y) = \tanh(ax^T y+b) \quad (6)$$

$$\text{RBF: } K(x, y) = \exp(-\gamma\|x-y\|2) \quad (7)$$

where, x^T = Transpose of x.

2.5.5 K-Nearest Neighbour:

Both Classification and Regression problems are solved using this algorithm but it is often used in classification problems in the real-life applications [11]. Classification is done for a data point by k- nearest neighbours by voting [20]. Distance functions are used to calculate the most probable neighbour to which the data point will be classified. Euclidean, Manhattan, Minkowski and Hamming distance are the distance functions majorly used. If the value of k = 1, then the data point is classified to the class of the nearest neighbour. Choosing the value of k is an important factor while performing k-NN modelling [21].

2.5.6 Neural Network based MLP Classifier

A Perceptron is one of the most efficient classifiers as it classifies input by separating two targets with a straight line [22] [23]. The working formula for Perceptron is: $y = w * x + b$. where, y is the output vector, w is the weight vector, x is the in-input vector and b are a bias. A Multilayer Perceptron (MLP) is an artificial neural network [24] [25]. MLP is composed of two or more perceptron. MLP is composed of the following layers.

There is the first layer where the input is given. There is a layer at the end which predicts the output based on the input by first layer. And in between those two layers, a random number of hidden layers are there that plays the most crucial role in predicting the output efficiently [26].

2.6 MEASURE USED FOR EVALUATION CLASSIFICATION MODELS

2.6.1 Classification Accuracy:

Classification Accuracy is the one of the most important measures for model evaluation. It is the ratio of number of accurate predictions to the total number of inputs [2] [5].

$$\text{Accuracy} = (TP+TN) / (TP+TN+FP+FN) \quad (8)$$

2.6.2 Confusion Matrix

Confusion Matrix gives us a matrix as output and describes the total performance of the model [5]. There are 4 major parameters:

Table.1. Confusion Matrix

		Actual Value	
Predicted Value		Yes	No
	Yes	True Positive	False Positive
	No	False Negative	True Negative

Confusion Matrix gives important results that are necessary for other types of metrics [4].

2.6.3 F1 Score:

Classifiers test accuracy is given by F1 score. It is the Harmonic Mean between precision(P) and recall(R) [5]. F1 score lies in range of 0 to 1 inclusive both. It provides information about the precision of the classifier (Number of data points correctly classified). High precision but lower recall, gives high accuracy. More the F1 Score, better is the performance of the model. It can be expressed as:

$$F1=2*((P*R)/(P+R)) \quad (9)$$

$$\text{Precision: } P=TP/(TP+FP) \quad (10)$$

$$\text{Recall: } R=TP/(TP+FN) \quad (11)$$

3. DATASETS USED

3.1 DATASET A: PIMA INDIAN DIABETES DATASET

Dataset A was obtained from UCI Repository. It contains 768 cases of different patients and all patients are females of age more than 21 years. The target variable has values 0 for patients with no diabetes and 1 for patients suffering from diabetes. The target variable has 500 cases of diabetic patients and 268 cases of patients with no diabetes. Pima Indian population who are residing near Phoenix, Arizona were taken for this research. Those females have been under continuous observation since the year 1965 by the National Institute of Diabetes and Digestive and Kidney Diseases because people of that place were facing this disease a lot. Each people of that community was kept under observation and regular tests including oral glucose tolerance test over a span of 5 years [5].

3.2 DATASET B: DIABETES 130-US HOSPITALS FOR YEARS 1999-2008 DATASET

Dataset B was also obtained from UCI Repository. Analysis of the database was done for data points where these factors were satisfied:

- Inpatient encounter which means a hospital admission of a patient was done.
- Diabetic encounter which means in cases where any stage or type of diabetes was registered to the system as a diagnosis.
- The patient stayed at hospital admitted for at least 1 day and a maximum 14 days.
- Patients have undergone Laboratory tests when they were admitted.
- Administration of the patient's medication was done.

All the patients having any kind of Diabetic encounter along with any other disease were also considered. 101,766 encounters were identified to fulfil all of the mentioned five criteria and were used in further modelling. Some medical experts performed attributes selection from all attributes of the attributes that were related to diabetic conditions. After going through the database, the total of 51 features were extracted describing the diabetic encounters, including diagnoses, diabetic medications, number of visits in the year preceding encounter, and payer information [2].

3.3 NORMALIZATION

Normalization is a method commonly applied as part of data preparation for ML. In normalization some statistical methods are applied to re-scale the numeric data columns in a common range keeping in mind that the difference in the range of values doesn't change. All datasets do not require normalization. In this technique the value is re-scaled and brought to a range of 0 to 1. Any instance can take a highest value of 1 and a lowest value of 0. Normalization can be fruitful method to use when we do not know the distribution of the data or when we know the distribution

is not Gaussian. When we know that our data is Gaussian then Standardization (set mean value=0 and standard deviation=1 of the attribute) is done [6]. Both the dataset was subjected to normalisation.

3.4 DATA VALIDATION

The Pima Indians dataset was splitted into two sections that is for training and testing purpose. The method used for this purpose is K Fold Cross validation. From the 130-US hospitals dataset, 80% of the samples were used for training and 20% for testing.

4. DIABETES DATASET A CLASSIFICATION USING VARIOUS MACHINE LEARNING ALGORITMS

4.1 EXPLORATORY DATA ANALYSIS

Exploratory Data Analysis can be defined as the most initial steps and investigations performed on the data so as get insights of data, spot anomalies and to check assumptions with the help of statistical methods and graphs. After observing the Pima Indians Diabetes dataset using various techniques, we found some values of some features such as Blood Pressure, Glucose, Skin Thickness, BMI and Insulin to be 0(zero) which is practically not possible in real life scenario. So, it was necessary to remove those impractical values. We replaced those values with mean or median as suitable.

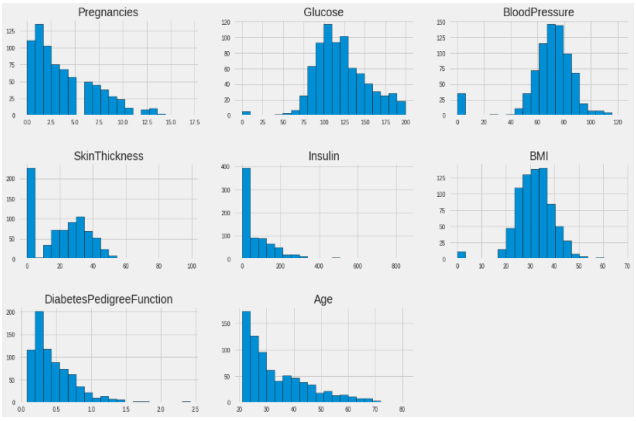

Fig.1. Distribution of Dataset

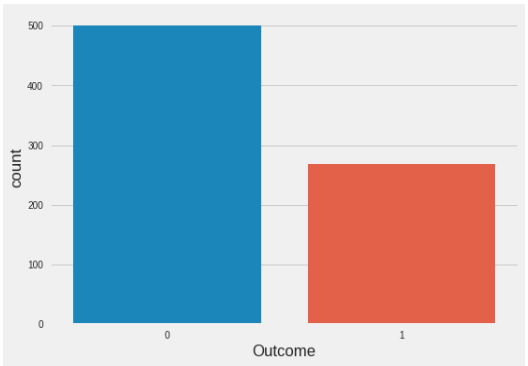

Fig.2. Count of Diabetic and Non-Diabetic patients

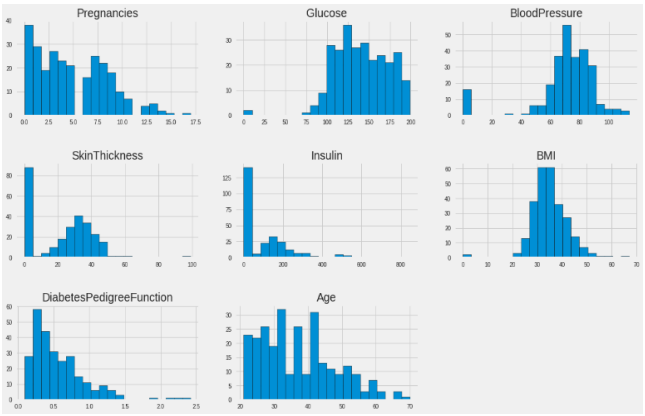

Fig.3. Distribution for Diabetic Cases

4.2 FEATURE SELECTION

4.2.1 Correlation Matrix using Heat Map:

Correlation value states how much the features are similar to each other or the target variable. Features with high correlation are more linearly dependent and hence have almost the same effect on the dependent variable. So, when two features have high correlation, we can drop one of the two features. Heat map makes it easy to identify which features are highly correlated to the other variables, we plotted the heat map using the seaborn library of Python. From the Fig.4 we found the correlation between two features. If two features have high correlation, then we can eliminate one of the features. This help in speeding up the Machine Learning algorithm and also increases the accuracy of the prediction [7]. Here, we concluded that no two features are highly correlated and we cannot eliminate any of the features.

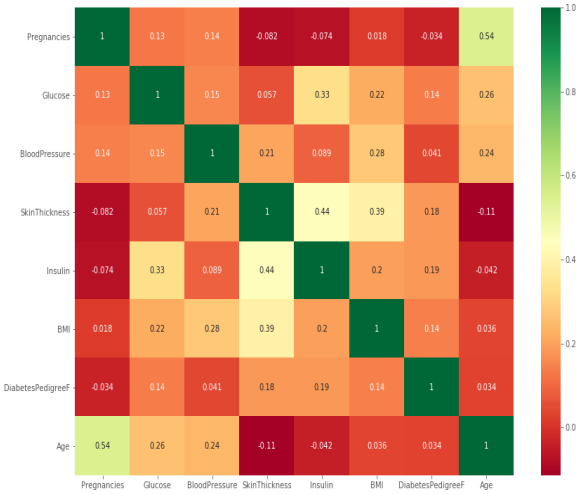

Fig.4. Correlation Matrix using Heat Map

4.2.2 Feature Importance using Tree Based Model:

Here, we used Random Forest Classifier and its class feature_importances_ to get the importance of each feature. The result of this is presented in Fig.5. We found that no feature has zero or negligible importance. Hence, no feature was eliminated from the dataset.

```
Glucose                  0.242098
BMI                      0.172574
Age                      0.135220
DiabetesPedigreeF        0.128324
BloodPressure            0.092903
Pregnancies              0.086774
SkinThickness            0.073109
Insulin                  0.068999
dtype: float64
```

Fig.5. Importance of each feature

4.3 EVALUATION AND RESULTS

The following accuracy was obtained using the different algorithms as shown in Table 1 below.

Table.1. Prediction Accuracy using different Algorithms

Machine Learning Algorithm	Prediction Accuracy Percentage
Logistic Regression	76.87
K-Nearest Neighbor	74.58
Decision Tree	68.73
Random Forest	72.79
XGBoost	74.57
Support Vector Machine (SVM)	76.54

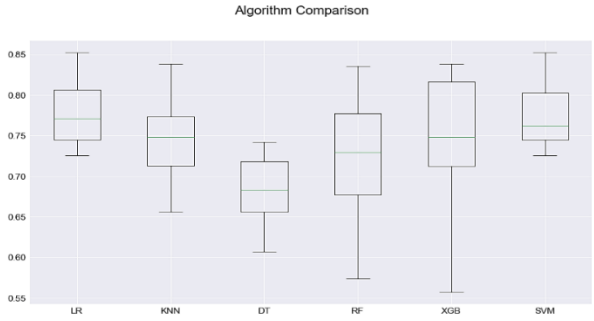

Fig.6. Algorithm Accuracy Comparison using Boxplot

4.4 RESULT ANALYSIS

It is found that Logistic Regression Classifier yields the maximum Accuracy percentage of 76.87% for this Pima Indians Dataset.

5. DIABETES "DATASET B"CLASSIFICATION USING VARIOUS MACHINE LEARNING ALGORITHMS

5.1 EXPLORATORY DATA ANALYSIS

After doing some initial data analysis it was found that some features have high percentage of null values.

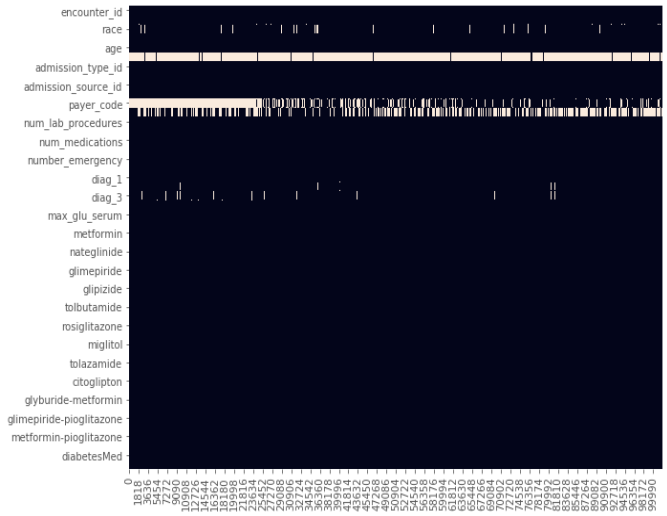

Fig.7. Features null values plot

The white places in the plot in Fig.7 represents the null values present in the corresponding feature. From Fig.8, the percentage of null values in the feature columns weight, prayer_code, medical_specialty is found to be very high. So, these three columns were dropped from the dataset as it will not be helpful in getting the result. On further data analysis, the columns diag_1, diag_2 and diag_3 is splitted into further more columns based on the idc9 code values corresponding to different diseases which was obtained from Table.2. The values in Table.2 were given in the description.pdf file of the UCI Repository of the dataset https://archive.ics.uci.edu/ml/datasets/diabetes+130us+hospitals+for+years+1999-2008 [2].

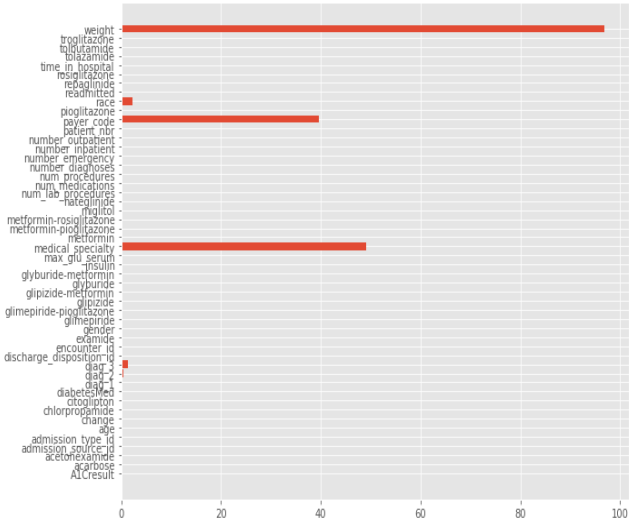

Fig.8. Percentage of Null Values

Table.2. Grouping of Diseases based on different Idc9 Codes in the Diagnosis Columns (diag_1, diag_2 and diag_3) [2].

Group Name	IDC9 CODES
Circulatory Diseases	390–459, 785
Respiratory Diseases	460–519, 786
Digestive Diseases	520–579, 787

Diabetes	250.xx
Injury	800–999
Musculoskeletal Issues	710–739
Genitourinary	580–629, 788
Neoplasms	140–239
Other (17.3%)	780, 781, 784, 790–799 240–279, without 250 680–709, 782 001–139 290-319 E–V 280–289 320–359 630–679 360–389 740–759

After doing some more feature analysis, operations were performed on the dataset to make the dataset more homogeneous so that the Machine can learn the data and give most efficient results. The original dataset had 51 features including the target variable 'readmitted'. But after performing the EDA on the dataset as discussed above the features increased to 72. Now, from this prepared dataset of 72 features, the further methods were applied in order to accomplish our goal of making an efficient Machine Learning model.

5.2 FEATURE SELECTION

5.2.1 Correlation Matrix using Heat Map:

Correlation Matrix is a very clear indicator of the correlation between two different features as we discussed earlier too. It helps us by providing clear visualizations of the correlation. The Correlation Matrix (Fig.9) obtained was not very clear in terms of the exact value of the correlation as the number of features were high. But from the colour indicators of the Heat Map, the visualizations were done effectively and the following was concluded for feature selection using this method:

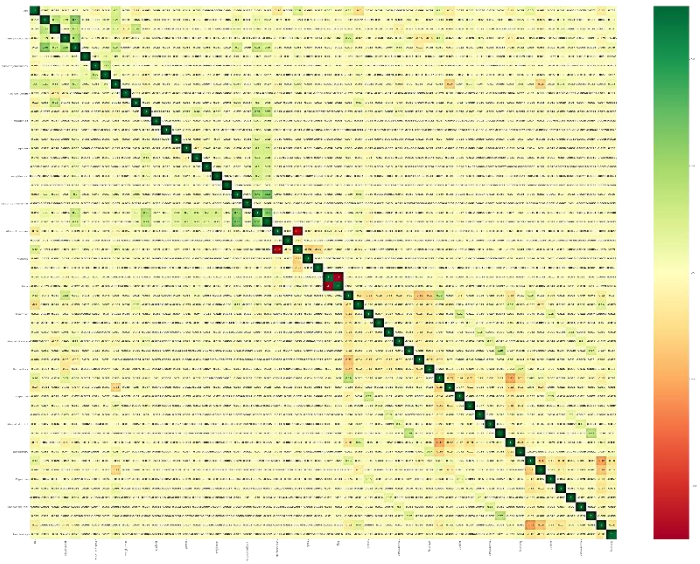

Fig.9. Correlation Matrix using Heat Map for Dataset B

From the correlation matrix, it was found that no two features were highly correlated to each other. So, on the basis of this deduction from the plot none of the features could be eliminated from the prepared dataset.

5.2.2 Feature Importance using Tree Based Model:

Here also, we used Random Forest Classifier and its class feature_importances_ to get the importance of each feature.

```
number_inpatient          2.303089e-02
num_lab_procedures        1.073292e-02
num_medications           1.034794e-02
number_diagnoses          7.070677e-03
time_in_hospital          6.929725e-03
number_emergency          6.324179e-03
age                       6.314246e-03
number_outpatient         4.940263e-03
num_procedures            4.826775e-03
insulin                   2.807402e-03
A1Cresult                 2.107800e-03
diabetesMed               1.701640e-03
```

Fig.10. Top features by its importance

```
chlorpropamide            3.800226e-05
tolazamide                1.879809e-05
miglitol                  1.736049e-05
glipizide-metformin       7.681090e-06
tolbutamide               7.209686e-06
troglitazone              1.493415e-06
acetohexamide             5.055298e-07
metformin-pioglitazone    3.760627e-07
metformin-rosiglitazone   0.000000e+00
glimepiride-pioglitazone  0.000000e+00
citoglipton               0.000000e+00
examide                   0.000000e+00
Length: 71, dtype: float64
```

Fig.11. Least/Zero importance features

So, the least/zero importance features as shown in Fig.11 were dropped from the dataset as these were of no significance in building our ML model. The following accuracy and f1_scoree was obtained using the different algorithms as shown in Table.3 below.

Table.3. Accuracy and F1_Score of different Classifier

Index	Classifier Name	F1_score	Accuracy
0	Logistic Regression	0.598719	0.614726
1	Random Forest	0.581261	0.590047
2	Neural Network based MLP Classifier	0.612418	0.618805
3	Decision Tree	0.544388	0.544157
4	K-Nearest Neighbour	0.545450	0.547012
5	Linear SVC	0.337394	0.478075

It was found that Neural Network based yields the maximum Accuracy percentage of 61.88% for 130-US Hospitals dataset. KNN underperforms in diabetes classification due to its sensitivity to irrelevant features and the curse of dimensionality, which can degrade accuracy. XGBoost might struggle if the model is not properly tuned, leading to overfitting or poor generalization. Both models require careful feature selection and hyperparameter optimization to achieve optimal performance in complex datasets.

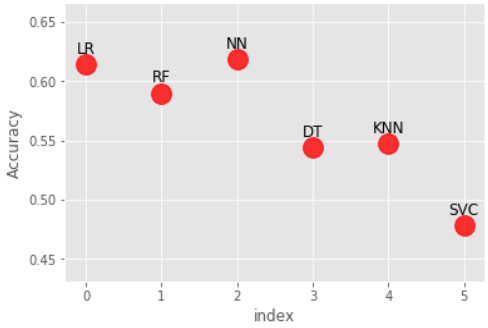

Fig.12. Accuracy Comparison of different Algorithms

6. CONCLUSION AND FUTURE PROSPECTS

In this paper, application of various Machine Learning algorithms is studied on the Pima Indians dataset and the 130-US Hospitals dataset. The performance of all the models were recorded and compared. Various empirical studies with Exploratory Data Analysis have been conducted to evaluate the efficacy of all the models. Classification Accuracy and F1 Score measures are adapted for comparing the performance of all the models from the results it can be concluded that ha Logistic Regression based model out performed all other ML models in classifying Dataset A whereas NN based MLP Classifier was found to be superior for dataset B. This work indicates that Logistic Regression and MLP Classifier can be better used for Diabetes study. However, in future various Interpretable ML models can be designed for healthcare study and better ML algorithms can be proposed to design Health Recommender Systems.

REFERENCES

[1] F. Asnicar, A.M. Thomas, A. Passerini, L. Waldron and N. Segata, "Machine Learning for Microbiologists", *Nature Reviews Microbiology*, Vol. 22, No. 4, pp. 191-205, 2024.

[2] B. Strack, J.P. De Shazo, C. Gennings and J.N. Clore, "Impact of HbA1c Measurement on Hospital Readmission Rates: Analysis of 70,000 Clinical Database Patient Records", *BioMed Research International*, Vol. 2014, No. 1, pp. 781670-781685, 2014.

[3] S.A. Lule, S.B. Kushitor, C.S. Grijalva-Eternod and M.K. Kushitor, "The Contextual Awareness, Response and Evaluation (CARE) Diabetes Project: Study Design for a Quantitative Survey of Diabetes Prevalence and Non-Communicable Disease Risk in Ga Mashie, Accra, Ghana", *Global Health Action*, Vol. 17, No. 1, pp. 2297513-2297521, 2024.

[4] S.J. Pilla, R. Jalalzai, O. Tang, N.L. Schoenborn, C.M. Boyd, M.P. Bancks and N.M. Maruthur, "A National Survey of Physicians Views on the Importance and Implementation of Deintensifying Diabetes Medications", *Journal of General Internal Medicine*, Vol. 39, No. 6, pp. 992-1001, 2024.

[5] F. Mercaldo, V. Nardone and A. Santone, "Diabetes Mellitus Affected Patients Classification and Diagnosis through Machine Learning Techniques", *Procedia Computer Science*, Vol. 112, pp. 2519-2528, 2017.

[6] I. Kavakiotis, O. Tsave, A. Salifoglou, N. Maglaveras, I. Vlahavas and I. Chouvarda, "Machine Learning and Data Mining Methods in Diabetes Research", *Computational and Structural Biotechnology Journal*, Vol. 15, pp. 104-116, 2017.

[7] H. Kaur and V. Kumari, "Predictive Modelling and Analytics for Diabetes using A Machine Learning Approach", *Applied Computing and Informatics*, Vol. 18, No. 1-2, pp. 90-100, 2022.

[8] H. Maheshwari, P. Goswami and I. Rana, "A Comparative Study of Different Machine Learning Tools", *International Journal of Computer Sciences and Engineering*, Vol. 7, No. 4, pp. 184-190, 2019.

[9] V.R. Patel and R.G. Mehta, "Impact of Outlier Removal and Normalization Approach in Modified K-means Clustering Algorithm", *International Journal of Computer Science*, Vol. 8, No. 5, pp.331-340, 2011.

[10] J. Huang, N. Huang, L. Zhang and H. Xu, "A Method for Feature Selection Based on the Correlation Analysis", *Proceedings of International Conference on Measurement Information and Control*, Vol. 1, pp. 529-532, 2012.

[11] S. Khalid, T. Khalil and S. Nasreen, "A Survey of Feature Selection and Feature Extraction Techniques in Machine Learning", *Proceedings of IEEE International Conference on Science and Information*, pp. 372-378. 2014.

[12] A. Das, "*Logistic Regression*", Springer, 2024.

[13] N.B. Nanda and A. Parikh, "Experimental Analysis of K-Nearest Neighbor, Decision Tree, Naive Baye, Support Vector Machine, Logistic Regression and Random Forest Classifiers with Combined Classifier Approach for NIDS", *International Journal of Computational Science and Engineering*, Vol. 6, No. 9, pp. 940-943, 2018.

[14] T.M. Oshiro, P.S. Perez and J.A. Baranauskas, "How many Trees in a Random Forest?", *Proceedings of International Conference on Machine Learning and Data Mining in Pattern Recognition*, pp. 154-168, 2012.

[15] R. Singh and N. Srivastava, "Assessing the Impact of Student Employability using Decision Tree Classifier in Education 4.0: An Analysis of Key Factors and Predictive Model Development", *Proceedings of IEEE International Conference on Architecture and Technological Advancements of Education 4.0*, pp. 178-198, 2024.

[16] K. Pramilarani and P.V. Kumari, "Cost Based Random Forest Classifier for Intrusion Detection System in Internet of Things", *Applied Soft Computing*, Vol. 151, pp. 111125-111135, 2024.

[17] S. Tong and D. Koller, "Support Vector Machine Active Learning with Applications to Text Classification, *Journal of Machine Learning Research*, Vol. 2, pp. 45-66, 2001.

[18] S.S. Guddanti, A. Padhye, A. Prabhakar and S. Tayur, "Pneumonia Detection by Binary Classification: Classical, Quantum, and Hybrid Approaches for Support Vector Machine (SVM)", *Frontiers in Computer Science*, Vol. 5, pp. 1286657-1286667, 2024.

[19] S. Palei, R.K. Lenka, S.R. Mallick, S. A. Sinha, S.S. Biswal, S.S. Rath and S. Saxena, "Secure and Decentralized Apple Leaf Disease Identification using DL Integration Models", *Proceedings of International Conference on Emerging Systems and Intelligent Computing*, pp. 68-73, 2024.

[20] M. Kubara and K. Kopczewska, "Akaike Information Criterion in Choosing the Optimal K-Nearest Neighbours of the Spatial Weight Matrix", *Spatial Economic Analysis*, Vol. 19, No. 1, pp. 73-91, 2024.

[21] Q. Hu, D. Yu and Z. Xie, "Neighborhood Classifiers", *Expert Systems with Applications*, Vol. 34, No. 2, pp. 866-876, 2008.

[22] Y. LeCun, Y. Bengio and G. Hinton, "Deep Learning", *Nature*, Vol. 521, No. 7553, pp. 436-444, 2015.

[23] M. Riedmiller, "Advanced Supervised Learning in Multi-Layer Perceptrons-from Backpropagation to Adaptive Learning Algorithms", *Computer Standards and Interfaces*, Vol. 16, No. 3, pp. 265-278, 1994.

[24] L.K. Hansen and P. Salamon, "Neural Network Ensembles", *IEEE Transactions on Pattern Analysis and Machine Intelligence*, Vol. 12, No. 10, pp. 993-1001, 1990.

[25] T. Windeatt, "Accuracy/Diversity and Ensemble MLP Classifier Design", *IEEE Transactions on Neural Networks*, Vol. 17, No. 5, pp. 1194-1211, 2006.

[26] K. Teler, M. Skowron and Orlowska T. Kowalska, "Verification of the MLP Network-Based Current Sensor Fault Classifier for Vector-Controlled AC Motor Drives", *Bulletin of the Polish Academy of Sciences Technical Sciences*, Vol. 56, pp. 1-12, 2006.