ISBN: 978-93-6773-974-7

A COMPARATIVE MACHINE LEARNING APPROACH FOR DISEASE PREDICTION

ARUN SINGH YADAV
RESEARCH SCHOLAR, DEPARTMENT OF COMPUTER SCIENCE,
UNIVERSITY OF LUCKNOW, LUCKNOW
arun.ai.lkouniv@gmail.com

KEYWORDS

MACHINE LEARNING, DISEASE PREDICTION, CLASSIFICATION, PREPROCESSING, LOGISTIC REGRESSION, ANN, NAÏVE BAYES, SVM, DECISION TREE

ABSTRACT

Machine Learning (ML) models are becoming robust and more accurate now a days as the rapid increase of amount and quality of training data. Researchers are proposing complex models for real life problems to achieve higher accuracy which requires high computing and resources. This all depends on problem type, its complexity and the domain. In the context of the healthcare disease diagnosis, detection and prediction is still a challenge. Diseases like type II diabetes mellitus (T2DM) has risen dramatically all around the world. If it is not diagnosed early, may cause other serious complications like kidney failure, heart failure, blindness etc. The whole experiment is done on Pima Indian Diabetes Dataset (PIDDS) in two stages A and B. The main objective of this study is to review the accuracy of the different machine learning classification algorithms and analyze their efficiency in predictions. Another important objective is to justify that every time one need not to go for complex algorithms even the simplest one can solve your problem effectively. We have also applied preprocessing methods (imputation, feature selection, scaling, discretization etc.) to improve the classification accuracy. The algorithms selected for this problem are Logistic regression (NB), Naïve Bayes (NB), Artificial Neural Networks (ANN), Support Vector Machine (SVM) and Decision Tree (DT). LR provided the best accuracy and rest of the model are very close to each other.

1. INTRODUCTION

Artificial Intelligence is a major branch of computer science that tries to make computers more intelligent. Learning is the key requirement for any kind of intelligent behavior. Researchers are agreed that without learning there is no intelligence. Therefore, machine learning is becoming the most rapidly developing subfield of AI research. These intelligent algorithms were from the very beginning designed and used to analyze medical, clinical information[1].Machine learning algorithms analyze the historical data or cases and extract the useful and hidden patterns from the dataset for prediction and diagnosis[2][3].One of the major challenges with medical data, it is non-linear, heterogeneous and noisy[4]. So that information needs to be preprocessed to get the better result. Diabetes is a serious health problem in which the amount of sugar content cannot be regulated. Type I diabetes is caused when the human body refused to produce insulin. Type II diabetes makes the human body insulin resistance that causes other serious complications. So, the early and timely diagnosis of diabetes may prevent serious complications. The various machine learning based system has been developed in recent years to predict diabetes[5][6] and still scientist and medical expert evolving new and intelligent algorithms and proved that machine learning algorithms[7][8] performed better in disease diagnosing. The capability to work on large, heterogeneous data taken from different sources and keep improving the model performance by adding the background details to make it a more powerful tool[9]. The only objective of these developed system is to improve the accuracy that leads to the correct prediction of the disease. The main objective of this study is to do a comparative study of different supervised ML algorithms. We will investigate their logic, assumptions, feature selection, preprocessing impact, etc. And will show that the less complex algorithms can do better on less complex problems. The rest of the paper is organized as follows the Section II includes the related work and limitations of the previous system. The adopted methodology for this study included in Section III that contains data collection, data-preprocessing and the classification task for disease prediction. The last section is included the comparison and discussion on the produced results by the various classifier.

2. REVIEW OF LITERATURE

Machine learning is the problem of induction where general rules are learned from specific observed data from the domain. It is infeasible to know what representation or what algorithm is to best learn from the data on the given problem beforehand,

without knowing the problem so well you probably don't need machine learning to begin with. The machine learning model allows a section of preprocessing which removes irrelevant information from the data sources. The removal of unwanted data must be done very carefully by understanding the nature of data and the correlation of different features. The logistic regression method compares the relationship between a dependent and one or more independent variables. These variables usually continuous in nature. The logistic regression predicts the value of a dependent variable using probability. The outliers surely impact the prediction accuracy. The author used the distance-based outlier detection as a preprocessing method and proposed a modified prediction model for diabetes type ii prediction. the model achieved 79% of accuracy by using the sigmoid function[10] but after applying the Neuro based weight activation function to calculate bipolar sigmoid the accuracy reached at 90.4%[10]. The impact of preprocessing techniques like feature selection, missing values imputation and reducing class imbalance improves the classifier prediction of risk of 30-days hospital readmission for diabetes patient[11]. A very slight improvement can be seen in the Naïve Bayes model after applying preprocessing technique as compared to logistic regression and decision tree[11].

But this study also shows that the impact of these schemes varies by techniques and problem formulation. The problem of the highly skewed dataset can be overcome using subsampling, but the class imbalance problem cannot produce a good prediction model. N. Barakat and his team [12] proposed a hybrid diabetes prediction model using SVM classifier. The author has used K-means clustering algorithms for the preprocessing scheme to handle the class imbalance problem. Total five clusters are derived from the dataset and from every cluster positive samples are taken based on Euclidian distance. After that final dataset is divided into training and test dataset. Here the SVM provides a promising tool for diabetes prediction with 94% of accuracy. A. A. Al Jarullah[13] has used the J48 decision tree classifier on the modified dataset (pre-processed data). After applying the unsupervised k-means clustering for class imbalance problem and numerical discretization to make small groups of each attribute. The author achieved 78.17 % of accuracy. But the decision tree can do better and W. Chen et al.[14] has used k-means and 10-fold cross-validation technique for data pre-processing. The author significantly improved the performance of the decision tree model. With this dataset, the author achieved 90.04% accuracy on PIDD dataset. The outlier problem may produce the wrong result. The author R. Ramezaniet al.[15] and the team used multiple imputation methods for missing value treatment and OT for dimensionality reduction. This modified dataset applied to the hybrid model LANFIS (Logistic Adaptive Network-

based Fuzzy Inference System). This model has achieved 88.05% accuracy. Sometimes the uncorrelated variables reduce the performance of any learning model, so finding uncorrelated attributes means the principal components. M. K. M. Dhomse Kanchan B.[7] and the team used the PCA as a pre-processing scheme. The modified dataset applied on classifiers where SVM outperform after applying PCA. One of the closest work can be seen where the author has used the PCA and some other unsupervised ml methods for pre-processing. this pre-processed dataset then applied on ANN classifier which predicts diabetes with 92.28% accuracy[16]. Model selection for the problem is the biggest challenge where even the less complex models can do the better prediction but here the quality of data plays the big role. The author H. Wu et al.[17] and the team has done the excellent work on data using feature selection approach with correlation check and k-means clustering. They prepared the data so well that even the less complex models like logistic regression classified the diabetic positive and negative patient with 95.42 % accuracy. Naïve Bayes always worked better for imbalanced and missing data[18]. It fairly achieved the 76.3% accuracy after applying k-means and weka tool filtering approach.

3. EXPERIMENTAL SETUP

In this study we have used the famous PIMA INDIAN DIABETES DATASET(PIDD)[19]. Pima Indian is a group of native Americans living in the Southern Arizona. Due to some genetic issue they take the poor diet of carbohydrate. But in recent years they moved towards the processed food rather than traditional agriculture food with minimum physical activity. This sudden change in habit and food make them the highest prevalence to type-II diabetes which makes them a reason for research. This database was taken from the UCI machine learning library[20].This is a benchmark for comparing methods and widely adopted free dataset for research purpose[21] in machine learning community. The experimental setup for this study is divided into two stages. The first stage deals with data-preprocessing(A) methods as we have seen in the literature review and previous study[22] that the data preprocessing has improved the results in a drastic way. The preprocessed dataset of the first stage forms the input for the second stage classification(B) where the 5 ML methods make the predictions for diabetes. All the experiment has done in this study on jupyter notebook[23] using python programming language. Here I python compiler is used to run python programs. The methodology includes the data collection and analyzing the nature, the preprocessing methods and the predictions. The model proposed for this study are as follows:

4. DATA PREPROCESSING

The PIDD contains 768 records of pregnant females with 8 characteristics and one more column for the outcome. Each attribute is assigned the numeric value. In the dataset 65.10% (500 females) are non-diabetic (represented with value 0) and 34.90% (268 females) have diabetes (represented with value 1). The main attributes that are taken are listed below:

S.No	Parameter	Description	Data Type
1	PREGNANT	Number of time women get pregnant	Numeric
2	PGLUCOSE	Plasma glucose concentration measured using a 2-hour oral glucose tolerance test in mm Hg	Numeric
3	DBP	Diastolic blood pressure	Numeric
4	INSULIN	Two-hour serum insulin in muU/ml	Numeric
5	TSFT	Triceps skin fold thickness in mm	Numeric
6	BMI	Body mass index in mm^2	Numeric
7	DPF	Diabetes pedigree function	Numeric
8	AGE	Age of the patient	Numeric
9	OUTCOME	Patient with diabetes on set within five years(0 or 1)	Nominal

TABLE 1.1 THE PIMA INDIAN DIABETES DATASET WITH A DESCRIPTION

FIGURE 1.1 SAMPLE DATASET OF PIMA INDIAN DIABETIC DATASET BEFORE PREPROCESSING

	PREGNANT	PGLUCOSE	DBP	TSFT	INSULIN	BMI	DPF	AGE	OUTCOME
0	6	148	72	35	0	33.6	0.627	50	1
1	1	85	66	29	0	26.6	0.351	31	0
2	8	183	64	0	0	23.3	0.672	32	1
3	1	89	66	23	94	28.1	0.167	21	0
4	0	137	40	35	168	43.1	2.288	33	1

The initial investigation of the dataset suggests that it is a supervised classification problem. The PIDD contains several inconsistencies in it as the metadata shows no missing values but figure2 exhibit biologically implausible zero values. Which suggest that metadata is incorrect and must be treated as missing values. Some of the previous published studies have overlooked this and directly used them as recorded. But this was the serious concern because INSULIN variable has more than 40% values are zero. After that researchers start treating them as missing data and have published several studies. The occurrences of zero value in different variables are as follows:

TABLE 1.2 OCCURRENCES OF ZERO IN DIFFERENT VARIABLES

S.No.	Variable	No. of Zero
1	PREGNANT	111
2	PGLUCOSE	5
3	DBP	35
4	TSFT	227
5	INSULIN	374
6	BMI	11
7	DPF	0
8	AGE	0

It is important to note that there may be cases where we can not sure that the presence of zero should be treated as missing or not, just in case of variable PREGNANT (number of times a woman get pregnant). It can be zero times or more than one both cases can be considered but treat it as non-missing is more relevant. Missing data can seriously distort the correlation between the variables. In case of BMI and TFST both variables used to measure obesity and must be highly correlated, but the computed correlation coefficient recorded 0.393 which is weak positive correlation. After removing the record of zero instances of TFST yields correlation coefficient 0.632(highly positive). Instead of removing the missing instances we have calculated the feature importance on the sample with no missing values using Recursive Feature Elimination (RFE). To get the more confidence on feature selection k-fold cross validation with Stratified k-fold is used.

The top performing features has chosen based on Recursive Feature Elimination with Cross Validation (RFECV) result are PREGNANT, PGLUCOSE, BMI and DPF. The selected features have very fewer missing values which has replaced by the

mean. After selecting above features we have used complete dataset for the experiment. After that we have applied Standard Scaler method to scale the features at unit variance for efficient learning. Now the preprocessed dataset is ready for the Stage B for the predictions.

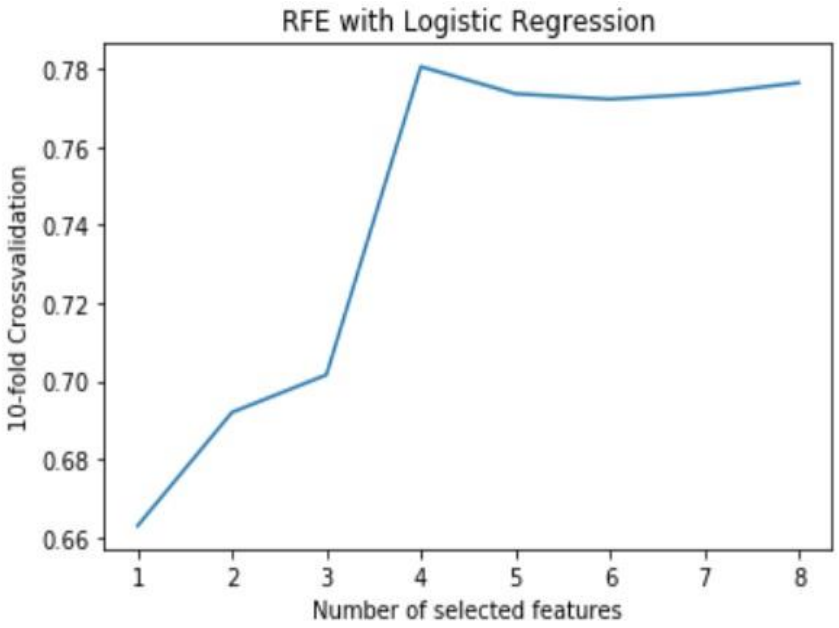

FIGURE 1.2 THE MOST SUITABLE & NUMBER OF FEATURES FOR PREDICTIONS

5. CLASSIFICATION

Classification is the task of assigning the new observation to the class to which they most likely to belong i.e. close to the accuracy, based on the classification model built from the labeled training data. For e.g., A good classifier can predict the condition of patient in the future based on various symptoms and other parameters.
The classification can binary and multilevel. When only two target classes are there in the problem, it is known as binary classification. For example, whether the patient has type-2 diabetes or not? But in multilevel classification there must be more than one target class present in the problem statement. For example, a patient admitted in the ICU has a low, medium and high risk of mortality. The dataset taken for this study is a binary classification problem.

In machine learning approach the actual dataset is divided into two parts. The first part of the data (training data) is used to build the classification model by train it and the second part(test data)validates the model accuracy. But splitting of data must be done very carefully else the problem of overfitting and underfitting will be arises. In this study, we have used train_test_split() method of Scikit-Learn library of python. through this function, you can divide the dataset into the different ratio. But 80/20 (train/test) rule is mostly used in the studies. The classification algorithms used in this study are given below.

6. LOGISTIC REGRESSION (LR)

It is a supervised machine learning algorithm borrowed from the traditional statistics which uses a Logistic function called sigmoid function g(z) that takes any value (independent variables) and predict the discrete categories (dependent variables) between 0 and 1. But this model can be extended to multiclass classification using OvR technique. As it is borrowed from the linear regression, so the z value is similar as linear regression:

$$z = \theta_0 + \theta_1 x_1 + \theta_2 x_2 + \theta_3 x_3 + \cdots$$

$$h(\theta) = g(z)$$

$$g(z) = \frac{1}{1 + e^{-z}}$$

The $h(\theta)$ means to p(y=1|x), i.e. the probability of positive event that we want to predict, for example, the probability that the patient has type II diabetes, given features x. So, the inverse probability, not having disease $p(y=0 \mid x)=1-h(\theta)$. Logistic regression uses cross entropy as a loss functions due to non-linear sigmoid function at the end. The cost function will use two equations as given below:

$$J(\theta) = \frac{1}{m}\sum \text{cost}(y`, y)$$

$$\text{cost}(y`, y) = -\log(1 - y`) \text{ if } y = 0$$

$$\text{cost}(y`, y) = -\log(y`) \text{ if } y = 1$$

In this experiment we have used Grid Search with k-fold cross validation to find the best parameters for the LR. Parameter used with LR are given as follows:

Parameters	Values
C	1
Penalty	L1
Solver	newton-cg

TABLE 1.3 PARAMETERS USED IN LR AS RETURNED BY GRID SEARCH

6.1 ARTIFICIAL NEURAL NETWORK (ANN)

In this study we have used Multilayer Perceptron (MLP) model of Artificial neural Network. MLP is a supervised learning algorithm which learns from the training set of the given data for the function f(.):Rm==>Ro. Here m represents the number of features given as an input vector whereas o denotes the number of features for the output vector. it learns the non-linear function approximation for regression or classification problem from the given independent variable X=x1,x2,x3….. and dependent variable Y. We have used MLP classifier for our problem. For training it uses gradient descent where gradients are calculated using backpropagation. In the classification process it minimizes cross-entropy loss function, giving a vector of probability estimate p(y|x) per sample x. Two-layer feed forward backpropagation neural network is employed for the experiment in this paper. Grid search was utilized for optimal parameter setting of ANN. Parameters used in this experiment are given below in the table:

TABLE 1.3 PARAMETERS USED FOR ANN SUGGESTED BY GRIDSEARCH

Parameters	Levels
Learning rate	Constant
Hidden layers	2
Activation	Relu
Maximum Iterations	500

As we have used four best features to train the model so that input layer comprises with four neurons. Each neuron represents a unique feature. One hidden layer was used with five hundred neurons as set in the maximum iterations. Similarly, the result were obtained from other hidden layer with the constant initial learning rate 0.001 and activation function relu.

6.2 NAÏVE BAYES (NB)

It is a probabilistic method that applies Bayes theorem. It calculates the probability of a given record belonging to a specific class. It assumes that given the class, features are statistically independent of each other. This assumption is called as class conditional independence which greatly simplifies the learning process. It is a generative method that generate the data from the assumptions and distributions and then uses this prior knowledge to predict the unseen data. It performs better on less training data despite of naive assumptions. NB is always the best choice for quick and dirty implementation and considered to be the benchmark. In this experiment we have used Gaussian Naive Bayes to predict likelihood. We have not used grid search for NB because it has nothing to tune.

6.3 SUPPORT VECTOR MACHINE (SVM)

Support vector classifier is also called maximum margin classifier because it creates the maximum margin hyperplane. to achieve this the decision boundary defined to maximize the margin between the positive and negative classes. The window functions or kernels are responsible to convert the inputs into required format. SVM have different types of kernels according to problem like linear, non-linear, polynomial, radial basis function (RBF) and sigmoid etc. It returns the inner product of two points in a suitable feature space and thus can work well with high dimensional dataset. In this experiment RBF the most popular kernel is used. Gamma and C parameters are tuned to get the optimal values to achieve higher accuracy.

Parameters	Levels
Kernal	Rbf
Gamma	0.05
Regularization (C)	12

TABLE 1.4 OPTIMAL PARAMETER COMBINATION USED IN SVM USING GRID SEARCH

6.4 DECISION TREE (DT)

It constructs a hierarchical tree-like structure of the given training data. It divides the training data on the value of a feature. Its goal is to create a model that predicts the value of a target variable by learning decision rules inferred from the features. In this experiment we have used the CART (classification and regression tree) algorithm of

decision tree because training space have only numerical values. CART creates the binary tree using the features and threshold that yields the maximum information gain using Gini index at each node. We have used Decision Tree classifier from sklearn library that contains fourteen different parameters, but we tune only two parameters that are max_depth and min_samples_split to control the size and complexity of the tree. The optimal parameters used in the model are given in the table below:

Parameters	Values
Maximum depth of the model	3
Minimum samples to split	2

TABLE 1.6

7. EVALUATION MEASURES AND RESULT

Accuracy, sensitivity and specificity matrices are used in this experiment to evaluate the performance of predictions of the model. If the training space is properly balanced, then accuracy measure is enough to evaluate the model performance. But in this experiment the target variable is imbalanced i.e. 34.9% are diabetic and 65.1% are non-diabetic patients that's why precision, recall and F-score measures has used. To calculate all these measures confusion matrix is needed that are True Positive, False Positive, True Negative, False Negative. The formulation of the measures is given below in the table:

Matric	Formula
Precision(P)	TP/(TP+FP) & TN/(TN+FN)
Recall(R)	TP/(TP+FN) & TN/(TN+FP)
F1-Score	2*P*R/(P+R)
Accuracy	(TP+TN)/(TP+TN+FP+FN)

TABLE 1.7 MODEL EVALUATION MEASURES

To compare the performance of the experimented model in the study, we have used ROC-AUC curve. All the test has conducted on the discussed experimental setup and only the best results are taken of the discussed model evaluation matrices. The results of each prediction model are reported in table 7 and the comparative chart of the model performance is given in figure 3. The results produced by each model are satisfactory with the average accuracy of 80% while the best being achieved around

84% by the LR. From the recorded values in the table 7, LR has identified the diabetic patients with the high recall of 84% but at the same time model has classified the non-diabetic patients with 84% precision. The computed harmonic mean(F1-Score) of the precision and recall for LR is also 84%.

Model	Precision	Recall	F1-Score	Accuracy
LR	0.84	0.84	0.84	0.837
ANN	0.81	0.81	0.81	0.817
NB	0.80	0.80	0.81	0.805
SVM	0.80	0.80	0.80	0.798
DT	0.80	0.81	0.80	0.805

TABLE 1.8 USED EVALUATION MEASURES AND THE OBTAINED RESULT

The performance of ANN is very close to best performing prediction model and has achieved 81% accuracy. ANN is more complex than LR, it is difficult to train, takes time, overfitting issue with small dataset and needs large dataset to produce optimal result. In our case ANN could perform better if we have more training examples and balanced dataset. The Gaussian Naïve Bayes(NB) and the Decision Tree(DT) predicted with 80% accuracy but DT predicted the diabetic patients better than NB. The worst performance is given by SVM with 79% of accuracy. The SVM performs worse with small dataset this is because the data points near the support vectors (decision boundary) may not be true representation of classification decision boundary and thus creates the false maximum margin hyperplane.

8. CONCLUSION

This experimental study aims to do the comparative analysis of different ML models for predicting Type II diabetic patients. As we have shown in the literature review that many complex ML models has accurately predicted the Type II diabetic and non-diabetic patients with greater accuracy. But we hypothesize that even the simplest ML model can do better than complex models, if we proper examine the problem type and apply the suitable preprocessing techniques. In previous studies authors have trimmed the dataset to treat the inconsistencies but, in this experiment, we have taken complete dataset. We have applied the RFECV method on complete dataset and found that the features that contains the many missing values have minimal impact on the prediction accuracy and top four features are giving the best result. After excluding these features preprocessed dataset applied on different ML

model, LR outperforms on all used model evaluation measures. While others complex model performs approximately same on all measures. We have also observed that feature selection method on the few dimensions (in our case 8 independent features) has contributed to improve the model accuracy and has helped to avoid the serious concerns like multi collinearity. This study tries to establish the fact that not everytime we need to with highly complex models and even the less complex models can give the better accuracy. But this is not true in all respect and depends on the nature of data, its quality, volume etc. It is also possible that complex models can give the better result by going the deep dive in the problem set and its inconsistencies.

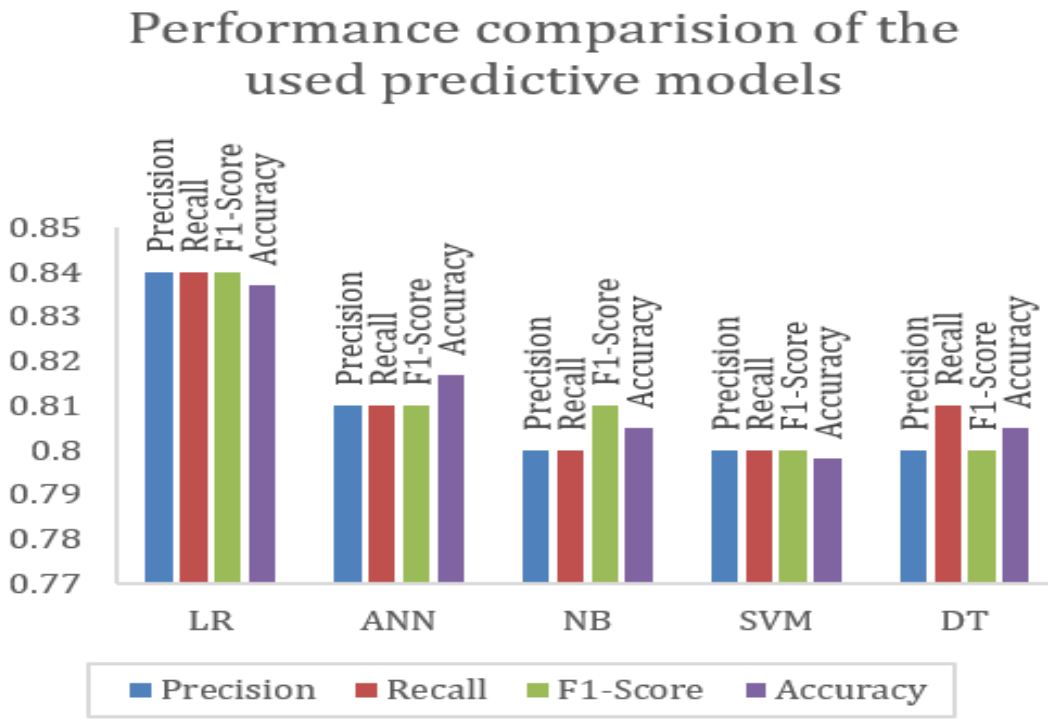

FIGURE 1.3 COMPARISON OF DIFFERENT MODELS

9. REFERENCES

- G. D. Magoulas and A. Prentza, “Machine Learning in Medical Applications,” Mach. Learn. Its Appl., vol. 2049, pp. 300–307, 2001.
- A. J. Frandsen, “Machine Learning for Disease Prediction,” p. Paper 5975, 2016.
- I. Kononenko, “Machine learning for medical diagnosis: history, state of the art and perspective.,” Artif. Intell. Med., vol. 23, no. 1, pp. 89–109, 2001.
- E. Menasalvas and C. Gonzalo-Martin, “Challenges of Medical Text and Image Processing: Machine Learning Approaches,” Springer, Cham, 2016, pp. 221–242.
- S. Habibi, M. Ahmadi, and S. Alizadeh, “Type 2 Diabetes Mellitus Screening and Risk Factors Using Decision Tree: Results of Data Mining,” Glob. J. Health Sci., vol. 7, no. 5, pp. 304–310, Sep. 2015.
- B. Farran, A. M. Channanath, K. Behbehani, and T. A. Thanaraj, “Predictive models to assess risk of type 2 diabetes, hypertension and comorbidity: machine-learning algorithms and validation using national health data from Kuwait—a cohort study,” BMJ Open, vol. 3, no. 5, p. e002457, May 2013.
- M. K. M. Dhomse Kanchan B., “Study of Machine Learning Algorithms for Special Disease Prediction using Principal of Component Analysis,” 2016 Int. Conf. Glob. Trends Signal Process. Inf. Comput. Commun., pp. 5–10, 2016.
- I. Kavakiotis, O. Tsave, A. Salifoglou, N. Maglaveras, I. Vlahavas, and I. Chouvarda, “Machine Learning and Data Mining Methods in Diabetes Research,” Comput. Struct. Biotechnol. J., vol. 15, pp. 104–116, 2017.
- S. Gambhir, S. K. Malik, and Y. Kumar, “Role of Soft Computing Approaches in HealthCare Domain : A Mini Review,” J. Med. Syst., 2016.
- M. Nirmala Devi, A. A. Balamurugan, and M. Reshma Kris, “Developing a modified logistic regression model for diabetes mellitus and identifying the0 important factors of type II DM,” Indian J. Sci. Technol., vol. 9, no. 4, pp. 1–8, 2016.
- R. Duggal, S. Shukla, S. Chandra, B. Shukla, and S. K. Khatri, “Impact of selected pre-processing techniques on prediction of risk of early readmission for diabetic patients in India,” Int. J. Diabetes Dev. Ctries., vol. 36, no. 4, pp. 469–476, 2016.
- N. H. Barakat, A. P. Bradley, and M. N. H. Barakat, “Intelligible support vector machines for diagnosis of diabetes mellitus.,” IEEE Trans. Inf. Technol. Biomed., vol. 14, no. 4, pp. 1114–1120, 2010.

- A. A. Al Jarullah, “Decision tree discovery for the diagnosis of type II diabetes,” 2011 Int. Conf. Innov. Inf. Technol., pp. 303–307, 2011.
- W. Chen, S. Chen, H. Zhang, and T. Wu, “A hybrid prediction model for type 2 diabetes using K-means and decision tree,” Proc. IEEE Int. Conf. Softw. Eng. Serv. Sci. ICSESS, vol. 2017-Novem, no. 61272399, 2018.
- R. Ramezani, M. Maadi, and S. M. Khatami, “A novel hybrid intelligent system with missing value imputation for diabetes diagnosis,” Alexandria Eng. J., 2016.
- M. Nilashi, O. Ibrahim, M. Dalvi, H. Ahmadi, and L. Shahmoradi, “Accuracy Improvement for Diabetes Disease Classification: A Case on a Public Medical Dataset,” Fuzzy Inf. Eng., vol. 9, no. 3, pp. 345–357, 2017.
- H. Wu, S. Yang, Z. Huang, J. He, and X. Wang, “Type 2 diabetes mellitus prediction model based on data mining,” Informatics Med. Unlocked, vol. 10, pp. 100–107, 2018.
- D. Sisodia and D. S. Sisodia, “Prediction of Diabetes using Classification Algorithms,” Procedia Comput. Sci., vol. 132, no. Iccids, pp. 1578–1585, 2018.
- C. O. Rep, “HHS Public Access,” vol. 4, no. 1, pp. 92–98, 2016.
- “PIMA INDIAN DIABETES DATASET,” UCI Machine Learning Repository, 1988. [Online]. Available: http://archive.ics.uci.edu/ml/datasets/Pima+Indians+Diabetes. [Accessed: 15-Apr-2018].
- A. Idri, H. Benhar, J. L. Fernández-Alemán, and I. Kadi, “A systematic map of medical data preprocessing in knowledge discovery,” Comput. Methods Programs Biomed., vol. 162, pp. 69–85, 2018.
- P. Misra and A. Yadav, “Impact of Preprocessing Methods on Healthcare Predictions,” SSRN Electron. J., Jan. 2019.
- D. Avila, M. Bussonnier, S. Corlay, Brian Granger, and J. Grout, “Juypter Notebook with Ipython,” 2014. [Online]. Available: http://jupyter.org/install. [Accessed: 18-May-2018].