A STUDY ON INDIANS DIABETES DATABASE WITH ANALYSIS AND PREDICTION USING DATA MINING AND MACHINE LEARNING CLASSIFICATION APPROACHES

Abstract

Machine learning classification constitutes a subset of both artificial intelligence and data science. It involves training models to categorize or classify data points into predefined classes based on their distinct attributes or features. The core objective of classification is to empower the model to discern patterns and correlations within the data, enabling it to correctly assign new and unseen data points to their appropriate classes. In this research, the Indian diabetes datasets consist of several medical predictors using different independent variables, and the outcomes are based on one target using a dependent variable, Outcome. Independent variables include Pregnancy, Glucose, blood pressure, Skin Thickness, Insulin, BMI, Diabetes pedigree function, Age, and Outcome in this research, analysis, and prediction using four different machine learning approaches, namely Linear Regression, SMOreg, Random Tree and REP Tree with accuracy parameters. Numerical illustrations were also provided to prove the results and discussion.

Keywords: Data Mining, Machine Learning, Decision Tree, Classifications and Diabetes.

Authors

G. K. Arun
Assistant Professor
PG Research
Department of Computer & Information Science
Arignar Anna Govt Arts College
Villupuram
Annamalai University
Annamalainagar, Tamil Nadu, India
arunnura2370@gmail.com

Dr. P. Rajesh
Assistant Professor
PG Department of Computer Science
Government Arts College
Chidambaram
(Deputed from the Department of Computer & Information Science
Annamalai University, Annamalai Nagar)
Tamil Nadu, India

I. INTRODUCTION

In classification scenarios, the input data encompasses a collection of features representing individual data points' quantifiable traits or characteristics. These features serve as the basis for the model to generate predictions about the category to which each data point belongs. The classes, in turn, signify the discrete categories or labels that the model strives to allocate to the data points. Constructing a classification model encompasses several pivotal stages: data collection and preparation, feature extraction and selection, model selection, model training, model evaluation, hyperparameter tuning, and model deployment.

Classification finds extensive utility in diverse domains, encompassing image recognition, natural language processing, fraud detection, medical diagnosis, sentiment analysis, and more. Its effectiveness hinges on factors such as the caliber and extent of the training data, algorithmic selection, and precise parameter tuning to attain accurate and dependable predictions. The term "Correctly Classified Instances" represents a concept used in evaluating machine learning models to assess their performance. Correctly Classified Instances are part of the overall model evaluation process. Incorrectly Classified Instances refer to the instances or data points in a machine learning model's evaluation or testing dataset that the model classifies incorrectly. In simpler terms, these are instances where the model's predictions do not align with the actual target or ground truth values.

In this literature survey, the corresponding authors present their study to explain conduct through machine learning applications and data mining algorithms concerning the prediction, complications, genetic background, and health conditions. This research demonstrates various aspects of ML algorithms to be used for analysis and forecasts. The research outcomes were nearly 85% using supervised learning approaches and 15% unsupervised. SVM ML approach is the most widely used algorithm, leading to new hypotheses testing targeting further investigation [1].

In a world environment, many DM and ML computerized algorithms are used to diagnose Diabetes with various drawbacks based on multiple medical tests and their measurements. There is only a need to provide some physical parameter values, and based on the provided information, the literature modeling techniques to detect the outcomes, namely the person who has Diabetes or is not using Neural Network with the support of MATLAB [2].

Various literature recently explains his research with different analyses and patterns to discover the outcomes, produce knowledge, and explain to the physician through the CP. Advanced tools in the CP allow the physician to prescribe personalized treatment plans and frequently quantify patient adherence [3]. The main goal of ML and DM is to discover various patterns in users' requirements and the outcomes to provide meaningful and valuable information for researchers and users. DM techniques are a familiar way to find proper methods to retrieve the patterns, and they help in the significant tasks of medical diagnosis and treatment. This project aims to mine the different relationships for the diabetes dataset for a better way of classification approaches [4].

In medical research, the predictive model uses short-term glucose homeostasis on ML approaches with test statistics and aims to prevent hypoglycemic and hyperglycemia using

daily measures. Data mining techniques are employed and proposed using explaining and predicting the long-term glucose control and the incidence of diabetic complications [5]. Different DM approaches help to analyze and predict diabetes detection and ultimately to improve the human health care of corresponding diabetes patients. This study clearly explains data mining methods applied to detect Diabetes data analysis and prediction of the disease [6].

The current research based on various medical-related issues using medical-related data is taken from Open source UCI warehousing, and it includes nine input parameters with a diabetes dataset and one outcome parameter, which is used to indicate whether the patient is affected by Diabetes or not [7]. Researchers explained his study and proposed a DM-related model to predict suitable planning for diabetes patients. This research considers 89 records of a different patient. This research finds 318 diabetes rows extracted using various DM and ML methods using ANFIS [8].

DM tasks include different techniques and functionalities. In this area, 57 papers were published between 2000 and 2017, and some clarifications using four research questionaries. The study explains the prediction mainly using the DM task with Neural Networks technique [9]. RF is a familiar machine learning decision tree algorithm that belongs to supervised learning methods. In these approaches, working principles are based on classification and regression. RF is generally called ensemble learning, which combines different classifiers to solve various problems with enhanced model performance. The Random Forests classifier, compared to others, is the best classifier capable of precisely classifying a massive amount of data. RF decision tree approaches mainly focus on learning procedures for classification and regression methods; it will create many decision trees and level of the tree at training time for outputs the class with classes output from single trees [10]. The researchers explain their research questions, namely the data mining techniques through classifications using various decision tree approaches es, data pre-processing, namely data transformation, how to find the accuracy using various test statistics in agriculture research [11] and [12].

II. METHODS AND BACKGROUND

1. **Linear Regression:** The statistical technique forecasts the connection between two variables by discovering the optimal straight line that most effectively aligns with the data points. It aids in ascertaining how alterations in one variable correspond to changes in another, proving valuable for predictions and trend recognition. The core idea of linear regression is to find the best-fitting straight line (also called the "regression line") through a scatterplot of data points. This line represents a linear equation of the form:

$$y = mx+b \quad \ldots (1)$$

Where y is the dependent variable (the one you want to predict or explain), x is the independent variable (the one you're using to make predictions or explanations), m is the slope of the line, representing how much, y changes will be a unit in x and b is the y-intercept, indicating the value of y when x is 0.

- **SMO:** SMO stands for "Sequential Minimal Optimization," an algorithm for training support vector machines (SVMs), machine learning models commonly used for classification and regression tasks. The SMO algorithm is particularly well-suited for solving the quadratic programming optimization problem that arises during the training of SVMs.

  Step 1. Initialization: Start with all the data points as potential support vectors and initialize the weights and biases of the SVM.
  Step 2. Selection of Two Lagrange Multipliers: In each iteration, the SMO algorithm selects two Lagrange multipliers (associated with the support vectors) to optimize.
  Step 3. Optimize the Pair of Lagrange Multipliers: Fix all the Lagrange multipliers except the selected two, and then optimize the pair chosen to satisfy certain constraints while maximizing a specific objective function.
  Step 4. Update the Model: After optimizing the selected pair of Lagrange multipliers, update the SVM model's weights and bias based on the new values of the Lagrange multipliers.
  Step 5. Convergence Checking: Check for convergence criteria to determine whether the algorithm should terminate.
  Step 6. Repeat: If convergence has not been reached, repeat steps 2 to 5 until it is step 1

- **Random Tree:** A "Random Tree" could refer to different things depending on the context. It might refer to a decision tree built using some form of randomness or a term used in a specific domain or framework. With more context, it's easier to provide a precise answer. However, I can offer a couple of interpretations that might be relevant:

  Step 1. Randomized Decision Tree: A Random Tree might refer to a decision tree constructed using randomness, like how a Random Forest uses random sampling of data and features.
  Step 2. Specific Framework: Depending on your machine learning or data analysis framework, "Random Tree" could be a term or concept introduced.

- **REPTree:** REPTree, short for "Reduced Error Pruning Tree," is a decision tree algorithm primarily used for classification tasks in machine learning. It is designed to create decision trees while incorporating a reduced-error pruning technique to avoid overfitting. The algorithm was introduced as a part of the WEKA machine-learning software. Here's how the REPTree algorithm works:

  Step 1. Tree Construction: REPTree follows a recursive approach to build a decision tree. It starts by selecting the best attribute to split the data into the metrics like information gain ratio.
  Step 2. Recursive Splitting: The algorithm examines potential attribute splits at each node and chooses the one that maximizes the selected splitting criterion.
  Step 3. Reduced Error Pruning: After the tree is fully grown, REPTree performs reduced-error pruning to eliminate branches that do not contribute significantly to the tree's accuracy.
  Step 4. Prediction: Once the tree is pruned, it can be used for making predictions.

2. **Accuracy Metrics:** The predictive model's error rate can be evaluated by applying several accuracy metrics in machine learning and statistics. The basic concept of accuracy evaluation in regression analysis is comparing the original target with the predicted one and using metrics like R-squared, MAE, MSE, and RMSE to explain the errors and predictive ability of the model [13]. The R-squared, MSE, MAE, and RMSE are metrics used to evaluate the prediction error rates and model performance in analysis and predictions [14] and [15]. R-squared (Coefficient of determination) represents the coefficient of how well the values fit compared to the original values. The values from 0 to 1 are interpreted as percentages. The higher the value is, the better the model is.

$$R^2 = 1 - \frac{\sum(y_i - \hat{y})^2}{\sum(y_i - \bar{y})^2} \quad \ldots (2)$$

MAE (Mean absolute error) represents the difference between the original and predicted values extracted by averaging the absolute difference over the data set.

$$MAE = \frac{1}{N}\sum_{i=1}^{N}|y_i - \hat{y}| \quad \ldots (3)$$

RMSE (Root Mean Squared Error) is the error rate by the square root of MSE.

$$RMSE = \sqrt{\frac{1}{N}\sum_{i=1}^{N}(y_i - \hat{y})^2} \quad \ldots (4)$$

III. NUMERICAL ILLUSTRATIONS

The dataset's main objective is to predict whether a patient has Diabetes or not diagnostically. Several constraints were placed on selecting this more extensive database—all patients in the female category with a year of at least 21. The datasets contain several medical predictor variables and one target variable, Outcome [16]. Predictor variables include various related fields, namely BMI, insulin level, age, etc. The dataset was collected from the NIDDKD and its available in Kaggle [17].

Table 1: Sample Data

Pregnancies	Glucose	Blood Pressure	Skin Thickness	Insulin	BMI	Diabetes Pedigree Function	Age	Outcome
8	183	64	0	0	23.3	0.672	32	1
1	89	66	23	94	28.1	0.167	21	0
0	137	40	35	168	43.1	2.288	33	1
5	116	74	0	0	25.6	0.201	30	0
3	78	50	32	88	31	0.248	26	1
10	115	0	0	0	35.3	0.134	29	0
2	197	70	45	543	30.5	0.158	53	1
8	125	96	0	0	0	0.232	54	1
4	110	92	0	0	37.6	0.191	30	0
10	168	74	0	0	38	0.537	34	1
10	139	80	0	0	27.1	1.441	57	0
1	189	60	23	846	30.1	0.398	59	1

5	166	72	19	175	25.8	0.587	51	1
7	100	0	0	0	30	0.484	32	1
0	118	84	47	230	45.8	0.551	31	1
7	107	74	0	0	29.6	0.254	31	1
1	103	30	38	83	43.3	0.183	33	0
1	115	70	30	96	34.6	0.529	32	1

Table 2: R2 Score or Correlation coefficient

Function and trees	Correlation Coefficient
Linear Regression	0.7322
SMOreg	0.6266
Random Tree	0.4552
REP Tree	0.7018

Table 3: Mean Absolute Error and Root Mean Squared Error

Function and trees	MAE	RMSE
Linear Regression	0.3366	0.4036
SMOreg	0.3248	0.4144
Random Tree	0.3268	0.5717
REP Tree	0.3182	0.4187

Table 4: Time taken to build the ML modeling (seconds)

Function and trees	Time Taken (seconds)
Linear Regression	0.2200
SMOreg	0.4000
Random Tree	0.0400
REP Tree	0.0900

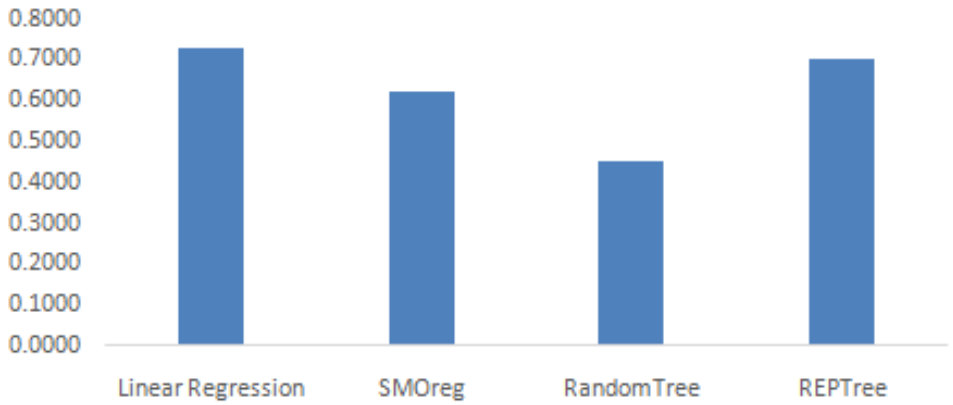

Figure 1: R2 Score or Correlation coefficient

Figure 2: Mean Absolute Error and Root Mean Squared Error

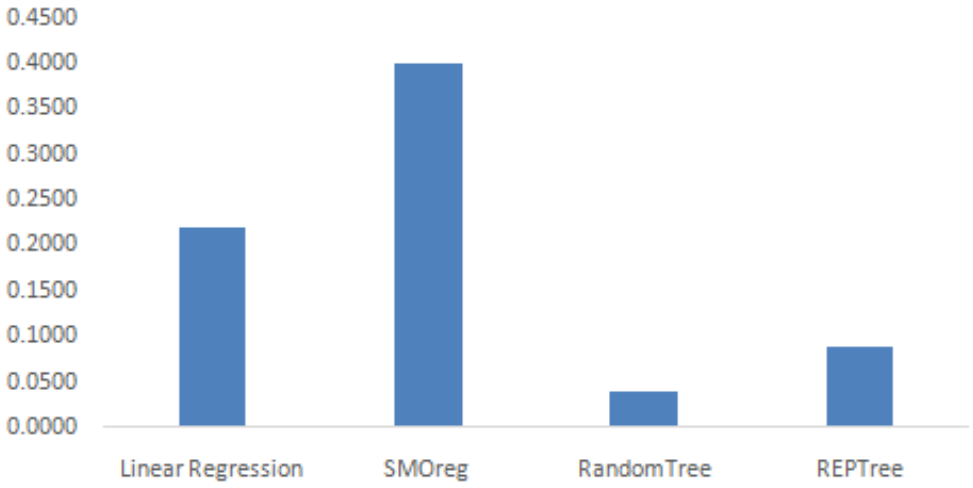

Figure 3: Time taken to build the ML model (seconds)

IV. RESULT AND DISCUSSION

This research focused on Indians Diabetes recommendation systems, including Pregnancy, Glucose, blood pressure, Skin Thickness, Insulin, BMI, Diabetes pedigree function, Age, and Outcome. The related sample dataset is shown in Table 1.

Based on Table 2 and Fig. 1, the R2 score is the most essential technique in machine learning, which is used to find the relationship between independent and dependent variables. In this case study, linear regression returns a robust positive correlation based on different parameters. REP Tree returns nearly 70%, which means linear regression produces strong positive correlations. The related results and discussions are shown in Table 2 and Figure 2.
Mean Absolute Error (MAE) is a metric commonly used for finding the measure and its accuracy of a predictive model, particularly in the context of regression tasks. In this case, all the weather and nutrient parameters have nearly 0 error rates for using MAE test statistics. Similarly, RMSE is another standard metric used to evaluate the predictive models, particularly in regression tasks. Like Mean Absolute Error (MAE), RMSE measures the accuracy of predictions. In these cases, the error rate is also nearly 0. The MAE and RMSE also returned almost 0—Table 3 and Figure 2 show the related results and discussions.

Time complexity is one of the essential parameters for analysis and prediction using machine learning approaches. In this case, for analysis of Diabetes recommendation systems, a Random tree takes very little time for research and prophecy, and the next position is the REP Tree. Similar results and discussion are shown in Table 4 and Figure 3.

Based on results and discussion, most ML approaches return better results with test statistics. However, it's essential to acknowledge the limitations of our study. Our analysis was constrained by the available datasets, which occasionally hindered a more nuanced exploration of certain factors. Furthermore, the study primarily focused on specific predictions for all the parameters. Urban context may need to be more generalizable to diverse geographical and cultural settings. In the future, consider other machine learning approaches with test statistics to improve accuracy and reduce the time complexity.

REFERENCES

[1] Kavakiotis, I., Tsave, O., Salifoglou, A., Maglaveras, N., Vlahavas, I. and Chouvarda, I., "Machine learning and data mining methods in Diabetes research," Computational and structural biotechnology journal, 15, 2017, pp.104-116.
[2] Kumari, S. and Singh, A., "A data mining approach for the diagnosis of diabetes mellitus," In 2013 7th International Conference on Intelligent Systems and Control (ISCO), 2013, pp. 373-375.
[3] Georga, E., Protopappas, V., Guillen, A., Fico, G., Ardigo, D., Arredondo, M.T., Exarchos, T.P., Polyzos, D. and Fotiadis, D.I., "Data mining for blood glucose prediction and knowledge discovery in diabetic patients: The METABO diabetes modeling and management system," In 2009 annual international conference of the IEEE Engineering in Medicine and Biology Society, 2009 pp. 5633-5636.
[4] Rajesh, K. and Sangeetha, V., "Application of data mining methods and techniques for diabetes diagnosis", International Journal of Engineering and Innovative Technology, 2(3), 2012.
[5] Georga, E.I., Protopappas, V.C., Mougiakakou, S.G. and Fotiadis, D.I., "Short-term vs. long-term analysis of diabetes data: Application of machine learning and data mining techniques," In 13th IEEE International Conference on BioInformatics and BioEngineering, 2013, pp. 1-4.
[6] Shivakumar, B.L. and Alby, S., "March. A survey on data-mining technologies for prediction and diagnosis of Diabetes," International Conference on Intelligent Computing Applications, 2014, pp. 167-173.

[7] Sanakal, R. and Jayakumari, T., "Prognosis of Diabetes using data mining approach-fuzzy C means clustering and support vector machine," International Journal of Computer Trends and Technology, 11(2), 2014, pp.94-98.
[8] Yıldırım, E.G., Karahoca, A. and Uçar, T., "Dosage planning for diabetes patients using data mining methods," Procedia Computer Science, 2011, pp.1374-1380.
[9] El Idrissi, T., Idri, A. and Bakkoury, Z., "Data mining techniques in Diabetes self-management: a systematic map", In Trends and Advances in Information Systems and Technologies, 2018, pp. 1142-1152.
[10] Rajesh, P. and Karthikeyan, M. 2017, "A comparative study of data mining algorithms for decision tree approaches using WEKA tool," Advances in Natural and Applied Sciences, 11(9), 2018, pp. 230-243.
[11] Rajesh, P. and M. Karthikeyan, "Prediction of Agriculture Growth and Level of Concentration in Paddy - A Stochastic Data Mining Approach," Advances in Intelligent Systems and Computing, Springer, Vol. 750, 2019, pp.127-139.
[12] Rajesh, P., Karthikeyan, M. and Arulpavai, R., "Data mining approaches to predict the factors that affect the groundwater level using a stochastic model", In AIP Conference Proceedings, Vol. 2177, No. 1, 2019, AIP Publishing.
[13] Smith, J.W., Everhart, J.E., Dickson, W.C., Knowler, W.C., & Johannes, R.S, "Using the ADAP learning algorithm to forecast the onset of diabetes mellitus," In Proceedings of the Symposium on Computer Applications and Medical Care, IEEE Computer Society Press,1988, pp. 261-265..
[14] Akusok, A., What is Mean Absolute Error (MAE)? Retrieved from https://machinelearningmastery.com/mean-absolute-error-mae-for-machine-learning, 2020.
[15] S. M. Hosseini, S. M. Hosseini, and M. R. Mehrabian, "Root mean square error (RMSE): A comprehensive review," International Journal of Applied Mathematics and Statistics, vol. 59(1), 2019, pp. 42–49.
[16] Chi, W. (2020). Relative Absolute Error (RAE) – Definition and Examples. Medium. https://medium.com/@wchi/relative-absolute-error-rae-definition-and-examples-e37a24c1b566
[17] https://www.kaggle.com/code/bhaktidhavade/logistic-regresstion-diabetes-dataset/input

