**A Study on Indians Diabetes Database with Analysis and Prediction Using Data Mining and Machine Learning Classification Approaches**

**G. K. Arun1\* and P. Rajesh2\***

1Assistant Professor, PG Research Department of Computer Science, Arignar Anna Govt Arts College, Villupuram, Tamil Nadu, India

Email: arunnura2370@gmail.com

2Assistant Professor, PG Department of Computer Science, Government Arts College, Chidambaram, Tamil Nadu, India

Email: rajeshdatamining@gmail.com

\*Department of Computer & Information Science, Annamalai University, Annamalainagar,

Tamil Nadu, India

**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 predictor using differernt 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 are also provided to prove the results and discussion.

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

**1.0 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. Calculating Correctly Classified Instances is 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.

**2.0 Literature Review**

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

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

Recently, various literatur explains his research with different analyis and patterns to discover the outcomes, produce knowledge, and explanations 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 outcoms to provide meaningful and useful information for the researchers and users. DM techniques is a familar way to find useful methods to retrieve the patterns, and they help in the significant tasks of medical diagnosis and treatment. This project aims to mine the differernt relationship for the diabetes dataset for better way of classification

approaches [4].

In medical research, the predictive model using short-term glucose homeostasis on ML approaches with test statistics and to aim for preventing 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 analysis and prediction in diabetes detection and ultimately to improve the human health care of corresponding diabetes patients. In this study clearly explain 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 include 9 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 explain his study and to proposed a DM related models to predict a suitable planning for diabetes patients. In this research consider 89 records of different patient . In this research, consider 318 diabetes rows extracted using various DM and ML methods using ANFIS

[8].

DM tasks include differert techniques and functionalities. In this area totally published 57 papers between the years of 2000 and 2017 and to retrive some clarifications using 4 research questionaries. The study clearlu explains the prediction was mostly used 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 performance of the model. The Random Forests classifier, compared to others, is the best classifier capable of precisely classifying a huge amount of data. RF decision tree approaches mainly focused learning procedure for classification and regression methods, it will be creating 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 his research questions, namely the data mining techniques through classifications using various decision tree approach

es, data pre-processing, namely data transformation, how to find the accuracy using various test statistics in the area of agriculture research [11] and [12].

**3.0 Methods and Background**

**3.1 Linear Regression**

Linear regression is a statistical technique employed to comprehend and forecast 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 … (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 for a unit change in x.

b is the y-intercept, indicating the value of y when x is 0.

**3.2 SMO**

SMO stands for "Sequential Minimal Optimization," an algorithm used 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.

1. **Initialization:** Start with all the data points as potential support vectors and initialize the weights and bias of the SVM.
2. **Selection of Two Lagrange Multipliers:** In each iteration, the SMO algorithm selects two Lagrange multipliers (associated with the support vectors) to optimize.
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.
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.
5. **Convergence Checking:** Check for convergence criteria to determine whether the algorithm should terminate.
6. **Repeat:** If convergence hasn't been reached, repeat steps 2 to 5 until it is.

**3.3 Random Tree**

A "Random Tree" could refer to different things depending on the context. It might refer to a decision tree that has been built using some form of randomness, or it could be a term used in a specific domain or framework. Without more context, it's challenging to provide a precise answer. However, I can offer a couple of interpretations that might be relevant:

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

**3.4 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:

1. **Tree Construction:** REPTree follows a recursive approach to build a decision tree. It starts by selecting the best attribute to split the data based on metrics like information gain or gain ratio.
2. **Recursive Splitting:** The algorithm examines potential attribute splits at each node and chooses the one that maximizes the selected splitting criterion.
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.
4. **Prediction:** Once the tree is pruned, it can be used for making predictions.

**4.0 Numerical Illustrations**

**4.1 Dataset**

dataset is originally from the National Institute of Diabetes and Digestive and Kidney Diseases. The objective of the dataset is to diagnostically predict whether a patient has diabetes, based on certain diagnostic measurements included in the dataset. Several constraints were placed on the selection of these instances from a larger database. All patients here are females at least 21 years old of Pima Indian heritage. The datasets consist of several medical predictor variables and one target variable, Outcome [13]. Predictor variables include the number of pregnancies the patient has had, their BMI, insulin level, age, and so on.

**Table 1: Sample Data**

|  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- |
| **Pregnancies** | **Glucose** | **Blood Pressure** | **Skin Thickness** | **Insulin** | **BMI** | **Diabetes****Pedigree****Function** | **Age** | **Outcome** |
| 6 | 148 | 72 | 35 | 0 | 33.6 | 0.627 | 50 | 1 |
| 1 | 85 | 66 | 29 | 0 | 26.6 | 0.351 | 31 | 0 |
| 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 (MAE) and Root Mean Squared Error (RMSE)**

|  |  |  |
| --- | --- | --- |
| **Function and trees** | **MAE** | **RMSE** |
| **Linear Regression****SMOreg****Random Tree****REP Tree** | 0.3366 | 0.4036 |
| 0.3248 | 0.4144 |
| 0.3268 | 0.5717 |
| 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 |

**Fig 1: R2 Score or Correlation coefficient**

**Fig 2: Mean Absolute Error and Root Mean Squared Error**

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

**5.0 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 indicated 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 strong positive correlation based on different parameters. REP Tree returns nearly 70%, which means linear regression returns strong positive correlations. The related results and discussions are shown in Table 2 and Fig. 2.

Mean Absolute Error (MAE) is a metric commonly used to measure the 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, the Root Mean Squared Error (RMSE) is another standard metric used to evaluate the performance of 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. Both the MAE and RMSE also returned almost 0. Table 3 and Fig. 2 show the related results and discussions.

Time complexity is one of the important 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 analysis and prediction, and the next position is the REP Tree. The similar results and discussion shown in Table 4 and Fig. 3.

**6.0 Conclusion and Further Studies**

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 not be fully generalizable to diverse geographical and cultural settings. In the future, consider other machine learning approaches with test statistics to improve the accuracy and reduce the time complexity.

**References**

1. Kavakiotis, I., Tsave, O., Salifoglou, A., Maglaveras, N., Vlahavas, I. and Chouvarda, I., 2017. Machine learning and data mining methods in diabetes research. Computational and structural biotechnology journal, 15, pp.104-116.
2. Kumari, S. and Singh, A., 2013, January. A data mining approach for the diagnosis of diabetes mellitus. In 2013 7th International Conference on Intelligent Systems and Control (ISCO) (pp. 373-375). IEEE.
3. Georga, E., Protopappas, V., Guillen, A., Fico, G., Ardigo, D., Arredondo, M.T., Exarchos, T.P., Polyzos, D. and Fotiadis, D.I., 2009, September. 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 (pp. 5633-5636). IEEE.
4. Rajesh, K. and Sangeetha, V., 2012. Application of data mining methods and techniques for diabetes diagnosis. International Journal of Engineering and Innovative Technology (IJEIT), 2(3).
5. Georga, E.I., Protopappas, V.C., Mougiakakou, S.G. and Fotiadis, D.I., 2013, November. 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 (pp. 1-4). IEEE.
6. Shivakumar, B.L. and Alby, S., 2014, March. A survey on data-mining technologies for prediction and diagnosis of diabetes. In 2014 International Conference on Intelligent Computing Applications (pp. 167-173). IEEE.
7. Sanakal, R. and Jayakumari, T., 2014. Prognosis of diabetes using data mining approach-fuzzy C means clustering and support vector machine. International Journal of Computer Trends and Technology, 11(2), pp.94-98.
8. Yıldırım, E.G., Karahoca, A. and Uçar, T., 2011. Dosage planning for diabetes patients using data mining methods. Procedia Computer Science, 3, pp.1374-1380.
9. El Idrissi, T., Idri, A. and Bakkoury, Z., 2018. Data mining techniques in diabetes self-management: a systematic map. In Trends and Advances in Information Systems and Technologies: Volume 2 6 (pp. 1142-1152). Springer International Publishing.
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), pp. 230-243.
11. Rajesh, P. and M. Karthikeyan, 2019. Prediction of Agriculture Growth and Level of Concentration in Paddy - A Stochastic Data Mining Approach. Advances in Intelligent Systems and Computing, Springer, Vol. 750, pp.127-139.
12. Rajesh, P., Karthikeyan, M. and Arulpavai, R., 2019, December. Data mining approaches to predict the factors that affect the groundwater level using stochastic model. In AIP Conference Proceedings, (Vol. 2177, No. 1). AIP Publishing.
13. Smith, J.W., Everhart, J.E., Dickson, W.C., Knowler, W.C., & Johannes, R.S. (1988). [Using the ADAP learning algorithm to forecast the onset of diabetes mellitus](http://rexa.info/paper/04587c10a7c92baa01948f71f2513d5928fe8e81). In Proceedings of the Symposium on Computer Applications and Medical Care (pp. 261-265). IEEE Computer Society Press.