International Journal of Advances in Engineering and Management (IJAEM)
Volume 4, Issue 12 Dec. 2022, pp: 885-890 www.ijaem.net ISSN: 2395-5252

Prediction of Diabetes using R

Omkar Kalange, Tejaswini Katale, Atharv Kale, Rushikesh Kahat, Juwairia Sayyed

Department of Computer Engineering, Vishwakarma Institute of Technology

Submitted: 18-12-2022 Accepted: 31-12-2022

ABSTRACT—Diabetes, a chronic disease which is caused due to continued high blood sugar levels in the human body. It is further classified into "Type1" and "Type2" based on the level of glucose in the body and also gestational diabetes (diabetes while pregnant). Currently diabetes is diagnosed using A1C, Fasting blood sugar test, Glucose tolerance test and Random blood sugar test. However, if detected early diabetes can be avoided. Detection of diabetes with Machine Learning and Deep learning techniques come into play to solve this issue. This research paper experiments and analyzes 3 Machine learning algorithms- Random Forest(RF), Decision tree and K-Nearest Neighbor(KNN) and also Upsampling, Feature Selection and Performance Metric (Precision and Recall). The data used in the dataset was procured from the Iraqi Society from the laboratory of Medical City Hospital (The specialized center for Endocrinology and Diabetes-Al-Kindy Teaching Hospital).The dataset consists of 11 risk factors. However, Upsampling, Feature Selection and Correlation Matrix helped to wave off some irrelevant factors.

Keywords: Machine Learning, Diabetes prediction, Regression analysis, KNN, Random Forest, Decision Tree ,Upsampling,Feature Selection, Precision, Recall.

I. INTRODUCTION

Diabetes is a disease that is threatening lives around the world today..The most common types of Diabetes are -Type1 , Type2 and gestational diabetes. Some of the factors include Age, High Blood Pressure , Weight , family history etc . The symptoms may include hunger , fatigue , high thirst , blurred vision , numbness etc [1]. In India's adult population, probably 72.96-million cases are of diabetes. The prevalence in urban areas ranged from 10.9% to 14.2%[9]. In rural India, the prevalence was 3.0-7.8%, from the population age group 20 years and above, with a much higher prevalence among individuals over the age of 50 (INDIAB Study)[9].More than 200 million people are infected and about a seven percent increase in the annual predominance of diabetes in the world [16]. K- Nearest Neighbor Algorithm is a simple and supervised algorithm which is used for both classification and regression models. Decision tree Algorithm is used for preparing a training model which is used to predict the outcomes . Random Forest is one of the best algorithms which is widely used for Classification and Regression analysis.Hence, this paper implements three prediction techniques as mentioned above also taking into consideration only significant factors from the dataset.For better results up-sampling, feature selection and data cleaning has been implemented.

II. DATASETDESCRIPTION

The Diabetes data is selected from the Iraqi Society from the laboratory of Medical City Hospital (The specialized center for Endocrinology and Diabetes-Al- Kindy Teaching Hospital).10 risk factors are included in the dataset also the patient's gender is taken into consideration.These characteristics are displayed in Table 1.The dataset consists of a total 1000 observations including 11 attributes. Dataset contains 2 Integer 2-Character and 8 Numeric attributes.

Table1
Diabetes Dataset Risk Factors

FEATURENUMBER	ATTRIBUTENAME	ATTRIBUTETYPE
1	Gender	Character
2	Age	Integer
3	Urea	Numeric
4	Cr	Integer
5	HbA1c	Numeric
6	Chol	Numeric
7	TG	Numeric

International Journal of Advances in Engineering and Management (IJAEM)
Volume 4, Issue 12 Dec. 2022, pp: 885-890 www.ijaem.net ISSN: 2395-5252

8	HDL	Numeric
9	LDL	Numeric
10	VLDL	Numeric
11	BMI	Numeric
12	Class	Charact0er

III. METHODOLOGY

The model proposed by the paper is divided into three main stages.The first stage is Data Processing which includes Data Cleaning, Typo Conversions and dividing the data into training and testing data.The second stage involves implementation of Machine learning models: upsampling, wrapper method and feature selection. Algorithms implemented are KNN, Decision Tree and Random Forest. The third and the final stage is to draw Accuracies, Precision and Recall values.

DATA PROCESSING:

To achieve the goal some data preprocessing is done on the given diabetes dataset[17].
It includes data cleaning which means removing duplicate values, converting categorical attributes to integer values to perform mathematical operations.

1. Data Cleaning: NA values, Duplicate data, outliers were removed from the dataset for better accuracies.
2. Typo Conversions :It refers to temporarily changing the datatype of the variable to carry out numeric operations on it .In the dataset ,Gender (M,F) was type casted into (0,1) .The outcomes (N,P,Y) which implies N- Don't have Diabetes, P- Possibility of having Diabetes , Y- have Diabetes ; were type casted into (1,2,3)
3. Training and Testing data: Training data refers to the initial dataset which is used to train your Machine Learning Model whereas Testing dataset refers to the evaluation of your model .The dataset is divided in 2 parts using split function in the ratio 0.7 for training and testing dataset.

To prevent the results to be inclined towards the majority class the following methods are used which would result in an equalization procedure.
1. Upsampling : It refers to training disproportionately the upper subset of majority class examples. The model being trained would be dominated by the majority class such as knn would predict the majority class more effectively than the minority class due to an imbalance dataset this would result in high value for sensitivity rate and low value for specificity rate. For the same the Up.sampling () method is implemented .

2. Feature Selection: Feature selection is the procedure of reducing the number of non-significant input variables when developing a predictive model for improving the performance of the model . By using the Boruta function under Boruta package a total of 4 unimportant features are found : Gender, Cr, HDL, Urea .
3. Wrapper Method : Boruta package used for Feature selection comes under Wrapper Algorithm. It helps to understand the mechanisms related to the variable of interest, rather than just building a black box predictive model with good prediction accuracy.

ALGORITHMS:

This research paper implements the following Supervised Learning Algorithms:
1. K-Nearest Neighbor : The K-Nearest Neighbor (KNN) method can be used to solve both regression and classification issues, while it is most commonly employed to tackle classification problems in business.Its main
benefit is the ease with which it may be translated and the little amount of time it takes to compute [2]. The selection of K's value is very important. Note that the K value is frequently odd in order to avoid ties [6]. To determine the distance from the point of interest to the point of the training data set it uses[17].

2. Decision tree: Decision trees are a type of supervised machine learning where the data is continuously split according to a certain parameter [17]. It uses nodes and branches, where the test on each attribute is represented
at the nodes, and the outcome of this procedure is represented at the branches, the class labels are represented at the leaf nodes.

3. Random Forest : This algorithm is self explanatory, it consists of many decision trees and utilizes ensemble learning which is a technique that combines multiple classifiers to provide solutions to complex problems.Random forests are ensemble learning methods for classification and regression that works by developing a huge number of decision trees at the time of training and yielding the class which is the method of
the classification or regression of the individual trees that are present in the forest[18].

TECHNIQUES TO EVALUATE MODEL'S EFFECTIVENESS.

1. Precision: It is one of the methods to determine the effectiveness of the model's performance.It refers to Positive Prediction made by the

International Journal of Advances in Engineering and Management (IJAEM)
Volume 4, Issue 12 Dec. 2022, pp: 885-890 www.ijaem.net ISSN: 2395-5252

model.procedure is represented at the branches, the class labels are represented at the leaf nodes.

TP(True Positive): Number of Correct predicted values.

FP(False Positive): Number of Incorrect predicted values positive class.

Precision=TP/(TP+FP)

2. Recall : Like precision, recall is also used to determine a model's performance.It refers to Positive Prediction made by the model. Higher the value of recall claims more the number of positive samples detected.It ranges from 0.01 to 1.0. TP(True Positive):Number of Correct predicted values.

IV. RESULTS

Results are inferred on the basis of 3 cases

C.1) Without Feature Selection and Upsampling

Algorithm	Accuracy	Precision	Recall
DecisionTree	0.9782609	N:1 P:0.6923 Y:0.9913	N:0.9411 P:1 Y:0.9828
KNN	0.9094203	N:0.7096 P:0.5 Y:0.9610	N:0.6875 P:0.5384 Y:0.9610
RandomForest	0.8949275	N:0.9473 P:0.2580 Y:0.9778	N:0.5625 P:0.6153 Y:0.9567

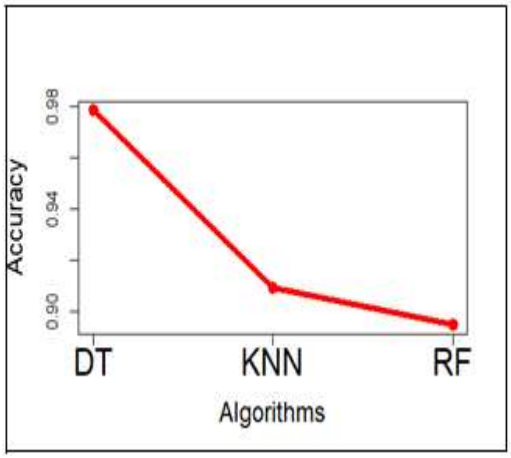

Fig 1.1- Accuracy without feature selection and upsampling

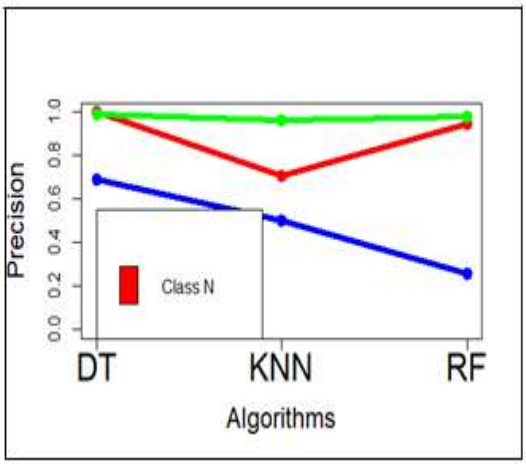

Fig 1.2- Precision without feature selection and upsampling.

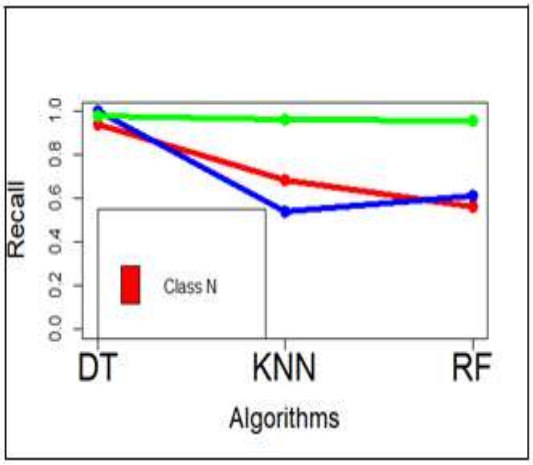

Fig 1.3- Recall without feature selection and upsampling

International Journal of Advances in Engineering and Management (IJAEM)
Volume 4, Issue 12 Dec. 2022, pp: 885-890 www.ijaem.net ISSN: 2395-5252

C.2) With first Upsampling and then Feature Selection

Algorithm	Accuracy	Precision	Recall
DecisionTree	0.9784483	N:1 P:1 Y:0.9353	N:0.9707 P:0.9666 Y:1
KNN	0.9698276	N:0.9583 P:0.9547 Y:1	N:0.9913 P:1 Y:0.9181
RandomForest	0.9827586	N:0.9957 P:0.9547 Y:1	N:1 P:1 Y:0.9482

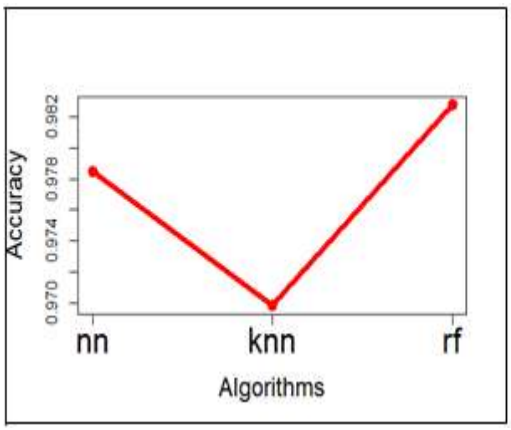

Fig 2.1- Accuracy with first Feature Selection and then Upsampling

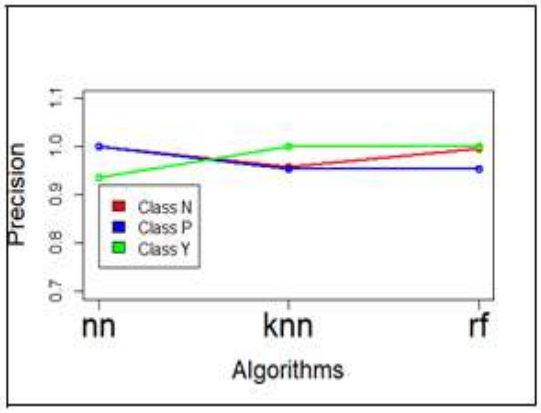

Fig 2.2- Precision with first Feature Selection and then Upsampling

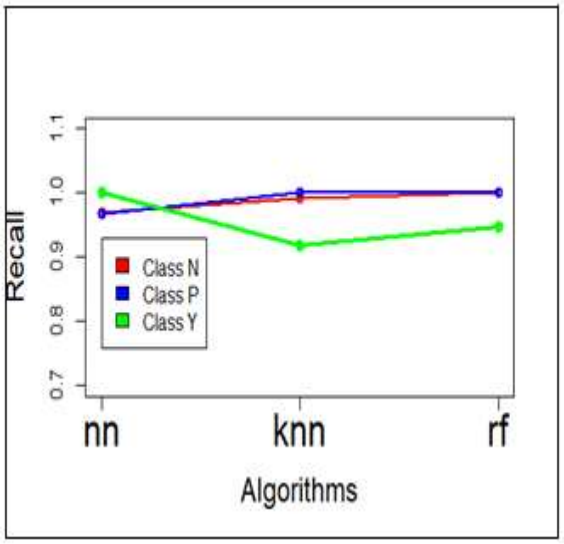

Fig 2.3- Recall with first Upsampling and then Feature Selection

C.3) With first Feature Selection and then Upsampling

Algorithm	Accuracy	Precision	Recall
Decision Tree	0.9760766	N:1 P:1 Y:0.9285	N:0.9720 P:0.9585 Y:1
KNN	0.9744817	N:0.9669 P:0.9585 Y:1	N:0.9808 P:1 Y:0.9428

International Journal of Advances in Engineering and Management (IJAEM)
Volume 4, Issue 12 Dec. 2022, pp: 885-890 www.ijaem.net ISSN: 2395-5252

RandomForest	0.9920255	N:0.9905 P:0.9857 Y:1	N:1 P:1 Y:0.9761

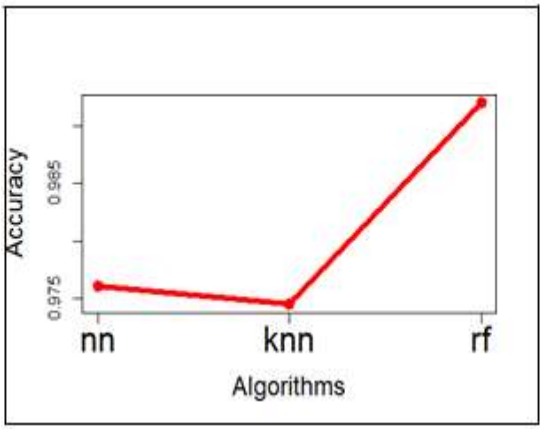

Fig 3.1- Accuracy with first Feature Selection and then Upsampling

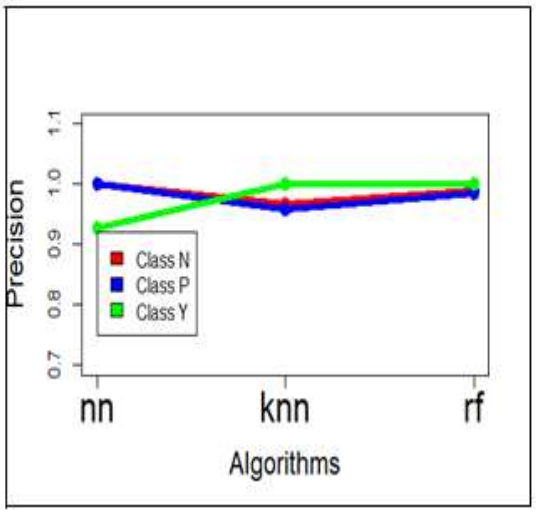

Fig 3.2- Precision with first Feature Selection and then Upsampling

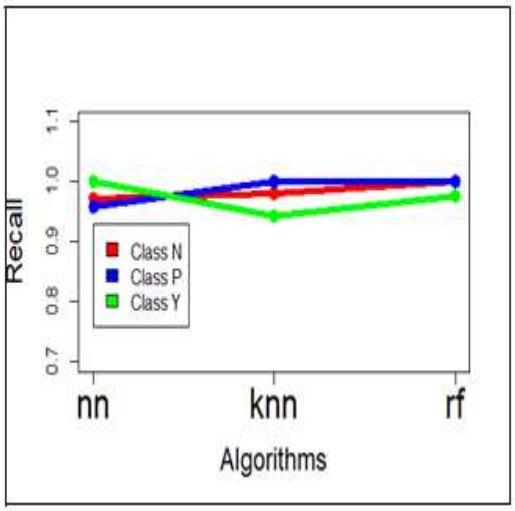

Fig 3.3- Recall with first Feature Selection and then Upsampling

The highest accuracy for all the algorithms is observed in the third model where feature selection is applied first and then upsampling is implemented In terms of other performance metrics it is observed that the precision and recall increases drastically in the second model, where first upsampling is applied and then feature selection with respect to the first model without upsampling or feature selection. In the third model where feature selection is implemented first and then upsampling a significant increase in all the three performance metrics is observed.

Sections	Highest Accuracy
Without Feature Selection and Upsampling	Decision Tree 0.9782609
With first Upsampling and then Feature Selection	Random Forest 0.9827586
With first Feature Selection and then Upsampling	Random Forest 0.9920255

V. CONCLUSION:

The detection and prediction of diabetes is collectively one of the most common medical problems in today's world and if not diagnosed in the early phase it can lead to a lot of other issues and health problems. The above use of algorithms as well model effectiveness techniques can serve as

a future scope for researchers.

REFERENCES:

[1]. Rashid, Ahlam (2020), "Diabetes Dataset", Mendeley Data, V1, doi: 10.17632/wj9rwkp9c2.1
[2]. Procedia Computer Science, Volume 167, 2020 -Prediction of Type 2 Diabetes using Machine Learning Classification Methods
[3]. 2020 IEEE International Conference on advances and development electrical and electronics Engineering (ICADE 2020) -Comparison of Different Machine Learning Models for diabetes detection
[4]. 2019 International Conference on Computing, Power and Communication Technologies (GUCON) Galgotias University, Greater Noida, UP, India. --Ensemble Learning on Diabetes Data Set and Early Diabetes Prediction
[5]. International Conference on Computational Intelligence and Data Science (ICCIDS 2018)-Prediction of Diabetes using Classification Algorithms
[6]. International Journal of Electrical and Computer Engineering (IJECE) Vol. 8, No. 5, October 2018, pp. 3966~3975 --A Comparative Analysis on the Evaluation of Classification Algorithms in the Prediction of Diabetes
[7]. (IJCSIT) International Journal of Computer Science and Information Technologies, Vol. 5 (4) , 2014, 5174-5178 -Prediction of Diabetes Using Bayesian Network
[8]. Machine Learning Tools for Long-Term Type 2 Diabetes Risk Prediction
[9]. Prediction of the Onset of Diabetes Using Artificial Neural Network and Pima Indians Diabetes Dataset
[10]. Predicting Diabetes in Healthy Population through Machine Learning
[11]. Machine Learning-Based Application for Predicting Risk of Type 2 Diabetes Mellitus (T2DM) in Saudi Arabia: A Retrospective Cross-Sectional Study
[12]. Received 26 January 2019, Revised 2 July 2019, Accepted 4 July 2019, Available online 9 July 2019. -A model for early prediction of diabetes
[13]. AINIT 2020 -Research on Diabetes Prediction Method Based on Machine Learning
[14]. Springer Nature Switzerland AG 2019 --Prediction and diagnosis of future diabetes risk: a machine learning approach
[15]. European Journal of Science and Technology Special Issue 24, pp. 53-59, April 2021 Copyright © 2021 EJOSAT -Diabetes Prediction Using Machine Learning Classification Algorithms
[16]. International Journal of Advanced Science and Technology -Diabetes Prediction Using Artificial Neural Network
[17]. International Journal of Scientific & Engineering Research Volume 12, Issue 3, March-2021 - DIABETES PREDICTION USING MACHINE LEARNING
[18]. 2019 International Conference on Computing, Power and Communication Technologies (GUCON) -Ensemble Learning on Diabetes Data Set and Early Diabetes Prediction