Ensembling Approach: Prediction on Diabetes in Medical Healthcare using Machine Learning

Astha Soni[1], Shobhana Kashyap[2]

[1]B.Tech Scholar in Department of Computer Engineering, Poornima Group of Institutions, Jaipur,[2]Assistant Professor in Department of Computer Engineering, Poornima Group of Institutions, Jaipur

[1]2014pgicsastha@poornima.org, [2]shobhana.kashyap@poornima.org

Abstract

Now a days, diabetes is one of the worst thing which occurs when the level of blood glucose becomes high which ultimately leads to our health problems such as kidney disease heart disease etc.According to the investigation of WHO (world health organization) the number of people with diabetes has been increased over the years. This research will portray how data related to diabetes can be leveraged to predict if person has diabetes or not. More specifically this research will focus on how machine learning can be utilized to predict disease such as diabetes. in this persuade diabetes dataset has been used, different machine learning algorithm has been executed on a given dataset to predict the accuracy for the analysis of patient that a person is having diabetes or not. AdaBoost classification trees, least squares support vector algorithm, naive based, random forest, support vector machine. These classification algorithms classify the person having diabetes or not on the basis of classification parameters like precision, recall, accuracy, true positive rate, false positive rate. Further, out of these five models best three models will be ensembled and predict the same parameters on the diabetic dataset.

Keywords: *Diabetes, Machine learning, classification, RF, SVM, AdaBoost, Ensembling approach*

1. Introduction

Machine learning is all about discovering new knowledge and possibilities. In some previous years machine learning has been used in various industries and research areas. There are various sectors which are in the development state using machine learning approach. Healthcare is one of the fastest growing sectors today and it is the core of a complete global transformation. Medical-Healthcare is one of those branches which are utilizing the potential of it. There are various initiatives which are being taken by healthcare sector. There is believed that machine learning is the life-saving technology that will transform healthcare. Machine learning in solution has as of late stood out as truly newsworthy. Google has built up a machine learning calculation to help recognize carcinogenic tumors on mammograms. Stanford is utilizing a profound learning calculation to recognize skin tumor. Obviously machine learning puts another bolt in the bunch of clinical basic leadership. There are several examples in medical healthcare simply shows its importance: Reduce readmissions, Prevent hospital acquired infections (HAIs), Reduce hospital Length-of-Stay (LOS), Reduce 1-year mortality. Machine learning is useful in various streams of medical and it able to reduce the effect of various diseases through its procedures. There are various examples which provides insight vision into the areas for continued innovation, likePersonalized Treatment/Behavioral Modification,Drug

Discovery/Manufacturing, Clinical Trial Research, Radiology and Radiotherapy, Smart Electronic Health Records, Epidemic Outbreak Prediction, Disease Identification/Diagnosis. Yearly, in view of better basic leadership, upgraded advancement and enhanced proficiency of research/clinical trials, and new device creation for doctors, buyers, back up plans, and controllers. Initially, our main goal is to match our capabilities. As machine learning is one of the most important powerful life-saving technologies.

2. Literature Survey

Diabetes mellitus (DM) is characterized as a gathering of metabolic issue applying critical weight on human health worldwide. Broad research in all parts of diabetes (determination, etiopathophysiology, treatment, and so on.) has prompted the age of colossal measures of information. The point of the present examination is to lead an orderly audit of the utilizations of machine learning, information mining systems and instruments in the field of diabetes look into with deference to an) Expectation and Analysis, b) Diabetic Complexities, c) Hereditary Foundation and Condition, and e) Human services and Administration with the principal classification giving off an impression of being the most mainstream. An extensive variety of machine learning calculations were utilized. When all is said in done, 85% of those utilized were portrayed by directed learning approaches and 15% by unsupervised ones, and all the more particularly, affiliation rules. Bolster vector machines (SVM) emerge as the best and widely utilized calculation. Concerning the sort of information, clinical datasets were fundamentally utilized. The title applications in the chose articles venture the convenience of separating significant learning prompting new theories focusing on more profound understanding and further examination in DM.

In this examination, a methodical exertion was made to distinguish and survey machine learning and information mining approaches connected on DM inquire about. To date, there is a huge work did in nearly all parts of DM investigate and particularly biomarker recognizable proof what's more, forecast determination. The appearance of biotechnology, with the immense measure of information delivered, alongside the expanding measure of EHRs is anticipated that would offer ascent to facilitate inside and out investigation toward determination, etiopathophysiology and treatment of DM through work of machine learning and data mining procedures in improved datasets that incorporate clinical and natural data.

Profound convolutional neural system (CNN) has been broadly connected to restorative imaging nowadays. Two specific papers concentrated on utilizing this propelled method for analysis of bosom (Sun et al., 2017) and lung diseases (Wang et al., 2017a), separately. Actually, we might have the capacity to apply the information exhibited in these papers to expand and get ready new analytic instruments for other sorts of tumor, i.e. skin, leukaemia, gastrointestinal, and so on.[3] The advances in data innovation have seen incredible advance on human services innovations in different areas these days. In any case, these new advancements have additionally made social insurance information substantially greater as well as significantly more troublesome to deal with and process. Also, in light of the fact that the information are made from an assortment of gadgets inside a brief timeframe traverse, the attributes of these information are that they are put away in various configurations also, made rapidly, which can, to a huge degree, be viewed as a major information issue. To give a more advantageous administration and condition of human services, this paper proposes a digital physical framework for persistent driven human services applications and administrations, called Health-CPS, based on cloud and huge information examination advances.

3. Data Set Description

This data set contains certain parameters which is useful to indicate the presence of Diabetes. This dataset contains Blood Pressure, BMI, Skin Thickness and Insulin, Diabetes pedigree, Glucose and age of people. By the measurement of these factors presence of Diabetes can be detected. Here the data entries are 768 in range.

Glucose	BloodPressure	SkinThickness	Insulin	BMI	DiabetesPedigreeFunction	Age	Outcome
148	72	35	0	33.6	0.627	50	1
85	66	29	0	26.6	0.351	31	0
183	64	0	0	23.3	0.672	32	1
89	66	23	94	28.1	0.167	21	0
137	40	35	168	43.1	2.288	33	1
116	74	0	0	25.6	0.201	30	0
78	50	32	88	31	0.248	26	1
115	0	0	0	35.3	0.134	29	0
197	70	45	543	30.5	0.158	53	1
125	96	0	0	0	0.232	54	1
110	92	0	0	37.6	0.191	30	0
168	74	0	0	38	0.537	34	1
139	80	0	0	27.1	1.441	57	0
189	60	23	846	30.1	0.398	59	1
166	72	19	175	25.8	0.587	51	1
100	0	0	0	30	0.484	32	1
118	84	47	230	45.8	0.551	31	1
107	74	0	0	29.6	0.254	31	1

Figure 1. Data-Set

4. Data-flow Diagram

Here, Data has been collected from “Pima Indians Diabetes Database” This dataset contains description of people having Diabetes. By measuring some factors, presence of Diabetes can be identified in a proper manner. Data Filtering and cleaning phase will detect the inaccurate data content from the dataset. Duplicate or irrelevant data will be cleaned and missing value will be filled. After it, in feature engineering process the gathered data will be transformed into features, this will improve performance of model. Model selection phase includes the selection process of model which performs best for the dataset. There are various evaluation methods like train and test training data using K-Fold cross validation. By following this whole procedure the best suitable algorithm will be achieved for the given dataset.

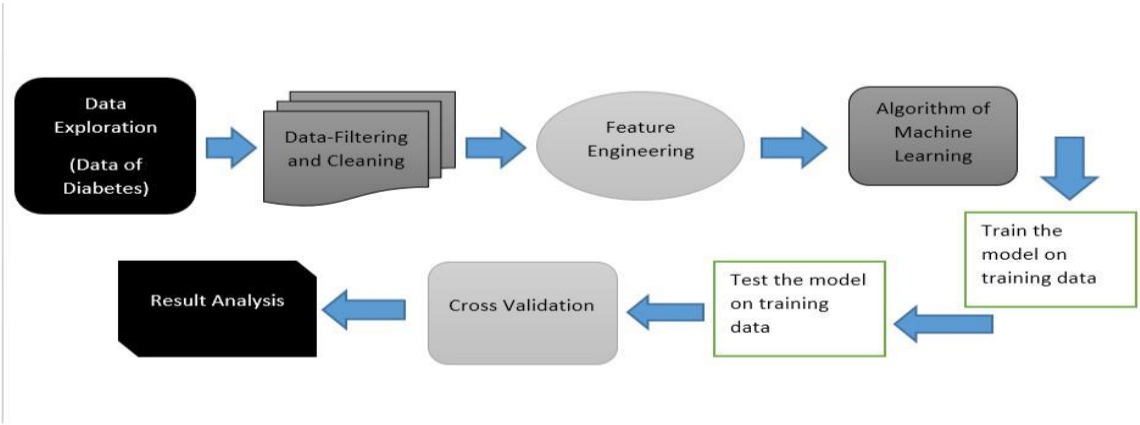

Figure 2. Data-Flow Diagram

5. Methodology

5.1 R Tool

R is free and open source software for statistical computing and graphics available for Linux, Windows and Mac OS platforms. R provides a wide variety of statistical models like Linear and nonlinear modeling, Classical statistical tests, Time-series analysis and Classification/Clustering/Regression Models. It is a well-developed, simple and effective programming language which includes conditionals, loops, user-defined recursive functions and I/O facilities.R is a dialect and condition for factual registering and illustrations.

5.2. Algorithm

Different classification algorithm has been applied on dataset like Partial Least Squares, Random Forest, Stochastic Gradient Boosting, Support Vector Machine with Radial Basis Function Kernel and Tree-Based Ensembles. Troupe strategies are methods that make different models and afterward consolidate them to deliver enhanced outcomes. Troupe techniques typically create more exact arrangements than a solitary model would. Incomplete Minimum Squares relapse (PLS) is a speedy, proficient and ideal for a model strategy in view of covariance. It is suggested in situations where the quantity of factors is high, and where it is likely that the logical factors are associated. Arbitrary backwoods or irregular choice timberlands are an outfit learning technique for characterization, relapse and different undertakings, that work by developing a huge number of decision trees at preparing time and yielding the class that is the method of the classes (arrangement).Gradient boosting is a standout amongst the most intense methods for building prescient models. Boosting is an outfit procedure in which the indicators are not made freely, but rather successively. In machine taking in, the (Gaussian) spiral premise work piece, or RBF bit, is a prominent bit work utilized as a part of different kernelized learning calculations. Specifically, it is generally utilized as a part of help vector machine arrangement.

S.no.	Algorithm name	Method name	Precision	Recall	F1 square	Accuracy
1	Partial Least Squares	pls	0.731	0.4	0.586	0.824
2	Random Forest	random-Forest	0.643	0.677	0.660	0.832
3	Stochastic Gradient Boosting	gbm,plyr	0.721	0.542	0.619	0.840
4	Support Vector Machine with Radial Basis Function Kernel	kernlab	0.721	0.513	0.600	0.836
5	Tree-Based Ensembles	nodeHarvest	0.740	0.549	0.631	0.846

Table 1.Classification Algorithms with Parameters

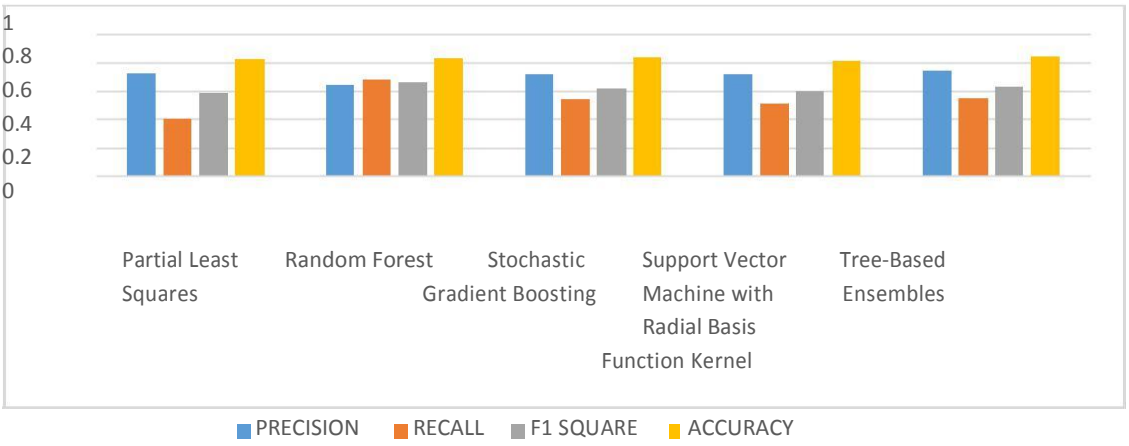

Figure 2 Graphical Representations of Classification Parameters

S.no.	Model name	Method Name	ACCURACY
1	Random Forest	random-Forest	0.73
2	Partial Least Squares	pls	0.80
3	Stochastic Gradient Boosting	gbm,plyr	0.79

Table 2. Best three Model on the basis of Accuracy

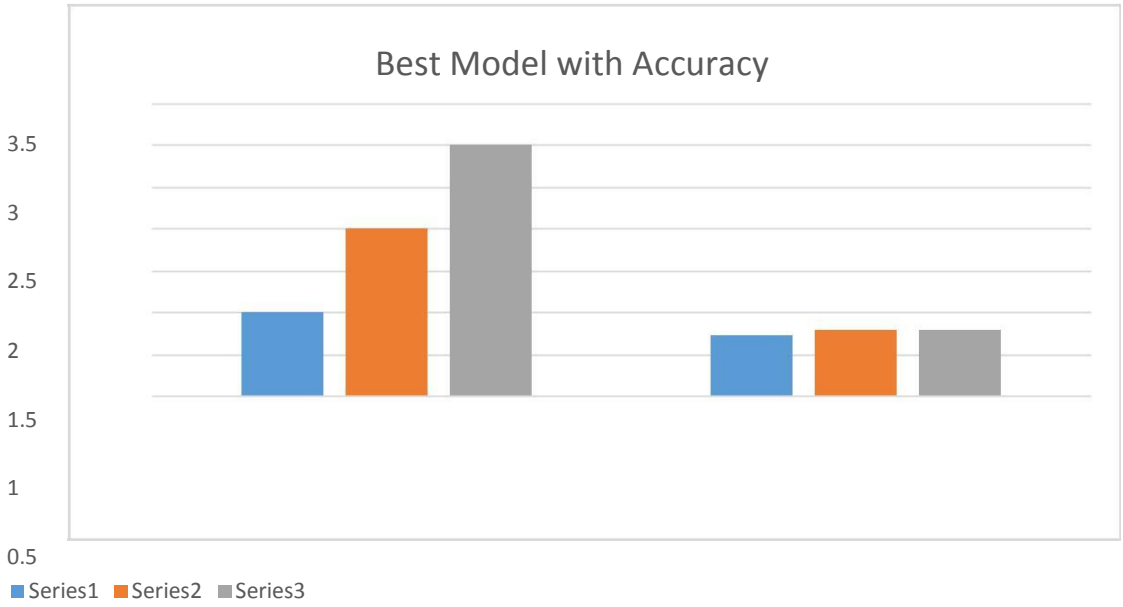

Figure 3. Graphical Representation of Best three Model

S.no.	Runs	PRECISION	RECALL	F1 SQUARE	ACCURACY
1	Run1	0.727	0.490	0.586	0.754
2	Run2	0.643	0.677	0.660	0.762
3	Run3	0.721	0.542	0.619	0.751
4	Run4	0.721	0.513	0.600	0.773
5	Run5	0.740	0.549	0.631	0.757
6	Run6	0.654	0.490	0.586	0. 761
7	Run7	0.721	0.542	0.619	0.772
8	Run8	0.643	0.677	0.660	0.763
9	Run9	0.427	0.492	0.586	0.783
10	Run10	0.721	0.542	0.619	0.757

Table 4. Ten-Fold Cross Validation

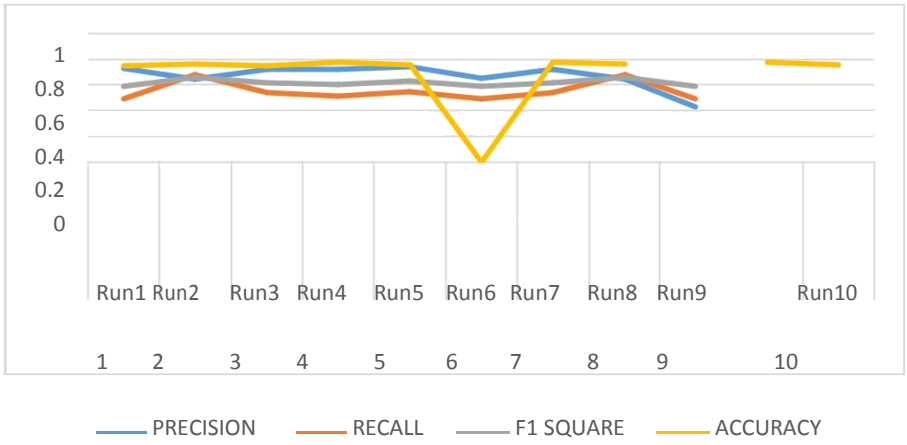

Figure 4.Ten-Fold cross Validation Graph

5. Conclusion

In this examination, a methodical exertion was made to recognize and audit machine learning and information mining approaches.DM is quickly rising as one of the best worldwide wellbeing difficulties of the 21st century. To date, there is a noteworthy work completed in nearly all parts of DM inquire.Here, a semantic approach is being done to identify the presence of Diabetes in human body. So many algorithms have been applied to get the higher percentage of accuracy. Some algorithms have been applied to the dataset; in those three algorithms are having higher accuracy. An ensembling algorithm is being made by merging of these three algorithms Random Forest, Partial Least Squares and Stochastic Gradient Boosting. After performing different run operations accuracy 78.3%has been gained which is having less complexity.

References

6.1. Journal Article

[1.] Craig P, Dieppe P, Macintyre S, Michie S, Nazareth I, Petticrew M. Developing and evaluating complex interventions: the new Medical Research Council guidance. Bmj. 2008 Sep 29;337:a1655.

[2.] Campbell R, Pound P, Pope C, Britten N, Pill R, Morgan M, Donovan J. Evaluating meta-ethnography: a synthesis of qualitative research on lay experiences of diabetes and diabetes care. Social science & medicine. 2003 Feb 1;56(4):671-84.

[3.] Evans JM, Donnelly LA, Emslie-Smith AM, Alessi DR, Morris AD. Metformin and reduced risk of cancer in diabetic patients. Bmj. 2005 Jun 2;330(7503):1304-5.

[4.] Kavakiotis I, Tsave O, Salifoglou A, Maglaveras N, Vlahavas I, Chouvarda I. Machine learning and data mining methods in diabetes research. Computational and structural biotechnology journal. 2017 Dec 31;15:104-16.

[5.] Wernick MN, Yang Y, Brankov JG, Yourganov G, Strother SC. Machine learning in medical imaging. IEEE signal processing magazine. 2010 Jul;27(4):25-38.

1.

6.2. Website

[1.] http://ieeexplore.ieee.org/document/7936635/

[2.] https://towardsdatascience.com/machine-learning-for-diabetes-562dd7df4d42

[3.] https://www.google.co.in/search?q=introduction+of+medical+field+diabetes&rlz=1C1C

[4.] https://www.slideshare.net/pacoid/data-workflows-for-machine-learning

[5.] http://www.healthcareitnews.com/news/machine-learning-101-healthcare-opportunities-are-endless