JETIR.ORG ISSN: 2349-5162 | ESTD Year : 2014 | Monthly Issue
JOURNAL OF EMERGING TECHNOLOGIES AND INNOVATIVE RESEARCH (JETIR)
An International Scholarly Open Access, Peer-reviewed, Refereed Journal

Comparison of three Machine learning Models for Diabetes Prediction

[1]Shikha Dewangan,[2]Shashank Girepunje
[1]M. Tech Scholar,[2]Assistant Professor,
Computer Science Department,
Kalinga University, Raipur,India

Abstract: Diabetes is a sickness caused on account of high glucose level in a human body. Diabetes ought not be disregarded on the off chance that it is untreated, Diabetes might cause a few significant issues in an individual like: heart related issues, kidney issue, pulse, eye harm and it can likewise influences different organs of human body. Diabetes can be controlled assuming it is anticipated before. To accomplish this objective this venture work we will do early expectation of Diabetes in a human body or a patient for a higher precision through applying, Various Machine Learning Techniques. AI methods Provide better outcome for forecast by developing models from datasets gathered from patients. In this work we will utilize Machine learning methods on a dataset to foresee diabetes. Which are Decision Tree (DT), Support Vector Machine (SVM) and Random Forest (RF). The precision is distinctive for each model when contrasted with different models. The Project work gives the precise or higher exactness model shows that the model is fit for anticipating diabetes successfully.

IndexTerms **- Diabetes, Machine, Learning, Prediction, Dataset**

I. INTRODUCTION

With the fast development and improvement of the nations, there has been a prominent example of expansion in inactive way of life and the most inescapable illness equivalent with such way of life is diabetes. To impel early identification of the illness, scientists have been directing out human blemishes and inefficacy due toward human constraints and blunders to appropriately dissect information and give vigorous outcomes that can be consumed and a trick can be created to address it for every person. Diabetes is harmful sicknesses on the planet. Diabetes caused due to stoutness or high blood glucose level, etc. It influences the chemical insulin, bringing about strange digestion of crabs and works on degree of sugar in the blood. Diabetes happens when body doesn't make sufficient insulin. As per (WHO) World Health Organization around 422 million individuals experiencing diabetes especially from low or inactive pay nations. What's more this could be expanded to 490 billion up to the time of 2030.

This work investigates forecast of diabetes by taking different characteristics connected with diabetes sickness. For this reason we utilize the Pima Indian Diabetes Dataset, we apply three Machine Learning Algorithms to anticipate illness utilizing given dataset. AI is a strategy that is utilized to prepare PCs or machines expressly. Different Machine Learning Techniques give proficient outcome to gather Knowledge by building three models utilizing SVM, Decesion tree and Random Forest classifiers from gathered dataset.

II. LITERATURE SURVEY

A. Prediction of Diabetes Using Machine Learning Algorithms in Healthcare:

This paper inspects the farsighted examination in clinical consideration; six particular AI estimations are used in this investigation work. For break down reason, a dataset of patient's clinical record is gained and six different AI computations are applied on the dataset. Execution and precision of the applied estimations is discussed and taken a gander at. Assessment of the particular AI techniques used in this survey uncovers which computation is the best for gauge of diabetes. These computations fuse SVM, KNN, LR, DT, RF and NB. Conjectures were made concerning diabetes on PIMA Indian dataset including 768 records. 8 credits were picked for getting ready and testing the insightful model. From the exploratory results obtained, it might be seen that SVM and KNN gives most critical precision for predicting diabetes.

B. Machine Learning in Predicting Diabetes in the Early Stage

In this survey, we used six customary AI models, including determined backslide, support vector machine, decision tree, self-assertive forest area, helping and neural association, to make an assumption model for diabetes end. Our data was from UCI Machine Learning Repository, which was accumulated by direct studies from the patients of the Sylhet Diabetes Hospital in

Sylhet, Bangladesh and upheld by a subject matter expert. We direct limit tuning on each model to tradeoff between the accuracy and unpredictability. The testing botch shows that unpredictable boondocks, helping and neural association should presentations than determined backslide, support vector machine and decision tree. The accuracy of neural association of the test dataset achieves 96%, which is the best model among these models for anticipating diabetes.

C. Early Detection of Diabetes Mellitus using Feature Selection and Fuzzy Support Vector Machine

The essential objective of this assessment is to utilize F-Score Feature Selection and Fuzzy Support Vector Machine in portraying and recognizing DM. Incorporate assurance is used to perceive the significant features in dataset. SVM is then used to set up the dataset to create the soft norms and Fuzzy derivation process is finally used to portray the outcome. The recently referenced technique is applied to the Pima Indian Diabetes (PID) dataset. The results show a promising accuracy of 89.02% in expecting patients with DM. Besides, the procedure taken gives a further developed count of Fuzzy guidelines while at this point staying aware of sufficient accuracy. We have in like manner seen that Fuzzy SVM classifier is reasonable to the extent setting up the data to deliver the Fuzzy norms, so the proposed Fuzzy Inference can be performed in a perfect world. The preliminary outcome shows a promising result with 89.02% accuracy which is comparable and might conceivably be redesigned in future work. A piece of the key entryways which can be feasible to work on the precision of this assessment is to accept gathering procedures or using inherited computations as an extraordinary estimation approach.

D. Prediction of Diabetes Using Various Feature Selection and Machine Learning Paradigms

This paper targets encouraging a classifier and differentiating different data mining systems considering their precision for the area of diabetes taking into account different incidental effects and components. The AI techniques were applied to the Diabetes enlightening assortment given by the Biostatistics program at Vanderbilt. The best accuracy (93.95%) was considered with the Genetic computation to be a component decision technique close by Random Forest for course of action. Thusly, Random Forest close by a Genetic Algorithm can be used for viable finding and assumption for diabetes.

III. EXPERIMENTAL SETUP

Dataset:

The dataset utilized in this review is the Pima Indian Diabetes (PID) dataset, which was initially came from the National Institute of Diabetes and Digestive and Kidney Diseases (www.niddk.nih.gov). This dataset has been utilized broadly to anticipate whether a patient has diabetes in view of eight analytic estimations displayed underneath:

```
In [2]: #Reading the dataset

        data = pd.read_csv('diabetes.csv')
        data.head(10)
```

Out[2]:

	Pregnancies	Glucose	BloodPressure	SkinThickness	Insulin	BMI	DiabetesPedigreeFunction	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
5	5	116	74	0	0	25.6	0.201	30	0
6	3	78	50	32	88	31.0	0.248	26	1
7	10	115	0	0	0	35.3	0.134	29	0
8	2	197	70	45	543	30.5	0.158	53	1
9	8	125	96	0	0	0.0	0.232	54	1

Figure 1: description of dataset

Machine Learning Models:

Support Vector Machine

A Support vector machine assembles a hyperplane or set of hyperplane in high layered space and it maps every one of the models in a guide and split the examples by an unmistakable hole which is pretty much as wide as could really be expected, and each side presents one classification. In this strategy, we should change the regularization boundary C, which controls the intricacy of the model. Greater C means more grounded punishment on the misclassification, which makes the model, is almost certain over fitting. The SVM classifier is utilized to group the information into specific number of classes. To break down the diabetes, it is exceptionally difficult to apply AI and information mining in each and every examination study. We will examine various procedures and apply on the dataset. We will attempt to create the effective outcome. The current improvement straightforwardly builds precision of arrangement and less execution time.

Decision Tree

It is a supervised learning strategy, which is utilized for taking care of characterization issues. Decesion tree is a procedure which iteratively breaks the given dataset into at least two example information. The objective of the strategy is to foresee the class worth of the objective variable. The Decesion tree will assist with isolating the informational index and constructs the Decesion model to anticipate the obscure class marks. A Decision tree can be built to both parallel and persistent factors. Choice tree ideally observes the root hub in light of the greatest entropy esteem. This gives choice tree a benefit of picking the most predictable speculation among the preparation dataset. A contribution to the Decision tree is a dataset, comprising of a few credits and cases esteems and result will be the choice model. Issues looked while building a Decision model are choosing the parting property, parts, halting rules, pruning, preparing test, quality and amount, the request for parts and so forth.

Random Forest

It is supervised learning, utilized for both order and Regression. The rationale behind the random forest is stowing strategy to make arbitrary example highlights. The distinction between the choice tree and the arbitrary woods is the most common way of tracking down the root hub and parting the element hub will run haphazardly. The Steps are given beneath

1. Load the information where it comprises of "m" highlights addressing the conduct of the dataset.
2. The preparing calculation of arbitrary woods is called bootstrap calculation or sacking procedure to choose n highlight haphazardly from m elements, for example to make arbitrary examples, this model trains the new example to out of sack sample(1/third of the information) used to decide the unprejudiced OOB blunder.
3. Calculate the hub d utilizing the best parted. Split the hub into sub-hubs.
4. Repeat the means, to observe n number of trees.
5. Calculate the all out number of votes of each tree for the foreseeing objective. The most noteworthy casted a ballot class is the last forecast of the random forest.

In Our Experiment Scikit learn and Python is utilized for the test. The over three calculations have been executed utilizing python libraries. The connection of the relative multitude of qualities is likewise displayed here:

```
In [24]: #check correlations

         data.corr()
```

Out[24]:

	Pregnancies	Glucose	BloodPressure	SkinThickness	Insulin	BMI	DiabetesPedigreeFunction	Age	Outcome
Pregnancies	1.000000	0.129459	0.141282	-0.081672	-0.073535	0.017683	-0.033523	0.544341	0.221898
Glucose	0.129459	1.000000	0.152590	0.057328	0.331357	0.221071	0.137337	0.263514	0.466581
BloodPressure	0.141282	0.152590	1.000000	0.207371	0.088933	0.281805	0.041265	0.239528	0.065068
SkinThickness	-0.081672	0.057328	0.207371	1.000000	0.436783	0.392573	0.183928	-0.113970	0.074752
Insulin	-0.073535	0.331357	0.088933	0.436783	1.000000	0.197859	0.185071	-0.042163	0.130548
BMI	0.017683	0.221071	0.281805	0.392573	0.197859	1.000000	0.140647	0.036242	0.292695
DiabetesPedigreeFunction	-0.033523	0.137337	0.041265	0.183928	0.185071	0.140647	1.000000	0.033561	0.173844
Age	0.544341	0.263514	0.239528	-0.113970	-0.042163	0.036242	0.033561	1.000000	0.238356
Outcome	0.221898	0.466581	0.065068	0.074752	0.130548	0.292695	0.173844	0.238356	1.000000

Figure 2: Correlation of Attributes of dataset

IV. RESULT AND DISCUSSION

In this work various advances were taken. The proposed approach utilizes diverse order and outfit techniques and carried out utilizing python. These strategies are standard Machine Learning techniques used to get the best exactness from information. In this work we utilized three AI calculations to accomplish better contrasted with others. The Accuracy Score of the carried out calculations are displayed here:

```
from sklearn import metrics

print("Accuracy_Score =", format(metrics.accuracy_score(y_test, predictions)))
```

```
Accuracy_Score = 0.7559055118110236
```

Figure 3: Accuracy core of Random Forest classifier

```
from sklearn import metrics

print("Accuracy Score =", format(metrics.accuracy_score(y_test,predictions)))
```

```
Accuracy Score = 0.7440944881889764
```

Figure 4: Accuracy score of Decesion tree Classifier

```
#Accuracy score for SVM
from sklearn import metrics

print("Accuracy Score =", format(metrics.accuracy_score(y_test, svc_pred)))
```

```
Accuracy Score = 0.6377952755905512
```

Figure 5: The Accuracy Score of SVM Classifier

The Following Table shows the Confusion Matrix for the Implemented Algorithms:

Table 1: Confusion Matrix data for Applied Algorithms

Algorithms	TP	FP	FN	TN
SVM	162	0	92	0
Decesion Tree	128	34	31	61
Random Forest	132	30	32	60

V. CONCLUSION

The fundamental point of this task was to plan and execute Diabetes Prediction Using Machine Learning Methods and Performance Analysis of that strategies and it has been accomplished effectively. The proposed approach utilizes different Machine Learning Algorithm which are SVM, Random Forest and Decision Tree. In after grouping calculation the Random Forest and Decesion Tree are showing the great Result. The Experimental outcomes can be asst medical services to take early expectation and settle on early choice to fix diabetes and save people life.

REFERENCES

[1] M. A. Sarwar, N. Kamal, W. Hamid and M. A. Shah, "Prediction of Diabetes Using Machine Learning Algorithms in Healthcare," *2018 24th International Conference on Automation and Computing (ICAC)*, 2018, pp. 1-6, doi: 10.23919/IConAC.2018.8748992.

[2] J. Ma, "Machine Learning in Predicting Diabetes in the Early Stage," *2020 2nd International Conference on Machine Learning, Big Data and Business Intelligence (MLBDBI)*, 2020, pp. 167-172, doi: 10.1109/MLBDBI51377.2020.00037.

[3] Lukmanto, Rian & Suharjito, Suharjito & Nugroho, Ariadi & Akbar, Habibullah. (2019). Early Detection of Diabetes Mellitus using Feature Selection and Fuzzy Support Vector Machine. Procedia Computer Science. 157. 46-54. 10.1016/j.procs.2019.08.140.

[4] Maniruzzaman, Md & Rahman, Md & Ahammed, Benojir & Abedin, Md. (2020). Classification and prediction of diabetes disease using machine learning paradigm. Health Information Science and Systems. 8. 1-14. 10.1007/s13755-019-0095-z.

[5] Beloufa F, Chikh MA. Design of fuzzy classifier for diabetes disease using Modified Artificial Bee Colony algorithm. Computer methods and programs in biomedicine. 2013;: p. 92-103.

[6] Alkaragole, Mohammed Layth Zubairi, and Sefer Kurnaz. "COMPARISON OF DATA MINING TECHNIQUES FOR PREDICTING DIABETES OR PREDIABETES BY RISK FACTORS." (2019).

[7] Debadri Dutta, Debpriyo Paul, Parthajeet Ghosh, "Analyzing Feature Importance's for Diabetes Prediction using Machine Learning". IEEE, pp 942-928, 2018.

[8] K.VijiyaKumar, B.Lavanya, I.Nirmala, S.Sofia Caroline, "Random Forest Algorithm for the Prediction of Diabetes ".Proceeding ofInternational Conference on Systems Computation Automation and Networking, 2019.

[9] Md. Faisal Faruque, Asaduzzaman, Iqbal H. Sarker, "Performance Analysis of Machine Learning Techniques to Predict Diabetes Mellitus". International Conference on Electrical, Computer and Communication Engineering (ECCE), 7-9 February, 2019.

[10] Tejas N. Joshi, Prof. Pramila M. Chawan, "Diabetes Prediction Using Machine Learning Techniques".Int. Journal of Engineering Research and Application, Vol. 8, Issue 1, (Part -II) January 2018, pp.-09-13