PREDICTING DIABETES

EXPOSYS PRROJECT

TEAM

1. Arjun P

   arjunprakash027@gmail.com

3rd Semester in B.TECH Artificial Intelligence & Data Science of Easwari Engineering College, Chennai

2. Harish Raj Y

   harishyraj@gmail.com

3rd Semester in B.TECH Artificial Intelligence & Data Science of Easwari Engineering College, Chennai

3. Naresh S

   nareshsakthi1511@gmail.com

3rd Semester in B.TECH Artificial Intelligence & Data Science of Easwari Engineering College, Chennai

4. Sushmeetha C.K

   sushmee.karthick2003@gmail.com

3rd Semester in B.TECH Artificial Intelligence & Data Science of Easwari Engineering College, Chennai

ABSTRACT: Diabetes is a disease caused due to high glucose level in a human body. Diabetes ought not be overlooked assuming it is untreated then Diabetes might cause some significant issues in an individual like: heart related issues, kidney issue, circulatory strain, eye harm and it can likewise influence different organs of human body. Diabetes can be controlled on the off chance that it is anticipated before. To accomplish this objective this task work we will do early expectation of Diabetes in a human body or a patient for a higher exactness through applying, Various Machine Learning Techniques. AI strategies Provide better outcome for forecast by constructing models from datasets gathered from patients. In this work we will utilize Machine Learning Classification and group procedures on a dataset to foresee diabetes. Which is K-Nearest Neighbour (KNN), Logistic Regression (LR), Decision Tree (DT), Support Vector Machine (SVM), Gradient Boosting (GB) and Random Forest (RF). The exactness is distinctive for each model when contrasted with different models. The Project work gives the exact or higher precision model shows that the model is capable of foreseeing diabetes successfully. Our Result shows that Random Forest accomplished higher exactness contrasted with other AI methods.

INTRODUCTION: Diabetes is toxic infections on the planet. Diabetes caused in view of stoutness or high blood glucose level, etc. It influences the chemical insulin, bringing about unusual 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 particularly from low or inactive pay nations. What's more, this could be expanded to 490 billion up to the extended time of 2030. Anyway, predominance of diabetes is found among different Countries like Canada, China, and India and so forth Populace of India is presently in excess of 100 million so the genuine number of diabetics in India is 40 million. Diabetes is significant reason for death on the planet. Early expectation of illness like diabetes can be controlled and save the human existence. To achieve this, this work investigates expectation of diabetes by taking different qualities identified with diabetes illness. For this reason, we utilize the Pima Indian Diabetes Dataset, we apply different Machine Learning order and group Techniques to foresee diabetes. AI Is a technique that is utilized to prepare PCs or machines unequivocally. Different Machine Learning Techniques give proficient outcome to gather Knowledge by building different characterization and outfit models from gathered dataset.

Such gathered information can be valuable to anticipate diabetes. Different procedures of Machine Learning can proficient to do expectation, but its extreme to pick best strategy. Consequently, for this reason we apply famous grouping and troupe techniques on dataset for forecast.

Logistic regression algorithm: Logistic regression is a statistical model that in its basic form uses a logistic function to model a binary dependent variable, although many more complex extensions exist. In regression analysis, logistic regression[1] (or logit regression) is estimating the parameters of a logistic model (a form of binary regression). Mathematically, a binary logistic model has a dependent variable with two possible values, such as pass/fail which is represented by an indicator variable, where the two values are labelled "0" and "1". In the logistic model, the log-odds (the logarithm of the odds) for the value labelled "1" is a linear combination of one or more independent variables ("predictors"); the independent variables can each be a binary variable (two classes, coded by an indicator variable) or a continuous variable (any real value). The corresponding probability of the value labelled "1" can vary between 0 (certainly the value "0") and 1 (certainly the value "1"), hence the labelling; the function that converts log-odds to probability is the logistic function, hence the name. The unit of measurement for the log-odds scale is called a *logit*, from ***log**istic un**it***, hence the alternative names. Analogous models with a different sigmoid function instead of the logistic function can also be used, such as the probit model; the defining characteristic of the logistic model is that increasing one of the independent variables multiplicatively scales the odds of the given outcome at a *constant* rate, with each independent variable having its own parameter; for a binary dependent variable this generalizes the odds ratio.

Algorithm: Main aim of logistic regression is to best fit which is responsible for describing the relationship between target and predictor variable. Logistic regression is a based on Linear regression model. Logistic regression model uses sigmoid function to predict probability of positive and neg- ative class.

Sigmoid function P = 1/1+e – (a+bx) Here P = probability, a and b = parameter of Model.

XGboost model: XGBoost is a decision-tree-based ensemble Machine Learning algorithm that uses a gradient boosting framework. In prediction problems involving unstructured data (images, text, etc.) artificial neural networks tend to outperform all other algorithms or frameworks.

Algorithm:

Regularization: It penalizes more complex models through both LASSO (L1) and Ridge (L2) regularization to prevent overfitting.

Sparsity Awareness: XGBoost naturally admits sparse features for inputs by automatically ‘learning’ best missing value depending on training loss and handles different types of sparsity patterns in the data more efficiently.

Weighted Quantile Sketch: XGBoost employs the distributed weighted Quantile Sketch algorithm to effectively find the optimal split points among weighted datasets.**Cross-**

validation: The algorithm comes with built-in cross-validation method at each iteration, taking away the need to explicitly program this search and to specify the exact number of boosting iterations required in a single run.

Proposed methodology: The information dataset to perform logistic regression is collected from Kaggle pima Indians diabetes dataset, which contains data which amounts nearly 768 entries about women who has/does not have diabeties.the column entries include 'Pregnancies', 'Glucose', 'BloodPressure', 'SkinThickness', 'Insulin','BMI', 'DiabetesPedigreeFunction', 'Age' and Outcome(has diabetes or not)

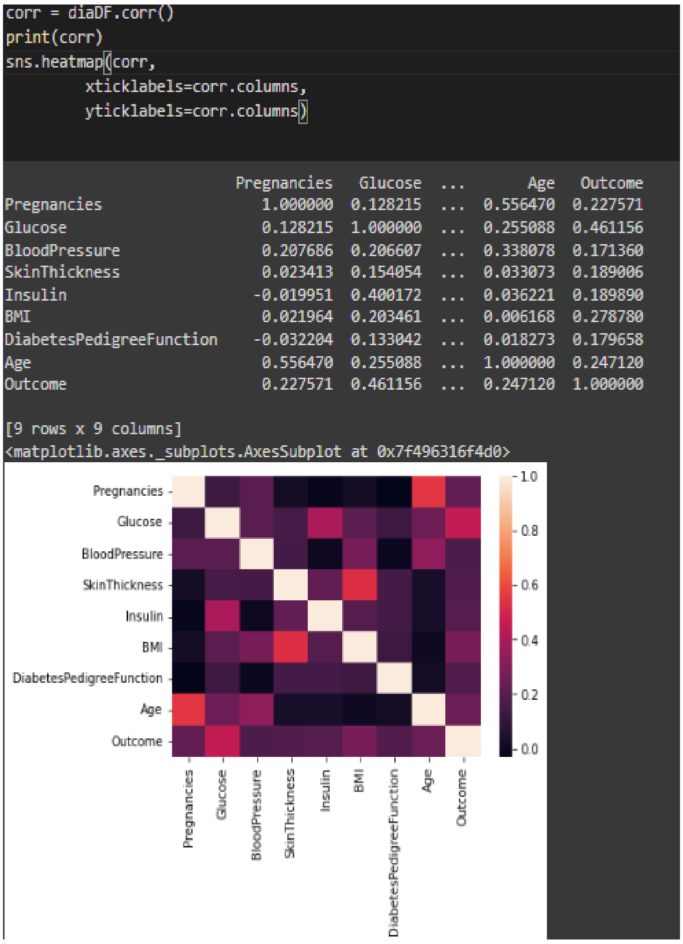
Firstly the data is visualized to find the correlation between the entries, then all the missing values(0 in our case) is replaced with mean value for all non zero terms.

We then use random undersampler to balance the amount of “0” and “1” in our output in effort to stop our model from biasing against one particular type of outcome

We then split the train data and test data in ratio of 8:2

We then train our data using logistic regression

Visualize the correlation between each entires:

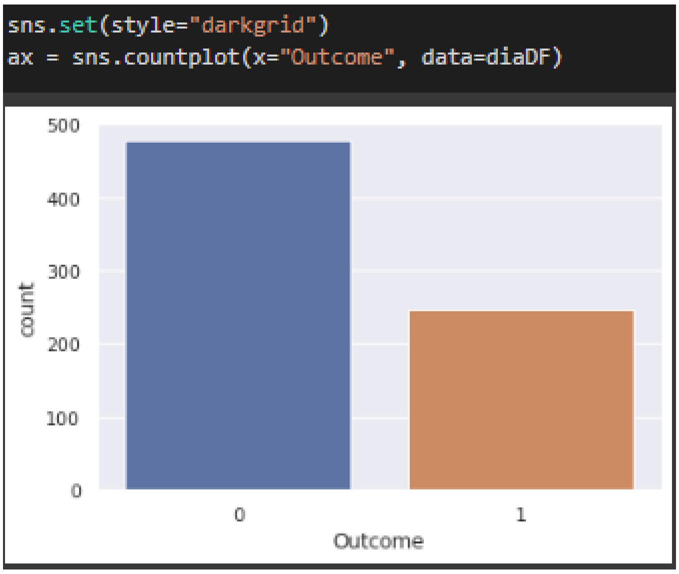


```
corr = diaDF.corr()
print(corr)
sns.heatmap(corr,
        xticklabels=corr.columns,
        yticklabels=corr.columns)
```

```
                          Pregnancies   Glucose  ...       Age   Outcome
Pregnancies                  1.000000  0.128215  ...  0.556470  0.227571
Glucose                      0.128215  1.000000  ...  0.255088  0.461156
BloodPressure                0.207686  0.206607  ...  0.338078  0.171360
SkinThickness                0.023413  0.154054  ...  0.033073  0.189006
Insulin                     -0.019951  0.400172  ...  0.036221  0.189890
BMI                          0.021964  0.203461  ...  0.006168  0.278780
DiabetesPedigreeFunction    -0.032204  0.133042  ...  0.018273  0.179658
Age                          0.556470  0.255088  ...  1.000000  0.247120
Outcome                      0.227571  0.461156  ...  0.247120  1.000000

[9 rows x 9 columns]
<matplotlib.axes._subplots.AxesSubplot at 0x7f496316f4d0>
```

Visualize total number of diabetes cases among the entries:



```
sns.set(style="darkgrid")
ax = sns.countplot(x="Outcome", data=diaDF)
```

Conclusion: The main aim of our project was to design and implement Diabetes Prediction Using Machine Learning Methods and Performance Analysis of that methods and it has been achieved successfully. Our proposed approach uses both Logistic regression and XGboost to predict diabetes with 75% accuracy. This model when deployed along with health data from doctors can save a lot of human lives

REFERENCES

1. Debadri Dutta, Debpriyo Paul, Parthajeet Ghosh, "Analyzing Feature Importances for Diabetes Prediction using Machine Learning". IEEE, pp 942-928, 2018.
2. K.VijiyaKumar, B.Lavanya, I.Nirmala, S.Sofia Caroline, "Random Forest Algorithm for the Prediction of Diabetes ".Proceeding of International Conference on Systems Compu- tation Automation and Networking, 2019.
3. Md. Faisal Faruque, Asaduzzaman, Iqbal H. Sarker, "Perfor- mance Analysis of Machine Learning Techniques to Predict Diabetes Mellitus". International Conference on Electrical, Computer and Communication Engineering (ECCE), 7-9 February, 2019.
4. Tejas N. Joshi, Prof. Pramila M. Chawan, "Diabetes Prediction Using Machine Learning Techniques".Int. Journal of Engineer- ing Research and Application, Vol. 8, Issue 1, (Part -II) Janu- ary 2018, pp.-09-13
5. Nonso Nnamoko, Abir Hussain, David England, "Predicting Diabetes Onset: an Ensemble Supervised Learning Approach ". IEEE Congress on Evolutionary Computation (CEC), 2018.
6. Deeraj Shetty, Kishor Rit, Sohail Shaikh, Nikita Patil, "Diabe- tes Disease Prediction Using Data Mining ".International Con- ference on Innovations in Information, Embedded and Com- munication Systems (ICIIECS), 2017.
7. Nahla B., Andrew et al,"Intelligible support vector machines for diagnosis of diabetes mellitus. Information Technology in Biomedicine", IEEE Transactions. 14, (July. 2010), 1114-20.
8. A.K., Dewangan, and P., Agrawal, Classification of Diabetes Mellitus Using Machine Learning Techniques, International Journal of Engineering and Applied Sciences, vol. 2, 2015.