Machine Learning Workflow on Diabetes Data

WARDA FIAZ
Department of computer science
Riphah International University
Lahore, Pakistan
wardafiaz4@gmail.com

KHADIJA TAHIR
Department of computer science
Riphah International University
Lahore, Pakistan
Khadijatahir65@gmail.com

SIDRA RANI
Department of computer science
Riphah International University
Lahore, Pakistan
17mcs1556@gmail.com

MUHHMAD UMAIR ANWAR
Punjab Safe Cities Authority Government Of Punjab
Lahore, Pakistan
m.umair09@live.com

TAYYAB WAQAR
Department of computer science
Riphah International University
Lahore, Pakistan
maliktayyabwaqar@gmail.com

ABSTRACT—
Diabetes is a condition that prevents the body from producing blood sugar, also called sugar. Without gradual and careful management, diabetes can increase the incidence of sugar in the blood, creating a risk factor for stroke and coronary artery disease. . Diabetes is divided into two types. Type 1 and Type 2. Type 1 is the type that occurs when the body neglects insulin production. Type 2, corn and diabetes affect the way your body uses insulin. Despite all the insulin action, the body is not the same as type 1, but the cells of the body do not respond as successfully as once.
In this paper, we are doing prediction of diabetes. We are using following algorithms for prediction of diabetes. The algorithms are SVC, KNN, DT, LR, GB, GNB and RF. We find out the best accuracy after comparison of all algorithm's result. For this purpose, we used python language and Jupiter notebook.

Keywords*: GB, KNN, SVC, DT, LR, GNB, classification, diabetes prediction, work flow.*

I. INTRODUCTION

The term diabetes was generally brought about by the Apolovinis Memphis in 250 BC. Diabetes was first utilized in English, in a clinical book around 1425, as diabetes. In 1675, Thomas Willis included "mellitus" in the word diabetes. Here is a quick conclusion about the sweetness of mint.

Diabetes, like sugar, is a condition that blocks the body's ability to make blood sugar. After a while, too much glucose in the blood can lead to serious health problems such as coronary heart disease, nerve damage, eye problems, and heart disease. Find ways to prevent or treat diabetes. Without stimulants, care is taken and diabetes can stimulate the progression of blood sugar, causing serious side effects, including stroke and infection. [6]

In the United States, the number of people diagnosed with diabetes over the age of 18 is 32 million. This figure is 27.9 to 32.7% of the general population. Without careful and well-controlled clinical trials, diabetes can cause blood sugar, which leads to a risk of complications including stroke and coronary arteries. There are several types of diabetes that can occur, and coping with the condition depends on the type. People who are overweight or obese are not people with diabetes. In fact, some are available from a young age. Type 1 diabetes and type 2 diabetes: This type, also known as juvenile diabetes, occurs when the body does not produce insulin. People with type 1 diabetes are insulin dependent, so they need to be injected daily to stay healthy. [7]

In this paper we will predict the diabetes and on behalf of results we examine that which treatment prefer for a person. We will build the models, create feature engineering, and apply different models for find best results. In this paper, step by step all processes will be done and examine the accuracy of models for further analysis.

I am using Python because it is simple and easy to learn code that highlights the connection, and in this way it reduces maintenance costs. Python supports modules and bundles to promote free programming and code reuse. Python clients and common libraries are freely available on the source or platform for a single platform and can be opened for use.

II. LITRATURE REVIEW

(Kavakiotis et al., 2017) various algorithms has utilized for prediction diabetes, using the machine learning techniques like as RF, SVM, or LR etc. [1]

(Muhammad Shahbaz, Karim Keshavjee, Sajida Perveen, Aziz Guergachi,) Our study developed

sensibly great models with better to arrange The J48 leaf selection uses packaging to produce three years of sugar from the Canadian public. The data used in this study came from Canada. Evaluation of the results revealed that the Adaboost troupe method was not eliminated as an independent J48 decision point. [2]
(Dheeraj Shetty, Kishor ,Nikita Patils Rit Sohail Shaikh) prediction for the illness will finished usign assistance the Bayesian calculation or KNN calculation or break down it at taking different traits the diabetes. Augmentation the precision of the calculations. Thus, working on some more ascribes which is utilized to handle the diabetes much more. [3]
(Uswa Ali Zia, Dr. Naeem Khan) using Boot strapping resampling technique for increase accuracy and then applying, k-Nearest Neighbors. Decision tree, Naive Bayes, after result DT has the highest accuracy (94%) then other. [4]
(Fikirte Girma, Woldemichael, Sumitra Menaria) used back propagation algorithms to find out the accuracy and prediction of diabetes. He used j48, SVM, and Naive Bayes classifiers. Examine 83% is best accuracy and further he will experiment on incremental of algorithm accuracy. [5]

Problem statement

Diabetes is a disorder that occurs when blood sugar levels rise. For example, heart illnesses, kidney ailment, and so forth. Diabetes is crest fundamentally as of utilization of exceptionally handled food, awful utilization propensities, and so on.
In this paper, our aim is to find prediction of diabetes and compare different algorithms according accuracy.
Questions:

- What is prediction of diabetes?
- Which algorithms is best according accuracy?
- What is significance of python during analysis?

III. USED METHOLOGY

For analysis of algorithms and prediction of diabetes following steps are done:

A. Data preparation:

It is the most important and difficult task for stating any experiments. It is still hard to find out the suitable dataset for any experiment because if there is any unsuitable dataset then it may be possible the result will be wrong and prediction will be useless. But in this paper, we are using the PIDD provided by UCI machine learning repository for prediction and find out the accuracy of models.

B. Data Expolation::

When experienced with an informational collection, first we ought to dissect and "become acquainted with" the informational index. This progression is important to acclimate with the information, to increase some comprehension of the expected highlights and to check whether information cleaning is required.
First we will import libraries and import our dataset.

```
In [26]: %matplotlib inline
         import pandas as pd
         import numpy as np
         import matplotlib.pyplot as plt
         import seaborn as sns
         import sklearn

         # Import necessary modules

         from sklearn.model_selection import KFold
         from sklearn.model_selection import train_test_split
         from sklearn.metrics import mean_squared_error
         diabetes = pd.read_csv('diabetes.csv')
         diabetes.columns

Out[26]: Index(['Pregnancies', 'Glucose', 'BloodPressure', 'SkinThickness', 'Insulin',
                'BMI', 'DiabetesPedigreeFunction', 'Age', 'Outcome'],
               dtype='object')
```

Figure 1 Dataset Attributes

We will examine the dataset using head () method

```
In [27]: diabetes.head()
```

Out[27]:

	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

Figure 2 Dataset Columns

In this dataset, 768 rows and 9 columns. ‘Outcome’ is the outcome which shows the prediction. 1 means people with diabetes and 0 means people without diabetes.

```
In [28]: diabetes.groupby('Outcome').size()

Out[28]: Outcome
         0    500
         1    268
         dtype: int64
```

Figure 3 Outcome of Dataset

C. Data Cleansing:

The next step of machine learning analysis is cleansing. It is critical step because it can make or break model. If the model building is best the result will be better.

Some process cleansing steps are:

- Remove duplication or irrelevant data
- Null or missing values
- Unexpected outliers

Observation of missing value in dataset:

```
In [30]: diabetes.isnull().sum()
         diabetes.isna().sum()

Out[30]: Pregnancies                 0
         Glucose                     0
         BloodPressure               0
         SkinThickness               0
         Insulin                     0
         BMI                         0
         DiabetesPedigreeFunction    0
         Age                         0
         Outcome                     0
         dtype: int64
```

Figure 4 Null values in dataset

Unexpected outlier

While observing the histogram we can recognize that there are a few exceptions in certain sections. We will additionally dissect those exceptions and figure out what we can do about them.

D. Feature Engineering::

Feature engineering is the way toward changing the accumulated information into feature that better speak to the issue that we are attempting to settle to the model, to improve its presentation and exactness.

It makes more input features from the current features and further more joins a few features to create progressively instinctive features to take care of to the model.

E. Model Selection::

Model selection is also known as algorithms selection. It is most energizing step and important for ML.

In this analysis, first we will calculate the accuracy of classifier models with default parameters to determine which model give us the best accuracy for diabetes dataset. We are using following classification models for observing the better performance.

- SVC
- GNB
- Gradient Boost
- Random Forest
- KNN
- Logistic Regression

We will import all the libraries of these models and then use evaluate methods. We will apply all algorithms in this step and find out the better accuracy for diabetes dataset.

Now we will apply algorithms for generate results:

Evaluation Methods

During evaluation may be it is possible that the model will be over fit. So, avoiding this issue we will not use same data for test and train. To avoid this problem we will use two precautions

- Train / Test data
- K-Fold Cross-Validation

Test and train- this strategy split the informational collection into two segments: a train set and a testing set. The train utilized for model prepare. Furthermore, testing set utilized the test model, and assess the precision.

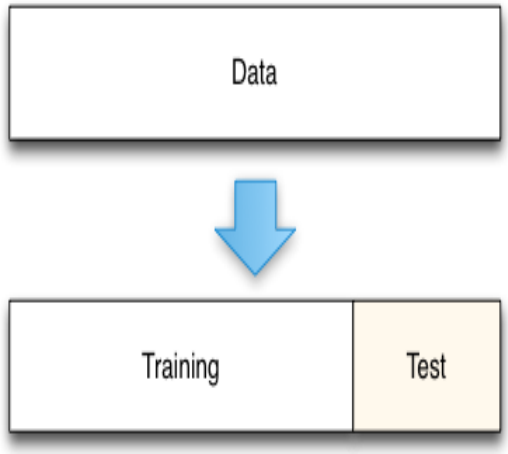
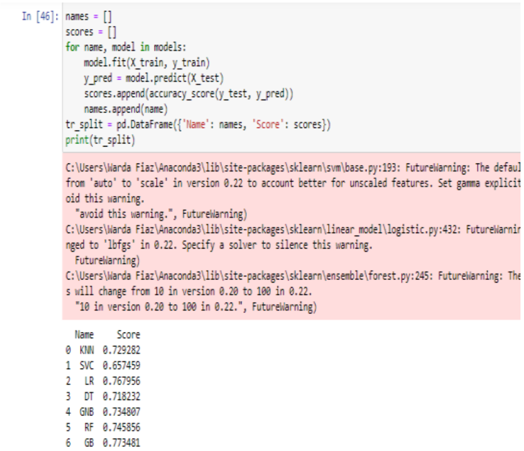


Figure 5 Test and Train Data

We will fit each model in a loop and observe the accuracy of all model using the “accuracy score”. The blow fig shows the data division in two parts.



```
In [46]: names = []
         scores = []
         for name, model in models:
             model.fit(X_train, y_train)
             y_pred = model.predict(X_test)
             scores.append(accuracy_score(y_test, y_pred))
             names.append(name)
         tr_split = pd.DataFrame({'Name': names, 'Score': scores})
         print(tr_split)

C:\Users\Warda Fiaz\Anaconda3\lib\site-packages\sklearn\svm\base.py:193: FutureWarning: The defaul
from 'auto' to 'scale' in version 0.22 to account better for unscaled features. Set gamma explicit
oid this warning.
  "avoid this warning.", FutureWarning)
C:\Users\Warda Fiaz\Anaconda3\lib\site-packages\sklearn\linear_model\logistic.py:432: FutureWarnir
nged to 'lbfgs' in 0.22. Specify a solver to silence this warning.
  FutureWarning)
C:\Users\Warda Fiaz\Anaconda3\lib\site-packages\sklearn\ensemble\forest.py:245: FutureWarning: The
s will change from 10 in version 0.20 to 100 in 0.22.
  "10 in version 0.20 to 100 in 0.22.", FutureWarning)

  Name     Score
0  KNN  0.729282
1  SVC  0.657459
2   LR  0.767956
3   DT  0.718232
4  GNB  0.734807
5   RF  0.745856
6   GB  0.773481
```

Figure 6 Accuracy using test and train spit

K-Fold Cross-Validation- This technique split dataset into K equivalent segments ("folds"), at that point utilize 1 overlap as the testing set and the association of different creases as the preparation set. At that point the model is tried for exactness. The procedure will follow the above advances K times, utilizing various overlays the testing set all the time. The normal testing precision of the procedure is the trying exactness.

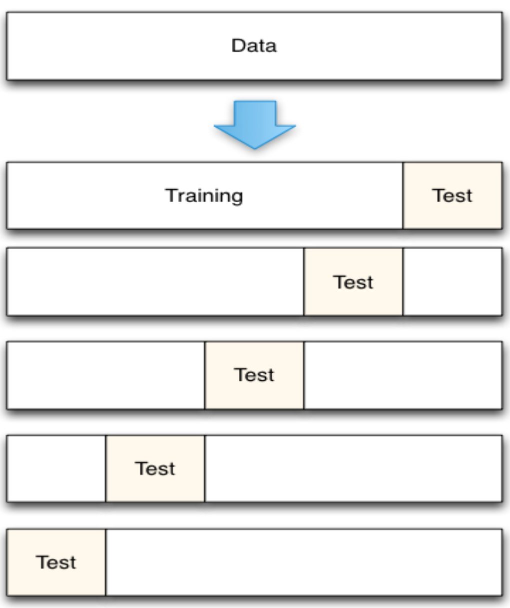
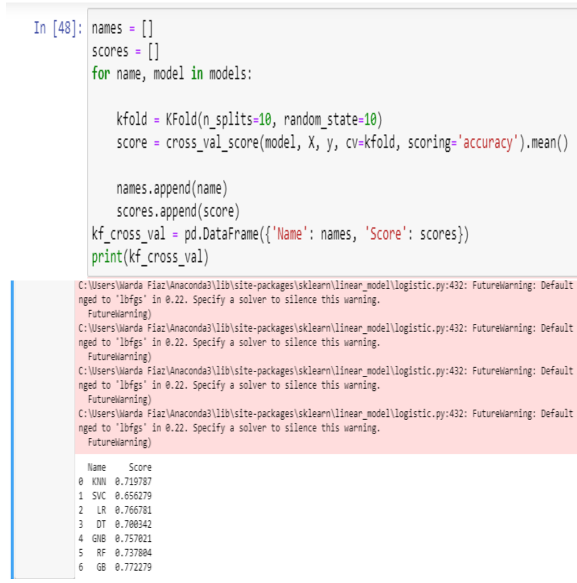


Figure 7 K-Fold cross-validation

We will push ahead with K-Fold cross approval as it is progressively exact and utilize the information effectively. We will prepare the models utilizing 10 crease cross approval and figure the mean precision of the models. "cross_val_score" gives its own preparation and exactness computation interface.



```
In [48]: names = []
         scores = []
         for name, model in models:

             kfold = KFold(n_splits=10, random_state=10)
             score = cross_val_score(model, X, y, cv=kfold, scoring='accuracy').mean()

             names.append(name)
             scores.append(score)
         kf_cross_val = pd.DataFrame({'Name': names, 'Score': scores})
         print(kf_cross_val)

C:\Users\Warda Fiaz\Anaconda3\lib\site-packages\sklearn\linear_model\logistic.py:432: FutureWarning: Default s
nged to 'lbfgs' in 0.22. Specify a solver to silence this warning.
  FutureWarning)
C:\Users\Warda Fiaz\Anaconda3\lib\site-packages\sklearn\linear_model\logistic.py:432: FutureWarning: Default s
nged to 'lbfgs' in 0.22. Specify a solver to silence this warning.
  FutureWarning)
C:\Users\Warda Fiaz\Anaconda3\lib\site-packages\sklearn\linear_model\logistic.py:432: FutureWarning: Default s
nged to 'lbfgs' in 0.22. Specify a solver to silence this warning.
  FutureWarning)
C:\Users\Warda Fiaz\Anaconda3\lib\site-packages\sklearn\linear_model\logistic.py:432: FutureWarning: Default s
nged to 'lbfgs' in 0.22. Specify a solver to silence this warning.
  FutureWarning)

  Name     Score
0  KNN  0.719787
1  SVC  0.656279
2   LR  0.766781
3   DT  0.700342
4  GNB  0.757021
5   RF  0.737804
6   GB  0.772279
```

Figure 8 K-Fold Cross-Validation

IV. RESULT AND DISSCUSION

Diabetes, like sugar, is a condition that blocks the body's ability to make blood sugar.
In paper, we have done machine learning on dataset. We have used different classifier models to evaluate the better results for prediction.

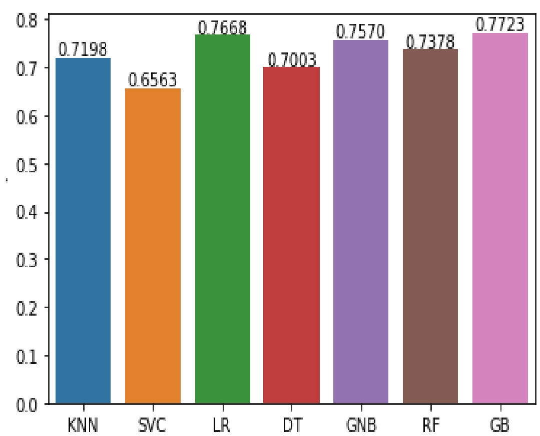


Figure 9 Classifier Accuracy

In this graph, we can see that Gradient Boosting, Logistic Regression, Random Forest and Gaussian Naive Bayes are better than other. But all of these we can see that Gradient Boosting (77%) is much better than the other all classifiers. So, we can used GB for better prediction on diabetes dataset.

V. CONCLUSION AND FUTURE WORK

In paper, we discussed about concepts of the machine learning work flow steps such as data cleansing, null values, data exploration and model

selection. We also used Scikit learn library for model selection. We discuss about the diabetes and its type. We have used many classifiers for find out the best accuracy.

After applying the model we observe that the GB classifier is better than all.

In future, we will add more feature engineering factors and also used deep learning for observe which is better for prediction of diabetes.

VI. REFERENCES

[1] Kohavi, R. (1995). "A study of cross-validation and bootstrap for accuracy estimation and model selection," in Proceedings of the 14th International Joint Conference on Artificial Intelligence, Montreal.

[2] Sajida Perveena, Muhammad Shahbaz, Aziz Guergachi, Karim Keshavjee "Performance Analysis of Data Mining Classification Techniques to Predict Diabetes" Symposium on Data Mining Applications, SDMA2016, 30 March 2016, Riyadh, Saudi Arabia.

[3] Dheeraj Shetty, Kishor Rit,Sohail Shaikh,Nikita Patils "Diabetes Disease Prediction Using Data Mining" International Conference on Innovations in Information, Embedded and Communication Systems(ICIIE CS).

[4] Uswa Ali Zia, Dr. Naeem Khan "Predicting Diabetes in Medical Datasets Using Machine Learning Techniques" International Journal of Scientific & Engineering Research (IJSER).

[5] Fikirte Girma, Woldemichael, Sumitra Menaria "Prediction of Diabetes Using Data Mining Techniques" International Conference on Trends in Electronics and Informatics (ICOEI).

[6] Kaiyang Qu , Yamei Luo , Dehui Yin , Ying Ju and Hua Tang "predicting diabetes mellitus with machine learning techniques" Frontiers in Genetics Received: 29 July 2018 Accepted: 12 October 2018 Published: 06 November 2018.

[7] https://www.medicalnewstoday.com/articles/323627