Diabetics prediction and cause analysis in United States

Akshay Puthanmari Murali , Sravya Arutla, Shwetha Panampilly
Indiana University Bloomington (Data Mining Project)]

ABSTRACT

The abstract of this project is to predict diabetics and analyze the basic cause in the geographical boundaries of the United States. The main motivation of this topic came from the fact that Diabetics is one of the most leading cause of death in US with being the #1 in kidney failure , lower limb amputation and adult blindness. We took dataset from Kaggle and analyzed more than 20 factors to achieve our results (approx 80%). With our future works like outliers removal , we are expecting to increase our accuracy even further.

Keywords : *Diabetes,Decision Tree Classifier,Multinomial Logistic Regression,Forward Feature Selection,Neural Networks.*

1. INTRODUCTION

The Flow of our project as visually depicted in Fig 1.1, we used the dataset from Kaggle and our dataset had 200K + instances with 21 features. Upon getting the data input , did Exploratory Data Analysis and Data preprocessing to understand the data. We then did the pre-processing to eliminate all the redundant and missing elements. Followed by Dimensionality Reduction to avoid the curse of Dimensionality , then we did Data Modeling with 3 different methods and later evaluated our progress.

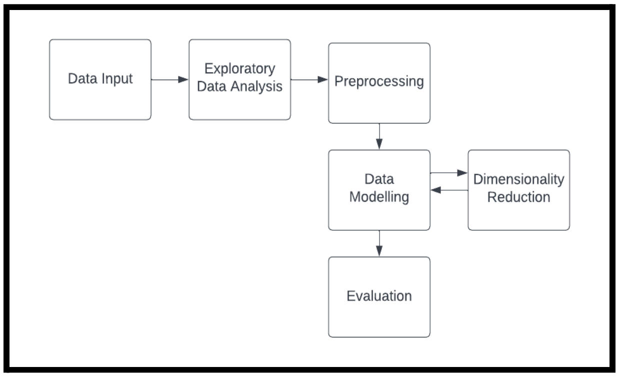
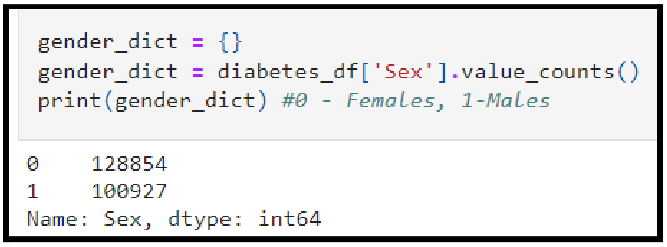


Fig 1.1 : WorkFlow Visualization.

2. DATA ANALYSIS AND PREPROCESSING

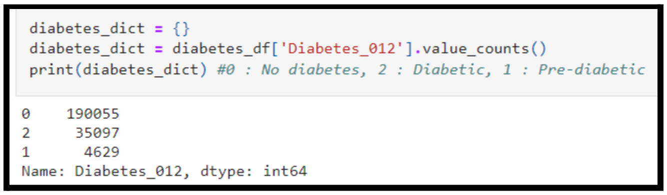
With our dataset we had a lot of features which let us perform some of the data Analysis that we have quoted in the presentation as Interesting find. We Analyzed some of the features like male female ratio, Different types of Diabetics and so on as shown in the figure.



```
gender_dict = {}
gender_dict = diabetes_df['Sex'].value_counts()
print(gender_dict) #0 - Females, 1-Males
```

```
0    128854
1    100927
Name: Sex, dtype: int64
```

Fig 2.1 : Male and Female Ratio.

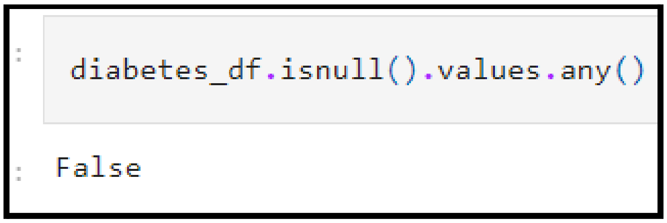


```
diabetes_dict = {}
diabetes_dict = diabetes_df['Diabetes_012'].value_counts()
print(diabetes_dict) #0 : No diabetes, 2 : Diabetic, 1 : Pre-diabetic
```

```
0    190055
2     35097
1      4629
Name: Diabetes_012, dtype: int64
```

Fig 2.2 : Types of Diabetics

Data Preprocessing had the following process to eliminate all unwanted data which could throw off our accuracy.
i) Drop Duplicates
ii) Drop Null Values
iii) Check for Outliers and Elimination (Future Works).



```
diabetes_df.isnull().values.any()
```

```
False
```

Fig 2.3 : Null Value Check after Data Preprocessing.

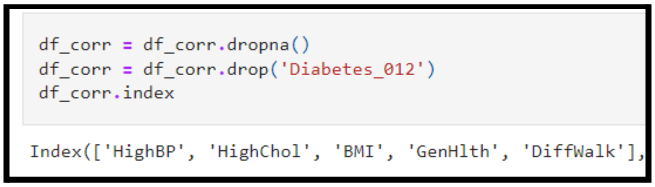
3. Dimensionality Reduction

One of the major issues with our Dataset is the sheer size and the factors associated with it. We had 200k + dataset with 21 Features affecting our scores and accuracy. In order to overcome this curse of Dimensionality and overfitting we came up with 2 Different Models to reduce the features.

i) Correlation
ii) Forward Feature Selection.

3.1 Correlation Features Selection

The correlation Feature is a straight up approach in which we check the relation between the diabetics value and the current feature , come up with a **Pearson** value for the said features and thus we perform an operation with them.
based on our Approach we found **['HighBP', 'HighChol', 'BMI', 'GenHlth', 'DiffWalk']** these elements to have a higher correlation value and took them to model the algorithm.



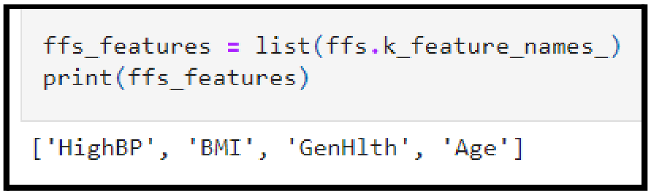
```
df_corr = df_corr.dropna()
df_corr = df_corr.drop('Diabetes_012')
df_corr.index
```

```
Index(['HighBP', 'HighChol', 'BMI', 'GenHlth', 'DiffWalk'],
```

Fig 3.1 : Correlation Outcome.

3.2 Forward Feature Selection

The forward feature selection was a better approach and always gave us better accuracy . We used the following site [1], to gain more input. Basically in forward feature selection we take a particular parameter and check the correlation , if it has the highest value in the correlation, we take the next set of features to see if the accuracy increases , we add the feature to our list else ignore it. The values we got from Forward Feature Selection are **['HighBP', 'BMI', 'GenHlth', 'Age']**



```
ffs_features = list(ffs.k_feature_names_)
print(ffs_features)
```

```
['HighBP', 'BMI', 'GenHlth', 'Age']
```

Fig 3.2 : Forward Feature Selection Outcome.

4. Data Modeling

All the feature variables in our dataset are numeric, hence the classification algorithms in Supervised machine learning are used. We have chosen the main algorithms as Decision Tree Classifier, Multinomial Logistic Regression and as an additional task , we have tried to implement the concepts of neural networks like MLPClassifier as an additional task.

4.1 Decision Tree Classifier

Decision Tree Classifier is a probabilistic prediction model which starts from the root attribute. Out of all the available independent features, the root attribute could be anything and related to this, the child nodes are calculated accordingly using the statistical approach.The attributes are selected using the measures such as Entropy, Information Gain, Gini index etc. The sklearn's DecisionTreeClassifier calculates the best suited root feature, with the remaining nodes being the children until the leaf nodes are reached. Hence this model gives slightly varied results whenever it's trained each time, Also this model has the drawback of the overfitting problem.

4.2 Multinomial Logistic Regression

The independent variable 'Diabetes_012' has three class labels 0,1 and 2, with respect to no diabetic, pre-diabetic and diabetic respectively. Multinomial logistic regression is the preferred classification algorithm due to the factor that we are classifying more than 2 items. The major advantage with Multinomial Logistic Regression is that it is uniform and its value doesn't change with respect to repetition of Code Run. We used sklearn's Logistic Regression by using the multi_class label as multinomial as mentioned in [2].

4.3 Neural Networks-Mlp Classifier

The concept of neural networks comes under deep learning which inturn is the subset of the Machine Learning concept. Neural networks involve the concept of hidden layers apart from the input and output layers. A neural network learns the data, the weights , the strength and the connection between the above-mentioned three layers allowing it to come up with more accurate predictions.

A Multilayer Perceptron, a classification algorithm in a neural network used to train the model. MLPClassifier from sklearn uses the parameters called hidden_layer_sizes, in order to set the number of layers and the number of nodes in the neural network. The parameters reference is taken from [3] and [4].

5. RESULTS

The above discussed supervised machine learning algorithms are implemented to train the model for the different cases mentioned below. The training dataset consists of the 80% from the diabetes dataset while the testing dataset remains 20%.

Case 1 : Training the models with all the 21 independent features available in dataset

Case 2: Training the models with the main Features obtained from the Forward features selection method.

Case 3: Training the models with the highly correlated features from the correlation method.

5.1. Decision Tree Classifier

The Decision Tree Classifier algorithm is implemented in order to train our dataset in three different cases. In the case 1) it is seen that as there are many features involved its accuracy is around 74% but when its features are reduced in the other 2 cases, the accuracy almost remains the same i.e. around 82%. The results for these are shown in below figure 5.1 and table 5.1.

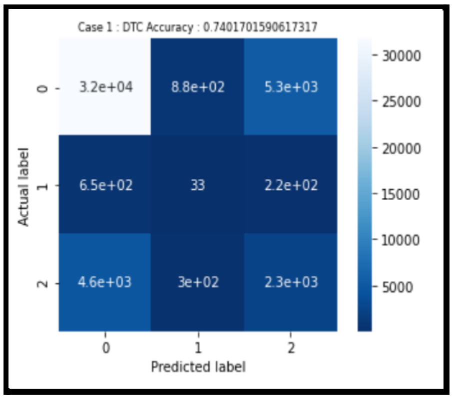


Fig 5.1.1 : Case (1) General Accuracy

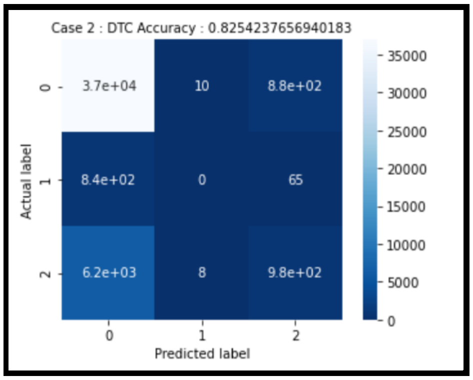


Fig 5.1.2 : Case (2) Correlation Accuracy

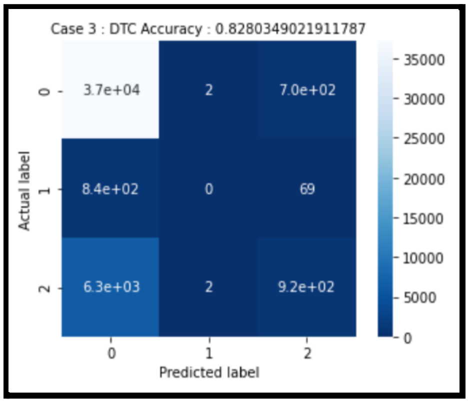


Fig 5.1.3 : Case (3) FFS Accuracy

5.1. DTC : Confusion Matrix for all the 3 cases 1,2 and 3

Types	Decision Tree Classifier
Case 1	0.7401
Case 2	0.8254
Case 3	0.8280

Table5.1: DTC Accuracy Values

5.2. Multinomial Logistic Regression

The multinomial logistic regression method is used in all three cases for the same set of cases as mentioned in the section 5.1. It is quietly evident from the results mentioned below that the accuracy of the model remained the same in all the three cases implemented, stating that this model could be the best fit model for our project.

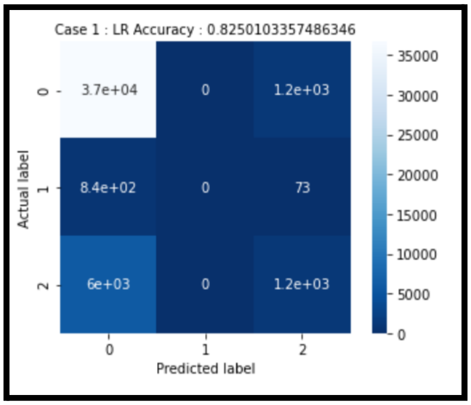


Fig 5.2.1 : Case (1) General Accuracy

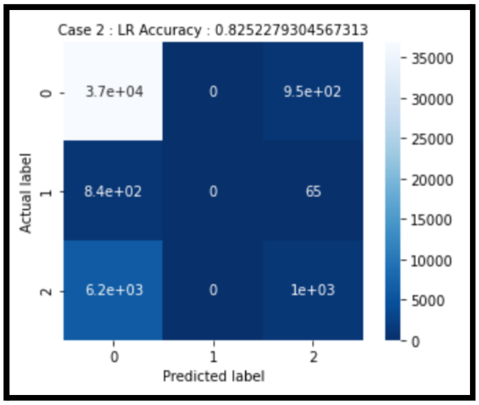


Fig 5.2.2 : Case (2) Correlation Accuracy

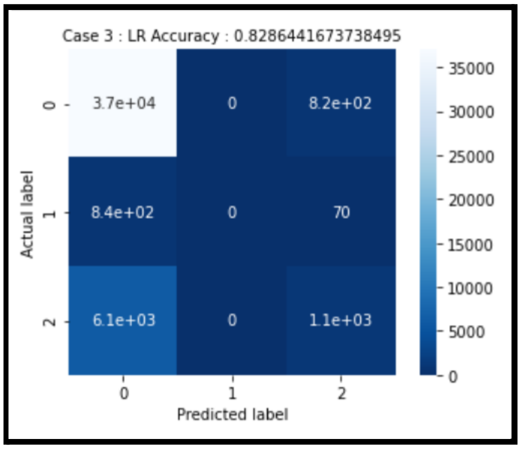


Fig 5.2.3 : Case (3) FFS Accuracy

Types	Multinomial Logistic Regression
Case 1	0.8250
Case 2	0.8252
Case 3	0.8286

Table 5.2: LR Accuracy Values

5.3. Neural Networks - MLPClassifier

This method is implemented as an additional task, just to check if there is any other change in the accuracy obtained when deep learning modeling is used. The neural network from MLPClassififer uses the 8 hidden layers to train our dataset with the rest of the parameters as the defaulted ones as mentioned in [2] and [3]. It is seen from the results that there is no significant change in the accuracy for all the three cases is almost the same i.e. 83% just like the case with the usage of multinomial logistic regression as shown in Figures 5.3 and table 5.3.

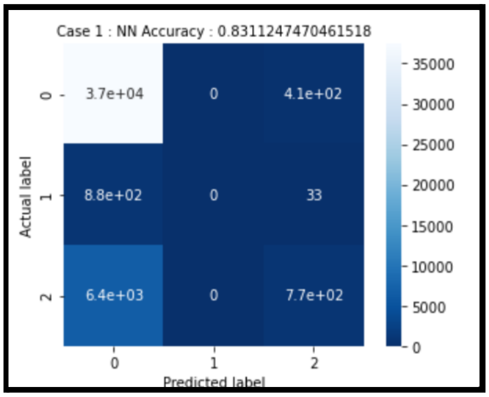


Fig 5.3.1 : Case (1) General Accuracy

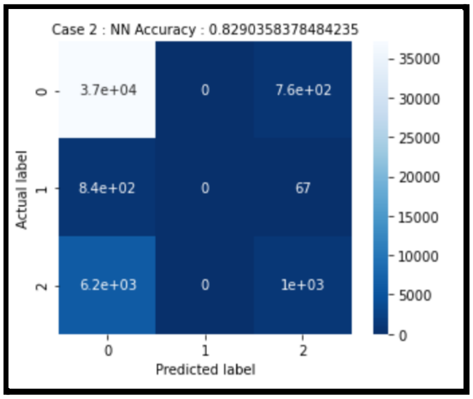


Fig 5.3.2 : Case (2) Correlation Accuracy

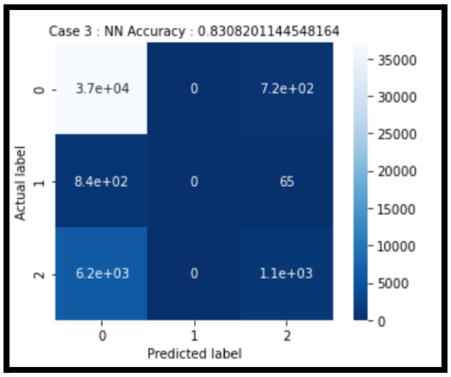


Fig 5.3.3 : Case (3) FFS Accuracy

Types	Multinomial Logistic Regression
Case 1	0.8250
Case 2	0.8252
Case 3	0.8286

Table 5.3 MLPCLassifier Accuracy Values

6. DISCUSSION

With the Result section we have come to the conclusion sections. To talk more about the results we can say that Decision Tree though not very accurate gave an approximate accuracy between 70 - 80 which with the help of Correlation and Feature Selection increases.

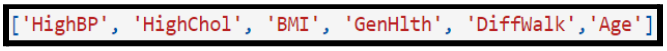
Since Decision Tree had a variation on every run we switched to Multinomial Logistic Regression and Logistic Regression was pretty stable giving almost the same value every time. Even on Multinomial Logistic Regression we could see the advantage of Correlation and Feature Selection. We later tried Neural Network to see if we can get any further improvements on our accuracy and pretty much as expected there wasn't much of a difference.

With this we can conclude that the project has successfully analyzed the major causes and invokes the detection of Diabetics in the United States.

7. FUTURE WORKS

The entire project is implemented considering that there are no outliers in the taken dataset. We have decided to implement the modeling of the data with the above mentioned supervised machine learning algorithms by handling the outliers as a part of future work. We are expecting the accuracy to be improved when these outliers are handled from the dataset.

For the features obtained by both the dimensionality reduction methods as shown in Fig 7.1, we have checked the presence of outliers by using the seaborn's boxplots as depicted in Fig 7.2. It is quite evident that the 'BMI' feature has the most outliers for about 5638 with the maximum and minimum values being 98 and 45 respectively as in Fig 7.3. The outliers are found using the InterQuartile Range technique as mentioned in [6].



```
['HighBP', 'HighChol', 'BMI', 'GenHlth', 'DiffWalk','Age']
```

Fig 7.1: Feature's list

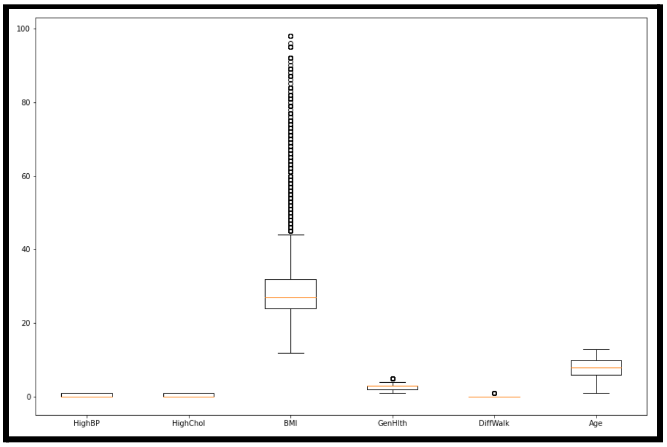
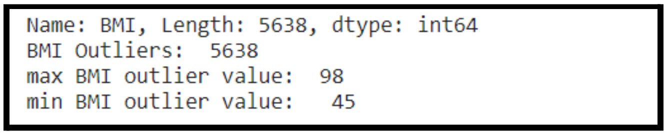


Fig 7.2. Boxplot for all the Features



```
Name: BMI, Length: 5638, dtype: int64
BMI Outliers:  5638
max BMI outlier value:  98
min BMI outlier value:   45
```

Fig 7.3. Outliers details for 'BMI'

8. RELATION TO PRIOR WORK

There have been many works done on diabetes prediction over the years.Most of the papers we had come across had mostly done for diabetes prediction on a whole. We have decided to improve upon it by just taking the United States into consideration.

1)Model for early prediction of diabetes:This paper was published in 2019 in Science Direct. In this paper the attribute selection has been performed by principal component analysis. K means clustering, ANN and Random Forest techniques were implemented for the prediction of diabetes.A relation was found between BMI and glucose level which was extracted by the Apriori method.

In our project we have performed feature selection by Forward feature selection and we have obtained correlations between attributes such as Education and high blood pressure and income and heart attack.

2) [5] This paper uses different classification algorithms such as back propagation, SVM ,j48 and Naive bayes.

Comparison with our project:We use different algorithms for predictions.

9. RELATED WORK AND REFERENCES.

There have been a lot of interesting papers done in this field.Although the core premise remains the same, there are various different approaches.Added the references to each number in the reference section.

[1]https://www.analyticsvidhya.com/blog/2021/04/forward-feature-selection-and-its-implementation/ this was used to study and implement Forward Feature Selection.

[2]https://scikit-learn.org/stable/modules/generated/sklearn.linear_model.LogisticRegression.html

[3]https://analyticsindiamag.com/a-beginners-guide-to-scikit-learns-mlpclassifier/ ,

[4]https://scikit-learn.org/stable/modules/generated/sklearn.neural_network.MLPClassifier.html - Used for sklearn MLPClassifer

[5]Fikirte Girma Woldemicheal, Sumitra Menria,”Prediction of Diabetes Using Data Mining Techniques’,*IEEE Xplore*,2018 International Conference

[6]https://towardsdev.com/outlier-detection-using-iqr-method-and-box-plot-in-python-82e1e15232bd to check the outliers in the "BMI" feature.

[7]https://www.kaggle.com/datasets/alexteboul/diabetes-health-indicators-dataset - Our Dataset for the project.

[8]https://www.kaggle.com/code/shohanursobuj/exploratory-data-analysis-eda-with-reports - Reference for our EDA.