University of Colorado **Boulder**

Diabetes Severity Prediction System Using ML Algorithms

Rohith Reddy Peddi
Bhuvanendra Putchala
Sai Srinivas Puranam
Sai Gowtham Dasappa Ravindra

Department of Data Science
University of Boulder
Colorado, USA

ABSTRACT

In the recent times, diabetes has become one of the most chronic diseases in the world. According to WHO 2014 report, around 422 million people worldwide are suffering from diabetes. If there is a chance for early diagnosis, then the risk factor of diabetes will be decreased significantly. There are lot of research work that had been done in this area and built some prediction models and provided the solution. But the existing methods gave low accuracy and consumed more time for dataset trainin. The main objective of this project is to predict the severity of diabetes with the best accuracy and precision for the dataset. Our proposed work consists of two phases: pre-processing and disease severity prediciton. In the first phase, pre-processing is performed for the BRFSS dataset and in the second phase ML algorithms are used for the disease severity prediction and the best accuracy has been taken into consideration. In this paper we are going to discuss about the brief analysis of the dataset and comparison of various ML algorithms and select which gives the best accuracy and predict whether the patient has diabetes and the severity of the disease.

TABLE OF CONTENTS

LIST OF FIGURES

LIST OF ABBREVATIONS

1. BMI: Body Mass Index
2. BP: Blood Pressure
3. ML: Machine Learning
4. EDA: Exploratory Data Analysis
5. FE: Feature Engineering
6. LR: Logistic Regression
7. DT: Decision Tree
8. RF: Random Forest
9. ROC: Receiver Operating Characteristic
10. AUC-ROC: Area Under the Receiver Operating Characteristic Curve

1. Introduction

Diabetes has been a problem for many years in the past which has been creating problems for daily on-going human life. Diabetes creates a long term metabolic problems like high blood sugar which is considered to be one of the threat for many individuals across the world . Diabetes doesn't create any impact on any individual until and unless it causes initial problems like fatigue or dizziness, later creates heart problems.

Diabetes could be treated properly if the disease is detected correctly on time. We can stop the immediate symptoms on individual by identifying individuals and modifying there treatment accordingly. By using machine learning and many more techniques in data mining we can identify the person who is at risk on time. Data mining is the process of mining large data from repositories to extract hidden relations and patterns within data. These insights can help us to make machine learning models which could be efficient in predicting the data. These machine learning models can help us extract hidden patterns which cannot be detected by humans with the help of abundant information available as open patient records. Severity of diabetes can be predicted by training machine learning models help extract useful information, which helps medical professional to treat in early stage(Cheng et al., 2020a). With the help of these results, we can identify diabetes and provide customized treatment for every individual.

We are investigating machine learning to build diabetes severity prediction models that correctly predict the intricate web of health factors that cause diabetes with the aid of CDC's BRFSS dataset. Furthermore, our research expands upon this earlier work. Utilizing machine learning algorithms like decision trees, logistic regression, random forests, and XG boost, we hope to develop a dependable and strong diabetes severity prediction system. Our mission is to create a reliable system that will empower medical professionals, point them in the direction of the best course of action, and ultimately provide better patient care.

This research shows how machine learning will enable early and preventive intervention in the future by helping medical professionals identify patients who are at high risk of serious complications(Liu & Wang, 2021). Lessening the effects of diabetes and any potential complications will improve patient well-being and quality of life. Personalized treatment programs that are based on each patient's needs and the course of the disease will maximize treatment outcomes. However, there are certain obstacles along the way. Obstacles include those related to data access, model improvement, and ethical considerations. Nevertheless, diabetes will no longer cast a shadow over our future—instead, it will highlight the creative and machine learning possibilities that lie ahead.

We are committed to transforming the diabetes care landscape, using technology and data to shift it from an inactive battlefield into a proactive space that promotes personalized management and health. The following sections covers: Related Work, Proposed Approach, Methodology, Conclusion and Future Scope.

1.1 Problem statement

Millions of individuals worldwide are impacted by diabetes, a silent pandemic that carries a long-term risk of complications and reduced quality of life. It is characterized by high blood sugar levels and is often not diagnosed until symptoms such as fatigue and cardiovascular disease appear. Although treatable, early detection is critical to prevent progression and its devastating consequences. Many companies around the world have researched this area, built predictive models, and implemented solutions. However, existing methods had low accuracy and required additional time to train the dataset.

Our research revolves around this major concern: "Can we develop a highly precise, accurate and reliable model to detect diabetes severity at early stages?"

1.2 Research Aim and Objectives

Our research aims and Objectives:

1. This research addresses the critical need for an accurate and proactive diabetes severity prediction system.

2. We aim to leverage the power of machine learning and analyze intricate relationships within real-world data i.e., utilizing CDC's BRFSS survey data to unveil hidden patterns and risk factors.

3. Through advanced algorithms and models, we seek to predict diabetes severity with remarkable accuracy, empowering healthcare professionals to:

- Identify high-risk individuals before complications arise.
- Design personalized treatment plans tailored to individual needs and disease progression.
- Implement proactive interventions to mitigate diabetic effects and improve patient well-being.

2. Literature Review

2.1 Related work

In the recent years many researchers are focusing on machine learning algorithms (Zhang et al., 2022)in the healthcare industry. The use of ML algorithms has proven a significant change in the performance of the treatment(Al-Janabi et al., 2021a). In this section we are going to discuss about the related for diabetes severity, and the techniques that can be diagnosed using machine learning algorithms.

Although accurately predicting diabetes severity remains challenging. (Rodríguez et al., 2016a)conducted a review that found some limitations such as small sample sizes, data imbalance. Overcoming these challenges is essential, for improving the reliability and practicality of machine learning models in predicting diabetes severity within settings.

Feature selection is a critical component of model deployment, which has been rigorously evaluated in terms of diabetes severity prediction, understanding. In their research work, they have compared different methods of selecting the features for analysis which includes feature elimination(Warraich et al., 2019a). This has proven that it is important to understand how feature selection impacts the model performance.

Another research paper suggests the importance of data pre-processing techniques such as data normalization and shows the improvement of accuracy in of severity in diabetes prediction models. (Patterson et al., 2018)This study has showed the potential of data pre-processing in improving the performance of the model.

Selecting the best ML algorithm is also a crucial factor in predicting the severity of diabetes. For instance, (Wang et al., 2017)has compared different models such as SVM, decision tree and random forest, neural networks in diabetes progression. From the above inferences we can say that understanding the differences between these algorithms is crucial when selecting the method for predicting diabetes severity.

In a recent study, it suggests that it can predict diabetes mellitus and its severity level using machine learning machine learning techniques(Zohair et al., 2023a). This model uses multi-class classification, and it distinguishes difference between diabetic patients. In this paper they have used a technique called model fusion by combining logistic regression, decision tree.

Although certain studies have shown precision it's essential to consider the aspects of each study such, as the tasks, data characteristics and evaluation metrics used, in order to make a meaningful comparison. A significant hurdle in this field is the nature of data sources and prediction tasks which makes it challenging to compare different models. Additionally, more research and development are needed to explore factors, like model interpretability and their ability to generalize across populations.

2.2 Proposed work

We have conducted quantitative research in this case. Since our goal is to predict the severity of the disease, we have used quantitative method. In our work, the focus is on determining the patient's severity on the diabetes disease, along with the presence and absence of it.

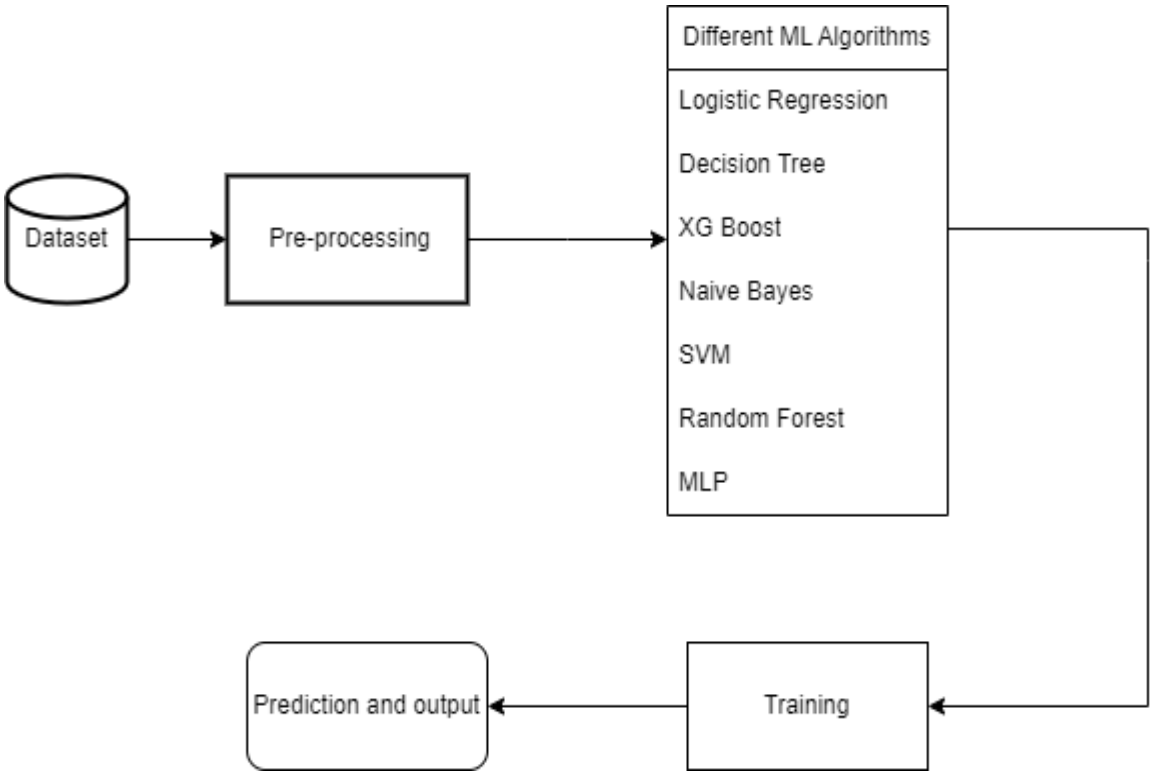

Fig 1: Proposed architecture

We utilized machine learning algorithms. The selection of these algorithms was based on their performance with a focus, on those that displayed levels of accuracy. Initially we performed pre-processing on the data, which involved cleaning it up by addressing missing values and implementing feature scaling(Rodríguez et al., 2016b). After that the dataset had imbalances for class classification, so we employed an under-sampling technique to balance it out. Following this step, we divided the data into test and train sets. Then we proceeded to train each algorithm on the dataset.

3. Data Exploration and Analysis

3.1 Data pre-processing

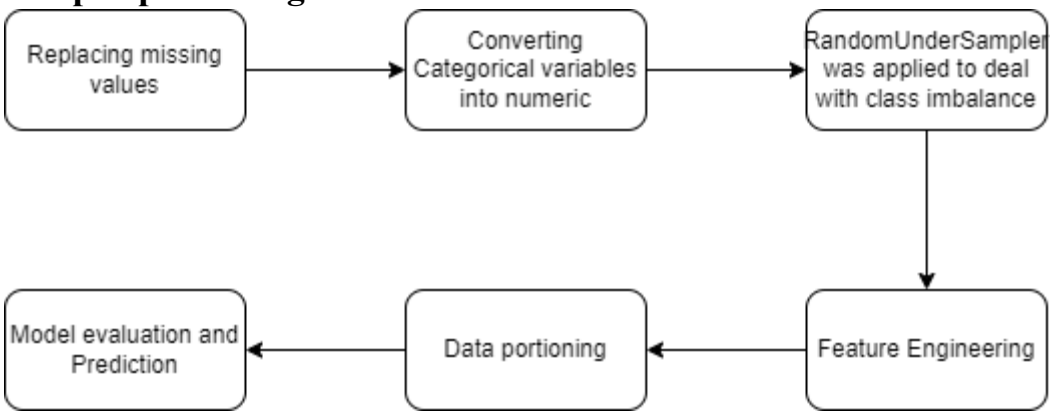

Fig 2: Pre-processing stages

3.1.1 Data cleaning

The null values in the column sleeping time are being replaced by the mean of the sleeping time column. Number of days physical health was not good, and number of days mental health was not good had null values which are replaced by the mean of those two columns. More than 25 days values are separated into very unhealthy group and below 5 days are grouped into very healthy. Columns like smoking, deaf/ blind, had stroke , asthma , had cancer etc were having null values , not sure , refused to say and these are converted to not sure value.

3.1.2 Data Transformation

Our initial data which we've taken from CDS website was in XPT format we have transformed the XPT data to csv format using 'Hmisc' package from r environment. Data related to height and weight are in various metrics (example: kilo grams, pounds, lbs , meters , feet etc.) . We transformed the data into centimetres and weight in kilograms for better analysis purposes. Body mass index column which is called BMI was modified based on the updated values of height and weight of each person. New column was created and added to the data frame.

We have collected geospatial data of united states of America's states, and we have merged into main data frame of latitudes and longitudes data based on state values in main data frame by creating two new columns.

3.1.3 Handling Categorical Data

Majority of our data had values which are in categorical format and had numerical values. We used a schema that had details about data description for all of these categories.

Columns like how was your general health, how was your physical health had range of numerical values from 1 to 5 , so we have changed these values ranging from excellent to poor . Health care access, deaf/blind ,exercise done in last month , are you pregnant , did you had any heart stroke in last one month , do you have asthma , have you had any cancer , have

you had any depressive disorder or are you having any depressive disorder columns had numerical values 1 and 2 which we have modified based on schema notebook to Yes or No for better understanding and for better processing .

Did you have any insurance plan column had numerical values from 1 to 10 and these are modified to ('employer', 'private', 'Medicare', 'Medigap', 'Medicaid', 'Children´s Health Insurance Program (CHIP)', 'Military related health care: TRICARE (CHAMPUS) / VA health care / CHAMP- VA', 'Indian Health Service', 'State sponsored health plan', 'Other government program').

When did you last visit doctor had values from 1 to 4 range 1 year, 1 to 2 years , 2 to 5 years and 5 plus years .

Income column was modified and grouped to range ('Less than 10,000', '10,000_to_15,000', '15,000_to_20,000', '20,000_to_25,000', '25,000_to_35,000', '35,000_to_50,000', '50,000_to_75,000', '75,000_to_100,000', '100,000_to_150,000', '150,000_to_200,000').

Finally diabetic column in the dataset which was used as target column was modified to ('Yes', 'Yes_during_pregnancy', 'No', 'pre_diabetes_or_boderline') for better understanding

3.1.4 Data Imputation

Height column in the data contained null values and these null values are filled using the mean value of the height column. Similarly, weight columns also has null values and these are also filled using the mean value.Sleeping time which describes the number of hours a person sleeps in a day had values like”77”, “99” which meant “Don’t know/Not Sure” and “Refused”. These values are replaced with null values and are further dropped as they had very low count in our data set.

In other columns we had values like Don’t know/Not Sure and Refused as categorical values, so we have merged these values based on other values’ mode in columns after analysing using other columns and these values had very low frequency in data set.

3.2 Exploratory Data Analysis (EDA)

We analysed number of people in various categories in diabetes column. We came to know that people who are not having diabetes are female than male. Male and female suffering from diabetes is compared together and we came to know that there is no huge difference between count to female and male. We came to know that female population are getting infected during their pregnancy. Also, there are almost equal amount of female and male those are detected in prediabetes.

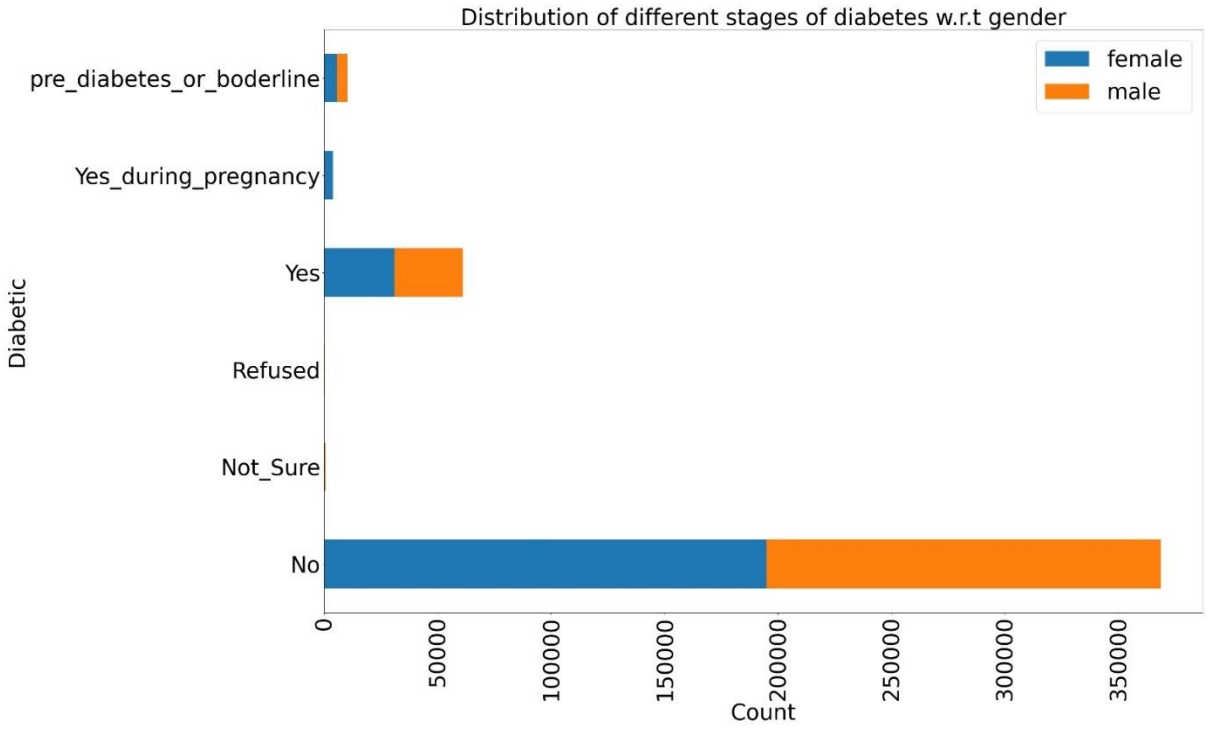

Fig 3: Distribution of different stages of diabetes with respective to gender

Secondly, we analysed the age group that are being infected. We came to know that people who are in age groups of 50 to 59 are not subjected to diabetes. 70 to 79 age group are the highest number of people that are being affected with diabetes. Age group of 18 to 29 are the having diabetes in their border stage. Age group of 30 to 39 are the most that are being affected by diabetes during their pregnancy.

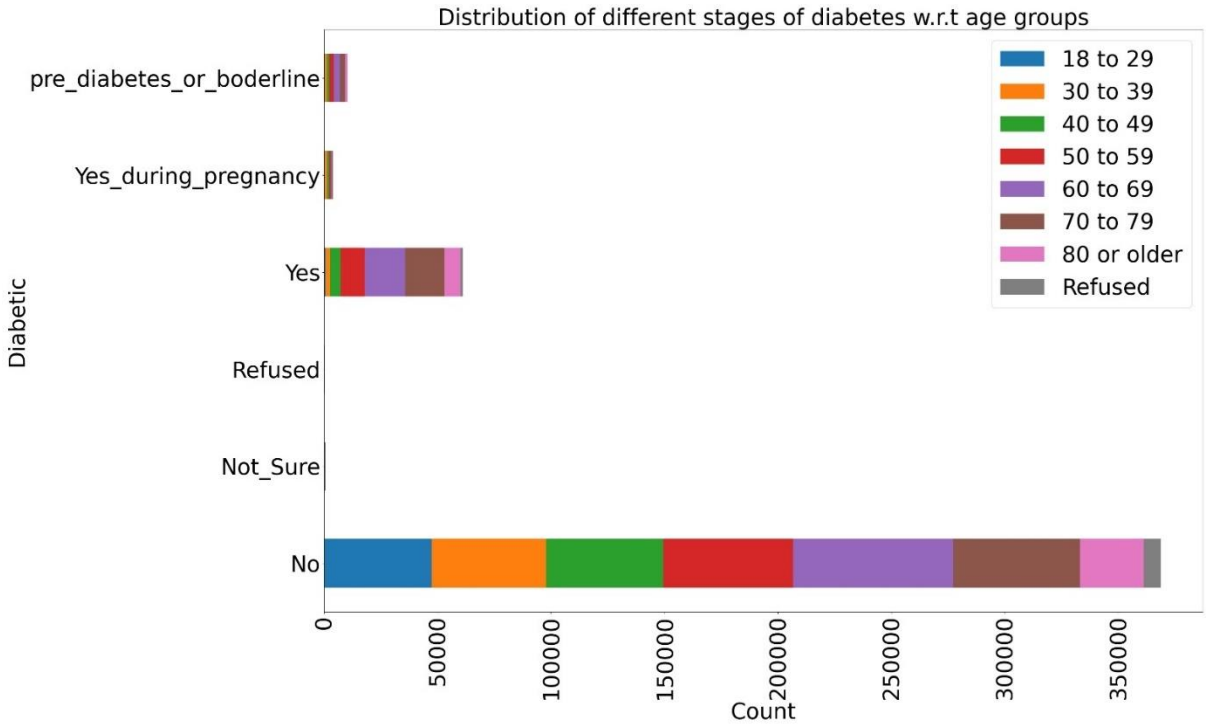

Fig 4: Distribution of different stages of diabetes with respective to age groups

Later we have analysed mean number of days people's physical health was not good and their mental health was not good. We came to know that people who are not sure whether they are having diabetes or not are having more than 7 days that their mental health was not good and more that 6 days their mental health was not good. People who are having diabetes are having a greater number of unhealthy physical days than their mental days before they were detected with diabetes. People who were detected pregnant are having similar number of unhealthy physical and mental days. People who are in border line of diabetes are having almost similar number of unhealthy days.

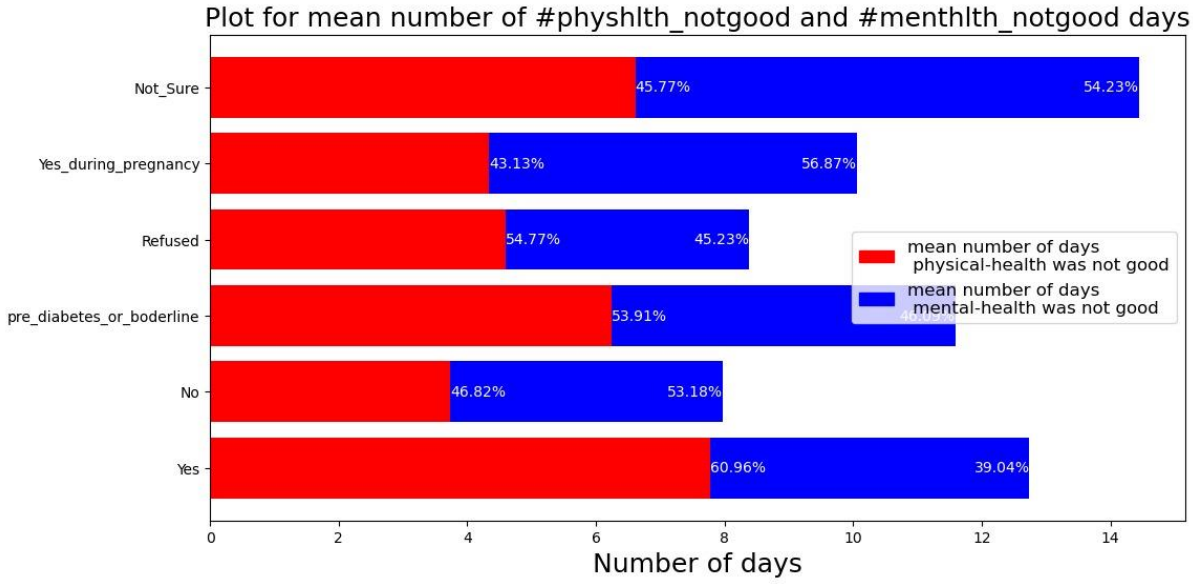

Fig 5: Plot for mean number of physical health not good and mental health not good days

We have analysed trends in health care access by income groups. We got to know that for up to 22500 incomes rage the count of number of people who can afford medical health care is rising and from 22500 income rage till 42500 income group the count of people who can afford rose dramatically. Later the count of number of people count dropped from 62500 till 175000 and after that group the line got stabilized. For people who can't afford medical health care the count kept increasing till income of 3000 and later it dropped slowly till 200000 income group.

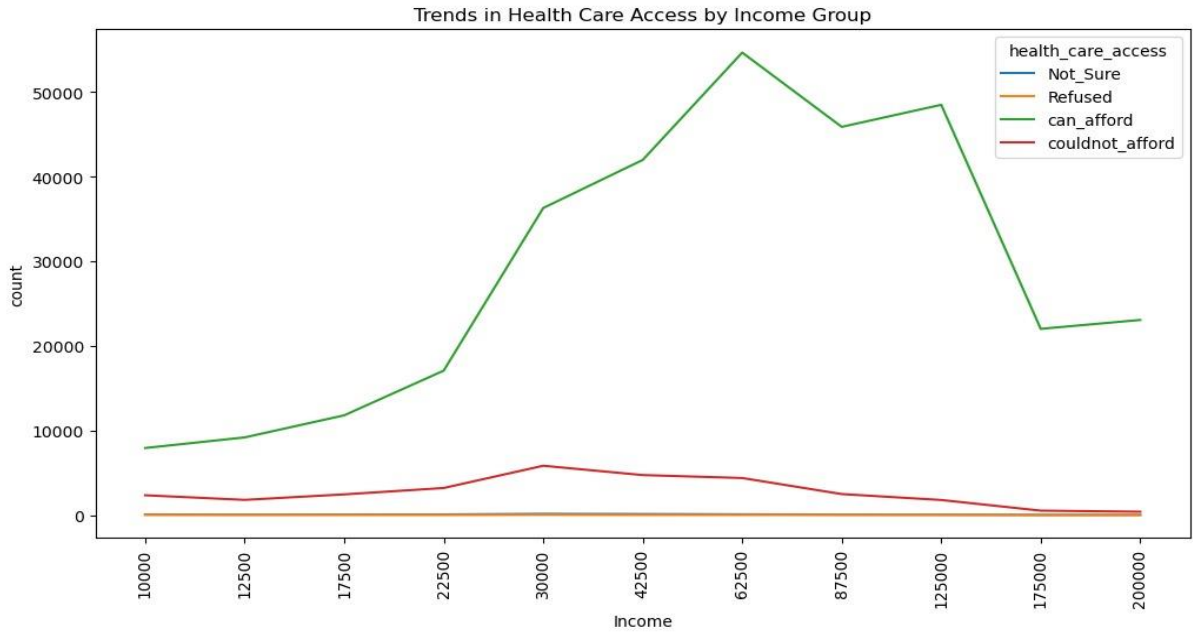

Fig 6: Line plot for trends in health care access by income group

Using geospatial analysis, we came to know that the number of people who are more effected with diabetes are on east coast compared to west coast.

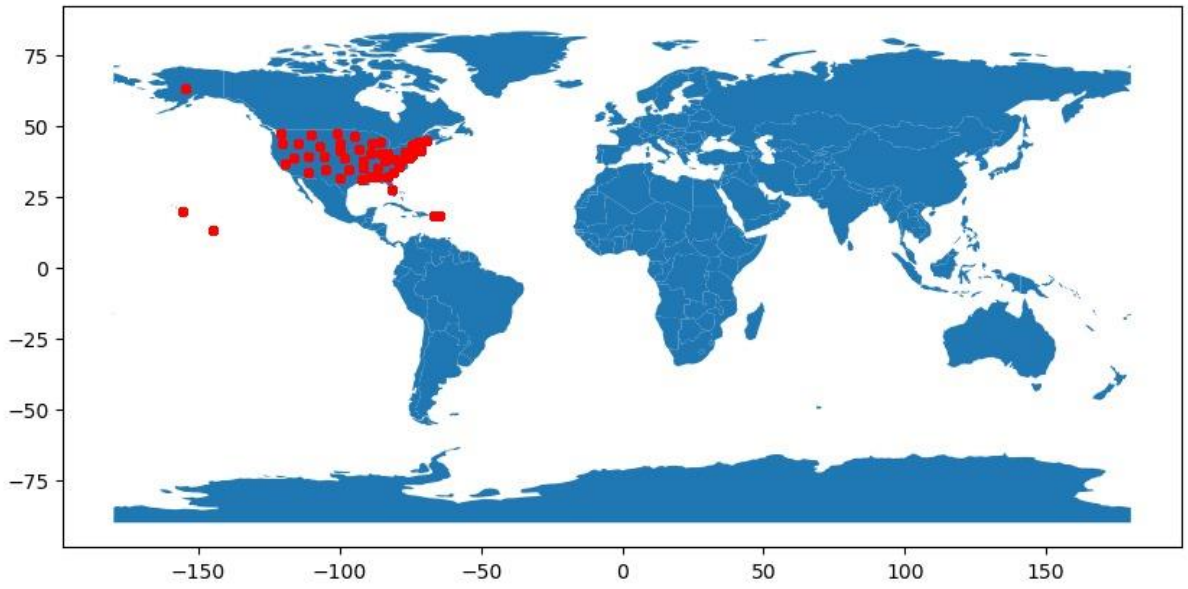

Fig 7: Geospatial analysis for dataset

3.3 Hypothesis Testing

We wanted to do perform some test using statistical methods which is used to make inferences about population parameters based on a sample of data. The process involves formulating a null hypothesis (H0) and an alternative hypothesis (H1), collecting and analysing data, and drawing conclusions about the population based on the observed results. After going through the Exploratory Data analysis, we derived two hypothesis and tested out those hypotheses using ANOVA test.

Hypothesis1:

H0: Null hypothesis is there is no significant difference in average sleep duration between individuals with and without diabetes.

H1: Alternate hypothesis is Individuals with diabetes have a significantly different average sleep duration compared to those without diabetes.

Hypothesis2:

H0: Null hypothesis is there is no significant association between having diabetes and reporting a certain number of mentally unhealthy days.

H1: Alternate hypothesis is the number of mentally unhealthy days is significantly different between individuals with and without diabetes.

Using ANOVA test we tested the above hypothesis and for hypothesis 1 we got the p-value> 0.05 which resembles that there is no significant difference between sleeping time and diabetic while for hypothesis 2 we got a p-value<0.05 which resemble that there is a significant difference between mentally unhealthy days with people with and without diabetes.

4. Methodology

4.1 Predictive data mining

Generally, when we train machine learning models on imbalanced dataset, the models may overfit and tend to learn on majority class only while leaving behind or struggle to identify trends from the minority class as the data is less for minority class. One of the techniques to handle imbalanced datasets are Undersampling and Oversampling. With these techniques we can equalize the datapoints in both classes so that models would recognize/learn perfectly and there would not be any bias/ class imbalance.

In order to use Under sampling we need to make sure we have high rate of class imbalance i.e in the order of 80-20 ration because under sampling may drastically reduce the size of the data and we end up in information loss(Zohair et al., 2023b). In our case we had almost 85-15 ratio and we used the under-sampling method where we got 30k as the data shape for our model training.

While under sampling has its benefits it is important to consider the loss of information and its impact, on the generalization of the model. Under sampling is one of strategies used to address class imbalance, in predictive modelling, so to achieve better accuracy from classification methods and to reduce the overfitting issues with the imbalance we used random sampling, SMOTE to overcome this issue and achieved better results.

4.2 Modelling

4.2.1 Logistic Regression

Logistic regression is frequently known as Generalized Linear model, it is a statistical modelling technique which is mainly used for binary classification and probability estimation problems. It is similar to linear regression, but it predicts True or False, type of problems. It predicts the probability that an observation belongs to a particular category. Logistic regression can handle both continuous and categorical data(Warraich et al., 2019b). Unlike regression, which fits a line using squares logistic regression relies on maximum likelihood estimation. By calculating the likelihood, for curves logistic regression selects the curve with the likelihood to perform classification tasks. 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(Balpande & Wajgi, 2017). Logistic regression model uses sigmoid function to predict probability of positive and negative class(Cheng et al., 2020b).

$$ln(p\ /\ (1\ -\ p))\ =\ \beta_0\ +\ \beta_1 X_1\ +\ \beta_2 X_2 + \ldots + \beta_{\mathrm{k}} X_{\mathrm{k}}$$

$$[\backslash\mathrm{logit}(p) = \ln\left(\frac{p}{1-p}\right) = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \cdots + \beta_n x_n]$$

Here, ln represents the natural logarithm, and p is the probability of the event occurring.

For instance, the Trauma and Injury Severity Score (TRISS) is a used method to estimate the chances of mortality, in patients with injuries. Initially logistic regression was employed to

develop this score. Logistic regression has also been utilized in fields, including engineering, where it helps determine the likelihood of process or product failure. In marketing it aids in predicting customer behaviours like purchasing tendencies or subscription cancellations. Additionally logistic regression finds application in economics to forecast labour force participation and in business, for predicting mortgage default rates among homeowners.

4.2.2 Decision Tree

Decision is a hierarchical type of structure which consists of root and leaf node. Leaf nodes are the final variables that are defined as final class labels that shows outcome of the problem. Decision trees have evolved from these algorithms like ID3, CART and random forest. These algorithms work on the criteria such as information gain and gini index based on feature nodes. Mainly, ID3 algorithm is used for construction of decision tree. This algorithm tells that the first step of creating a decision tree is to select the right attributes for splitting of tree(Al-Janabi et al., 2021b). Splitting of tree can be done by measuring the information gain. It calculates the total entropy value from the top to bottom node of the tree(Aljumah et al., 2013). If the information gain value is higher for a attribute then that attribute is used for construction of decision tree.

Hyper parameter tuning is necessary for decision tree to improve its overall performance. In our proposed model(Kibria & Matin, 2022) we have tuned various parameters, at last we got the set, which has maximum depth and also have maximum features.

4.2.3 Random Forest

Random Forest algorithm can be used for both classifying data and for regression analysis. Random forest uses multiple decisions trees and creates one optimized model. This algorithm is very helpful for large data sets and predicts data with better accuracy compared to other models(Kavakiotis et al., 2017). Training process goes on by constructing multiple decision tress and predicts results based on most common predictions from all other trees.

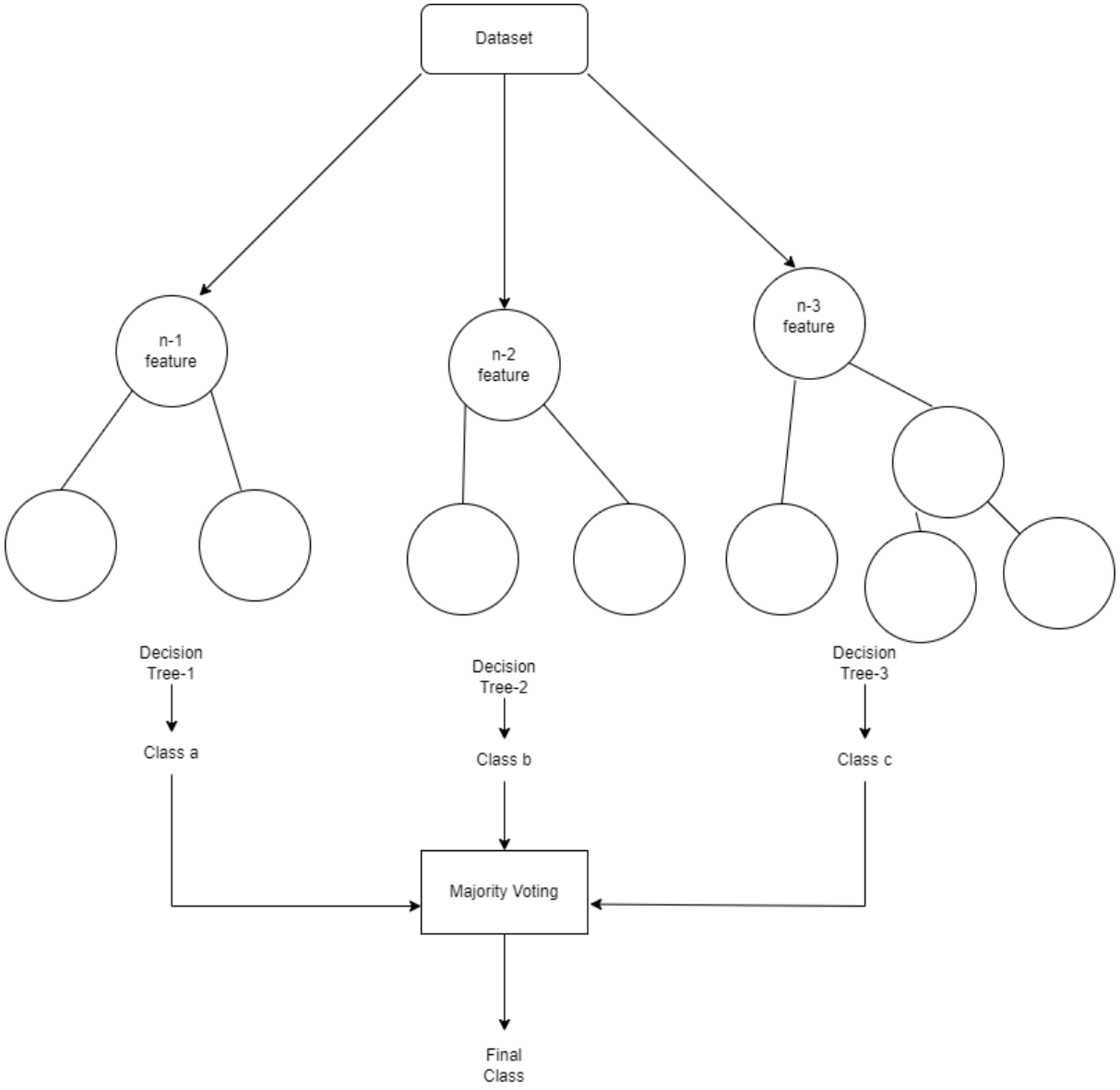

Fig8: Representation of Random Forest algorithm

Hyper parameter tuning can be very affective when machine learning models overfit on a given data set. Overfitting of data usually happens when the model relies more on training data an provides poor accuracy score on testing set. When model overfits model doesn't generalize well on all instances(Yogapriya & Dhanalakshmi, 2023). We have adjusted hyper parameters like max depth , max features for the model to better operate. Using these parameters we can make trees complex accordingly as more deep models tend to overfit again.

4.2.4 XG Boost

It's called extreme Gradient Boosting which is very powerful efficient technique which was widely used in area of machine learning. XGBoost is based on boosting technique where multiple weak learners are merged together to form strong learner. Various models will be applied on data one after another in sequential way where the error made by the previous model will be improved in the next model. XGBoost model has an objective function that contains regularization term and loose function which helps model to prevent overfitting and for minimizing errors(Gurka et al., 2017). Regularization term also helps to control the complexity of each and every tree's complexity. XGBoost assigns scores for every leaf in a decision tree other than class labels. Scores are used as strength for every instances. XGBoost allows for early stopping while training for model.

$$\text{Objective}(\Theta) = \sum_{i=1}^{n} l(y_i, \hat{y_i}) + \sum_{k=1}^{K} \Omega(f_k).$$

- Θ represents the set of model parameters.
- n is the number of training instances.
- $\ell(y_i, \hat{y}_i)$ is described as an individual instance loss function measuring differences between true labels and predicted labels for each instance.
- K represents the number of trees in an ensemble.
- $\Omega(f_k)$ is a regularization term for each tree in an ensemble.

$$l(y_i, \hat{y_i}) = -[y_i \log(\hat{p_i}) + (1 - y_i) \log(1 - \hat{p_i})]$$

where Pi is the predicted that i nth instance belongs to positive class.

4.2.5 Hyperparameter tuning:

Hyperparameters are the arguments given to the model to control output behaviour, when the arguments are not given model assumes default values and provides the results(Narasimha & Dhanalakshmi, 2022). This may cause wrong results or poor results as the model is not appropriately trained on the data. Hyperparameter tuning is done in the following way using below 2 popular methods:

1. Select the model on which the data needs to be trained on.
2. Define a set of arguments for the particular model as a parameter grid.
3. Pick a method for searching best parameters from grid.
4. Select cross validation which is nothing but let's say we decided 4 fold cv, then the data gets divided into 4 groups out of which one would be hold out for testing this happens for 4 iterations and average metrics of all 4 iterations is used to analyse the model.

GridSearchCV:

```
params = {
    'max_depth': [2, 3, 5, 10, 20],
    'min_samples_leaf': [5, 10, 20, 50, 100],
    'criterion': ["gini", "entropy"]
}

grid_search = GridSearchCV(estimator=DecisionTreeClassifier(),
                           param_grid=params,
                           cv=4, n_jobs=-1, verbose=1, scoring = "accuracy")
grid_search.fit(X_train, y_train)
clf_dec = grid_search.best_estimator_
```

Fig 9: Code snippet for GridSearch CV

GridsearchCV is a basic method to perform parameter tuning, it considers each and every combination of the parameters from parameter grid and builds the model. From these iterations best performing model is selected. As it fits the model for each permutation this method is a computationally expensive one when dealt with higher sizes of parameter grids.

RandomSearchCV:

```
params = {
    'n_estimators': [5,20,50,100],
    'max_features': ['auto', 'sqrt'],
    'max_depth': [int(x) for x in np.linspace(10, 120, num = 12)],
    'min_samples_split': [2, 6, 10],
    'min_samples_leaf': [1, 3, 4],
    'bootstrap': [True, False]
}

rf_random = RandomizedSearchCV(estimator =  RandomForestClassifier(),param_distributions = params,
       n_iter = 100, cv = 5, verbose=2, random_state=35, n_jobs = -1)
```

Fig 10: Code snippet for RandomSearchCV

In order to overcome the computation problem from GridsearchCV we use RandomSearchCV method for bigger parameter grids. Instead of fitting the model with all possible combinations of hyper parameters it selects a sample based on the specified distributions. In this way the time computation reduces and would expect faster results.

4.3 Model Building

This is the important phase of the project which involves building model for predicting the severity of diabetes(Tsujimoto et al., 2015). In this we have implemented various ML algorithms which are discussed in the above sections.

Procedure for model building:

Step 1: Import necessary libraries and dataset.

Step 2: Do Pre-processing.

Step 3: Perform data partitioning with 80 percent for training and 20 percent for testing set.

Step 4: Select the ML algorithm.

Step 5: Build the model based on training set.

Step 6: Test the model based on testing set.

Step 7: Perform evaluation using confusion matrix and classification report.

Step 8: Perform hyper-parameter tuning.

Step 9: After analysing various metrics given by various algorithms, finalize the best fit algorithm.

5. Results and discussion

We have collated all the model evaluation metrics in a tabular format shown above. We performed 3 iterations of model training and found following insights and learnings from the models.

1. Pre – Sampling:

Initially we trained all the four models on the entire dataset and checked for the performance metrics and we came to know that because of class imbalance we saw there is huge bias in the model and the model is overfitting on the data, due to which the metrics like precision, f1 score and recall came out to be very low(Bodapati & Balaji, 2023). In order to overcome overfitting we used sampling techniques like random sampling, SMOTE etc,.

2. Post – Sampling:

To overcome the bias problem we sampled the data and with sampling there were interesting results. As we have shown there was great improvement in the metrics precision, f1 score and recall that means we removed the overfitting effect, but there was a dip in the accuracy of the model. We applied hyperparamter tuning on top of the sampled data to check the performance metrics.

3. Post – Hyperparamater tuning:

For each of the 4 models we applied gridseachCV and RandomSearchCV (parameter tuning techniques) to extract best set of parameters for the data. Post getting the best parameters from models we trained and tested our models using these parameters and checked again for accuracy and other metrics, we saw that there is slight increase in the accuracy which we lost in the earlier step. Metrics like precision, f1 score and recall are also doing better.

The comparison table is as follows:

Algorithm used	Sampling technique	Accuracy	Precision	F1 Score	Recall
Decision tree	Pre-Sampling	75.62%	30.17%	31.13%	32.25%
	Post-Sampling	62.49%	59.35%	59.18%	59.05%
	Post Hyper-Parameter Tuning	70.53%	67.89%	68.09%	68.30%
Logistic Regression	Pre-Sampling	83.38%	56.49%	21.45%	13.24%
	Post-Sampling	71.80%	69.69%	69.12%	68.56%
	Post Hyper-Parameter Tuning	71.74%	69.66%	69.04%	68.43%
Random Forest	Pre-Sampling	82.80%	50.10%	20.88%	13.19%
	Post-Sampling	70.76%	68.13%	68.34%	71.82%
	Post Hyper-Parameter Tuning	69.16%	69.16%	69.58%	70.01%
XGBoost	Pre-Sampling	83.48%	51.60%	25.41%	16.43%
	Post-Sampling	71.79%	69.01%	69.65%	70.30%
	Post Hyper-Parameter Tuning	72.21%	69.41%	70.13%	70.86%

Table 1: Represents about comparison of classification reports of different ML algorithms.

From this exercise we can say that XGBoost is performing better in all aspects when compared to other models like Random Forest, Decision tree and logistic regression.

Confusion matrix:
Post – Sampling Hyper parameter tuning (XGboost):

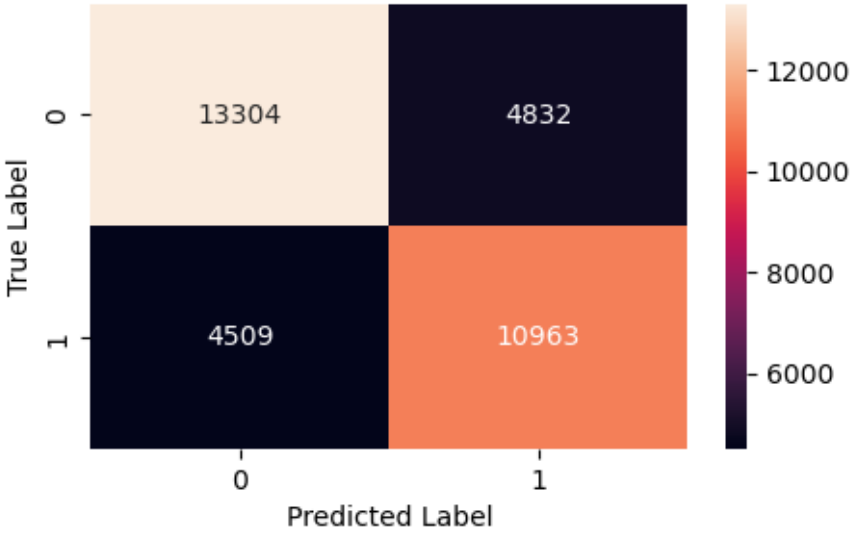

Fig 11: Confusion matrix for XG Boost

Post – Sampling Hyper parameter tuning (Decision tree):

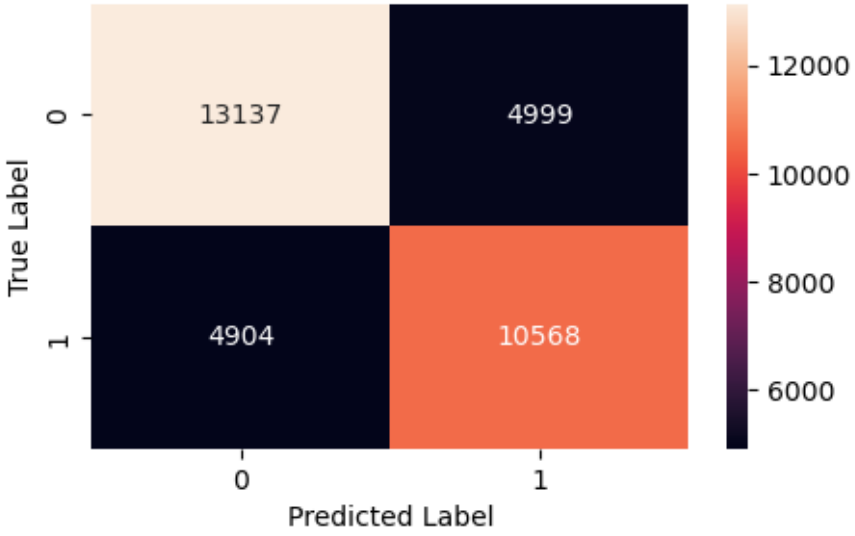

Fig 12: Confusion matrix for Decision tree

Post – Sampling Hyper parameter tuning (Random forest):

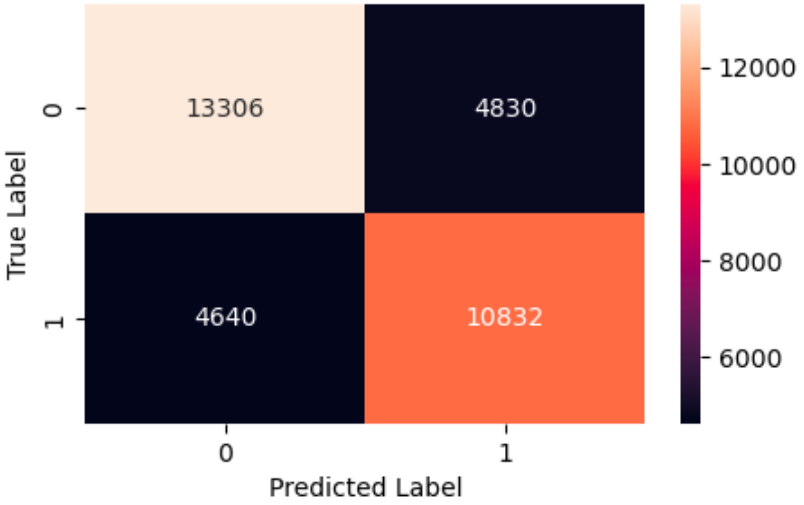

Fig 13: Confusion matrix for Random forest

Post – Sampling Hyper parameter tuning(Logistic regression):

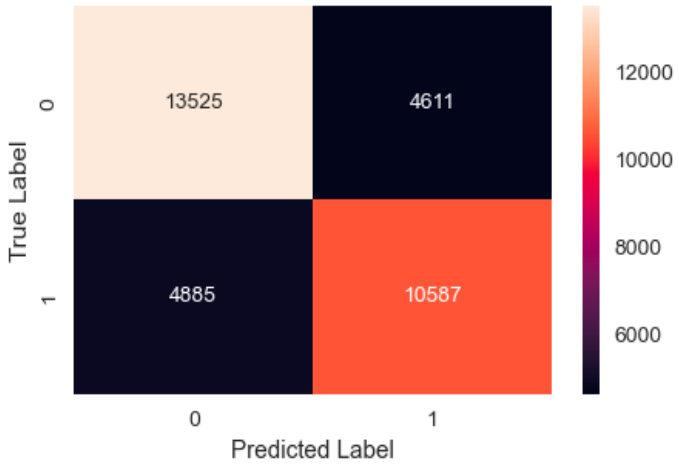

Fig 14: Confusion matrix for Logistic regression

AUC – ROC:
AUC – ROC curve is nothing but one of the measurement metric for the classification problem. ROC curve is plotted using true positive rate(sensitivity) and false positive rate(1-specificity) with true positive rate on y axis. This plot helps us in understanding the performance of the model at various thresholds(threshold is nothing but a value which classifies the labels). AUC is nothing but the area under the roc curve which ranges from 0-1. From our results using 3 different models xgboost, Randomforest and Decision tree we say that Xgboost and Raandomforest performs well and is giving best performance at different points of false positive rates while Decision Tree has lower performance but as the false positive rate increases gradually true positive rate is also increasing.

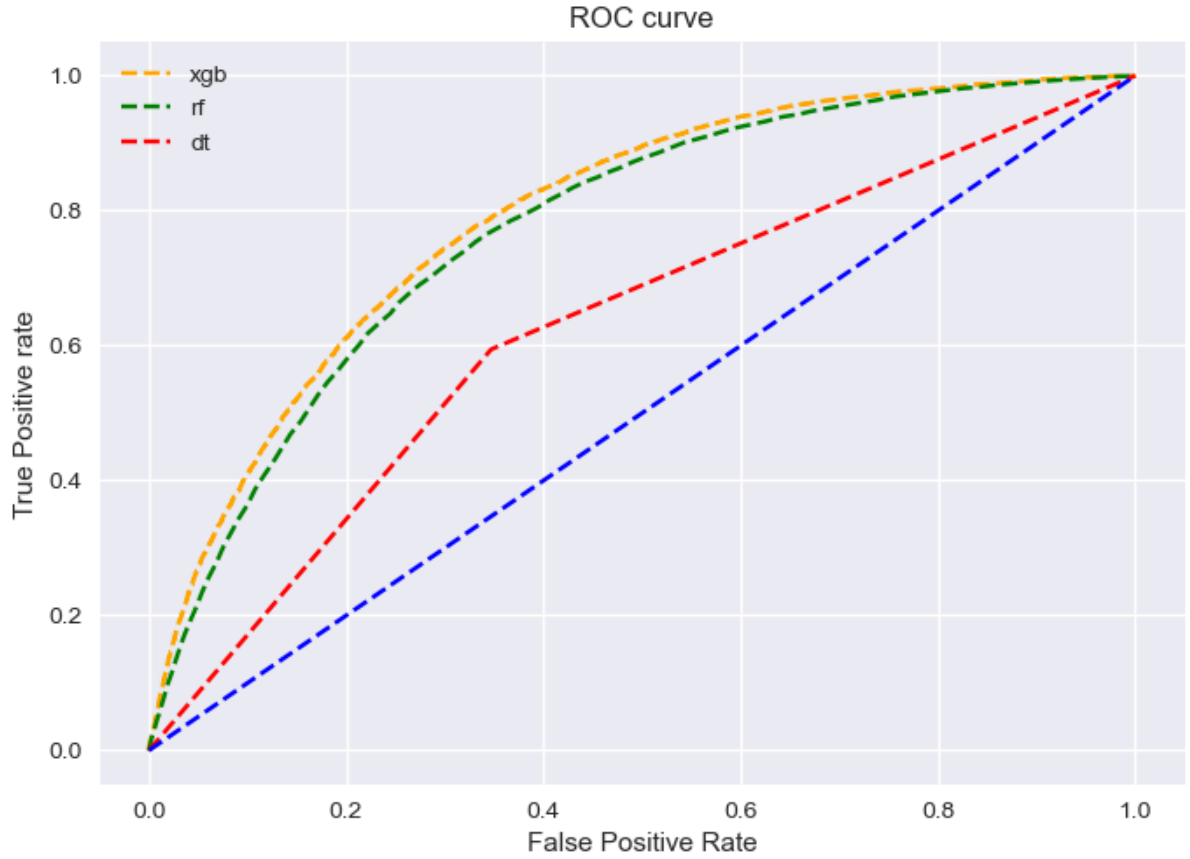

Fig 15: Shows ROC curve for implemented ML algorithms

6. Conclusion

The overall project aims to find the relationship or hidden patterns between the health conditions of an individual with respect to diabetes risk using the CDC's BRFSS survey data. An attempt is made to use the knowledge of predictive machine learning classification algorithms to predict the risk of diabetes using the health and mental conditions of individuals. We were able to achieve an accuracy of ~72% using XGboost algorithm, by this we can say that having the physical and mental health conditions of a person we could potentially predict the chance of diabetic attack.

7. Limitations and Future Scope

Limitations:

As the data was taken from a survey there could be a chance that we might not capture the full information with respect to a person as individuals may not always provide appropriate information.

The data was huge and in the raw format, it took us sometime to understand the data and do appropriate data cleaning tasks to extract quality data from it.

As the problem was on predicting the chronic disease and it is known that lot of people may not have that disease/symptom, which lead to class imbalance.

The data looks mostly related to health conditions, but incorporating SDOH (socio-economic determinants of health) data could support the analysis and predictions better.

Future work:

We extend our work in collecting the SDOH data and incorporating it to our present data and look for the trends and predictions.

This work would be extended to include potentially complex ML algortihms to improve the performance metrics.

Automating the entire process using MLOps in a way that as on when we get updated data. We would just upload the data to existing data frame, retrain models and refresh the model scores.

Further, this work is going to support in developing a website where user answers a survey on various questions related to health factors so that the data collection is automated and yielding quicker results.

8. References

Al-Janabi, A., Littlewood, Z., Griffiths, C. E. M., Hunter, H. J. A., Chinoy, H., Moriarty, C., Yiu, Z. Z. N., & Warren, R. B. (2021a). Antibody responses to single-dose SARS-CoV-2 vaccination in patients receiving immunomodulators for immune-mediated inflammatory disease. *British Journal of Dermatology*, *185*(3), 646–648. https://doi.org/10.1111/bjd.20479

Al-Janabi, A., Littlewood, Z., Griffiths, C. E. M., Hunter, H. J. A., Chinoy, H., Moriarty, C., Yiu, Z. Z. N., & Warren, R. B. (2021b). Antibody responses to single-dose SARS-CoV-2 vaccination in patients receiving immunomodulators for immune-mediated inflammatory disease. *British Journal of Dermatology*, *185*(3), 646–648. https://doi.org/10.1111/bjd.20479

Aljumah, A. A., Ahamad, M. G., & Siddiqui, M. K. (2013). Application of data mining: Diabetes health care in young and old patients. *Journal of King Saud University - Computer and Information Sciences*, *25*(2), 127–136. https://doi.org/10.1016/j.jksuci.2012.10.003

Balpande, V. R., & Wajgi, R. D. (2017). Prediction and severity estimation of diabetes using data mining technique. *2017 International Conference on Innovative Mechanisms for Industry Applications (ICIMIA)*, 576–580. https://doi.org/10.1109/ICIMIA.2017.7975526

Bodapati, J. D., & Balaji, B. B. (2023). Self-adaptive stacking ensemble approach with attention based deep neural network models for diabetic retinopathy severity prediction. *Multimedia Tools and Applications*. https://doi.org/10.1007/s11042-023-15120-7

Cheng, H.-Y., Jian, S.-W., Liu, D.-P., Ng, T.-C., Huang, W.-T., & Lin, H.-H. (2020a). Contact Tracing Assessment of COVID-19 Transmission Dynamics in Taiwan and Risk at Different Exposure Periods Before and After Symptom Onset. *JAMA Internal Medicine*, *180*(9), 1156. https://doi.org/10.1001/jamainternmed.2020.2020

Cheng, H.-Y., Jian, S.-W., Liu, D.-P., Ng, T.-C., Huang, W.-T., & Lin, H.-H. (2020b). Contact Tracing Assessment of COVID-19 Transmission Dynamics in Taiwan and Risk at Different Exposure Periods Before and After Symptom Onset. *JAMA Internal Medicine*, *180*(9), 1156. https://doi.org/10.1001/jamainternmed.2020.2020

Gurka, M. J., Golden, S. H., Musani, S. K., Sims, M., Vishnu, A., Guo, Y., Cardel, M., Pearson, T. A., & DeBoer, M. D. (2017). Independent associations between a metabolic syndrome severity score and future diabetes by sex and race: the Atherosclerosis Risk In Communities Study and Jackson Heart Study. *Diabetologia*, *60*(7), 1261–1270. https://doi.org/10.1007/s00125-017-4267-6

Kavakiotis, I., Tsave, O., Salifoglou, A., Maglaveras, N., Vlahavas, I., & Chouvarda, I. (2017). Machine Learning and Data Mining Methods in Diabetes Research. *Computational and Structural Biotechnology Journal*, *15*, 104–116. https://doi.org/10.1016/j.csbj.2016.12.005

Kibria, H. B., & Matin, A. (2022). The severity prediction of the binary and multi-class cardiovascular disease − A machine learning-based fusion approach. *Computational Biology and Chemistry*, *98*, 107672. https://doi.org/10.1016/j.compbiolchem.2022.107672

Liu, Q., & Wang, Z. (2021). Perceived stress of the COVID-19 pandemic and adolescents' depression symptoms: The moderating role of character strengths. *Personality and Individual Differences*, *182*, 111062. https://doi.org/10.1016/j.paid.2021.111062

Narasimha, V., & Dhanalakshmi, M. (2022). *Severity and Risk Predictions of Diabetes on COVID-19 Using Machine Learning Techniques* (pp. 195–208). https://doi.org/10.1007/978-981-19-3045-4_21

Patterson, R., McNamara, E., Tainio, M., de Sá, T. H., Smith, A. D., Sharp, S. J., Edwards, P., Woodcock, J., Brage, S., & Wijndaele, K. (2018). Sedentary behaviour and risk of all-cause, cardiovascular and cancer mortality, and incident type 2 diabetes: a systematic review and dose response meta-analysis. *European Journal of Epidemiology*, *33*(9), 811–829. https://doi.org/10.1007/s10654-018-0380-1

Rodríguez, A. J., Nunes, V. dos S., Mastronardi, C. A., Neeman, T., & Paz-Filho, G. J. (2016a). Association between circulating adipocytokine concentrations and microvascular complications in patients with type 2 diabetes mellitus: A systematic review and meta-analysis of controlled cross-sectional studies. *Journal of Diabetes and Its Complications*, *30*(2), 357–367. https://doi.org/10.1016/j.jdiacomp.2015.11.004

Rodríguez, A. J., Nunes, V. dos S., Mastronardi, C. A., Neeman, T., & Paz-Filho, G. J. (2016b). Association between circulating adipocytokine concentrations and microvascular complications in patients with type 2 diabetes mellitus: A systematic review and meta-analysis of controlled cross-sectional studies. *Journal of Diabetes and Its Complications*, *30*(2), 357–367. https://doi.org/10.1016/j.jdiacomp.2015.11.004

Tsujimoto, T., Yamamoto-Honda, R., Kajio, H., Kishimoto, M., Noto, H., Hachiya, R., Kimura, A., Kakei, M., & Noda, M. (2015). Prediction of 90-day mortality in patients without diabetes by severe hypoglycemia: blood glucose level as a novel marker of severity of underlying disease. *Acta Diabetologica*, *52*(2), 307–314. https://doi.org/10.1007/s00592-014-0640-9

Wang, X., Shen, Y., Wang, S., Li, S., Zhang, W., Liu, X., Lai, L., Pei, J., & Li, H. (2017). PharmMapper 2017 update: a web server for potential drug target identification with a comprehensive target pharmacophore database. *Nucleic Acids Research*, *45*(W1), W356–W360. https://doi.org/10.1093/nar/gkx374

Warraich, H. J., O'Connor, C. M., Yang, H., Granger, B. B., Johnson, K. S., Mark, D. B., Anstrom, K. J., Patel, C. B., Steinhauser, K. E., Tulsky, J. A., Taylor, D. H., Rogers, J. G., & Mentz, R. J. (2019a). African Americans With Advanced Heart Failure Are More Likely to Die in a Health Care Facility Than at Home or in Hospice: An Analysis From the PAL-HF Trial. *Journal of Cardiac Failure*, *25*(8), 693–694. https://doi.org/10.1016/j.cardfail.2019.05.013

Warraich, H. J., O'Connor, C. M., Yang, H., Granger, B. B., Johnson, K. S., Mark, D. B., Anstrom, K. J., Patel, C. B., Steinhauser, K. E., Tulsky, J. A., Taylor, D. H., Rogers, J. G., & Mentz, R. J. (2019b). African Americans With Advanced Heart Failure Are More Likely to Die in a Health Care Facility Than at Home or in Hospice: An Analysis From the PAL-HF Trial. *Journal of Cardiac Failure*, *25*(8), 693–694. https://doi.org/10.1016/j.cardfail.2019.05.013

Yogapriya, J., & Dhanalakshmi. (2023). Classification and Prediction of Diabetic Retinopathy Severity using Transfer Learning Algorithm. *2023 7th International Conference on Computing Methodologies and Communication (ICCMC)*, 217–220. https://doi.org/10.1109/ICCMC56507.2023.10083510

Zhang, Q., Li, X., Liu, X., Liu, S., Zhang, M., Liu, Y., Zhu, C., & Wang, K. (2022). Correction: Zhang et al. The Effect of Non-Invasive Brain Stimulation on the Downregulation of Negative Emotions: A

Meta-Analysis. Brain Sci. 2022, 12, 786. *Brain Sciences*, *12*(8), 1107. https://doi.org/10.3390/brainsci12081107

Zohair, M., Chandra, R., Tiwari, S., & Agarwal, S. (2023a). A model fusion approach for severity prediction of diabetes with respect to binary and multiclass classification. *International Journal of Information Technology*. https://doi.org/10.1007/s41870-023-01463-9

Zohair, M., Chandra, R., Tiwari, S., & Agarwal, S. (2023b). A model fusion approach for severity prediction of diabetes with respect to binary and multiclass classification. *International Journal of Information Technology*. https://doi.org/10.1007/s41870-023-01463-9

YouTube Link: https://youtu.be/euHZx-huso4

GitHub Link: https://github.com/Sai-Srinivasa-S-P/Data-Mining.git