**DIABETES PREDICTION**

|  |  |  |
| --- | --- | --- |
|  |  |  |

**Abstract**

The remarkable advancements in biotechnology and public healthcare infrastructures have led to a momentous production of critical and sensitive healthcare data. By applying intelligent data analysis techniques, many interesting patterns are identified for the early and onset detection and prevention of several fatal diseases. Diabetes mellitus is an extremely life-threatening disease because it contributes to other lethal diseases, i.e., heart, kidney, and nerve damage. In this paper, a machine learning based approach has been proposed for the classification, early-stage identification, and prediction of diabetes. Furthermore, it also presents an IoT-based hypothetical diabetes monitoring system for a healthy and affected person to monitor his blood glucose (BG) level. For diabetes classification, three different classifiers have been employed, i.e., random forest (RF), multilayer perceptron (MLP), and logistic regression (LR). For predictive analysis, we have employed long short-term memory (LSTM), moving averages (MA), and linear regression (LR). For experimental evaluation, a benchmark PIMA Indian Diabetes dataset is used. During the analysis, it is observed that MLP outperforms other classifiers with 86.08% of accuracy and LSTM improves the significant prediction with 87.26% accuracy of diabetes. Moreover, a comparative analysis of the proposed approach is also performed with existing state-of-the-art techniques, demonstrating the adaptability of the proposed approach in many public healthcare applications.

**Key words:**

Machine Learning , dataset , Preprocessing , Postprocessing, Random Forest Classifier ,Logistic Regression Classifier ,K-Nearest Neighbor Classifier, Support Vector Machine Classifier, feature selection, accuracy.

# **Introduction**

With the increasing power of computer technology, companies and institutions can nowadays store large amounts of data at reduced cost. The amount of available data is increasing exponentially and cheap disk storage makes it easy to store data that previously was thrown away. There is a huge amount of information locked up in databases that is potentially important but has not yet been explored. The growing size and complexity of the databases makes it hard to analyse the data manually, so it is important to have automated systems to support the process. Hence there is the need of computational tools able to treat these large amounts of data and extract valuable information.

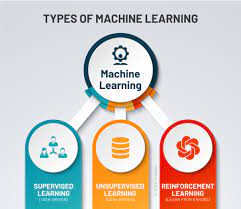
In this context, Data Mining provides automated systems capable of processing large amounts of data that are already present in databases. Data Mining is used to automatically extract important patterns and trends from databases seeking regularities or patterns that can reveal the structure of the data and answer business problems. Data Mining includes learning techniques that fall into the field of Machine learning. The growth of databases in recent years brings data mining at the forefront of new business technologies.

Public health is a fundamental concern for protecting and preventing the community from health hazard diseases Governments are spending a considerable amount of their gross domestic product (GDP) for the welfare of the public, and initiatives such as vaccination have prolonged the life expectancy of people However, for the last many years, there has been a considerable emergence of chronic and genetic diseases affecting public health. Diabetes mellitus is one of the extremely life-threatening diseases because it contributes to other lethal diseases, i.e., heart, kidney, and nerve damage

Diabetes is a metabolic disorder that impairs an individual’s body to process blood glucose, known as blood sugar. This disease is characterized by hyperglycemia resulting from defects in insulin secretion, insulin action, or both An absolute deficiency of insulin secretion causes type 1 diabetes (T1D). Diabetes drastically spreads due to the patient’s inability to use the produced insulin. It is called type 2 diabetes (T2D) Both types are increasing rapidly, but the ratio of increase in T2D is higher than T1D. 90 to 95% of cases of diabetes are of T2D.

## **What are the different types of Machine Learning?**

Machine learning is a subset of AI, which enables the machine to automatically learn from data, improve performance from past experiences, and make predictions. Machine learning contains a set of algorithms that work on a huge amount of data. Data is fed to these algorithms to train them, and on the basis of training, they build the model & perform a specific task. These ML algorithms help to solve different business problems like Regression, Classification, Forecasting, Clustering, and Associations, etc.



Based on the methods and way of learning, machine learning is divided into mainly four types, which are:

1. Supervised Machine Learning
2. Unsupervised Machine Learning
3. Reinforcement Learning

## **1. Supervised Machine Learning**

As its name suggests, Supervised Machine Learning is based on supervision. It means in the supervised learning technique, we train the machines using the "labelled" dataset, and based on the training, the machine predicts the output. Here, the labelled data specifies that some of the inputs are already mapped to the output. More preciously, we can say; first, we train the machine with the input and corresponding output, and then we ask the machine to predict the output using the test dataset.

**2. Unsupervised Machine Learning**

Unsupervised Learning is different from the Supervised learning technique; as its name suggests, there is no need for supervision. It means, in unsupervised machine learning, the machine is trained using the un labeled dataset, and the machine predicts the output without any supervision. unsupervised learning, the models are trained with the data that is neither classified nor labelled, and the model acts on that data without any supervision. **The main aim of the unsupervised learning algorithm is to group or categories the unsorted dataset according to the similarities, patterns, and differences.** Machines are instructed to find the hidden patterns from the input dataset.

**3.Reinforcement Learning**

Reinforcement learning works on a feedback-based process, in which an AI agent (A software component) automatically explore its surrounding by hitting & trail, taking action, learning from experiences, and improving its performanc

## **Benefits of Using Machine Learning in Diabetes Health Indicators Data Set**

1. **Helps in Maintain Accurate Data**

Earlier, keeping records of everything was challenging and time-consuming. But, all thanks to technologies like Machine learning, it has made it easier to maintain proper health records. It helps keep the entry and records, and most of all it- saves time, effort, and money. With evolving technologies, [**Machine learning-based tools**](https://www.hdatasystems.com/ai-ml-development) help in treatment from ground level with the clinical practice diagnosis and recommendations. It is one of the significant machine learning application cases in the health insurance sector.

1. **Forecast of Sudden Outbreak**

Machine learning not just supports the present issue but also lets you forecast the problems. In a situation such as epidemics across the globe can be predicted with Machine Learning.  In today’s circumstances, the expert has to obtain the enormous amount of data that is managed from the website data, present-time social media updates, and others. It helps to verify this data and foretell that everything from sickness outbreaks to severe infectious diseases

.

1. **Surgeries perform by Robots**

Latest technology like Machine Learning lets machines and devices do their work. One of the contributions of Machine Learning in the [**healthcare industry would be robotic surgery.**](https://www.hdatasystems.com/case-study/healthcare-automation) The application has become promising to many experts. It can be divided into four subcategories: Surgical workflow pattern, developments of the robotic surgical supplies, surgical skill evaluation, and automated suturing. Though, it can operate with the algorithm given by a human. It acts like another hand to execute.

1. **Helps in Detecting Illness and Analysis**

One of the significant benefits of machine learning in healthcare is the classification and analysis of infections and illnesses. It made it more manageable as it was difficult to diagnose. It can involve anything from tumors that are challenging to find at the time of the beginning stage to other transmitted diseases.

**5)Manufacturing and Managing of Medicine**

Discovering medicine development methods in the initial stage is one of the benefits of machine learning in healthcare. It also involves the team of Research and Development- which operates technologies such as next-generation order and accuracy medication. It can even help in discovering alternative ways for the healing of multifactorial disorders.

## **About Industry (Diabetes Prediction)**

Health care systems have played a central role in the public health response to the growing problem of diabetes and its complications. During the 1990s, managed care organizations (MCOs) began seeking system-level approaches to improve diabetes outcomes and control costs in covered populations. Although previous clinical trials had demonstrated that several clinical interventions could reduce complication rates and possibly control costs, these findings were not being systematically applied

Performance-reporting initiatives, such as the National Committee on Quality Assurance's Diabetes Quality Improvement Program, led MCOs to develop disease management programs that used diabetes registries, internal performance monitoring and feedback, physician and patient reminder systems, case management, and provider incentives to improve quality . Simultaneously, MCOs introduced cost-containment strategies, including utilization review, preauthorization requirements, cost-related incentives, and patient cost-sharing .

### **AI / ML Role in Diabetes Prediction**

Machine Learning is a sub-set of artificial intelligence where computer algorithms are used to autonomously learn from data. Machine learning (ML) is getting more and more attention and is becoming increasingly popular in many other industries. Within the insurance industry, there is more application of ML regarding the claims.

# **Diabetes Prediction**

Research in the field has identified the following as **important risk factors** for diabetes and other chronic illnesses like heart disease

**High Blood Pressure** : Adults who have been told they have high blood pressure by a doctor, nurse.

**High Cholesterol :**

* Have you EVER been told by a doctor, nurse or other health professional that your blood cholesterol is high?
* Cholesterol check within past five years

**BMI** : Body Mass Index (BMI)

**Smoking** : Have you smoked at least 100 cigarettes in your entire life? [Note: 5 packs = 100 cigarettes]

**Other Chronic Health Conditions :**

* (Ever told) you had a stroke.
* Respondents that have ever reported having coronary heart disease (CHD) or myocardial infarction (MI)

**Physical Activity** :

* Adults who reported doing physical activity or exercise during the past 30 days other than their regular job

**Diet :**

* Consume Fruit 1 or more times per day
* Consume Vegetables 1 or more times per day

**Alcohol Consumption**

* Heavy drinkers (adult men having more than 14 drinks per week and adult women having more than 7 drinks per week)

**Health Care**

* Do you have any kind of health care coverage, including health insurance, prepaid plans such as HMOs, or government plans such as Medicare, or Indian Health Service?
* Was there a time in the past 12 months when you needed to see a doctor but could not because of cost?

**Health General and Mental Health**

* Would you say that in general your health is: --> GENHLTH
* Now thinking about your mental health, which includes stress, depression, and problems with emotions, for how many days during the past 30 days was your mental health not good? --> MENTHLTH
* Now thinking about your physical health, which includes physical illness and injury, for how many days during the past 30 days was your physical health not good?
* Do you have serious difficulty walking or climbing stairs?

**Demographics**

* Indicate sex of respondent.
* Fourteen-level age category
* What is the highest grade or year of school you completed?
* Is your annual household income from all sources: (If respondent refuses at any income level, code "Refused.")

## **Main Drivers for Diabetes Prediction**

Predictive modelling allows for simultaneous consideration of many variables and quantification of their overall effect. When a large number of records are analysed, patterns regarding the conditions of the patients that drive loss development begin to emerge. Some of the attributes that influence the target variable Diabetes\_011 are:

The attribute of age has been useful to classify in five categories.

|  |  |
| --- | --- |
| Age(Years) | Age Bins |
| ≤30 | Youngest |
| 31–40 | Younger |
| 41–50 | Middle aged |
| 51–60 | Older |
| ≥61 | Oldest |

Blood glucose concentration in patients who do not have diabetes is different from patients with diabetes.

Glucose values have been divided into 5 categories.

|  |  |
| --- | --- |
| Glucose | Glucose Bins |
| ≤60 | Very Low |
| 61–80 | Low |
| 81–140 | Normal |
| 141–180 | Early Diabetes |
| ≥181 | Diabetes |

A [strong association](https://www.sciencedirect.com/topics/computer-science/strongest-association) has been found between healthy and diabetic patients regarding their blood pressure levels Blood pressure has been divided into five different categories.

|  |  |
| --- | --- |
| Blood Pressure | Diastolic Blood Pressure Bins |
| <61 | Very low |
| 61–75 | Low |
| 75–90 | Normal |
| 91–100 | High |
| >100 | Hypertension |

The relationship between BMI and diabetes prevalence is consistent. The prevalence of diabetes and obesity is increasing concurrently worldwide. Furthermore, previous studies have shown that BMI is the most important risk factor for type 2 diabetes . BMI values have been categorized into five classes.

|  |  |
| --- | --- |
| BMI | BMI Bins |
| <19 | Starvation |
| 19–24 | Normal |
| 25–30 | Overweight |
| 31–4 | Obese |
| >40 | Very Obese |

## **Project - Data Link**

The project data has taken from Kaggle and the link is <https://www.kaggle.com/datasets/alexteboul/diabetes-health-indicators-dataset>

# **AI / ML Modelling and Results**

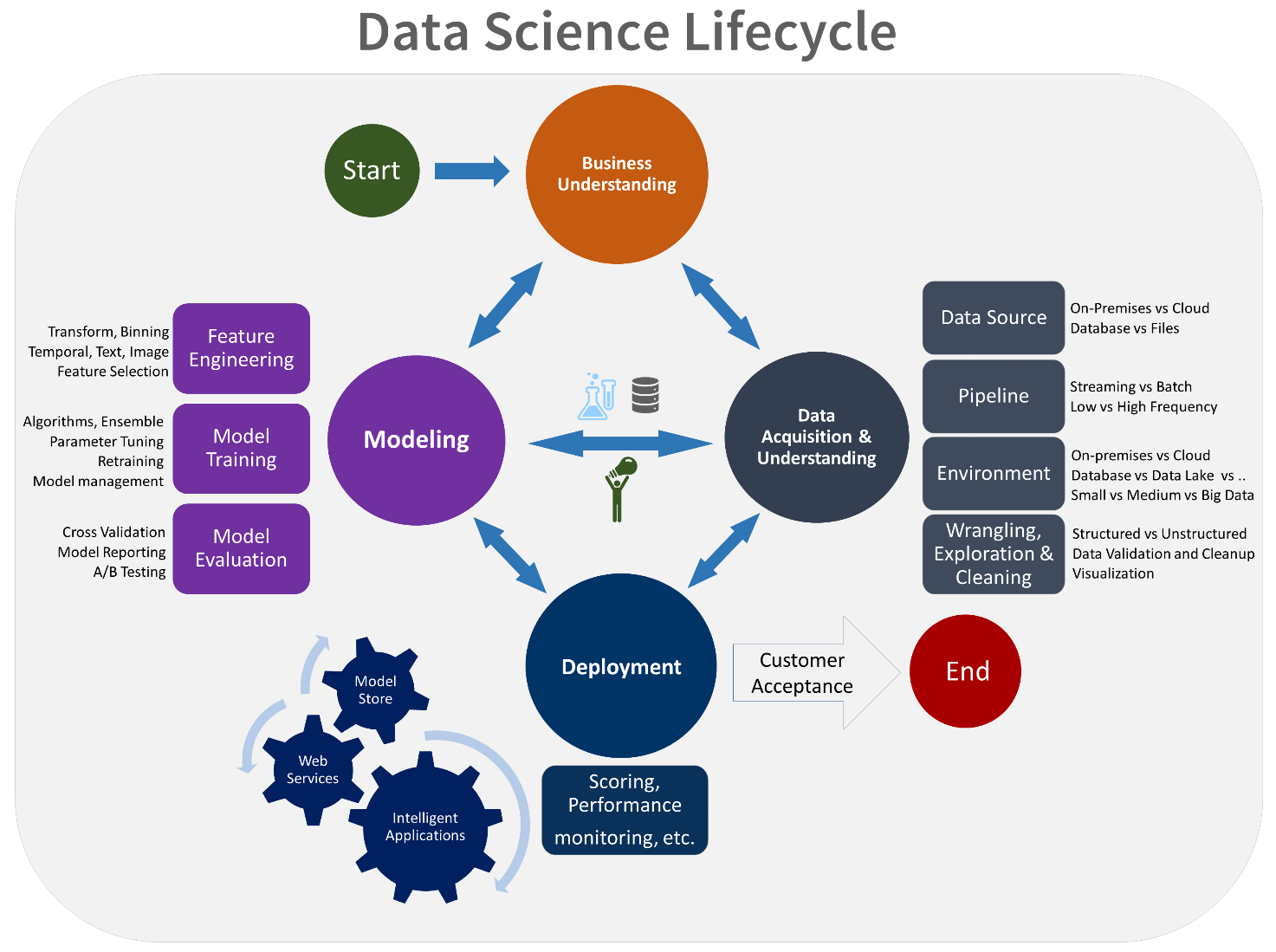
## **Your Problem of Statement**

The Behavioral Risk Factor Surveillance System (BRFSS) is a health-related telephone survey that is collected annually by the CDC. Each year, the survey collects responses from over 400,000 Americans on health-related risk behaviors, chronic health conditions, and the use of preventative services. It has been conducted every year since 1984. For this project, a csv of the dataset available on Kaggle for the year 2015 was used. This original dataset contains responses from 441,455 individuals and has 330 features. These features are either questions directly asked of participants, or calculated variables based on individual participant responses.

This dataset contains file:

1. diabetes \_ 012 \_ health \_ indicators \_ BRFSS2015.csv is a clean dataset of 253,680 survey responses to the CDC's BRFSS2015. The target variable Diabetes\_012 has 3 classes. 0 is for no diabetes or only during pregnancy, 1 is for prediabetes, and 2 is for diabetes. There is class imbalance in this dataset. This dataset has 21 feature variables

## **Data Science Project Life Cycle**

Data Science is a multidisciplinary field of study that combines programming skills, domain expertise and knowledge of statistics and mathematics to extract useful insights and knowledge from data.

### **Data Exploratory Analysis**

Exploratory data analysis has been done on the data to look for relationship and correlation between different variables and to understand how they impact or target variable.

The exploratory analysis is done for Auto Quote / Policy Conversionwith different parameters and all the charts are presented in **Appendices 6.2 - List of charts (6.2.1 to 6.2.9)**

### **Data Pre-processing**

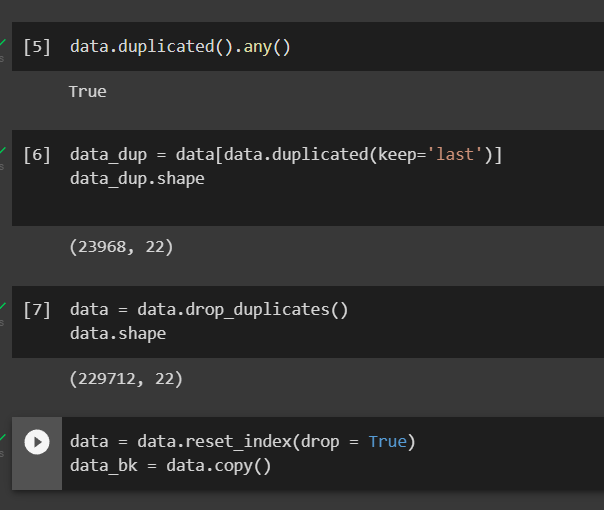
The first stage of the pipeline involves data mining methods and techniques for converting raw patient records to an acceptable format for training and testing machine learning models. In this stage, the raw data of patients was extracted from the NHANES database to be represented as records in the preprocessing step. The preprocessing stage also converted any undecipherable values (errors in datatypes and standard formatting) from the database to null representations.

### **Check the Duplicate and low variation data**

Generally data from industries and from any sources consists a lot of duplicates which decrease the accuracy of model and the data may be either structured or unstructured

Structured Data – here, duplicates very much come with the territory. In this situation you’ve also likely got a lot of implicit ambiguity in your problem.

Unstructured Data– here duplicates are weird. They’re drastically less common than in the structured space, and importantly much more problematic. They typically represent strange edge cases, ETL issues, or other aberrations in data processing.



### **Identify and address the missing variables**

In real world data, there are some instances where a particular element is absent because of various reasons, such as, corrupt data, failure to load the information, or incomplete extraction. [Handling](https://analyticsindiamag.com/get-started-preparing-data-machine-learning/) the missing values is one of the greatest challenges faced by analysts, because making the right decision on how to handle it generates robust data models. Let us look at different ways of imputing the missing values.

### **1. Deleting Rows**

This method commonly used to handle the null values. Here, we either delete a particular row if it has null value for a particular feature and a particular column if it has more than 70-75% of missing values. This method is advised only when there are enough samples in the data set.

## **2. Replacing With Mean/Median/Mode**

This strategy can be applied on a feature which has numeric data like the age of a person or the ticket fare. We can calculate the mean, median or mode of the feature and replace it with the missing values. This is an approximation which can add variance to the data set. But the loss of the data can be negated by this method which yields better results compared to removal of rows and columns. Replacing with the above three approximations are a statistical approach of handling the missing values.

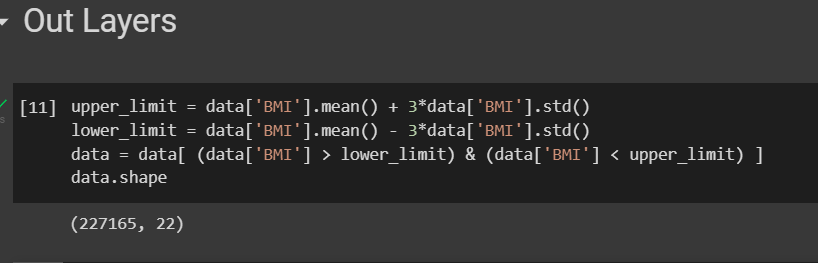
In this project the data null values are not addressed due to the absemce of null values.

### **Handling of Outliers**

### Outliers are those data points which differs significantly from other observations present in given dataset. It can occur because of variability in measurement and due to misinterpretation in filling data points.

**Most common causes of outliers on a data set:**

* ***Data Entry Errors:*** Human errors such as errors caused during data collection, recording, or entry can cause outliers in data.
* ***Measurement Error*(instrument errors)**: It is the most common source of outliers. This is caused when the measurement instrument used turns out to be faulty.
* ***Experimental errors*** (data extraction or experiment planning/executing errors)
* ***Intentional***(dummy outliers made to test detection methods)
* ***Data processing errors*** (data manipulation or data set unintended mutations)
* ***Sampling errors*** (extracting or mixing data from wrong or various sources)



### **Categorical data and Encoding Techniques**

Categorical data is a type of data that is used to group information with similar characteristics, while numerical data is a type of data that expresses information in the form of numbers.

**Why do we need encoding?**

* Most machine learning algorithms cannot handle categorical variables unless we convert them to numerical values
* Many algorithm’s performances even vary based upon how the categorical variables are encoded

**Categorical variables can be divided into two categories:**

* Nominal: no particular order
* Ordinal: there is some order between values

We will also refer to a cheat sheet that shows when to use which type of encoding.

## **Label Encoding or Ordinal Encoding**

We use this categorical data encoding technique when the categorical feature is ordinal. In this case, retaining the order is important. Hence encoding should reflect the sequence.

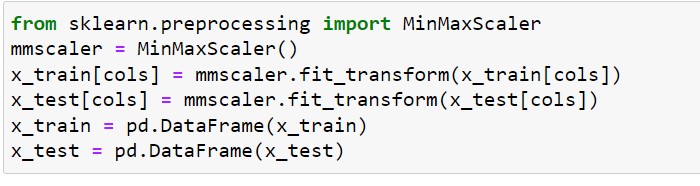
In Label encoding, each label is converted into an integer value. We will create a variable that contains the categories representing the education qualification of a person.

In this project the data is encoded no techniques are used.

### **Feature Scaling**

In Data Processing, we try to change the data in such a way that the model can process it without any problems. And Feature Scaling is one such process in which we transform the data into a better version. Feature Scaling is done to normalize the features in the dataset into a finite range. There are several ways to do feature scaling. I will be discussing the top 5 of the most commonly used feature scaling techniques.

1. Absolute Maximum Scaling
2. Min-Max Scaling
3. Normalization
4. Standardization



### **Selection of Dependent and Independent variables**

The dependent or target variable here isClaimed Target which tells us a particular policy holder has filed a claim or not the target variable is selected based on our business problem and what we are trying to predict.

The independent variables are selected after doing exploratory data analysis and we used Boruta to select which variables are most affecting our target variable.

### **Data Sampling Methods**

The data we have is highly unbalanced data so we used some sampling methods which are used to balance the target variable so we our model will be developed with good accuracy and precision. We used three Sampling methods

### **Stratified sampling**

Stratified sampling randomly selects data points from majority class so they will be equal to the data points in the minority class. So, after the sampling both the class will have same no of observations.

It can be performed using strata function from the library sampling.

### **Simple random sampling**

Simple random sampling is a sampling technique where a set percentage of the data is selected randomly. It is generally done to reduce bias in the dataset which can occur if data is selected manually without randomizing the dataset.

We used this method to split the dataset into train dataset which contains 70% of the total data and test dataset with the remaining 30% of the data.

### **Models Used for Development**

We built our predictive models by using the following ten algorithms

### **Logistic Regression Algorithm**

Logistic uses logit link function to convert the likelihood values to probabilities so we can get a good estimate on the probability of a particular observation to be positive class or negative class.The also gives us p-value of the variables which tells us about significance of each independent variable.

### **Random Forest Algorithm**

Random forest is an algorithm that consists of many decision trees. It was first developedby Leo Breiman and Adele Cutler. The idea behind it is to build several trees,to have the instance classified by each tree, and to give a "vote" at each class.The model uses a "bagging" approach and the random selection of features to build acollection of decision trees with controlled variance. The instance's class is to the classwith the highest number of votes, the class that occurs the most within the leaf in whichthe instance is placed.

The error of the forest depends on:

* Trees correlation: the higher the correlation, the higher the forest error rate.
* The strength of each tree in the forest. A strong tree is a tree with low error. By

using trees that classify the instances with low error the error rate of the forest

decreases.

### **Decision Tree Classifier**

It is a tool that has applications spanning several different areas. Decision trees can be used for classification as well as regression problems. The name itself suggests that it uses a flowchart like a tree structure to show the predictions that result from a series of feature-based splits. It starts with a root node and ends with a decision made by leaves.

### **Navies Bayes(GuassianNB) Classifier**

Navies Bayes law, which is used to determine the probability of a hypothesis with prior knowledge. It depends on the conditional probability.

The formula for is given as:

P(A|B) = P(B|A) \* P(A)

P(B)

where,

P(A|B) is Posterior probability: Probability of hypothesis A on the observed event B

P(B|A) is Likelihood probability: Probability of the evidence given that the probability of a hypothesis is true.

P(A) is Prior Probability: Probability of hypothesis before observing the evidence.

P(B) is Marginal Probability: Probability of Evidence.

There are three types of Naive Bayes Model, which are given below:

**Gaussian:** The Gaussian model assumes that features follow a normal distribution. This means if predictors take continuous values instead of discrete, then the model assumes that these values are sampled from the Gaussian distribution.

### **Extra Tree Classifier**

An extra tree is an algorithm used for classification and regression tasks. It works by randomly selecting a subset of features and then training a Decision tree on them. The tree is then pruned only to contain the most important features for making predictions. The Extra tree algorithm is considered an efficient and accurate machine learning method. It has outperformed other popular methods such as support vector machines and random forests. This article will discuss how the Extra Tree algorithm works and how it differs from the Random forest algorithm. We will also implement it on regression and classification datasets using Python.

### **Support Vector Machine**

* Support Vector Machines (SVM) classify data by separating the classes with a boundary, i.e. a line or multi-dimensional hyperplane. Optimization ensures that the widest boundary separation of classes is achieved. While SVM often outperforms logistic regression, the computational complexity of the model results in long training durations for model development

### **KNN Algorithm**

* K-Nearest Neighbour is one of the simplest Machine Learning algorithms based on Supervised Learning technique.
* K-NN algorithm assumes the similarity between the new case/data and available cases and put the new case into the category that is most similar to the available categories.
* K-NN algorithm stores all the available data and classifies a new data point based on the similarity. This means when new data appears then it can be easily classified into a well suite category by using K- NN algorithm.
* K-NN algorithm can be used for Regression as well as for Classification but mostly it is used for the Classification problems.
* K-NN is a **non-parametric algorithm**, which means it does not make any assumption on underlying data.
* It is also called a **lazy learner algorithm** because it does not learn from the training set immediately instead it stores the dataset and at the time of classification, it performs an action on the dataset.
* KNN algorithm at the training phase just stores the dataset and when it gets new data, then it classifies that data into a category that is much similar to the new data.

### **XGBoost Classifier Algorithm**

Gradient Boosted Trees (GBT) is also an ensemble prediction model based on decision trees. In contrast to Random Forest, this model successively builds decision trees using gradient descent in order to minimize a loss function. A final prediction is made using a weighted majority vote of all of the decision trees. We consider an implementation of gradient boosting, XGBoost, which is optimized for speed and performance.

### **Bagging Classifier**

While decision trees are one of the most easily interpretable models, they exhibit highly variable behavior. Consider a single training dataset that we randomly split into two parts. Now, let’s use each part to train a decision tree in order to obtain two models. When we fit both these models, they would yield different results. Decision trees are said to be associated with high variance due to this behavior. Bagging or boosting aggregation helps to reduce the variance in any learner. Several decision trees which are generated in parallel, form the base learners of bagging technique. Data sampled with replacement is fed to these learners for training. The final prediction is the averaged output from all the learners.

### **LightGbm Classifier Algorithm**

Light Gradient Boosted Machine, or **LightGBM** for short, is an open-source library that provides an efficient and effective implementation of the gradient boosting algorithm. LightGBM extends the gradient boosting algorithm by adding a type of automatic feature selection as well as focusing on boosting examples with larger gradients. This can result in a dramatic speedup of training and improved predictive performance. As such, LightGBM has become a de facto algorithm for machine learning competitions when working with tabular data for regression and classification predictive modeling tasks.

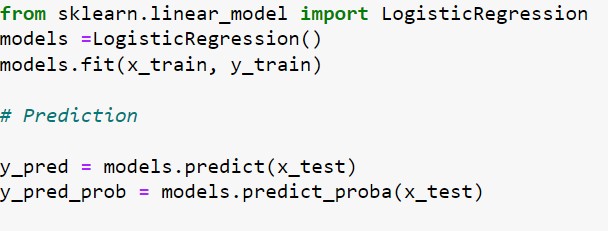
## **AI / ML ModelsAnalysis and Final Results**

We used our train dataset to build the above models and used our test data to check the accuracy and performance of our models.

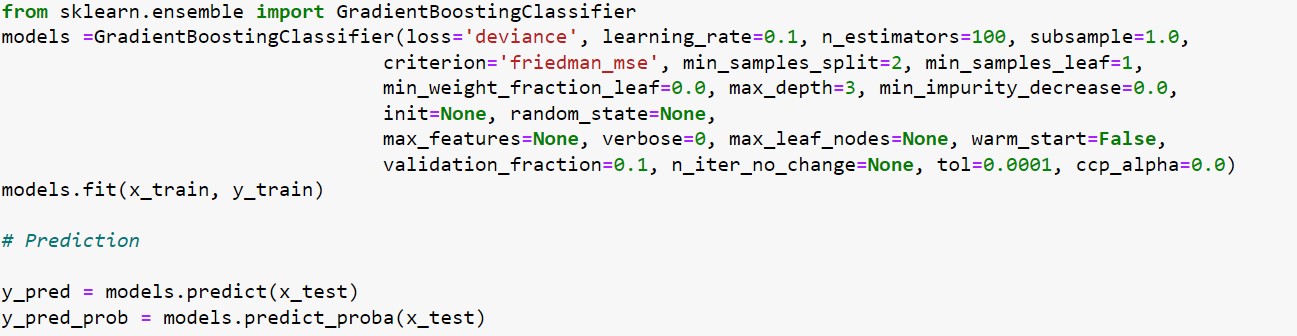
We used confusion matrix to check accuracy, Precision, Recall and F1 score of our models and compare and select the best model for given auto dataset of size ~ 272252 policies.

### **Logistic Regression Python code**

* The Python code for models with stratified sampling technique as follows:
* The Python code for models with simple random sampling technique as follows:

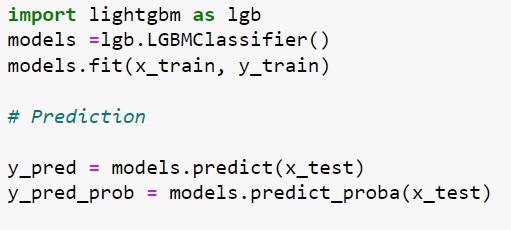


### **Gradient Boosting Classifier Python Code**

* The Python code for models with stratified sampling technique as follows:
* The Python code for models with simple random sampling technique as follows:

### **LGBM Classifier Python code**

* The Python code for models with stratified sampling technique as follows
* The Python code for models with simple random sampling technique as follows:



**Stratified Sampling**:

Random Forest model performance is good, by considering the confused matrix, highest accuracy (1.0) &good F1 score (1.0).This is because random forest uses bootstrap aggregation which can reduce bias and variance in the data and can leadsto good predictions withclaims dataset.

**Simple Random Sampling**:

Artificial Neural Networks / Random Forest are out performed Logistic Regression model, by considering the confused matrix, highest accuracy (1.0) &good F1 score (1.0).This is because Artificial Neural Networks havehidden and complex patterns between different variables and can leads to good predictions with claims dataset.

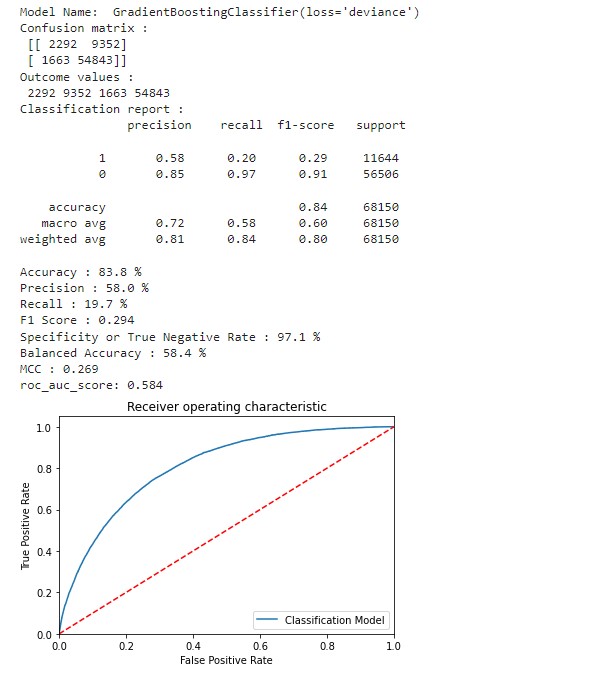
# **Conclusions and Future work**

The model results in the following order by considering the model accuracy, F1 score and RoC AUC score.

1. **Gradient Boosting Classifier** with Stratified and Random Sampling
2. **LGBM Classifier** with Simple Random Sampling
3. **Logistic Regression** with Simple Random Sampling

We recommend model - **Gradient Boosting Classifier** with Stratified and Random Sampling technique as a best fit for the give n BI claims dataset. We considered Random Forest because it uses bootstrap aggregation which can reduce bias and variance in the data and can leadsto good predictions with claims dataset.

The future work to evaluate the “Other Types Claims” in auto Insurance by using classification methods.

****

# **References**

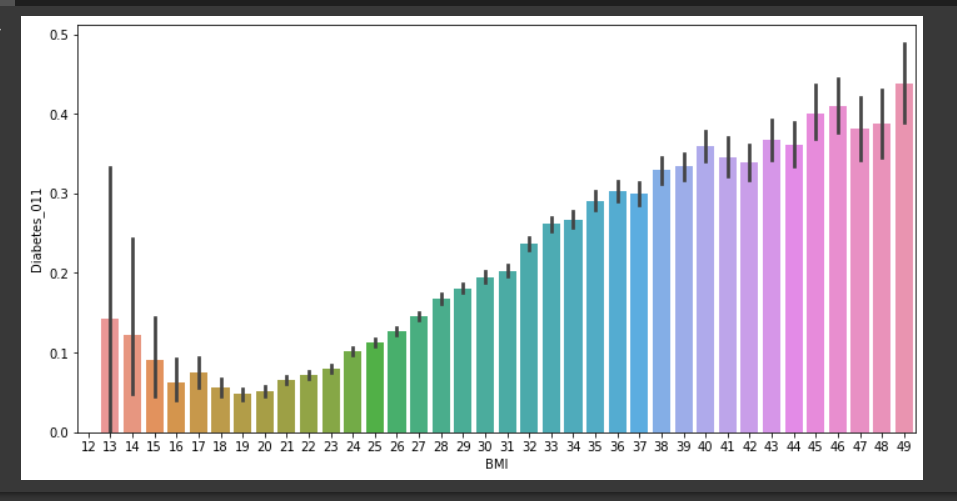
* + - 1. Nakahara T, Hyogo H, Yoneda M, Sumida Y, Sumida Y, Fujii H et al (2013) Type 2 diabetes mellitus is associated with the fibrosis severity in patients with nonalcoholic fatty liver disease in a large retrospective cohort of Japanese patients. J Gastroenterol 49(11):1477–1484. <https://doi.org/10.1007/s00535-013-0911-1>
      2. Solanki P, Baldaniya D, Jogani D, Chaudhary B, Shah M, Kshirsagar A (2021) Artificial intelligence: new age of transformation in petroleum upstream. Pet Res (in press). <https://doi.org/10.1016/j.ptlrs.2021.07.002>
      3. Duplaga M (2004) The impact of information technology on quality of healthcare services. In: Bubak M, van Albada GD, Sloot PMA, Dongarra J (eds) Computational science - ICCS 2004. 4th international conference, Kraków, Poland, June 2004. Lecture notes in computer science, vol 3039. Springer, Berlin, Heidelberg, pp 1118-1125. <https://doi.org/10.1007/978-3-540-25944-2_145>
      4. Lassi M, Sonnenwald DH (2010) Identifying factors that may impact the adoption and use of a social science collaboratory: a synthesis of previous research.Inf Res15(3)
      5. Bonfiglio S (2012) The role of ICT in a healthcare moving from “clinical-centric” to “patient-centric”. In: Donnelly M, Paggetti C, Nugent C, Mokhtari M (eds) Impact analysis of solutions for chronic disease prevention and management. 10th international conference on smart homes and health telematics, June 2012. Lecture notes in computer science, vol 7251. Springer, Berlin, Heidelberg, pp 250-253. <https://doi.org/10.1007/978-3-642-30779-9_37>
      6. Poston RS, Reynolds RB, Gillenson ML (2006) Technology solutions for improving accuracy and availability of healthcare records. Inf Syst Manag 24(1):59–71. <https://doi.org/10.1080/10580530601038097>

# **Appendices**

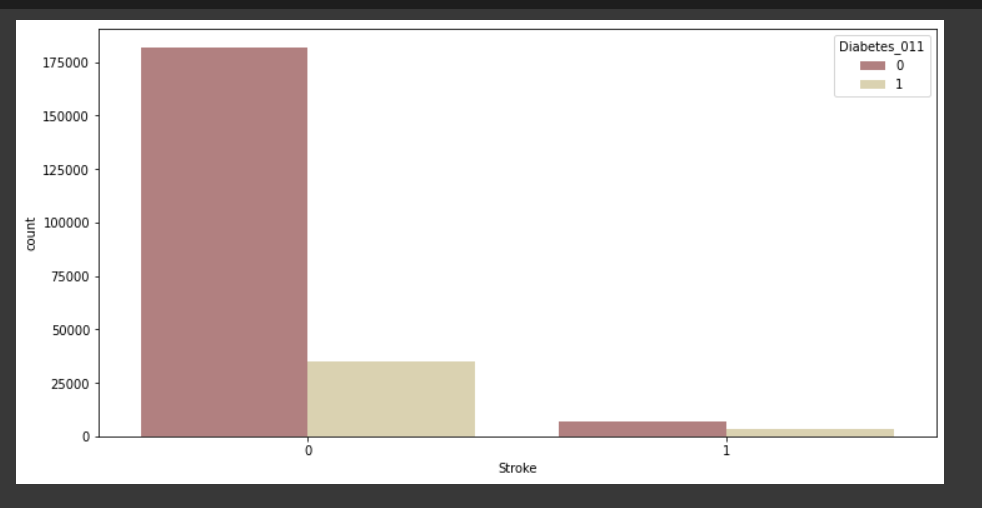
## **Python code Results**

## **List of Charts**

### **Chart 01: Diabetes\_011 Vs BMI**



### **Chart 02: Diabetes\_011 Vs Stroke**



### **Chart 03:** **Diabetes\_011 Vs Age**

