MSc in Data Science

School of Computing, Science and Engineering

MSc Dissertation

TOPIC

Early Diabetes Prediction in Machine Learning

Author: Samuel Obetta

Supervisor: Ali Dan

2021

Early Diabetes Prediction in Machine Learning

Abstract

Diabetes is one of the fastest growing chronic life-threatening diseases that have already affected 422 million people worldwide according to the report of World Health Organization (WHO), in 2018. Due to the presence of a relatively long asymptomatic phase, early detection of diabetes is always desired for a clinically meaningful outcome. Around 50% of all people suffering from diabetes are undiagnosed because of its long-term asymptomatic phase. Due to advancements in the field of computer science many researchers and scientists are trying their best to come up with an AI solution that detects diabetes early in patients. For this purpose, a similar study has been carried out in this paper, where advance machine learning technique such as an H20AutoML has been implemented to build a classification model that predicts diabetes given all the particular parameters such as as age, gender, polyuria, polydipsia, Sudden weight loss, weakness, polyphagia, etc., and the seed=12, and max model=20 were the parameters given to the H20AutoML, and it fitted the dataset on the models including, GBM (Gradient Boosting Machine), DRF (Distributed Uplift Random Forest), Deep Learning, GLM (Generalized Linear Model), etc., In this study, GBM (Gradient Boosting Machine) outperformed all others with an Area Under the Curve (AUC) of 1.0, Area Under the Curve Precision-Recall (AUCPR) of 1.0, Mean Per-Class Error of 0.0, and Log Loss of 0.00045, which is excellent and most satisfactory compared to earlier studies. This model can help people in detection of diabetes early to counter the disease, and help to prevent it.

Acknowledgement

Thanks to my dissertation supervisor, **Mr Ali Dan** for his patience and feedback, I couldn't have written this much better. It would have been impossible for me to undertake this journey without the expertise and knowledge I received from my lecturers, **Prof. Mo Saraee**, **Dr Kaveh Kiani, Dr Judita Preiss** and the rest of them that I could not mention their names. I would also like to acknowledge the generous support that my school, the University of Salford, has provided.

In addition, I am grateful to all my classmates, especially those under the same supervisor as me, who helped me with editing, provided feedback late at night, and provided moral support. I also want to thank the librarians, research assistants, and study participants from the university.

Specifically, I would like to thank my sponsor, **Pastor Chijioke Okonkwo**, for his tireless efforts to help me achieve my dreams.

Lastly, I wouldn't be able to envision life without my family, especially my parents, late **Mr Christopher Obetta** and **Mrs Virginia Obetta**, along with my spouse, **Mrs Fidelia Obetta**, and children, **Chidera Obetta, Toochukwu Obetta**, and **Munachimso Obetta**.The belief they have in me has motivated and kept me going throughout this process.

Table of Contents

List of Figures

List of Tables

List of Abbreviations

AUC	Area Under the Curve
AUCPR	Area Under the Curve Precision-Recall
GBM	Gradient Boosting Machine
DRF	Distributed Uplift Random Forest
Deep Learning	Deep Learning
GLM	Generalized Linear Model
RMSE	Root Mean Squared Error
MSE	Mean Squared Error

Chapter 1: Introduction

1.1 Background Study

Diabetes is one of the fastest growing chronic life-threatening diseases that have already affected 422 million people worldwide according to the report of World Health Organization (WHO), in 2018. Due to the presence of a relatively long asymptomatic phase, early detection of diabetes is always desired for a clinically meaningful outcome. Around 50% of all people suffering from diabetes are undiagnosed because of its long-term asymptomatic phase. Due to advancements in the field of computer science many researchers and scientists are trying their best to come up with an AI solution that detects diabetes early in patients. As no algorithm is every perfect, therefore, more quantitative and qualitative models are being developed every day to find the solution to this problem because when patients would realize that they are about to have diabetes. Hopefully, they would try to refrain from things that make people diabetic and do things that make a person feel health and strong.

In this study, a machine learning model would be developed that detects early diabetes detection using an advance Machine learning technique called AutoML. This technique implements several advance models and evaluates them using key performance measures. The dataset has been chosen from Kaggle.com, which provides genuine datasets for the implementation of Data Science, and Machine Learning projects in AI. In my study, I would be focusing more on identifying important features that lead to the development of diabetes in the patients using metrics like age, weakness, polyphagia, itching and much more. By doing so, I hope to find a reliable algorithm that has the ability to produce better results which could be used by hospitals, and especially doctors to predict diabetes in patients earlier. This will help people fight diabetes and more no. of people will be cured.

1.2 Problem Statement

The aim of this research is to predict early stages of diabetes in people at any age by using AutoML technique in Machine Learning.

1.3 Motivation and Aim

Diabetes is perhaps of the most quickly developing ongoing sickness, which has impacted huge number of individuals all over the planet. Its analysis, expectation, appropriate fix, and the executives are pivotal. Therefore, a predictive Machine Learning model would be developed to predict diabetes in patients through use of AutoML technique in Machine Learning.

1.4 Objective

Diabetes is quite possibly of the most lethal and constant illness which cause an expansion in glucose. Assuming diabetes stays untreated and unidentified numerous challenges might emerge because of that. The monotonous work is in recognizing the cycle which brings about visiting the centre and counselling the specialist. Be that as it may, this drawn-out work has been settled with the ascent in the methodologies utilized by AI.

Following are the objectives of my project:

- Find an appropriate dataset that appropriate data of various patients with relevant features concerning diabetes to predict the class attribute: Diabetes (Larxel, 2021).
- Perform Exploratory Data Analysis (EDA) on the dataset by finding relationships or correlation between various features of the dataset to know how various features affect diabetes.
- Perform data pre-processing and cleaning to get the dataset into a unified format by removing outliers, missing values, converting categorical features, etc. This makes the dataset ready for the implementation of models.
- Use and Implement AutoML (Automated Machine Learning) to build, evaluate, and find the best model that predicts the diabetes in patients early.
- Declare the best model in the study that fits the dataset well and helps predicts the early diabetes in patients with higher accuracy.

1.5 Adopted Approach

Supervised Learning, Unsupervised Learning, Semi-Supervised Learning, and Reinforced Learning are some of the four categories in machine learning. Depending on the sort of data involved, each category has its own approach to issue solving. Therefore, let's talk about different types of learning before explaining supervised learning because in my project, the dataset for the purpose of study was labelled data, so supervised learning would be implemented.

Supervised learning

In this type of learning, labelled dataset is given to the machine, which predicts the desired output after processing it. The main point to remember is that the dataset entering into the machine has features that are very similar to the output dataset. Essentially, an algorithm establishes a cause-and-effect relationship between variables. Finally, the algorithm generates an idea for how the dataset functions and the relationship between input and output. Machine algorithms will continue to improve their cognitive ability to discover new patterns and relationships after training themselves on new data.

Following Stages of Supervised Machine Learning would be employed:

- Exploratory Data Analysis
- Data Preprocessing & Cleaning
- Models Implementation: ***AutoML***
- Evaluation & Results
- Conclusion

1.6 Dissertations Structure

Dissertation’s structure depends on project according to the field of study, organizations involved, laid down principals or laws, etc., let’s discuss the structure being followed in this project:

I. **Introduction**

In this chapter, background of the project shall be discussed in detail, including problem statement, aims & objectives, project timeline, adopted approach, and an overview of dissertation structure.

II. Literature Review and Methodology

In Literature Review, studies conducted earlier in the field of diabetes detection in machine learning shall be discussed including, dataset used, models implemented, and what performance measures were used to find the best model. Analysis shall be performed on their project's importance, framework and methodology adopted, and Machine learning models' accuracy to bring novelty to the project.

III. Data Underacting, Data preparation and EDA

In this chapter, Data under study would be studied in details, for example, variables involved in the dataset, their data types, and what each type means. Then, EDA (Exploratory Data Analysis) shall be performed using visualization, such as graphs, and maps to understand the relationship of various features associated with diabetes, which would help in knowing more about the data, and helps to understand the problem that's being solved. Finally, Data pre-processing would be performed to clean and unify the dataset to make sure the dataset fits perfectly to Machine Learning models, and brings an outstanding higher accuracy.

IV. Analysis – Modelling, Evaluation, and Deploying

In this chapter, various problems faced during experimentation shall be discussed in relation to diabetes detection in Machine Learning, and analysis on the machine learning models shall be discussed to understand how things went during implementation. In addition, tools, libraries, frameworks used to implement would be justified to make the analysis and experimentation more well-grounded, which would be followed by results and discussion.

V. Critical Evaluation

In this chapter, critical evaluation of the experimentation (implementing Machine Learning models) would be performed. Moreover, results of experimentation would also be discussed and compared using performance measures such as, AUC, AUCPR, Log Loss, etc.,

VI. Conclusions and Future Work

In conclusion, the study of early diabetes detection in Machine Learning would reach its conclusion by answering to various questions related to computer science. In addition, a Machine learning would be built that would be able to predict diabetes disorder in patients at an early stage with the highest accuracy score provided all the implemented models build in this study. Finally, future work shall be discussed, where, I would provide details to how the study could be improved in the future using Machine learning techniques. This helps future researchers and scientists to undergo studies in the future with objective to improve performance or change way to implement the performance.

VII. References

In this section, all relevant citations would be referenced, where every word, sentence, paragraph, or idea taken from other sources would be mentioned. In this study, APA referencing style shall be implemented which is acceptable by university.

Chapter 2: Literature Review

2.1.1 A machine learning model for early prediction of diabetes

Alam, T., M., et al (2019) applied three machine learning techniques on diabetes dataset by Pima Indians. RF(random forest), Naïve Bayesian algorithms, and KNN were used by them. A learning model, then, made the prediction either the patient has diabetes or not. There were only 268 diabetic patients in dataset. The following three rules were generated through provided datasets.

First rule was if BMI is equal to Obesity, then patient has diabetes. Second rule also says yes when Glucose is equal to Diabetes. And the last rule describes that by interjecting above both rules will show Yes. The first of the three models that the researcher utilised is (ANN), which comprises of numerous nodes connected to one another. Additionally, the input is processed numerous times throughout training so that the network can modify and improve the weights. The second method uses a collection of tree predictors known as the (RF) random forest method, which has flexible nature, quick, learning algorithm. The third technique is known as K-means clustering, which groups comparable objects together according to their shared traits.

There has been comparison of the suggested models. K-means clustering provided 73.6% exactness, ANN provided 75.7% precision, and random forest provided 74.7% precision, ANN performed better than other approaches.

Aparametric model, whereas other measurable approaches are parametric and necessitate more thorough understanding of the data. Its capacity to depict the non-linear relationship between the factors during examination gives ANN an edge over other factual techniques. To sum up, the results showed a solid affiliation of glucose and body mass index with diabetes. The impediment of the study is that an organized dataset was selected whereas, in the future, unstructured data will be taken into account, and similar algorithms will be used to other medical areas for the predictions of other illnesses including various cancer types.

2.1.2 Early detection of Diabetes mellitus 2 using Machine Learning

Kopitar, L., et el., (2020) conducted a study for an early detection of diabetes type 2 and risk factors. In their study, they used different models such as Glmnet, RF, XGBoost, LightGBM to know the results.

Initially, their dataset consisted of EHRs from twenty-seven thousand and fifty adult individuals without prior diagnosis of T2DM. Moreover, there were some variables such as fruit and vegetable consumption, antihypertensive drug treatment and family history, but these variables were removed from the data set. Additionally, they created various subsets (T12, T18, T6, T30, and T24) and eliminated the variable "Date" in each of them in order to recreate the newly received information. Finally, 58 factors were used to train and test four forecast models: Glmnet, LightGBM, XGBoost, and Arbitrary Timberland (RF). Additionally, each machine learning technique may be used while modifying the parameter values to create the prediction model and their functionality.

In this way, we attempted to configure the parameters so that the execution and computing complexity would be as flexible as possible. As both strategies improve the prediction models, the RMSE of Glmnet is 0.859 and LightGBM exhibits RME of 0.846 improved soundness in positioning the significant factors compared to conceptually comparable XGBoost based models.

In conclusion, their findings revealed no clinically meaningful enhancement in machine learning-based predictive performance over conventional regression models. FPGL demonstrated certain benefits over the more straightforward paradigm. Approaches like LightGBM produce findings that are significantly more stable than those of other methods by observing the stability of ranking variables based on the relative relevance of factors. In some clinical settings, regression-based prediction models may be a preferable alternative to those currently employed in clinical practise. According to the study's findings, all investigated approaches significantly improve in terms of AUC, AUPRC, and RMSE as data collection volume rises.

Future research should look into how various ensemble construction methods can be implemented. In this situation, stacking and combining various prediction models can be an option. However, these systems make it more difficult to understand the results. This ought to bolster the doctor's judgement in this case.

2.1.3 Predicting diabetes Mellitus with Machine Learning

Zou, Q., et al., (2018) used a neural community, a selection tree, and random woodland regions to forecast diabetes mellitus. The statistics for clinic physical examinations in Luzhou, China, make up the dataset. Although, there are numerous feature selection methods, PCA and maximum relevance with minimal redundancy were chosen (mRMR). By using mRMR, the functions' pairwise correlations are assured to be decreased or the maximum Euclidean distances between them are ensured. The biggest pertinent requirements typically augment the very minimum redundancy requirements. First, for better generalisation, the consultant target phenotype for the mRMR characteristic set could be expanded. Second, by employing a smaller set of mRMR feature values, we may effectively replicate the precise region generated by a bigger normal feature set. Additionally, original various signs ae reduced into one or more complete signs by the PCA method. The few comprehensive symptoms can repeat the vast majority of the data considered by the individual indications, and they will no longer be connected to one another.

Furthermore, RF outperforms the other 3 classifiers while using the Luzhou body examination dataset. We came to know that Luzhou dataset J48 outperformed the others, having accuracy of more than 0.76. Only the use of blood glucose tolerance isn't always done correctly in the Pima Indians dataset. They knew that there are three indicators that can be used to identify diabetes mellitus: random blood glucose, fasting blood glucose, and blood glucose tolerance.

Ultimately, the good outcome for the Luzhou dataset is 0.8084, and the Pima Indians performed well with 0.7721, which may indicate that device learning can be used to predict diabetes. However,

identifying the right features, classifiers, and data mining technique are very important. Facts prevent us from identifying the kind of diabetes, so in the future, we plan to forecast it while also examining the percentage of each indicator, which may help predict diabetes more accurately. They plan to predict the kind of diabetes in the future and analyse the percentage of each indicator to determine if this may improve diabetes prediction accuracy.

2.1.4 Machine Learning Prediction Models for Gestational Diabetes Mellitus

(GDM) is an anomaly of metabolic system characterised by a carbohydrate intolerance of variable severity at some stage in pregnancy. It poses a harmful effect on the health of the mother as well as on the child's health. Therefore, it is imperative to utilize machine learning to assist human beings to make an initial judgment approximately diabetes mellitus in line with their everyday body's examination records, and it may also assist doctors.

Zhang, Z., et al., (2022) conducted the study to perform a meta-analysis and evaluation of published prognostic models for predicting the complications of GDM. The program used in their research was the Meta-DiSc software program (version 1.4), whereas PROBAST was used to identify the risks of ML models. Additionally, four reliable digital databases were looked for research that advance ML prediction fashions for GDM in the population instead of high-risk people.

As a result, one of the widely used ML techniques, logistic regression accomplished a typical pooled AUROC of 0.8151, while non–logistic regression fashions did better, with an overall pooled AUROC of 0.8891. Moreover, fasting blood glucose, own family history of diabetes, maternal age, and BMI were the four most generally used features of models established using various feature choice strategies. In the future, the significance of excellent assessments and unified diagnostic standards must be further emphasized.

2.1.5 Analysis and prediction of diabetes diseases using machine learning algorithm

Diabetes disease is typically known as diabetes mellitus (DM). It is a set of metabolic illnesses characterised by excessive blood sugar levels. and either insufficient or ineffective insulin production, or due to the body cells' inefficient response to insulin, or by way of both causes. It has fundamentally three sorts that affect nearly all ages of human beings named as DM1, DM2, and Gestational Diabetes. Therefore, early detection and prevention of diabetes is necessary to prevent early death and preserve human lifestyles. Machine learning algorithms possess a wide range of capabilities for prediction and categorization.

Joshi, R., and Alehegn, M., (2017) used several algorithms to predict diabetes were KNN, Naive Bayes, Random Forest, and J48, as shown in figure 4. The best and most widely used method is RF because it has the best performance and prediction accuracy and is simple to use with large data and high dimensionality. The nearest accurate prediction of the test data is used to store the training data in KNN.

The hybrid model was chosen for this study because, according to their research, it offers superior performance and accuracy over the single model. Furthermore, machine learning approaches have different power in diverse data sets. The ensemble algorithm was more accurate than the solo algorithm. Decision trees produced great accuracy in the majority of investigations. hybrid devices Weka and java are technologies to forecast diabetes datasets.

Chapter 3: Research Methodology

3.1 Software & Tools

In this project, Python, and Jupyter notebook has been used as programming language, and software for the purpose of implementation. So, let's discuss these further to understand better.

3.1.1 Python

It is a computer programming language that has an English like syntax, easy to read, and is one of the most popularly used programming language in the fields of computer science especially in artificial intelligence, machine learning, and data science. It contains various libraries such as Pandas, Numpy, Sklearn, etc., that help to perform visualization, and computations easier on the dataset (Tech Vidhwan, no date).

Source: Wikipedia.org

*Figure **Error! No text of specified style in document.**.1 Python Logo*

Advantages:

- It improves productivity because less code does most of the jobs
- It is free and open-source programming language.
- It has vast library support as there are many different libraries such as Pandas, Numpy, Skleran, etc., which are helpful in implement any sort of functionality with less code.

3.1.2 Jupyter Notebook

Jupyter Notebook is a famous web-based interactive development environment software that helps to run the code, data, and text at the same time in a graphically appealing way that helps users to visualize the data, perform computations, implement any programming code in an easy way that is friendly to user (Kumaraswamy, A., 2018).

Source: Wikipedia.Org

*Figure **Error! No text of specified style in document.**.2 Jupyter Notebook*

Advantages:

- **Everything in one place –** all data, code, images, videos, etc., are in one place
- **Sharing** – all files are in JSON format; therefore, sharing is very easy.
- **NB Convert** – NB Convert is a special feature available in Jupyter's Notebook that converts the notebooks into other formats such as HTML and PDF.
- **Customization** – it is very easy to customize
- **Interactive Code and Data Exploration** – it provides very important controls for exploring data interactively.

3.2 Machine Learning Life Cycle

Machine learning (ML) is a type of artificial intelligence (AI) that allows software applications to become more accurate at predicting outcomes without being explicitly programmed to do so. Machine learning algorithms use historical data as input to predict new output values. The definition of classical Machine Learning emerges from the way an algorithm learns and how it improves the performance of

the model. A general life cycle of Machine Learning includes, Gathering Data, Data Preparation, Data Wrangling, Analyse Data, Train Model, Test Model, Deployment. However, I would perform all of these steps in the following stages:

Source: Javapoint.com

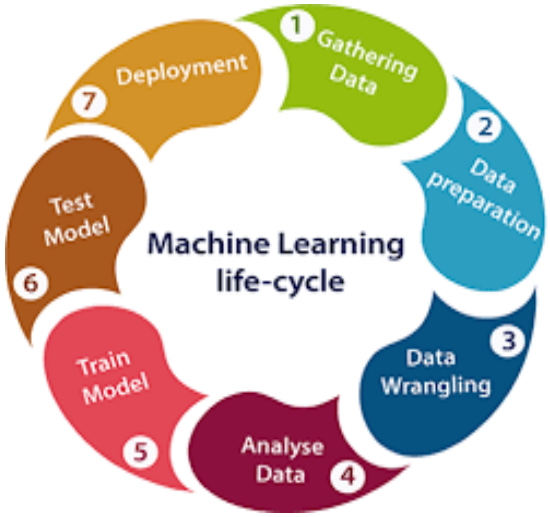


*Figure **Error! No text of specified style in document.**.3 Machine Learning Life Cycle*

3.2.1 Exploratory Data Analysis

It is better that we try to understand the data first, and get as many insights as possible before actually experimenting on the dataset). This is possible with the help of variable visualization libraries such as Pandas, Numpy, Seaborn, Plotly, etc. that are used for data manipulation and analysis, mathematical operations, and graphical visualizations - Uni-variate (observation of two variables with respect to target feature), Bi-variate (observation of two variables with respect to target feature), and multi-variate analysis (this involves observing two or more number of features). During this phase, we learn the relations, and correlations between different features of the dataset, and understand the significance of each important feature that directly affects the outcome of this project. Therefore, all experiments conducted in Exploratory Data Analysis (EDA) are driven towards answers to the questions related to the aims, and objectives of the project.

3.2.2 Data Pre-processing & Cleaning

The second stage after EDA is Data Pre-Processing and Cleaning where bases on the analysis performed on the dataset, data pre-processing and cleaning is performed which is the process building a suitable model by preparing the raw data (MonkeyLearn, no date). It helps to make the dataset in a unified format that is helpful and efficient in achieving the aims and objectives of the study. In fig. 3.4, data pre-processing is being shows in a concise, and understandable way. These are just the steps performed in data pre-processing, however, some other advance techniques, and tools can also be used in data pre-processing.

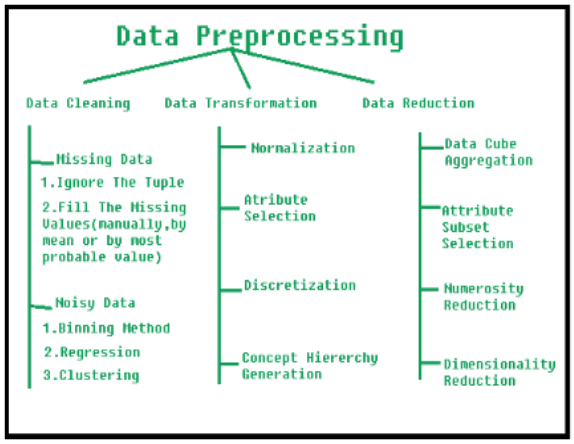


*Figure **Error! No text of specified style in document.**.4 Data Pre-Processing*

Following steps were involved in Data Pre-Processing:

3.2.2.1 **Data Cleaning**

Irrelevant and missing information can be part of the data, so data cleaning is performed to get rid of missing, irrelevant, or redundant information (GeeksforGeeks, 2021).

When some parts of the features/variables in the dataset contains missing, null, or redundant values, then cleaning is performed through following ways:

Ignore Tuples - Ignoring tuples involves removing the tuples from the dataset as part of the cleaning process.

Filling Missing Values - Missing values are filled with Mean, Median or Mode values.

3.2.2.2 Data Transformation

The implementation of Machine Learning models requires a suitable dataset that fits the model which is made possible through data transformation (GeeksforGeeks, 2021).

Let's discuss data transformation in detail for better understanding:

Normalization - In normalization, scaling of the data is performed that helps in converting the value ranges between (-1.0 to 1.0 or 0.0 to 1.0).

Attribute Selection - In this phase, new attributes are created from the existing attributes available in the dataset to form the basis for the implementation. It depends on the dataset, and way the you want to solve the problem using the desired features.

Discretization - Interval or conceptual values are set in the place of raw values of numeric attribute.

Concept Hierarchy - In concept hierarchy, lower-level hierarchy attributes are converted to high-level hierarchy attributes.

3.2.2.3 Data Reduction

Data analysis becomes harder when large number of rows are present in the data ranging to millions of thousand number of entries. Therefore, data reduction is required to get rid of this problem in order to get higher throughput, and avoid software and hardware costs and requirements (GeeksforGeeks, 2021).

Following techniques help to reduce the data in Machine Learning:

Dimensionality Reduction - In dimensionality reduction, the number of features present in the dataset are reduced to train the Machine Learning models.

Feature Selection Method - Important and significant features in the dataset are selected using the Feature selection techniques to increase the predictive accuracy of the model.

Matrix Factorization - The matrix of the dataset is divided into constituent sub parts by use of Matrix Factorization.

Manifold Learning - Manifold Learning is used to convert high dimensional data into low dimensional data for the purpose of visualization. This helps to reduce the number of dimensions.

3.2.3 AutoML (Automated Machine Learning)

Supervised Learning, Semi-Supervised Learning, Unsupervised Learning, and Reinforced Learning are some of the four categories in machine learning. Depending on the sort of data involved, each category has its own approach to issue solving. Therefore, let's talk about different types of learning before explaining supervised learning because in my project, the dataset we are working with is labelled data, so supervised learning is what we are aiming for.

3.2.3.1 Supervised learning

In this kind of learning, labelled dataset is given to the machine, which predicts the desired output after processing it. The main point to remember is that the dataset we enter into the machine has features that are very similar to the output dataset. Essentially, what an algorithm does is establish a relation of cause and effect among variables. Algorithm generates motive for how the dataset functions and the relationship of input to output. Machine algorithms will continue to improve their cognitive ability to discover new patterns and relationships after training themselves on new data.

3.2.3.2 Unsupervised learning

Unsupervised learning can work with unlabelled data, allowing it to work with larger and more complex datasets. Unsupervised learning creates relationships between two data points without the need for human input. Unsupervised learning, as opposed to supervised learning, creates hidden structures that allow the machine to work dynamically on predefined datasets. In other words, no human is required to supervise the process. The goal is to categories unsorted data based on collars, sizes, differences, similarities, and patterns.

3.2.3.3 Semi supervised learning

The phrase "semi-supervised" refers to a strategy that lays between supervised and unsupervised learning. It is the middle ground where soma data is labelled and most data is unlabelled. The main goal of introducing semi-supervised learning was to reduce the disadvantages of both supervised and unsupervised learning. Initially, similar records are clustered using an unsupervised learning algorithm, and it then allows for the labelling of unlabelled data into labelled data. It's due to the fact labelled data is a comparatively more expensive acquisition than unlabelled data.

3.2.3.4 Reinforced learning

It is essentially founded on the psychological idea of conditioning, which states that when an algorithm is placed in an environment with an interpreter and a reward system, it operates and gets better on its own. The interpreter determines whether or not the outcomes of each algorithm iteration are favourable. Additionally, algorithms are strengthened to iterate until the required output is attained in the case of unfavourable outcomes. On the other hand, when algorithms find the correct solution, the interpreter rewards them. Furthermore, reinforcement learning is used in a variety of fields, including game theory and operations research.

3.2.3.5 What is meant by Automated Machine Learning (AML)?

In this procedure machine learning (ML) models are utilised to real-world problems using automation. It focuses on automating the composition, selection, parameter estimation of models for machine learning. When the machine learning process is automated, it becomes more friendly for consumers and typically generates faster, more accurate results than when the algorithms are manually coded. Because of AutoML software packages, machine learning is becoming more accessible to organisations who do not have a professional data scientist or machine learning specialist. These platforms can be purchased from an outside provider or accessed through open-source repositories like GitHub.

3.2.3.6 How AutoML works?

AutoML typically streamlines every stage of the machine learning process, from processing raw datasets to installing an efficient machine-learning model. Handcrafted models are processed independently in each phase. It does look for and use the best machine learning algorithms to produce results. Transfer learning is the process by which models that have already been trained apply what they have discovered to fresh sets of data. AutoML can employ transfer learning to adapt current frameworks to solve brand-new issues.

In specific, below are a few of the machine learning steps that AutoML can optimize:

- Raw data processing
- Engineering and feature selection
- Model preference
- Hyper parameters and parameters' optimization
- Deployment with business and technological constraints in mind
- Choosing an Evaluation Metric
- Monitoring and problem resolution
- Results analysis

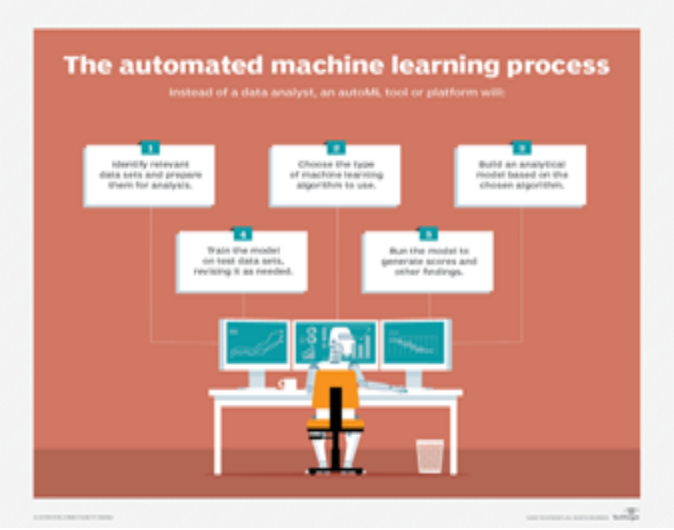


*Figure **Error! No text of specified style in document.**.5 AutoML (Automated Machine Learning)*

3.2.3.7 Importance of AutoML

AI and machine learning have been criticised for being "black boxes," because it is troublesome to reverse engineer using machine learning. Although they increase productivity and computing power to produce results, tracking how the algorithm delivered that output can be difficult. As a matter of fact, selecting the appropriate model for a given problem becomes more difficult, because predicting an outcome can sometimes be difficult whenever black box model is carried out. By making machine learning extra accessible, AutoML renders it less of a conundrum. This technique simplifies the machine learning procedures that apply the algorithm to real-world challenges. AutoML learns on its own in very less time than human would do. Conversely, a human will likely have to understand the internal working of algorithms and how it corresponds to the real world which are time and resource consuming.

3.2.3.8 Applications of AutoML

AutoML and traditional machine learning have similar use cases, whose examples can be seen as follows:

- **Finance fraud detection** - It might enhance the precision and accuracy of fraud detection models.
- **Healthcare research** – analysing major datasets and drawing conclusion.
- **Image recognition** - can be used to recognise faces.
- **Banking** - Finance and insurance risk assessment and management
- **Cyber Security** - It can be useful in cyber security for assessing, monitoring, and testing of risks. Moreover, it might be used to garner dynamic cybersecurity threats found within malware and spam.
- **Customer support** - where it can be used to improve the effectiveness of customer care personnel as well as sentiment analysis in bots.
- **Agriculture** - to speed up quality testing.

3.2.3.9 Advantages and Disadvantages of AutoML

Advantages

It reduces the amount of time needed to train learning models and streamlines and speeds the process of machine learning. A company can reserve their income by dedicating less of its resources to operating a faster, more effective machine learning process. Employers might spend less money on hiring specialists or training staff when the procedure is simplified. Additionally, it makes supervised learning a practical choice for a wider spectrum of businesses. In terms of performance, AutoML algorithms outperform hand-coded models.

Disadvantages

A fundamental obstacle is the tendency to see AutoML as a substitute for human skills. Similar to other automations, AutoML is made to accurately accomplish memorizing chores so that workers can focus on more complex or interesting jobs. Memorization duties like surveillance, analysis, and problem identification can all be automated to help them go faster. Despite the fact that a person is not engaged in the machine learning process, this model requires a human who will have to evaluate and oversee. We are still in the early stages of developing our tools, however, AutoML instead of replacing humans, might help scientist and employees.

3.2.3.10 Features of AutoML

- Google AutoML is a cloud-based machine learning automation tool developed by Google.
- Microsoft Azure Automated Machine Learning is a unique framework.
- Auto Keras, an open-source software library, was built at Texas A&M University's DATA lab.
- Auto-Sklearn is a commercially available freely available set of simple machine learning tools in Python that grew from and superseded Scikit learn. It is available on GitHub.
- Because they employ fewer resources than the other models, Auto-Sklearn and Azure are often seen as less cost effective. They rely substantially on previously seen data and well-known

designs, indicating that they do not require the whole collection of information to operate. This is accomplished through the use of classification and regression techniques.

3.2.4 Evaluation & Results

In machine learning, evaluation of the predictive model determines how accurately the model performs predictions. There are different ways, and methods to assess models depending upon the type of data being dealt with, and of course the type of classification being performed. As in this research, classification problem is being solved using AutoML technique, therefore, Auc, Log Loss, Auc Pr, and Mean per class error would be used to assess the performance of the predictive model. Therefore, critical evaluation of the experimentation (implementing Machine Learning models) would be performed with the results of experimentation and for each implemented models, to find the best model that fits the dataset with higher accuracy. In this way, I would be able to find a predictive model that predicts the diabetes in patients using the data available.

Chapter 4: Data Description, EDA, and Data Pre-Processing

In this chapter, dataset contents, data analysis or to be precise exploratory data analysis (EDA), and pre-processing performed on the dataset shall be discussed in detail. For dataset, importance or significance, and data type of the features or columns would be discussed to understand the importance of the dataset to set or map out a plan to further go deep into performing exploratory data analysis (EDA). In Exploratory Data Analysis, an investigation on the dataset would be performed critically to understand the data and its relationship with other variables of the data using different kinds of plots such as histogram to know the distributions, correlation heatmap graph to know the relationships, and others etc., to perform Univariate, Bivariate, and Multivariate analysis. Finally, in data pre-processing phase, pre-processing steps such as, removing outliers or missing values, or replacing them with median, mode, or mean values, etc., or conversion of categorical features into numerical ones to get the dataset ready for the implementation of Machine Learning models.

4.1 Data Description

*Table **Error! No text of specified style in document.**.1 Details of the features of the dataset*

Name	Data Type	Detail
Age	*Integer - 64bit*	*Ages of patients have been stored ranging from 20 to 65.*
Gender	*Object*	*Gender of the patients have been stored – Male/Female*
Ployuria	*Integer - 64bit*	*Whether the patient experienced excessive urination or not.*
Ploydipsia	*Integer - 64bit*	*Whether the patient experienced excessive thirst/excess drinking or not.*
Sudden_weight_loss	*Integer - 64bit*	*Whether patient had an episode of sudden weight loss or not.*
Weakness	*Integer - 64bit*	*Whether patient had an episode of feeling weak.*

Ployphagia	*Integer - 64bit*	*Whether patient had an episode of excessive/extreme hunger or not.*
Genital_Thrush	*Integer - 64bit*	*Whether patient had a yeast infection or not.*
Visual_Blurring	*Integer - 64bit*	*Whether patient had an episode of blurred vision.*
Itching	*Integer - 64bit*	*Whether patient had an episode of itch.*
Irritability	*Integer - 64bit*	*Whether patient had an episode of irritability.*
Detailed Healing	*Integer - 64bit*	*Whether patient had a noticed delayed healing when wounded.*
Partial_Paresis	*Integer – 64bit*	*Whether patient had an episode of weakening of a muscle/group of muscles or not.*
Muscle Stiffness	*Integer - 64bit*	*Whether patient had an episode of muscle stiffness.*
Alopecia	*Integer - 64bit*	*Whether patient experienced hair loss or not.*
Obesity	*Integer - 64bit*	*Whether patient can be considered obese or not using his body mass index.*
Class	*Integer - 64bit*	*Whether patient can be considered obese or not using his body mass index.*

4.1.1 Explanation of Features

Age – age is one of the most common attributes used everywhere in any organization of the world. In the dataset, age is an important factor because it represents talks much about patients in accordance the the particular disease being dealt with – Diabetes.

Gender – Gender is another most important feature in the dataset because it determines the masculinity and femininity of the person, being referred (Wikipedia, 2022).

Polyuria – Ployuria is known as excessive urination in human, and it is usually depending on the age and gender in human being. It is normal for any person to urinate 2 litres per day, however, more than

that is considered excessive urination. This problem is prevalent in diabetic patients. It has been seen that people have diabetic disorder to use more toilet than normal ones (Healthline, 2018).

Polydipsia – In diabetic patients, Polydipsia has been seen a common problem where people keep on drinking more and more water but still the thirst is not fulfilled (Dansinger, M., 2021).

Sudden_weight_loss – sudden weight loss in diabetic patients is also common because when insufficient insulin is provided to the patients of diabetic disorder. The body is not capable of providing enough glucose to the cells of the body, and therefore, the weight of the body starts to decreasing resulting in sudden weight loss (Diabetes, 2022).

Weakness – In diabetic patients, weakness occurs due to same reason of not getting enough glucose to the cells of the body. This results in weakness, and the person having diabetes gets tired very soon (Fletcher, J., and Wood, K., 2022).

Polyphagia - Polyphagia describes excessive hunger. Although we may all feel an increase in appetite in certain situations — such as after exercise or if we haven't eaten in a while — sometimes it can be a sign of an underlying condition. In people with diabetes, glucose can't enter cells to be used for energy (Healthline, 2019).

Genital_Thrush – Candida in the diabetic patients is grown because of sugar which occurs because of Yeast infections. A person having high sugar level results in having high sugar in saliva, sweat, and urine. This encourages the yeast to grow, that eventually ends up in a thrush (Dansinger, M., 2022).

Visual_Blurring – a diabetic patient’s eye lenses can be blurred due to swelling because of sudden change in the blood sugar levels from Low to normal. This results in blurred vision in diabetic patients, because of which they aren’t able to see things clearly. This goes back to normal when the sugar level in blood stabilizes, and the patient’s eye lenses become normal, and he/she is able to see things clearly (Cai, C., X., 2022).

Itching – Itching has been found a common symptom in diabetic patients where people have skin problems such as dry skin, yeast infection, and poor blood circulations. Most importantly, the lower legs feel a lot of itching because the blood is not circulated in the body normally. So, itching happens in diabetic patients (Diabetes, 2022).

Irritability – Mood swings is also a common problem found in diabetic patients, where people have high blood sugar levels feel irritable due to stress, pressure, and current condition of mind (mental health condition). Often times people say that diabetes affect the physical health of the human body, however, it also causes several other problems that are not inevitable (Healthline, 2022).

Delayed Healing – For normal human beings, when a wound happens to any part of the body, the antibodies react to it, and body itself starts healing the wound itself. However, this is not the case with people having diabetes because their wounds are not healed because due low blood circulation in the body, the body is not able to provide enough nutrients to the wound in turn making it difficult for the body to get heal by itself. As a result, people having diabetes or high blood sugar levels tend to heal slowly on wounds. Above all, diabetic patients also suffer from neuropathy that affects wound healing (Medical News Today, 2019).

Partial_Paresis – Diabetic disorder or high blood sugar level can affect every part of your body, in which one problem is slow digestion. When diabetes affects your stomach, it makes the digestion process slow, and prolongs the digestion of food – Food stays longer inside the body. To be precise, Vagus nerve is the particular term in human body that controls the stomach, and when it is damages, so happens paresis happens to the human body. This condition is also known as gastroparesis, which makes you feel vomit or queasy (WebMd, 2022).

Muscle Stiffness – Another serious problem that affects diabetic people is ‘Muscle Stiffness’, which includes joint pains, muscle pains, deformities, joint swelling, pinning inside body etc., all of this is

also called musculoskeletal problems that are unique to diabetic patients. Therefore, people with diabetes often complain about these kinds of problems, and feel irrigatable (Diabetes Journals, 2001).

Alopecia – Alopecia is a disease related to human hair, where people having Alopecia find their heads hair falling slowly resulting in patches of hair on the head, and other parts of the body. This disease is common in people having type 1 diabetic disorder. This is caused because immune system starts attacking the hair follicles (Healthline, 2022).

Obesity - Corpulence is the main gamble factor for type 2 diabetes. The Centres for Disease Control and Prevention report that 32% of white and 53% of people of colour are corpulent. Ladies with a weight record (BMI) of 30 kg/m2 have a 28 times more serious gamble of creating diabetes than do ladies of typical weight (Barnes, A., S., and Coulter, S., 2010).

Class – This is the target attribute of the dataset in which a person having the diabetes or not would be determined. Target feature comes among the most important feature in Machine learning that helps in determining the final result based on various features available in the dataset.

4.2 Exploratory Data Analysis

It is better to try to understand the data first, and get as many insights as possible before actually experimenting on the dataset). This is possible with the help of variable visualization libraries such as Pandas, Numpy, Seaborn, Plotly, etc. that are used for data manipulation and analysis, mathematical operations, and graphical visualizations - Uni-variate (observation of two variables with respect to target feature), Bi-variate (observation of two variables with respect to target feature), and multi-variate analysis (this involves observing two or more number of features). During this phase, we learn the relations, and correlations between different features of the dataset, and understand the significance of each important feature that directly affects the outcome of this project. Therefore, all experiments conducted in Exploratory Data Analysis (EDA) are driven towards answers to the questions related to the aims, and objectives of the project.

In fig. 4.1, various python libraries are being imported into Jupyter's notebook for the purpose of data analysis, and pre-preparation, which help in building the Machine Learning models to answer the questions of the study. For example, *Pandas* is a famous library, which helps to import the Excel or CSV files into the Jupyter's notebook and helps perform operations on the dataset very easily, *Seaborn* is a library which helps in plotting different kinds of visuals to perform the analysis on the dataset, and *Numpy* is an amazing library which helps in performing mathematical operations on the data. These all libraries will help us in conquest to answer the research questions, and will help us in achieving the goals until completion of the study.

```
import warnings
warnings.filterwarnings('ignore')
import numpy as np
import pandas as pd
from sklearn.preprocessing import MinMaxScaler
from sklearn.model_selection import train_test_split,cross_val_score
from sklearn.linear_model import LogisticRegression
from sklearn.metrics import  accuracy_score, f1_score, precision_score,confusion_matrix, recall_score, roc_auc_score
from xgboost import XGBClassifier
from sklearn.ensemble import RandomForestClassifier,AdaBoostClassifier
from sklearn.svm import SVC
import matplotlib.pyplot as plt
%matplotlib inline
from IPython.display import Image
```

*Figure **Error! No text of specified style in document.**.6 Importing Libraries*

Next, in fig. 4.2, the dataset is being read into the Jupyter's notebook using the read_csv() function of Pandas library, which helps us read the dataset, and makes it easy for us to perform different kinds of data manipulation operations. In fig. 9, diabetes_data.csv file is being loaded into jupyter's notebook using read_csv() function. This function reads the CSV (Comma Separated File) using read_csv() function, and stores it into a variable: df. This variable will be used to perform different kinds of computations because it contains the whole dataset.

```
df = pd.read_csv('diabetes_data.csv',sep=';')
df.head()
```

*Figure **Error! No text of specified style in document.**.7 Importing Diabetes Dataset: CSV File*

In fig. 4.3, the working of head() function displays the first five rows of the dataset. It can be seen that column names written at the top, i.e., age, gender, polyuria, polydipsia, sudden_weight_loss, weakness, polyphagia, genital_thrush, visual_blurring, itching, irritability, delayed_healing, etc., and also the field values below, which displays the information recorded in those columns. In the dataset, it can be seen that except for gender feature which is of object data type, all others are numerical features having only 0 or 1 values. However, age feature contains values in between 20 to 65 years of age. The most important feature in the dataset is the class column because it represents the labelled features which decides whether the person has the diabetes or not.

	age	gender	polyuria	polydipsia	sudden_weight_loss	weakness	polyphagia	genital_thrush	visual_blurring	itching	irritability	delayed_healing	partial_par
0	40	Male	0	1	0	1	0	0	0	1	0	1	
1	58	Male	0	0	0	1	0	0	1	0	0	0	
2	41	Male	1	0	0	1	1	0	0	1	0	1	
3	45	Male	0	0	1	1	1	1	0	1	0	1	
4	60	Male	1	1	1	1	1	0	1	1	1	1	

*Figure **Error! No text of specified style in document.**.8 Head() function of Pandas Library*

Next, another most important thing that matters in the dataset, whether it's being used in any field of study, the number of missing values present in the dataset. This is because if the dataset contains missing or null values, it could be a very critical problem with respect to the study being conducted. In Machine learning, if there are a greater number of Null values present in the dataset, the accuracy of the predictive models is affected greatly. Therefore, in every machine learning project, pre-processing steps are performed on the null values, where the values are either removed from the dataset, or mean, mode, or median values are replaced in place of them to handle the problem of missing values.

In fig. 4.4, it can be seen that Pandas libraries function isna().sum() is being used that takes the sum of null values in the dataset and prints the values Infront of the feature names as indicated. It can be seen that there are no null values present in the features: age, gender, polyuria, polydipsia,

sudden_weight_loss, weakness, polyphagia, genital_thrush, etc., of the dataset. Every feature has 0 number of missing values present in the dataset.

```
df.isna().sum()
```

```
age                 0
gender              0
polyuria            0
polydipsia          0
sudden_weight_loss  0
weakness            0
polyphagia          0
genital_thrush      0
visual_blurring     0
itching             0
irritability        0
delayed_healing     0
partial_paresis     0
muscle_stiffness    0
alopecia            0
obesity             0
class               0
dtype: int64
```

*Figure **Error! No text of specified style in document.**.9 Sum of null values in the dataset*

Now, it is also very important to read the general information regarding the dataset using Pandas info() function of the dataset. This function represents general information of the dataset and it's features in which it shows total number of entries – rows of the dataset, feature – number of columns of the dataset, memory usage, etc., so in diabetes dataset, there are 17 of total columns including, age at index 0, gender at index 1, polyuria at index 2, sudden_weight_loss at index 4, weakness at index 5, polyphagia at index 6, genital_thrush at index 7, visual_blurring at index 8, itching at index 9, irritability at index 10, delayed_healing at index 11, partial_paresis at index 12, muscle_stiffness at index 13, alopecia at index 14, obesity at index 15, and class at index 16. Total number of entries in the dataset are 520,

where gender has the ‘object’ data type, while all others have integer dataset of 64bit. Moreover, the total memory usage is 69.2+ KB.

```
df.info()
```

```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 520 entries, 0 to 519
Data columns (total 17 columns):
 #   Column              Non-Null Count  Dtype 
---  ------              --------------  ----- 
 0   age                 520 non-null    int64 
 1   gender              520 non-null    object
 2   polyuria            520 non-null    int64 
 3   polydipsia          520 non-null    int64 
 4   sudden_weight_loss  520 non-null    int64 
 5   weakness            520 non-null    int64 
 6   polyphagia          520 non-null    int64 
 7   genital_thrush      520 non-null    int64 
 8   visual_blurring     520 non-null    int64 
 9   itching             520 non-null    int64 
 10  irritability        520 non-null    int64 
 11  delayed_healing     520 non-null    int64 
 12  partial_paresis     520 non-null    int64 
 13  muscle_stiffness    520 non-null    int64 
 14  alopecia            520 non-null    int64 
 15  obesity             520 non-null    int64 
 16  class               520 non-null    int64 
dtypes: int64(16), object(1)
memory usage: 69.2+ KB
```

*Figure **Error! No text of specified style in document.**.10 info() function of the dataset*

Next, the most important thing that matters is the distribution of the class attribute in the dataset. For this purpose, Seaborn library is being to helps in plotting different kinds of visuals to perform the analysis on the dataset. In fig. 4.6, a countplot() function is being used using the sns variable of the Seaborn library, which displays the distribution of the class attributes. It is important to mention here that 0 indicates: no in diabetes, while 1 indicates: Yes in diabetes. In addition, 0 and 1 class is being shown in the x-axis, while the total number of values are being represented in the y-axis. In the count plot, blue represents 0 class label, while orange represents 1 class label. For 0 class label, two hundred records are present in the dataset, and for 1 class label, there are three hundred and twenty of total

entries are present. The distribution looks a bit imbalanced, therefore, in pre-processing phase, Smote library would be used to balance the dataset, which uses pipeline to under sample and over sample the dataset depending upon the amount of data available in the dataset.

```
import seaborn as sns
sns.countplot(df['class'],data=df)
```

```
<AxesSubplot:xlabel='class', ylabel='count'>
```

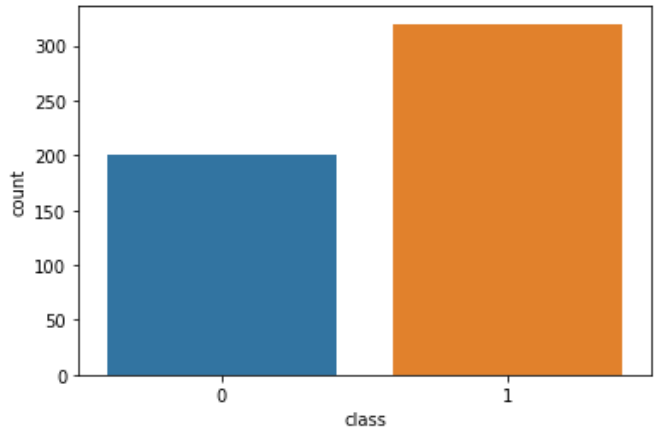


*Figure **Error! No text of specified style in document.**.11 Class Attribute: 0 or 1*

In addition, in fig. 4.7 the distribution of the target variable is being shown as a representation in pie chart showing the percentage of features, and a bar plot showing the total number of entries. In the pie-chart, positive class attribute values are being represented in the yellow colour, while negative class attribute values are being represented in the blue colour. In addition, the bar plot is being shown in blue colour. In the target variable, there are total of 62% positive class attributes, while 38% of negative class attributes. The dataset is imbalanced, whereas in the bar plot, 200 values are being represented for 0 or negative values, while 320 values are being represented for 1 or positive values.

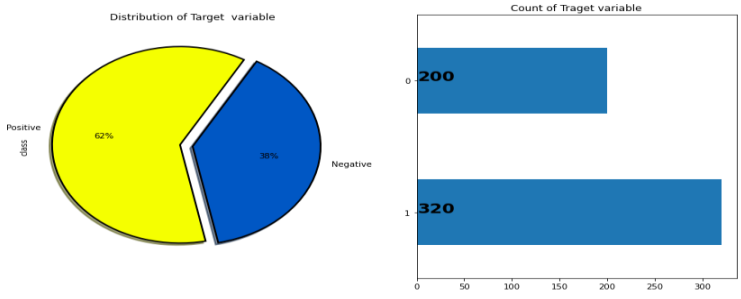


*Figure **Error! No text of specified style in document.**.12 Pie Chart and Histogram*

Similarly, in fig. 4.8, gender feature in the dataset is being shown with Male and Female with respect to the class attribute. On x-axis, 0 class label is being represented in the blue colour, while 1 class label is being represented in the orange colour. On y-axis, the total number of count values are being represented. In male people, it can be seen that, there are a greater number of 0 class label approximately close to 175, while 1 class label is less which is close to 140. In contrast, in females, there are a greater number of diabetic patients as compared to non-diabetic ones. The difference is very huge: females having diabetes have almost 175 entries, while females not having diabetes are close to 25 entries.

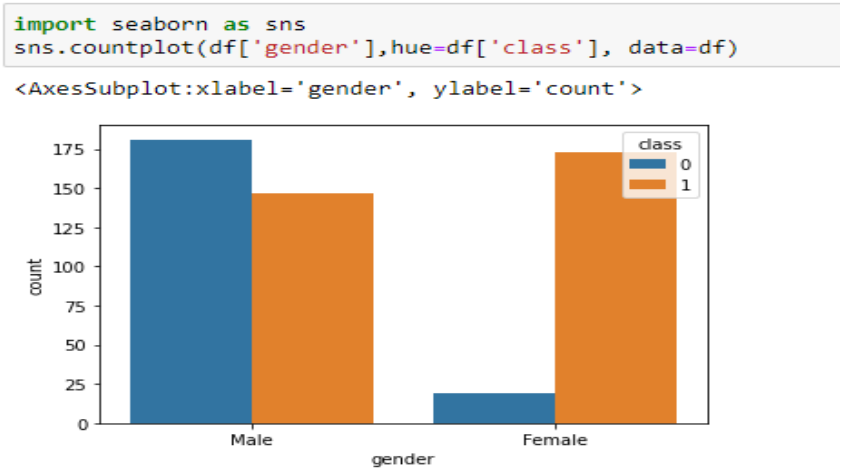


*Figure **Error! No text of specified style in document.**.13 Gender: Male and Female*

In fig. 4.9, a count plot of Sudden Weight Loss can easily be seen with class label 0 and 1. Sudden weight loss in diabetic patients is also common because when insufficient insulin is provided to the patients of diabetic disorder. The body is not capable of providing enough glucose to the cells of the body, and therefore, the weight of the body starts to decreasing resulting in sudden weight loss (Diabetes, 2022). On x-axis, 0 class label is being represented in the blue colour, while 1 class label is being represented in the orange colour. It can be seen from the count plot below that people were diabetic experienced a sudden weight loss as compared to those who weren't diabetic.

class	0	1
sudden_weight_loss		
0	85.500000	41.250000
1	14.500000	58.750000

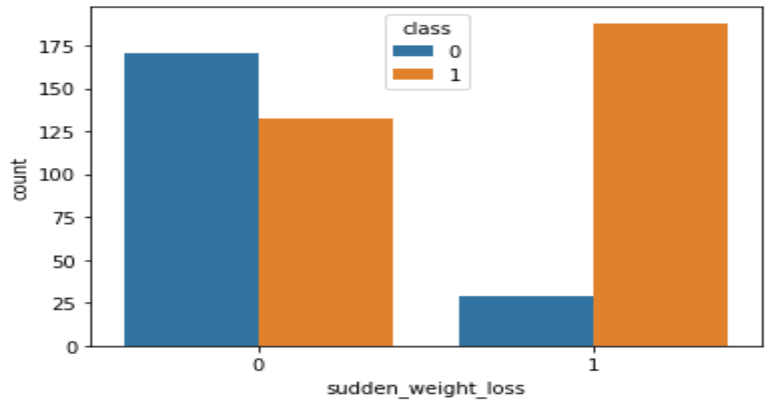


*Figure **Error! No text of specified style in document.**.14 Sudden Weight Loss: Class labels 0 and 1*

In fig. 4.10, the count plot of Polyphagia can be seen with class labels 0 or 1. - Polyphagia describes excessive hunger. Although we may all feel an increase in appetite in certain situations — such as after exercise or if we haven't eaten in a while — sometimes it can be a sign of an underlying condition. In people with diabetes, glucose can't enter cells to be used for energy (Healthline, no date). On x-axis, 0 class label is being represented in the blue colour, while 1 class label is being represented in the orange colour. It can be seen from the count plot below that people who had diabetes, the number of people having Polyphagia was more as compared to those not having diabetes. There were 59% of people having Polyphagia in diabetes, and 24% of people didn’t have polyphagia.

class	0	1
polyphagia		
0	76.000000	40.940000
1	24.000000	59.060000

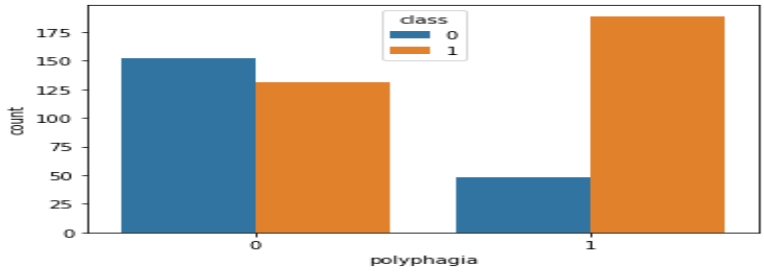


*Figure **Error! No text of specified style in document.**.15 Polyphagia: Class labels 0 and 1*

In fig. 4.11, the count plot of weakness can easily be seen with respect to class labels 0 and 1. Weakness occurs due to same reason of not getting enough glucose to the cells of the body. On x-axis, 0 class label is being represented in the blue colour, while 1 class label is being represented in the orange colour. In the count plot below, it can be seen that, almost equivalent amount of weakness was observed in people who didn't have diabetic problem. However, almost 50% people complained about having weakness for who those fighting through diabetes disease.

class	0	1
weakness		
0	56.500000	31.870000
1	43.500000	68.120000

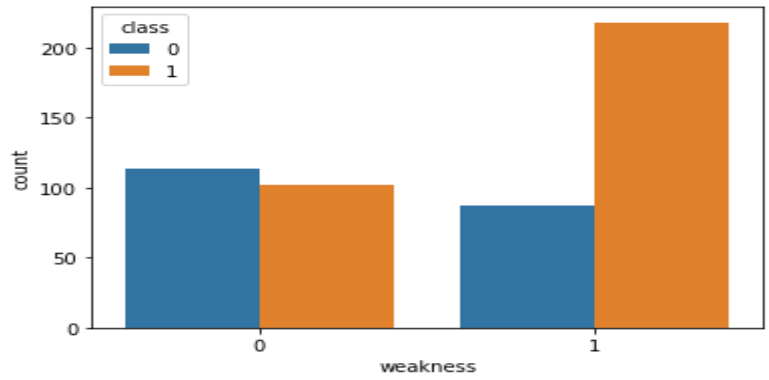


*Figure **Error! No text of specified style in document.**.16 Weakness: Class labels 0 or 1*

In fig. 4.12, the count plot of genital_thrush can be seen with class labels with 0 or 1. Candida in the diabetic patients is grown because of sugar which occurs because of Yeast infections. A person having high sugar level results in having high sugar in saliva, sweat, and urine. This encourages the yeast to grow, that eventually ends up in a thrush (Dansinger, M., 2022). On x-axis, 0 class label is being represented in the blue colour, while 1 class label is being represented in the orange colour. In the count plot below, it can be seen that people not having diabetes were found to have genital thrush more compared to people having diabetes disorder. In diabetic people, 74% of people didn't have genital thrush as compared to those non-diabetic patients, where only 25% of people had genital thrush. In contrast, 83% of people didn't have genital thrush who didn't have diabetes problem as compared to those who had genital thrush were 16%.

class	0	1
genital_thrush		
0	83.500000	74.060000
1	16.500000	25.940000

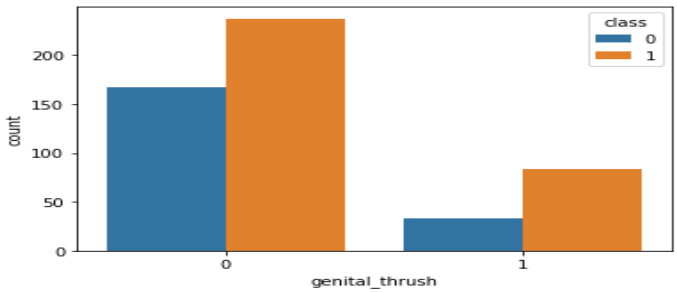


*Figure **Error! No text of specified style in document.**.17 Genital Thrush: Class label 0 or 1*

In fig. 4.13, the count plot of visual blurring can be seen with class labels 0 or 1. a diabetic patient's eye lenses can be blurred due to swelling because of sudden change in the blood sugar levels from Low to normal. This results in blurred vision in diabetic patients, because of which they aren't able to see things clearly (Cai, C., X., 2022). On x-axis, 0 class label is being represented in the blue colour, while 1 class label is being represented in the orange colour. In the count plot below, it can be seen that almost an equal amount of blurring was seen in people who didn't have diabetes, and a huge amount of visual blurring was seen in people who had been diagnosed with diabetes disorder. To be precise, people who were found diabetic were found having visual blurring 54% of the times, while 45% of times were found as not having visual blurring. Similarly, people who were not diabetic were found having visual blurring 71% of the times, and 29% of the times were not found having visual blurring.

class	0	1
visual_blurring		
0	71.000000	45.310000
1	29.000000	54.690000

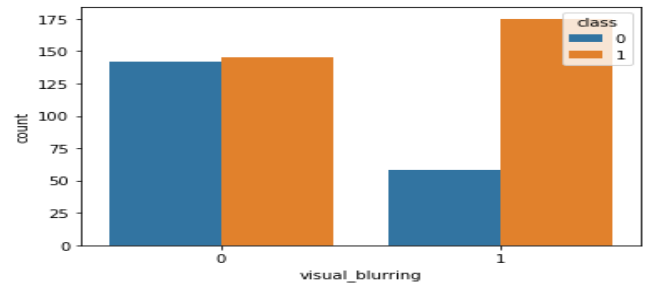


*Figure **Error! No text of specified style in document.**.18 Visual Blurring: Class label 0 or 1*

In fig. 4.14, the count plot of itching is being shown with class labels 0 or 1. Itching has been found a common symptom in diabetic patients where people have skin problems such as dry skin, yeast infection, and poor blood circulations (Diabetes, no date). On x-axis, 0 class label is being represented in the blue colour, while 1 class label is being represented in the orange colour. It can be seen from the count plot below that people who didn't have the problem of itching were found almost equal in both the classes of patients having diabetes (51%) and not having diabetes (50%). On the other hand, people who had been facing the problems of itching, were also almost equal in both classes of patients having diabetes (48%), and patients not having diabetes (49%).

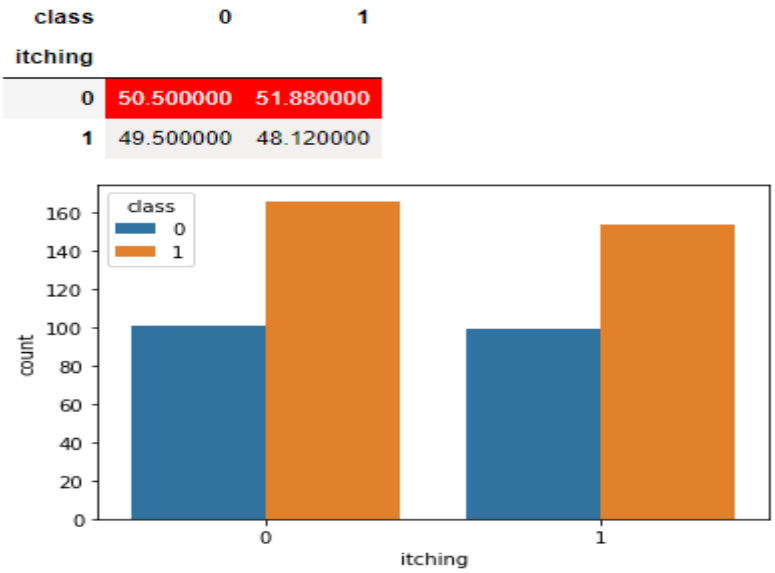


*Figure **Error! No text of specified style in document.**.19 Itching: Class label 0 or 1*

In fig. 4.15, the count plot of irritability is being shown with class labels 0 or 1. On x-axis, 0 class label is being represented in the blue colour, while 1 class label is being represented in the orange colour. In the count plot below, it can be seen that, people not having irritability were found 92% of the times not having diabetes, and people having diabetes were found 65% of the times having irritability problem. In contrast, people having diabetes were found to have 34% of the times having irritability problem, while only 8% of the times, they weren't having irritability problem.

class	0	1
irritability		
0	92.000000	65.620000
1	8.000000	34.380000

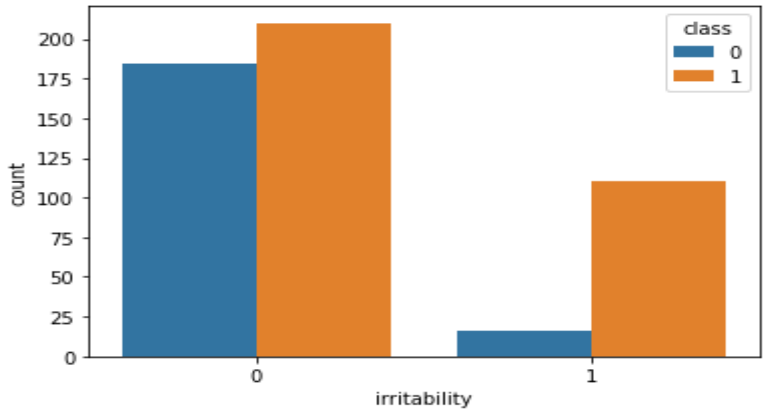


*Figure **Error! No text of specified style in document.**.20 Irritability: Class label 0 or 1*

In fig. 4.16, the count plot of delayed_healing is being shown with class labels 0 or 1. On x-axis, 0 class label is being represented in the blue colour, while 1 class label is being represented in the orange colour. In the count plot below, it can be seen that people having diabetes, were found having more cases of delayed healing as compared to those who didn't have diabetes problem.

class	0	1
delayed_healing		
0	57.000000	52.190000
1	43.000000	47.810000

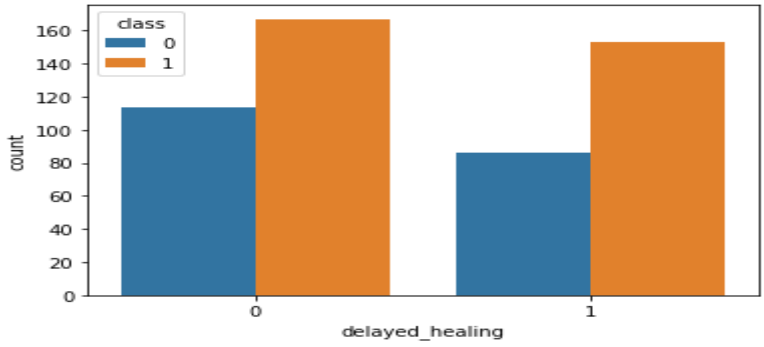


*Figure **Error! No text of specified style in document.**.21 Delayed Healing: Class label 0 or 1*

In fig. 4.17, the count plot of partial_paresis is being shown with class labels 0 or 1. Diabetic disorder or high blood sugar level can affect every part of your body, in which one problem is slow digestion. When diabetes affects your stomach, it makes the digestion process slow, and prolongs the digestion

of food – Food stays longer inside the body (WebMD, no date). On x-axis, 0 class label is being represented in the blue colour, while 1 class label is being represented in the orange colour. In the count plot below, it can be seen that people having diabetes disorder were found to have a greater number of people facing the problem of partial paresis. Almost 60% of people were found to have the problem, while 16% of people were not experiencing the problem. In addition, for people not having paresis, the diabetes patient was found lower as compared to non-diabetic patients.

class	0	1
partial_paresis		
0	84.000000	40.000000
1	16.000000	60.000000

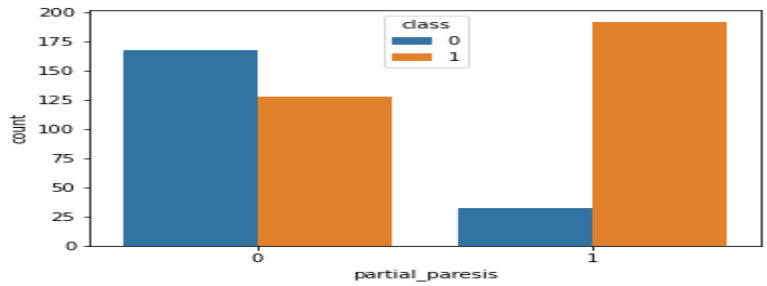


*Figure **Error! No text of specified style in document.**.22 Partial Paresis: Class label 0 or 1*

In fig. 4.18, the count plot of muscle_stiffness is being shown with class labels 0 or 1. On x-axis, 0 class label is being represented in the blue colour, while 1 class label is being represented in the orange colour. It can be seen from the graph below that people having diabetes experienced a greater number of cases reporting muscle stiffness 42%, while 30% of people didn't have the problem.

class	0	1
muscle_stiffness		
0	70.000000	57.810000
1	30.000000	42.190000

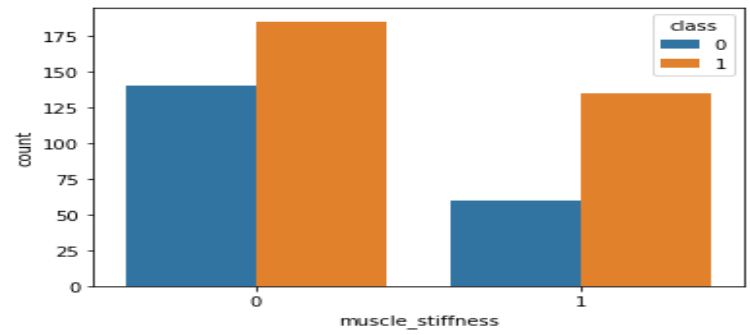


*Figure **Error! No text of specified style in document.**.23 Muscle Stiffness: Class label 0 or 1*

In fig. 4.19, the count plot of alopecia is being shown with class labels 0 or 1. Alopecia is a disease related to human hair, where people having Alopecia find their heads hair falling slowly resulting in patches of hair on the head, and other parts of the body (Healthline, no date). On x-axis, 0 class label is being represented in the blue colour, while 1 class label is being represented in the orange colour. It can be seen that people who didn't had alopecia, the diabetic patients were found more in numbers (75%), while people not having diabetes were found less in number. Similarly, people having alopecia, the diabetic patients were found almost 1% less in number as compared to those not having alopecia.

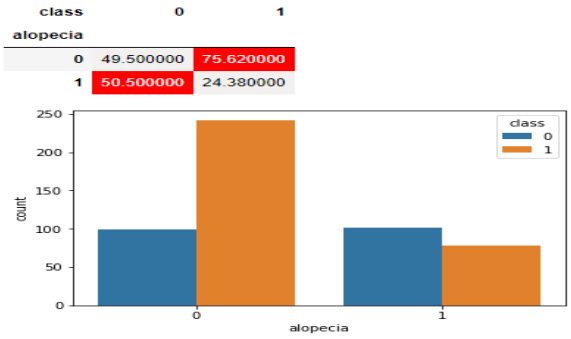


*Figure **Error! No text of specified style in document.**.24 Alopecia: Class label 0 or 1*

In fig. 4.20, the count plot of obesity is being shown with class labels 0 or 1. On x-axis, 0 class label is being represented in the blue colour, while 1 class label is being represented in the orange colour. It can be seen that diabetes was found more in people not being obesity, as compared to those where people were obese. 86% of people were found to have diabetes as compared to 13% of people not having diabetes in people having low fatty body.

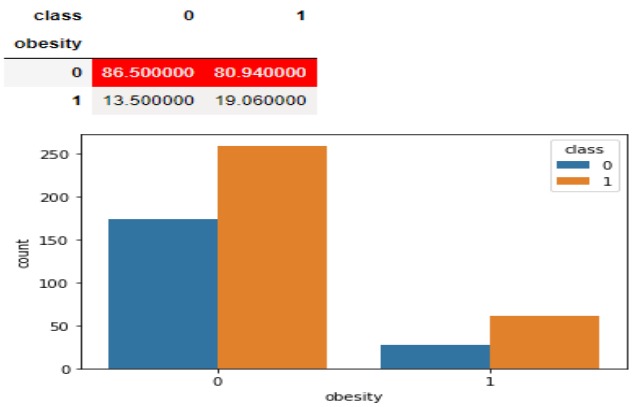


*Figure **Error! No text of specified style in document.**.25 Obesity: Class label 0 or 1*

4.3 Data pre-processing

Data pre-processing is the process of transforming the raw data and making it useful for the implementation of Machine Learning models. It is the basis necessity, and important phase that has to be carried out in Machine learning for the successful implementation of Machine Learning models. This is because it includes several important processes which are removing outliers present in the dataset, dealing with missing or redundant values, conversion of categorical features into numerical ones, splitting the dataset into testing and training, defining input and output as x and y. There are several libraries that provide useful functions that help to perform the data pre-processing on the dataset. All of the pre-processing is based on the EDA (Exploratory Data Analysis) because, there we are able to know the dataset, and features/values present in the dataset. Then, depending upon the type of project, the dataset is pre-processed for the model's implementation. Pre-processing helps the predictive models in many ways, and one of them is the increased accuracy of the predictive model. As in this project, a machine learning model is to be developed which predicts early diabetes in humans using machine learning, therefore, depending upon the Exploratory Data Analysis performed, the pre-processing of the dataset shall be performed.

4.3.1 Label Encoding

Label Encoding is pre-requisite and important in superivised learning approach before implementing Machine Learning models. It is used to convert Categorical features into numerical values or you can say machine readable form. This is because computers can only understand numerical values, and then can easily decide how the values would be operated. Therefore, to perform Label Encoding, Sklearn library has been used, and pre-processing has been imported to initialize Lable Encoder. A variable 'le' has been initialized using LableEncoder() function from pre-processing. Next, fit_transform() is the function that has been given a parameter called 'gender' from the data frame because this feature is of object data type, and in this study, it will be converted into numerical form to implement Machine Learning models. The whole line of code for performing label encoding can be seen in fig. 4.21.

```
from sklearn import preprocessing
le = preprocessing.LabelEncoder()
df['gender']=le.fit_transform(df['gender'])
```

*Figure **Error! No text of specified style in document.**.26 Label Encoder: Encoding of Categorical Features*

4.3.2 Smote Library: Oversampling & Under sampling

In EDA (Exploratory Data Analysis), it was observed that the dataset was imbalanced. There were 62% positive class attributes, and 38% negative class attributes in the target features. This poses a serious problem when a machine learning model would be developed based on this dataset because it is more inclined towards positive class attributes, and less inclined towards negative attributes. Therefore, there's a need to perform oversampling and undersampling of the data frame to distribute evenly the percentage of features availability.

In fig. 4.22, a code is being displayed that uses 'imblearn' library to perform both oversampling and undersampling of the dataset. SMOTE is being used for the purpose of over_sampling, while Random Under Sampler is being used for under sampling. In addition, Pipeline has been imported form 'imblearn' library to perform the sampling of the dataset. For Smote, a random state of 10000, has been chosen as a parameter, and it can be seen that pipeline has been used to use the function fit_resample() for the purpose of performing the sampling. Now, the dataset would be equally sampled, and it will have no problem in the predictive accuracy of the Machine Learning models.

```
# Oversample with SMOTE and random undersample for imbalanced dataset
from imblearn.over_sampling import SMOTE
from imblearn.under_sampling import RandomUnderSampler
from imblearn.pipeline import Pipeline
from matplotlib import pyplot
from numpy import where
# define pipeline
over = SMOTE(random_state=10000)

steps = [('over', over)]

pipeline = Pipeline(steps=steps)
# transform the dataset
X, y = pipeline.fit_resample(X, y)
```

*Figure **Error! No text of specified style in document.**.27 Smote Library: Oversampling and Under sampling*

4.3.3 Concatenating X and Y

Now, in my project, I'm implemented AutoML Technique for the purpose of model's implementation. Therefore, I'm concatenating 'X' and 'y' data frame with axis=1, and assigning it to the variable df. This is because, AutoML would be provided with the whole data frame, with a final.csv file. In fig. 4.23, it can be seen that concat() function of Pandas library is being used to concatenate both 'X', and 'y' data frame and df is being assigned.

```
df=pd.concat([X,y],axis=1)
```

*Figure **Error! No text of specified style in document.**.28 Concatenating X and y for AutoML*

4.3.4 Saving a Comma Separated File

Now, in fig. 4.24, it can be seen that 'final.csv' file is being generated by function to_csv() function. This file would be used by AutoML for the implementation of Machine Learning models.

```
df.to_csv('final.csv')
```

*Figure **Error! No text of specified style in document.**.29 Saving a Comma Separated File: file.csv*

4.3.5 Checking if Data is now Balanced

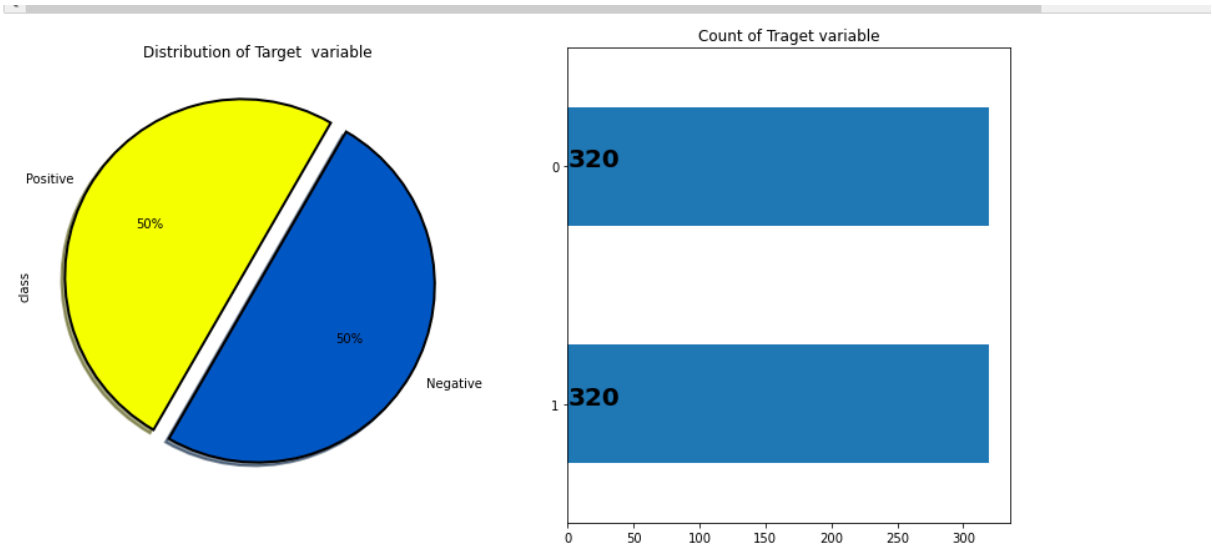
To confirm that data is now balanced, I used the script in the screen shot below to check it. From the pie and bar chart in the screen shot below, you could see that the data is balanced after executing SMOTE script and generating the final.csv file above. Since the data is now balanced, I will be able to implement AutoML technique by using H20AutoML from h20.automl library.

```
In [109]: # plotting to create pie chart and bar plot as subplots to check if data is now balanced
plt.figure(figsize=(14,7))
plt.subplot(121)
df["class"].value_counts().plot.pie(autopct = "%1.0f%%",colors = sns.color_palette("prism",7),startangle = 60,labels=["Posit
wedgeprops={"linewidth":2,"edgecolor":"k"},explode=[.1,0],shadow =True)
plt.title("Distribution of Target  variable")

plt.subplot(122)
ax = df["class"].value_counts().plot(kind="barh")

for i,j in enumerate(df["class"].value_counts().values):
    ax.text(.7,i,j,weight = "bold",fontsize=20)

plt.title("Count of Traget variable")
plt.show()
```



4.3.6 Importing H20AutoML and Parsing CSV File

Before implementing AutoML technique in Machine Learning, there's a need to import H20AutoML from h20.automl which is the library that helps in importing Comma Separated Values (CSV). It can be seen in fig. 4.25 that import_file() function is being used to import 'final.csv' file into the data, and it has successfully been imported.

```
from h2o.automl import H2OAutoML
```

```
data = h2o.import_file("final.csv")
```

```
Parse progress: |████████████████████████████████████████████████████████████████| (done) 100%
```

*Figure **Error! No text of specified style in document.**.30 Imporing H20 AutoML*

4.3.7 Defining Input and Output as x and y

Next, in fig. 4.26, input and output variables are being defined for the implementation of models. This is because in this study, I want machine learning models to predict the diabetes based on certain features that are available in the dataset. So, for this purpose, 'X' is a variable defined with having a data frame containing al features except 'class' features – target feature. Now, 'y' is another variable that has been defined but with only a 'class' data frame. This will help predictive models to split the

dataset for testing and training as well, and based on the features available in the ‘X’ variable, ‘y’ would be predicted.

```
x = train.columns
y = "class"
x.remove(y)
```

*Figure **Error! No text of specified style in document.**.31 Defining Input and Output as x and y*

4.3.8 Splitting Dataset: Training and Validation

Now, before implementing Machine Learning models that would be implemented using the AutoML technique, the dataset need to be split into training and validation sets, which would be carried out by split_frame() function with a ratio of 80%, and seed value of 1234. Therefore, now 80% of the dataset would be used for training, and remaining 20% would be used for testing and validation. The seed value is given to the split_frame() function because every time it must split the dataset the same way it did every time.

```
 # split into train and validation sets
train, valid = data.split_frame(ratios = [.8], seed = 1234)
```

*Figure **Error! No text of specified style in document.**.32 Splitting the Dataset into Testing and Training*

4.3.9 Converting asfactor() train[y] and valid[y]

H20AutoML requires training and validation sets to be converted into categorical features using asfactor() function. This function helps to convert the numerical features into categorical ones, and this how things work with H20AutoML.

```
train[y] = train[y].asfactor()
valid[y] = valid[y].asfactor()
```

*Figure **Error! No text of specified style in document.**.33 Converting asfactor() train[y] and valid[y]*

Chapter 5: Critical Evaluation

In this chapter, critical evaluation of the implemented approach shall be performed. In this study, AutoML approach has been implemented, which given a seed – parameters, automatically implements the Machine Learning models, and performs the evaluation based on the type of models implemented. In the earlier chapters, EDA (Exploratory Data Analysis), and Data Pre-Processing has been implemented on the dataset resulting in getting the dataset ready for the implementation of Machine Learning models. Therefore, let's now implement AutoML, and see critically evaluate the models.

In fig. 5.1, AutoML technique has been implemented using H20AutoML with Max models equal to 20, and Seed value equal to 12.

```
from h2o.automl import H2OAutoML

aml = H2OAutoML(max_models = 20, seed = 12)
aml.train(x =x, y = y, training_frame = train, validation_frame=valid)
```

```
AutoML progress: |
20:34:45.513: User specified a validation frame with cross-validation still enabled. Please note that the models will st
ill be validated using cross-validation only, the validation frame will be used to provide purely informative validation
metrics on the trained models.
20:34:45.518: AutoML: XGBoost is not available; skipping it.

| (done) 100%
Model Details
=============
H2OGradientBoostingEstimator :  Gradient Boosting Machine
Model Key:  GBM_grid_1_AutoML_3_20221001_203445_model_5


Model Summary:
```

	number_of_trees	number_of_internal_trees	model_size_in_bytes	min_depth	max_depth	mean_depth	min_leaves	max_leaves	mean_leaves
0	255.0	255.0	95315.0	5.0	12.0	8.819608	18.0	28.0	24.988235

*Figure **Error! No text of specified style in document.**.34 H20AutoML Implementation*

H20AutoML successfully implemented several important Machine Learning models, such as GBM (Gradient Boosting Machine), DRF (Distributed Uplift Random Forest), Deep Learning, GLM (Generalized Linear Model), etc., and some other versions of the them.

In fig. 5.2, it can be seen that H20AutoML's Leatherboard is being displayed showing different ids of models, Area Under the Curve (AUC), Log Loss, Area Under the Curve Precision-Recall (AUCPR), Mean Per Class Error, Root Mean Squared Error (RMSE), and Mean Squared Error (MSE) etc., and binary classification is being performed in this project, therefore, RMSE, and MSE, are not the

performance matrix to be discussed. Therefore, in this project, I would only use AUC, Log Loss, AUCPR, and Mean Per Class Error to declared the best model.

By analysing the Leaderboard below, it can be seen that, ***GBM_grid_1_AutoML_3_20221001_203445_model_5*** is at the top performed outstandingly with an Area under the curve of 99.9%, Log Loss of 0.015, Area Under the Curve Precison-Recall of 99.9%, and Mean Per Class Error of 0.003.

GBM_grid_1_AutoML_3_20221001_203445_model_2 scored the second spot with a higher AUC of 99.98%, less Logloss of 0.022, and mean per class error of 0.005. It seems that there's a slight difference in reading compared to the best model, however, the difference is not significant. Its just a few decimal points away.

StackedEnsemble_BestOfFamily_1_AutoML_3_20221001_203445 scored the third spot with a higher AUC of 99.98%, less Logloss of 0.019, and mean per class error of 0.005.

GBM_4_AutoML_3_20221001_203445 scored the 4th spot with a higher AUC of 99.98%, less Logloss of 0.023, and mean per class error of 0.005.

StackedEnsemble_AllModels_1_AutoML_3_20221001_203445 scored the 5th spot with a higher AUC of 99.97%, less Logloss of 0.020, and mean per class error of 0.005.

GBM_2_AutoML_3_20221001_203445 scored the 6th spot with a higher AUC of 99.97%, less Logloss of 0.025, and mean per class error of 0.007.

GBM_3_AutoML_3_20221001_203445 scored the 7th spot with a higher AUC of 99.96%, less Logloss of 0.028, and mean per class error of 0.007.

GBM_5_AutoML_3_20221001_203445 scored the 8th spot with a higher AUC of 99.93%, less Logloss of 0.034, and mean per class error of 0.013.

GBM_grid_1_AutoML_3_20221001_203445_model_1 scored the 9th spot with a higher AUC of 99.93%, less Logloss of 0.030, and mean per class error of 0.005.

GBM_grid_1_AutoML_3_20221001_203445_model_4 scored the 10th spot with a Log Loss of 0.034, AUC of 99.92%, and Mean Per Class Error of 0.011.

DRF_1_AutoML_3_20221001_203445 scored the 11th spot with a Log Loss of 0.066, AUC of 99.89%, and Mean Per Class Error of 0.013.

GBM_grid_1_AutoML_3_20221001_203445_model_3 scored the 12th spot with a Log Loss of 0.072, AUC of 99.84%, and Mean Per Class Error of 0.021.

GBM_1_AutoML_3_20221001_203445 scored the 13th spot with a Log Loss of 0.081, AUC of 99.72%, and Mean Per Class Error of 0.025.

DeepLearning_grid_2_AutoML_3_20221001_203445_model_1 scored the 14th spot with a Log Loss of 0.100, AUC of 99.49%, and Mean Per Class Error of 0.021.

DeepLearning_grid_3_AutoML_3_20221001_203445_model_1 scored the 15th spot with a Log Loss of 0.111, AUC of 99.43%, and Mean Per Class Error of 0.021.

DeepLearning_grid_1_AutoML_3_20221001_203445_model_1 scored the 15th spot with a Log Loss of 0.113, AUC of 99.28%, and Mean Per Class Error of 0.027.

Similarly, ***DeepLearning_grid_2_AutoML_3_20221001_203445_model_2*** was at 17th position with a Log Loss of 0.156, AUC of 99.15%, and Mean Per Class Error of 0.035.

XRT_1_AutoML_3_20221001_203445 was at 18th position with a Log Loss of 0.387, AUC of 98.07%, and Mean Per Class Error of 0.041.

The Log Loss error was seen higher at 14th model in **DeepLearning_grid_2_AutoML_3_20221001_203445_model_1** with a value of 0.10, AUC was

99.49%, while Mean Per Class Error was 0.021. After this model, an increase in the Log Loss Error was seen.

DeepLearning_grid_1_AutoML_3_20221001_203445_model_2 came at 19th position with a Log Loss of 0.177, AUC of 98.97%, and Mean Per Class Error of 0.038.

GLM_1_AutoML_3_20221001_203445 came at 20th position with a Log Loss of 0.131, AUC of 98.92%, and Mean Per Class Error of 0.044.

DeepLearning_grid_3_AutoML_3_20221001_203445_model_2 came at 21t1h position with a Log Loss of 0.235, AUC of 98.81%, and Mean Per Class Error of 0.050.

The Least accurate model was ***DeepLearning_1_AutoML_3_20221001_203445*** coming at the last with Log Loss of 0.155, AUC of 98.50%, and Mean Per Class Error of 0.058.

In addition, the detailed results of the implemention models by AutoML technique can be seen in the leader board in fig. 5.2.

```
lb = aml.leaderboard
lb.head(rows=lb.nrows)
```

model_id	auc	logloss	aucpr	mean_per_class_error	rmse	mse
GBM_grid_1_AutoML_3_20221001_203445_model_5	0.999878	0.0155183	0.999884	0.00384615	0.0648196	0.00420157
GBM_grid_1_AutoML_3_20221001_203445_model_2	0.999832	0.0220579	0.999839	0.00576923	0.0787731	0.00620521
StackedEnsemble_BestOfFamily_1_AutoML_3_20221001_203445	0.999802	0.0192801	0.999811	0.00576923	0.0754915	0.00569897
GBM_4_AutoML_3_20221001_203445	0.999802	0.0232723	0.999809	0.00589133	0.0842597	0.0070997
StackedEnsemble_AllModels_1_AutoML_3_20221001_203445	0.999756	0.0202361	0.999768	0.00583028	0.0767135	0.00588496
GBM_2_AutoML_3_20221001_203445	0.999741	0.0250783	0.999749	0.00781441	0.0875134	0.00765859
GBM_3_AutoML_3_20221001_203445	0.999695	0.0287106	0.999704	0.00787546	0.0938748	0.00881248
GBM_5_AutoML_3_20221001_203445	0.999389	0.0347114	0.999417	0.0136447	0.104574	0.0109357
GBM_grid_1_AutoML_3_20221001_203445_model_1	0.999374	0.0308421	0.999449	0.00576923	0.0882397	0.00778625
GBM_grid_1_AutoML_3_20221001_203445_model_4	0.999283	0.0346256	0.999338	0.0117827	0.0976915	0.00954363
DRF_1_AutoML_3_20221001_203445	0.998924	0.066757	0.999004	0.0137057	0.124782	0.0155707
GBM_grid_1_AutoML_3_20221001_203445_model_3	0.998413	0.0722812	0.998501	0.0215201	0.139494	0.0194587
GBM_1_AutoML_3_20221001_203445	0.997207	0.0810893	0.997489	0.0250611	0.146041	0.021328
DeepLearning_grid_2_AutoML_3_20221001_203445_model_1	0.994963	0.100726	0.996239	0.0214591	0.145101	0.0210542
DeepLearning_grid_3_AutoML_3_20221001_203445_model_1	0.994399	0.111846	0.995614	0.0214591	0.150428	0.0226286
DeepLearning_grid_1_AutoML_3_20221001_203445_model_1	0.992842	0.113741	0.994421	0.0271062	0.156894	0.0246158
DeepLearning_grid_2_AutoML_3_20221001_203445_model_2	0.991514	0.15612	0.993212	0.0351038	0.194502	0.0378309
XRT_1_AutoML_3_20221001_203445	0.990781	0.38734	0.992451	0.0411783	0.32746	0.10723
DeepLearning_grid_1_AutoML_3_20221001_203445_model_2	0.989713	0.177076	0.99155	0.0389499	0.189649	0.0359666
GLM_1_AutoML_3_20221001_203445	0.989217	0.131015	0.990899	0.0447802	0.193765	0.0375449
DeepLearning_grid_3_AutoML_3_20221001_203445_model_2	0.988194	0.235444	0.988636	0.0509158	0.221237	0.0489456
DeepLearning_1_AutoML_3_20221001_203445	0.98505	0.155371	0.988325	0.0587912	0.209732	0.0439875

*Figure **Error! No text of specified style in document.**.35 Leaderboard of AutoML*

5.1 Gradient Boosting Machine: Best Model

In fig. 5.3, the summary of the best model: Gradient Boosting Machine can be seen. This model has performed the best over all other models, and some details of the summary can be seen in the figure below. In the model, total number of trees implemented were 255, number of internal trees were also 255, model size in bytes was 95315.0, minimum depth was 5.0, maximum depth was 12.0, mean depth was 8.8, minimum leaves were 18.0, maximum leaves were 28, and mean leaves were 24.9.

```
Model Details
=============
H2OGradientBoostingEstimator :  Gradient Boosting Machine
Model Key:  GBM_grid_1_AutoML_3_20221001_203445_model_5


Model Summary:
```

	number_of_trees	number_of_internal_trees	model_size_in_bytes	min_depth	max_depth	mean_depth	min_leaves	max_leaves	mean_leaves
0	255.0	255.0	95315.0	5.0	12.0	8.819608	18.0	28.0	24.988235

*Figure **Error! No text***

of specified style in document..36 Gradient Boosting Machine: Summary

In fig. 5.4, the predictions of the Gradient Boosting Machine (GBM) on the validation set can be seen. It shows that the model performed excellently on validation set with Area Under the Curve (AUC) of 1.0, Area Under the Curve Precision-Recall (AUCPR) of 1.0, Mean Per-Class Error of 0.0, and Log Loss of 0.00045, which is excellent and most satisfactory.

```
preds = aml.predict(valid)
```

```
gbm prediction progress: |████████████████████████████████████████████████| (done) 100%
```

```
aml.leader.model_performance(valid)
```

```
ModelMetricsBinomial: gbm
** Reported on test data. **

MSE: 1.8901575491376075e-05
RMSE: 0.004347594218803783
LogLoss: 0.00045142923685749384
Mean Per-Class Error: 0.0
AUC: 1.0
AUCPR: 1.0
Gini: 1.0
```

*Figure **Error! No text of specified style in document.**.37 Predictions on Validation Set*

Now, let's see the confusion matrix of the Gradient Boosting Machine (GBM) in fig. 5.5, which has performed the highest over all other models. F1 score of the actual vs predicted values is 99.7%. The Confusion matrix is an N x N matrix used for evaluating the performance of a classification model, where N is the number of target classes. The matrix compares the actual target values with those predicted by the machine learning model. It says that 0 class representing people not having diabetes was predicted 68 times right, and 0 times wrong. Similarly, 1 class representing people having diabetes was predicted 60 times as right, while 0 times was predicted as wrong. The total rate of 0 class was 68, while 1 class had 60 rates.

Confusion Matrix (Act/Pred) for max f1 @ threshold = 0.997006481758634:

		0	1	Error	Rate
0	0	68.0	0.0	0.0	(0.0/68.0)
1	1	0.0	60.0	0.0	(0.0/60.0)
2	Total	68.0	60.0	0.0	(0.0/128.0)

*Figure **Error! No text of specified style in document.**.38 Confusion Matrix: Actual vs Predicted Values*

Chapter 6: Conclusion & Future Work

In this study, a machine learning predictive model was to be developed to detect early diabetes in the patients using some important features related to medical science such as age, gender, polyuria, polydipsia, Sudden weight loss, weakness, polyphagia, genital trush, visual blurring, itching, irritability, detailed healing, partical paresis, muscle stiffness, alopecia, obsesity, and class. These all features were found as part of the dataset found on Kaggle.com, which offers real-world dataset for machine learning and data science project. In this project, the main objective of the project was to predict and classify diabetes in patients using Machine Learning models. However, the technique used to implement Machine Learning model was AutoML, which is a technique that implements a number of Machine Learning models based on certain parameters provided, and evaluates them using the required parameters to find the best model that is able to perform predictions with higher accuracy.

Firstly, Exploratory Data Analysis (EDA) was performed on the dataset using various libraries, which is an investigation on the dataset would be performed critically to understand the data and its relationship with other variables of the data using different kinds of plots such as histogram to know the distributions, correlation heatmap graph to know the relationships, and others etc., to perform Univariate, Bivariate, and Multivariate analysis. In the dataset, all features were of type integer having 0 or 1 value, except for gender which was of object data type. Therefore, it wasn't possible to draw different various different kinds of plots for visualizations, and only countplot, piechart, and bar plots were used for the visualization. This helped us understand the significance of different features available in the dataset in relation to the class or target feature of the dataset.

Secondly, for the implementation of AutoML technique there's was a need to perform the pre-processing of the dataset to get the dataset ready for the implementation. Therefore, for this purpose, several pre-processing steps were performed including, encoding of categorical features, which converts the categorical features into numerical ones, performed over sampling and under sampling

using imblearns library, defined input and output as x and y, split the dataset for testing and validation, etc., that helped in successful implementation of the models using AutoML technique.

Finally, AutoML technique was implemented using H20AutoML, to successfully implement the technique on the project. Automated machine learning (AutoML) is the process of applying machine learning (ML) models to real-world problems using automation. More specifically, it automates the selection, composition and parameterization of machine learning models. The seed of 12 was given, and max 20 model were the parameters given to the H20AutoML, and it fitted the dataset on the models such as, GBM (Gradient Boosting Machine), DRF (Distributed Uplift Random Forest), Deep Learning, GLM (Generalized Linear Model), etc., and some other versions of the them, and performance measures were such as AUC, Log Loss, AUCPR, and Mean Per Class Error to declared the best model. In this study, GBM (Gradient Boosting Machine) outperformed all others with an Area Under the Curve (AUC) of 1.0, Area Under the Curve Precision-Recall (AUCPR) of 1.0, Mean Per-Class Error of 0.0, and Log Loss of 0.00045, which is excellent and most satisfactory compared to earlier studies. Therefore, Gradient Boosting Machine was declared as the best model in this study to predict diabetes prediction in Machine Learning. This model can help people in detection of diabetes early to counter the disease, and help to prevent it.

6.1 Objective Evaluation

In this study, AutoML was implemented in Machine Learning to build a predictive model that is able to classify people having diabetes or not depending upon the features available in the dataset. Fortunately, the objective of the study was fulfilled with a higher accuracy that wasn't possible in the previous studies conducted earlier.

Following objectives of the study were fulfilled:

- A prediction model (Gradient Boosting Machine) was developed using AutoML technique in Machine Learning with Area Under the Curve (AUC) of 1.0, Area Under the Curve Precision-

Recall (AUCPR) of 1.0, Mean Per-Class Error of 0.0, and Log Loss of 0.00045, which is excellent and most satisfactory.

- The dataset was analyzed using Exploratory Data Analysis (EDA) to find relationship and correlations between various features of the dataset. This helped to understand the type of data being dealt with, and helped to perform the data pre-processing.
- The dataset was pre-processed to make it ready for the implementation of Machine Learning models, which in this study was implemented using AutoML technique.
- Finally, in this study, Gradient Boosting Machine, was developed which performed excellently over all other models. This model can be useful for identifying early diabetes detection in patients.

6.2 Future Work

As diabetes detection in machine learning using AutoML was performed in this study, and exceptional results were achieved, however, future work can be enhanced using the following:

- Artificial Neural Networks to classify the diabetes prediction in Machine Learning.

Chapter 7: References

American Diabetes Association. (2020). classification and diagnosis of diabetes: Standards of medical care in diabetes_2020. Diabetes Care, vol. 43, no. 1, pp. S14_S31

Alam, T., M., et al, (2019). A model for early prediction of diabetes. Science Direct. Vol. 16. pp. 1-7. https://www.sciencedirect.com/science/article/pii/S2352914819300176

Barnes, A., S., and Coulter, S., (2010). The Epidemic of Obesity and Diabetes. NCBI. https://www.ncbi.nlm.nih.gov/pmc/articles/PMC3066828/#:~:text=Obesity%20is%20the%20leading%20risk,do%20women%20of%20normal%20weight.

Barreto, S., 2022. Real-Life Examples of Supervised Learning and Unsupervised Learning. Bealdung. pp.2-6. https://www.baeldung.com/cs/examples-supervised-unsupervised-learning

Bhandari, A., (2020). Everything You Should Know About Confusion Matrix for Machine Learning. Analytics Vidhya. https://www.analyticsvidhya.com/blog/2020/04/confusion-matrix-machine-learning/#:~:text=A%20Confusion%20matrix%20is%20an,by%20the%20machine%20learning%20model.

Bhandari, A., (2020). AUC-ROC Curve in Machine Learning Clearly Explained. Analytics Vidhya. Analytics Vidhya. https://www.analyticsvidhya.com/blog/2020/06/auc-roc-curve-machine-learning/#:~:text=The%20Area%20Under%20the%20Curve,the%20positive%20and%20negative%20classes.

Cindy Xinji Cai, (2022). Diabetes and Your Eyes: What You Need to Know. Hopkins Medicine. https://www.hopkinsmedicine.org/health/conditions-and-diseases/diabetes-and-your-eyes-what-you-need-to-know#:~:text=Another%20potential%20effect%20from%20diabetes,after%20your%20blood%20sugar%20stabilizes.

Chatterjee, C 2020, A Quick Introduction to KNN Algorithm, Great Learning. https://www.mygreatlearning.com/blog/knn-algorithm-introduction/

Diabetes Journals, (2001). Diabetes and Your Joints. https://diabetesjournals.org/clinical/article/19/3/136/2452/Diabetes-and-Your-Joints#:~:text=Symptoms%20of%20diabetes%2Drelated%20musculoskeletal,also%20affect%20people%20without%20diabetes.

Diabetes, (2022). Diabetes can affect every part of the body, including the skin. https://diabetes.org/diabetes/skin-complications#:~:text=call%20your%20doctor.-,Itching,lower%20parts%20of%20the%20legs.

Diabetes, (2022). Unexplained Weight Loss. https://www.diabetes.co.uk/symptoms/unexplained-weight-loss.html#:~:text=Diabetes%20and%20sudden%20weight%20loss,reduction%20in%20overall%20body%20weight.

Dansinger, M., (2021). What is polydipsia. WebMD. https://www.webmd.com/diabetes/polydipsia-thirsty#:~:text=Drinking%20plenty%20of%20water%20will,of%20time%20in%20the%20bathroom.

Dansinger, M., (2022). Diabetes and Thrush. Web MD. https://www.webmd.com/diabetes/diabetes-thrush#:~:text=Yeast%20infections%20are%20a%20particular,can%20end%20up%20with%20thrush.

Fletcher, J., and Wood, K., (2022). Why does diabetes cause fatigue. Medical News Today. https://www.medicalnewstoday.com/articles/323398#:~:text=Changes%20in%20blood%20sugar%20levels,-Diabetes%20affects%20the&text=In%20people%20with%20diabetes%2C%20the,do%20not%20get%20enough%20glucose.

G. Sparacino, F. Zanderigo, S. Corazza, A. Maran, A. Facchinetti, and C. Cobelli, ``Glucose concentration can be predicted ahead in time from continuous glucose monitoring sensor time-series," IEEE Trans. Biomed. Eng., vol. 54, no. 5, pp. 931_937, May 2007.

Geeks for Geeks, (2022). ML | Label Encoding of datasets in Python, https://www.geeksforgeeks.org/ml-label-encoding-of-datasets-in-python/

Gupta, M., 2022, ‘ML | Types of Learning – Supervised Learning ‘, geeksforgeeks, pp.-1-8, Available at: https://www.geeksforgeeks.org/ml-types-learning-supervised-learning/

Healthline, (2018). Excessive Urination Volume? https://www.healthline.com/health/urination-excessive-volume

Healthline, (2022). Can Diabetes Cause Mood Swings? https://www.healthline.com/health/diabetes/diabetes-mood-swings#mental-health

Healthline, (2022). How diabetes affects the hair growth cycle. https://www.healthline.com/health/does-diabetes-cause-hair-loss#:~:text=People%20with%20type%201%20diabetes,other%20parts%20of%20the%20body.

Healthline, (2022). Can Diabetes Cause Mood Swings? https://www.healthline.com/health/diabetes/diabetes-mood-swings#:~:text=Can%20Diabetes%20Cause%20Mood%20Swings%3F&text=People%20with%20diabetes%20may%20experience,mood%20and%20mental%20health%20too.

Healthline, (2019). What Are the 3 Ps of Diabetes? https://www.healthline.com/health/diabetes/3-ps-of-diabetes#:~:text=Polyphagia%20describes%20excessive%20hunger.,to%20be%20used%20for%20energy.

Healthline, (2022). Does Diabetes Cause Hair Loss? https://www.healthline.com/health/does-diabetes-cause-hair-loss#takeaway

Healthline, (2018). Excessive Urination Volume (Polyuria). https://www.healthline.com/health/urination-excessive-volume

Heidenreich, H., (2018). What are the types of machine learning?', Towards Data Science, pp.1-5, Available at: https://towardsdatascience.com/what-are-the-types-of-machine-learning-e2b9e5d1756f?gi=865d3d7e2f36

J. M. Norris, R. K. Johnson, and L. C. Stene (2020). Type 1 diabetes_Early life origins and changing epidemiology. Lancet Diabetes Endocrinol., vol. 8, no. 3, pp. 226_238

Joshi, R., and Alehegn, M., (2017). Analysis and prediction of diabetes diseases using machine learning algorithm', IRJET, vol. 04, pp. 426-432, Available at: https://www.irjet.net/archives/V4/i10/IRJET-V4I1077.pdf

Kopitar, L., et el., (2020). Early Detection of Type 2 Diabetes Mellitus Using Machine Learning-Based Prediction Models. Nature Research. vol.10. pp.1-5. Available at: https://www.nature.com/articles/s41598-020-68771-z

Lutkevich, B., (2020), Automated Machine Learning, Tech Target, https://www.techtarget.com/searchenterpriseai/definition/automated-machine-learning-AutoML

Larxel, (2021). Early Classification of Diabetes. Kaggle. Available at: https://www.kaggle.com/datasets/andrewmvd/early-diabetes-classification

Lutkevich, B., (2020). Automated Machine Learning (AutoML). https://www.techtarget.com/searchenterpriseai/definition/automated-machine-learning-AutoML

Masters in Data Science, (2022). What is Decision Tree. https://www.mastersindatascience.org/learning/machine-learning-algorithms/decision-tree/

Mbaabu, O., (2020). Introduction to Random Forest in Machine Learning', Section, pp. 4-9. https://www.section.io/engineering-education/introduction-to-random-forest-in-machine-learning/

Medical News Today, (2019). How does diabetes affect wound healing? https://www.medicalnewstoday.com/articles/320739#:~:text=Uncontrolled%20diabetes%20may%20also%20affect,which%20can%20affect%20wound%20healing.

Medical News Today, (2022). How does diabetes affect wound healing? https://www.medicalnewstoday.com/articles/320739#why-diabetes-affects-it

P. Dua, F. J. Doyle, and E. N. Pistikopoulos (2006). Model-based blood glucose control for type 1 diabetes via parametric programming," IEEE Trans.Biomed. Eng., vol. 53, no. 8, pp. 1478_1491

S. Guerra, et al. (2012). Enhancing the accuracy of subcutaneous glucose sensors: A real-timedeconvolution-based approach," IEEE Trans. Biomed. Eng., vol. 59, no. 6, pp. 1658_1669

Steen, D., (2020), Precision-Recall Curves, Medium, https://medium.com/@douglaspsteen/precision-recall-curves-d32e5b290248

WebMD, (2022). When Diabetes Causes Stomach Problems. https://www.webmd.com/diabetes/type-1-diabetes-guide/diabetes-and-gastroparesis

Wikipedia, (2022). Gender. https://en.wikipedia.org/wiki/Gender

Web MD, (2022). When Diabetes Causes Stomach Problems. https://www.webmd.com/diabetes/type-1-diabetes-guide/diabetes-and-gastroparesis#1

Zou, Q., et al., (2018). Predicting Diabetes Mellitus with Machine Learning Techniques. Frontiers In. vol. 2. pp. 12-19. Available at: https://www.frontiersin.org/articles/10.3389/fgene.2018.00515/full

Zhang, Z., et al., (2022). Machine Learning Prediction Models for Gestational Diabetes Mellitus. JMIR, vol. 24, pp. 33-39, Available at: https://www.jmir.org/2022/3/e26634/