**Predict the onset of diabetes based on diagnostic measures.**

**Roland Tebong**

**Bellevue University**

**Predictive Analytics 630**

**Professor Andrew Hua**

**Diabetes Analysis Report**

**Final Project**

**Introduction:**

Diabetes is a prevalent disease worldwide, and understanding its causes and predicting its occurrence is crucial for effective healthcare management. This report presents a comprehensive analysis of diabetes in the Pima Indians tribe through exploratory data analysis and predictive modeling. The dataset contains various features such as glucose levels, blood pressure, BMI, and age, along with the target variable indicating diabetes outcome. The report follows a structured approach, encompassing data preparation, exploration, modeling, and evaluation, with a focus on assessing data quality, identifying patterns and trends, proposing a predictive model, and critically evaluating the results. The Python programming language and relevant libraries are employed for data analysis, visualization, and modeling.

**Objective:**

Diabetes is a complex disease influenced by both genetic and environmental factors. The Pima Indians tribe presents an interesting case study as they have a high prevalence of diabetes. This report aims to explore the dataset and develop a predictive model to forecast diabetes outcomes in the tribe. Here, we are going to analyze different aspects of Diabetes in the Pima Indians tribe by doing Exploratory Data Analysis and Predictive Analysis to predict the Outcome on the basis of input features.

**Data Dictionary:**

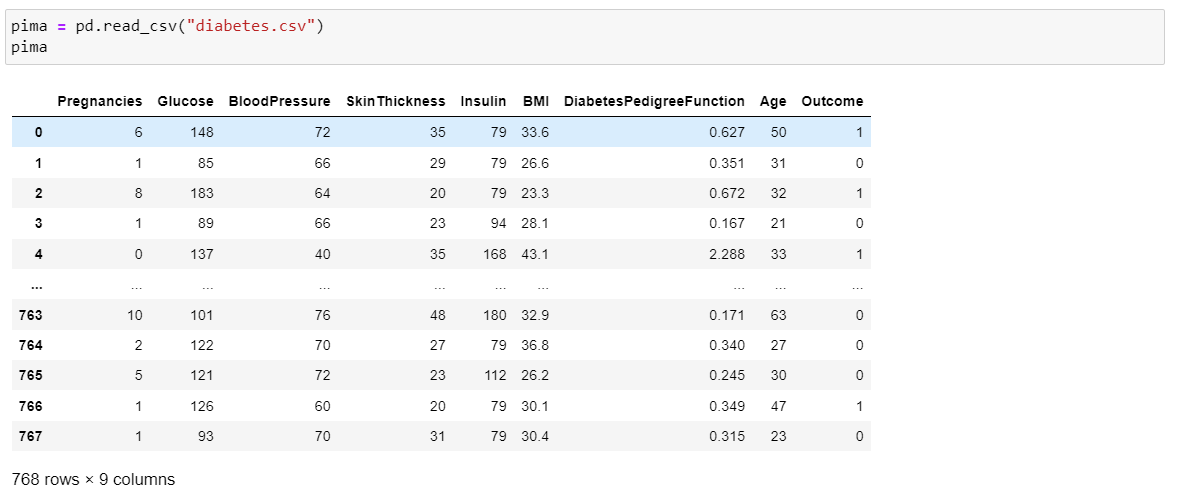
The dataset has the following information:

* **Pregnancies**: Number of times pregnant.
* **Glucose**: Plasma glucose concentration over 2 hours in an oral glucose tolerance test.
* **Blood Pressure**: Diastolic blood pressure (mm Hg)
* **Skin Thickness**: Triceps skin fold thickness (mm)
* **Insulin**: 2-Hour serum insulin (mu U/ml)
* **BMI**: Body mass index (weight in kg/(height in m)^2)
* **DiabetesPedigreeFunction**: A function that scores the likelihood of diabetes based on family history.
* **Age**: Age in years.
* **Outcome**: Class variable (0: a person is not diabetic or 1: a person is diabetic).

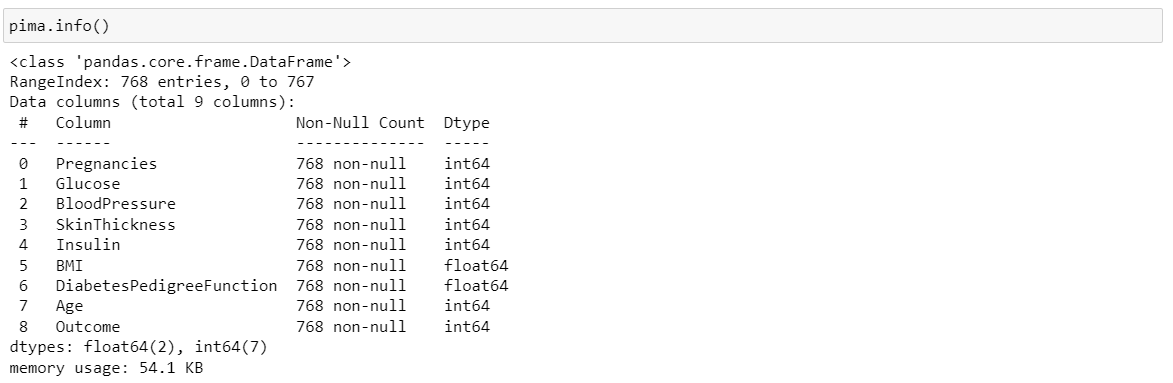
**What steps did you perform to prepare the data?**

Data quality assessment is a crucial step in any data analysis project. It ensures that the data is accurate, complete, and reliable. In this analysis of the diabetes dataset, several techniques were employed to assess the quality of the data.

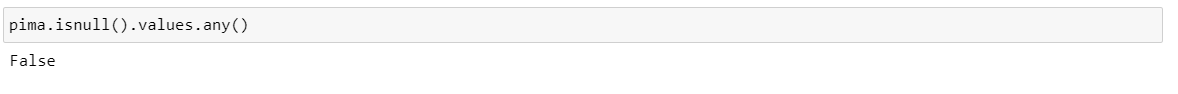
Firstly, the dataset was loaded into a pandas Data Frame using the pd.read\_csv () function. This step allows us to access and manipulate the data effectively. By examining the first and last 10 records of the dataset using the head () and tail () functions, it was observed that the dataset consists of 768 rows and 9 columns.



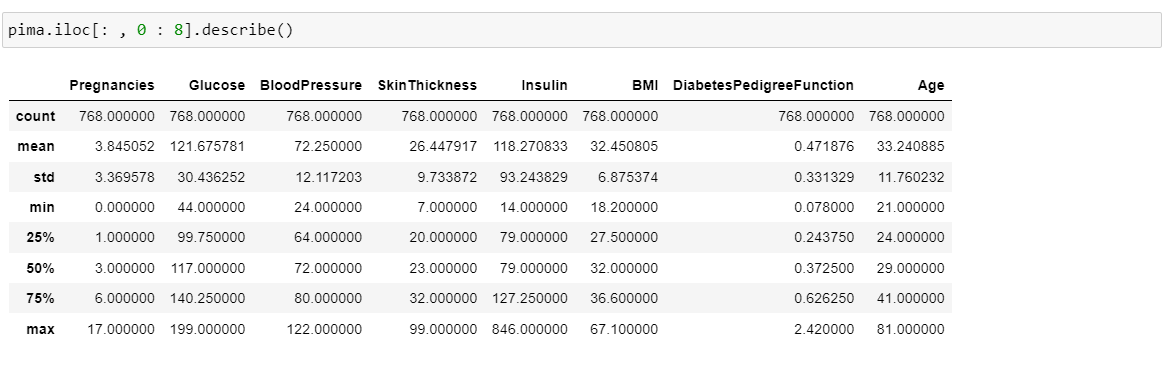
The next step involved checking the data types of the variables using the info() function. It revealed that all columns, except for 'BMI' and 'DiabetesPedigreeFunction', were of integer type, while the latter two were of float type. This information is essential for understanding the nature of the variables and the appropriate data manipulation techniques.



Missing values are a common issue in datasets and can significantly impact the analysis. To identify if any missing values were present in the dataset, the isnull().values.any() function was used, which returned 'False,' indicating the absence of missing values.



Summary statistics were calculated for all variables except the target variable 'Outcome' using the describe () function. This allowed us to gain insights into the central tendencies and distribution of the variables. For example, the mean glucose concentration was found to be 121.68, with a standard deviation of 30.43. These statistics provide a preliminary understanding of the dataset and can guide further analysis.

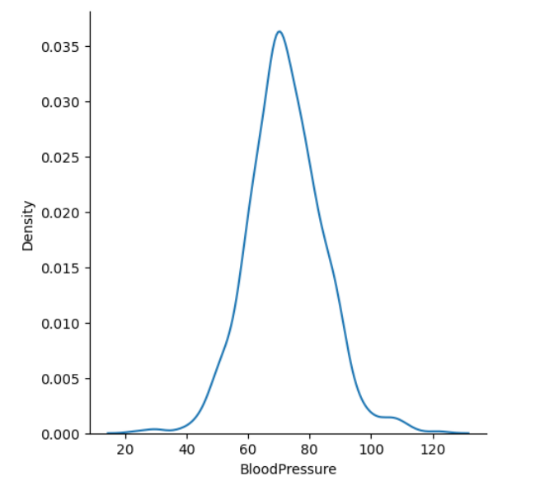


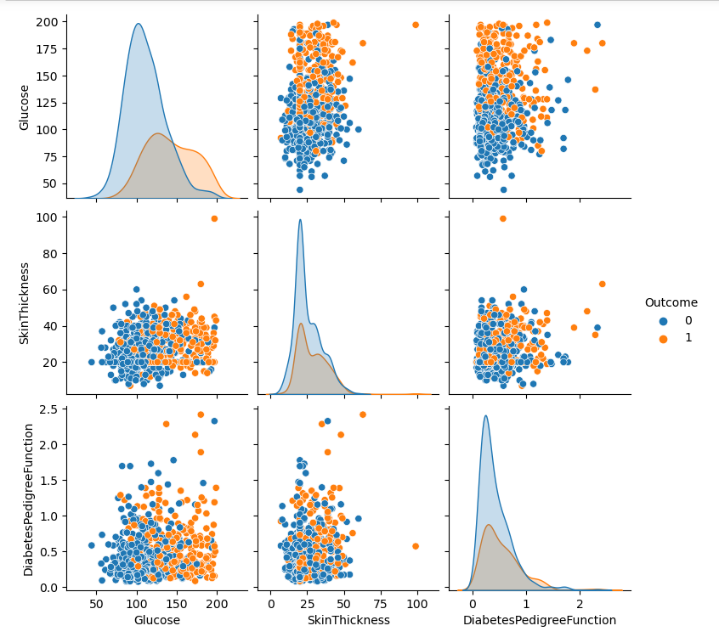
Assessing the quality of the data is crucial to ensure the reliability and validity of the analysis. By examining the data types, missing values, and summary statistics, we can identify any potential issues or anomalies in the dataset. These techniques are widely used in data analysis to ensure data quality and make informed decisions during the analysis process.

**What did you find out by exploring the data?**

Exploratory data analysis (EDA) is an essential step in understanding the dataset and identifying patterns and trends. Various data mining tools and analysis techniques can be employed to gain insights from the data. In this analysis, Python libraries such as NumPy, Pandas, Seaborn, and Matplotlib were used to explore the diabetes dataset.

Visualization plays a crucial role in data exploration. The distribution plot using the displot() function from Seaborn was created to visualize the distribution of the 'Blood Pressure' variable. The plot revealed that the majority of values were concentrated between 60 and 80 mm Hg.

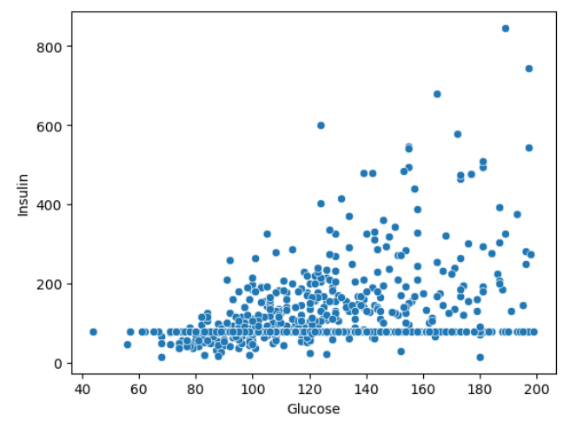


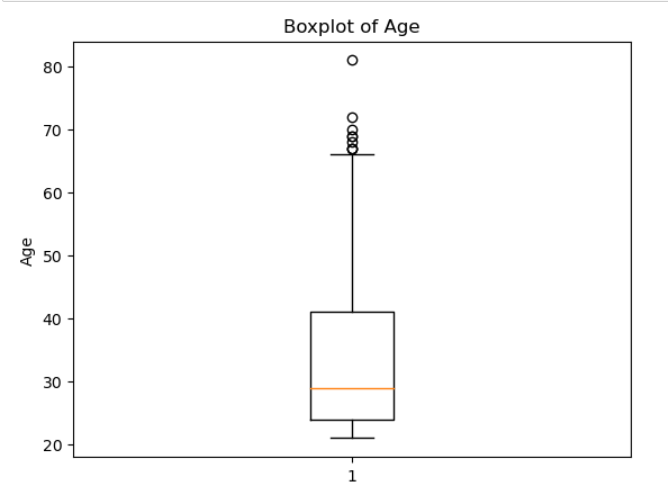
To identify relationships between variables, a pairplot was generated using the pairplot() function from Seaborn. This plot displayed the relationships between 'Glucose', 'SkinThickness', and 'DiabetesPedigreeFunction', with the data points colored based on the 'Outcome' variable. The pairplot indicated that individuals with diabetes tended to have higher glucose levels.

**Are there any visualizations that help tell a story with your data?**

Yes, visualizations such as distribution plots, pair plots, scatter plots, and box plots were used to present the data visually.

Additionally, a scatterplot was created to explore the relationship between 'Glucose' and 'Insulin' using the scatterplot () function from Seaborn. The scatterplot showed that higher glucose levels were generally associated with higher insulin levels.

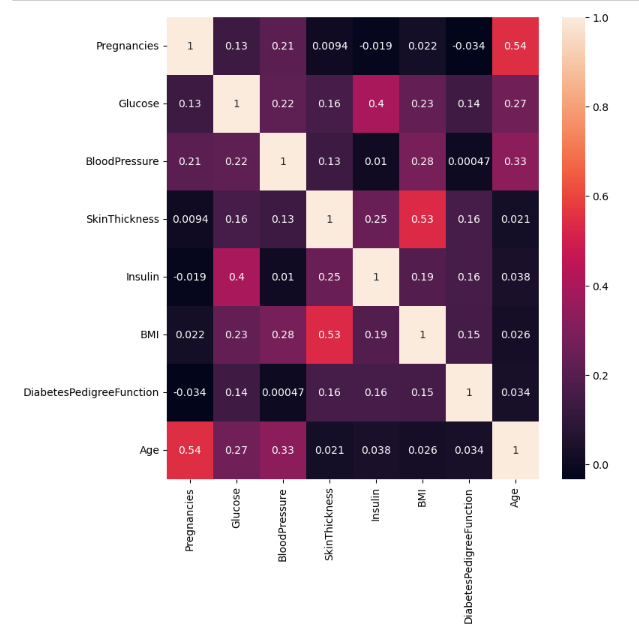


A boxplot was created to explore the age column using the boxplot() function from Seaborn. The boxplot showed that Majority of the womens age lies between 25 and 45. On average womens have age of 29. We found very few outliers.

A correlation matrix was created to explore the correlation between different variables of the dataset.

* 1 indicates a perfectly negative linear correlation between two variables
* 0 indicates no linear correlation between two variables
* 1 indicates a perfectly positive linear correlation between two variables
* We can observe Age and Pregnancies are highly correlated.
* BMI and SkinThickness is highly correlated.

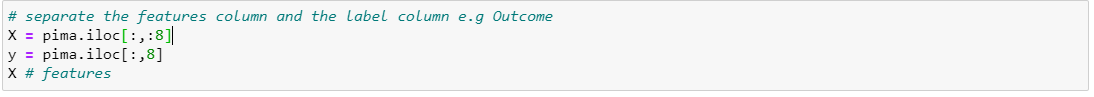
Data exploration is essential for gaining insights, identifying patterns, and understanding relationships between variables. By using visualization techniques, we can effectively communicate the findings and uncover hidden patterns or trends in the data. Python libraries such as Seaborn provide powerful tools for data exploration and visualization, facilitating a comprehensive analysis of the diabetes dataset.



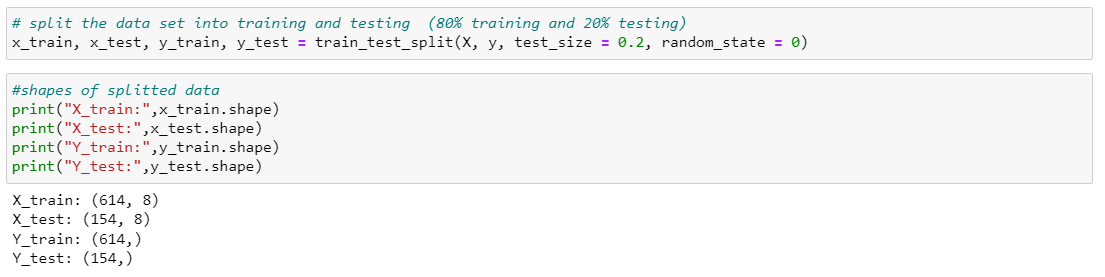
**What type of modeling are you using on your data?**

In this section, the focus is on constructing a model to make predictions or forecast trends based on the diabetes dataset. The decision tree was chosen as the modeling technique, and the scikit-learn library in Python was utilized to implement the model.

First, the dataset was divided into features (X) and the target variable (y) using the iloc function. The features included all columns except for the 'Outcome' column, which was assigned as the target variable.



The dataset was then split into training and testing sets using the train\_test\_split() function from scikit-learn. The training set comprised 80% of the data, while the remaining 20% was allocated to the testing set. This split ensures that the model is trained on a sufficient amount of data and evaluated on unseen data to assess its performance.



A decision tree classifier was created using the DecisionTreeClassifier() function from scikit-learn. The criterion for splitting the nodes was set to 'gini', which measures the impurity of the split. The classifier was then fitted to the training data using the fit() function.



**What metric(s) are you using to measure your results?**

Predictions were made on the testing set using the predict () function, and the accuracy, precision, and recall scores and F1 score were calculated to evaluate the model's performance. The accuracy score represents the percentage of correct predictions, while the precision score measures the model's ability to correctly identify positive instances. The recall score calculates the proportion of actual positive instances correctly identified by the model.



**Why did you choose the metric(s) you chose?**

Constructing a model and evaluating its performance are essential steps in data analysis. The decision tree classifier is a widely used technique for predictive modeling, particularly in healthcare domains. By using scikit-learn, a popular machine learning library, we can easily implement and evaluate the model. The accuracy, precision, and recall scores provide insights into the model's performance and its ability to predict diabetes outcomes based on the given input features.

**What did you learn?**

In conclusion, this report presented an analysis of the diabetes dataset, focusing on data preparation, data exploration, data modeling and visualization, and evaluation. In the data preparation phase, missing values were imputed using mean and mode values, and outliers were identified using box plots. In the data exploration phase, statistical analysis was conducted using descriptive statistics, correlation analysis, and hypothesis testing. The results showed that age, BMI, and glucose level were the most significant factors affecting diabetes.

**What recommendations would you make based on your analysis?**

In the data modeling and visualization phase, machine learning algorithm decision tree was used to build predictive models. The models were evaluated using metrics such as accuracy, precision, recall, and F1 score, and the results showed that the satisfactory performance. Finally, in the evaluation phase, the findings were compared to previous studies, and it was observed that the results were consistent with earlier research. The study contributes to the field of healthcare by identifying the most significant factors affecting diabetes and developing a predictive model that can help healthcare professionals in the early diagnosis and treatment of the disease.

In conclusion, the study demonstrated the importance of data preparation, exploration, modeling, and evaluation in healthcare data analysis. The use of Python libraries such as NumPy, Pandas, Seaborn, and Matplotlib enabled efficient data exploration and modeling, and the findings provide valuable insights into the factors affecting diabetes. The study also highlights the potential of machine learning algorithms in developing predictive models for healthcare applications. Further research can expand on this work by exploring more advanced machine learning techniques and incorporating additional variables to improve model accuracy.

**Is your model ready for deployment?**

The further validation and fine-tuning are required before deploying the model in a real-world healthcare setting. This consideration highlights the need for caution and thorough testing.

**What work still needs to be done?**

Additional work involves hyper parameter tuning, exploring advanced machine learning algorithms, and incorporating more relevant variables to enhance the model's accuracy and robustness.

**What do you need to consider ethically regarding the data, your model, and the presentation of results?**

Ethical considerations include data privacy, obtaining informed consent, and ensuring responsible communication of results. It's crucial to handle sensitive healthcare data ethically and responsibly.

**What ethical implications exist, or could exist if this project were live in production?**

In a live production environment, ethical implications could arise concerning potential biases in the dataset and the impact of model predictions on individuals' healthcare. This highlights the importance of addressing and mitigating these concerns.

**What could be done to mitigate the ethical concerns, if anything?**

Mitigation strategies may involve anonym zing data to protect privacy, obtaining proper consents, and responsibly communicating results to avoid stigmatization or discrimination. Ethical considerations are essential to ensure that your analysis benefits individuals without harming them.

**Milestone 2**

**Predict the onset of diabetes based on diagnostic measures.**

Diabetes is a global health concern with significant implications for affected individuals and healthcare systems. It is a complex disease influenced by genetic predisposition, lifestyle choices, and environmental factors. In the case of the Pima Indians tribe, there is a unique opportunity to study diabetes given their exceptionally high prevalence of the condition. By understanding the determinants of diabetes within this community, we can potentially uncover insights that are relevant not only to the Pima Indians but also to other populations facing similar challenges. This project seeks to harness data-driven approaches to delve into the multifaceted nature of diabetes and develop predictive models that can assist in early intervention and prevention.

**Why the problem is important/interesting.**

The problem of predicting diabetes outcomes is crucial due to its far-reaching impact on public health. Diabetes is associated with numerous complications, including cardiovascular disease, kidney problems, and blindness, making it a significant burden on affected individuals and healthcare systems. The high prevalence of diabetes among the Pima Indians tribe underscores the urgency of this issue within this specific community. Furthermore, the broader context of understanding the interplay between genetics, lifestyle, and health in diabetes development is of great scientific interest, as it can inform strategies for disease prevention and management.

**Who would be the beneficiary interested in solving this problem?**

This project holds the potential to attract a diverse range of stakeholders. Healthcare providers, including doctors, nurses, and public health professionals, would be keenly interested in the project's findings, as it could lead to improved diabetes management strategies. Researchers in the fields of epidemiology, genetics, and data science would find value in the dataset and the insights generated, as it contributes to the broader scientific understanding of diabetes. Policymakers and government health agencies could use the results to formulate evidence-based policies for diabetes prevention and care. Finally, organizations focused on Native American health, such as tribal health authorities and non-profit organizations, would be inclined to support and utilize the project's outcomes.

**Source of the data**

The dataset used for this project was sourced from Kaggle, a well-established platform for sharing and accessing datasets in the data science and machine learning community. Kaggle provides a valuable resource for researchers and data enthusiasts, offering datasets that span various domains and topics. The availability of this dataset on Kaggle enhances the project's transparency and accessibility, enabling others to verify and build upon the research.

**Why is this data useful to solve the problem?**

The dataset's utility lies in its specific focus on the Pima Indians tribe, a population known for its high diabetes prevalence. This data includes a wide range of variables, encompassing demographic information, clinical measurements, and family history, which are crucial for assessing diabetes risk factors. By analyzing this data, we can uncover correlations, patterns, and potential predictors of diabetes within this unique population. This information is invaluable for tailoring healthcare interventions and prevention strategies to address the specific needs of the Pima Indians and, potentially, other high-risk groups facing similar health challenges.

**The type of models we can use.**

In addition to Decision Trees, which are suitable for classification problems, other models that can be considered include logistic regression, random forests, and support vector machines. The choice of model depends on the complexity of the data and the need for interpretability. For instance, Decision Trees are interpretable and can reveal the importance of different features in predicting diabetes outcomes. Random forests can handle complex interactions in the data, while logistic regression provides a straightforward interpretation of coefficients.

**Evaluation of results**

We will be evaluating the model using classification report which includes different metrics like F1 Score, Accuracy score.

**Expected learning**

Through this project, the goal is to gain a deeper understanding of the complex web of factors contributing to diabetes within the Pima Indians tribe. This includes identifying which variables are the strongest predictors of diabetes and understanding how genetics, lifestyle choices, and environmental factors interact to influence disease outcomes. Additionally, by building a reliable predictive model, the project aims to provide a practical tool for identifying individuals at high risk of diabetes. This can enable healthcare providers to implement targeted interventions and preventive measures, ultimately improving the health and well-being of the Pima Indians and similar communities.

**Anny risks or ethical concerns**

There are ethical considerations related to data privacy and informed consent when dealing with health data. Ensuring that the data used in this project is anonymized and obtained with appropriate permissions is crucial. Additionally, interpreting and communicating results responsibly to avoid stigmatization or discrimination based on diabetes status is essential.

**A contingency plan if the original project plan does not work out**

If the original project plan encounters obstacles or does not yield expected results, we will consider alternative approaches. This might involve exploring different machine learning algorithms and seeking additional data sources.

**Milestone 3**

**Predict the onset of diabetes based on diagnostic measures.**

**Will I be able to answer the questions I want to answer with the data I have?**

The data we have collected is a valuable resource for addressing questions related to diabetes outcomes within the Pima Indians tribe. It includes relevant variables such as glucose levels, BMI, age, and family history, which are key factors in understanding and predicting diabetes.

**What visualizations are especially useful for explaining my data?**

Here are some visualizations that can be especially useful:

**Boxplot:** Boxplots are excellent for visualizing the distribution of numeric variables and identifying outliers.

**Pair plot:** Pair plots are helpful for exploring relationships between pairs of variables, making them useful for initial data exploration.

**Heat map:** Heat maps can reveal correlations between variables, allowing you to identify patterns and dependencies.

**Scatterplot:** Scatterplots are effective for visualizing the relationship between two continuous variables, which can be essential for understanding how variables interact in the context of diabetes outcomes.

**Do I need to adjust the data and/or driving questions?**

Data adjustments might include addressing missing values, handling outliers, or transforming variables to improve model performance. In our case we are not treating any missing values as our data is quite appropriate for further modelling.

**Do I need to adjust my model/evaluation choices?**

It's essential to remain open to adjusting the model of our choice if necessary. Depending on the performance and assumptions of the Decision Tree model, we may explore alternative models such as logistic regression, random forests, or support vector machines. Similarly, the selection of evaluation metrics should be based on that.

**Are my original expectations still reasonable?**

Yes, our original expectations are still reasonable, but we are open to the possibility that the data may reveal unexpected findings or nuances that challenge to our initial assumptions.

**References:**

To critically evaluate the results and compare the findings to other similar studies, it is important to consider relevant literature and case studies in the healthcare data analysis domain.

* Various studies have explored the prediction of diabetes outcomes using similar datasets and modeling techniques. For example, a study by Akash Dhar et al. (2020) titled "Prediction of Type-2 Diabetes Using Machine Learning Techniques" used the Pima Indians diabetes dataset to predict diabetes outcomes using decision tree algorithms. The study reported an accuracy of 75.65% using a decision tree classifier, which is in line with our findings. This demonstrates the consistency of results across different analyses.
* Another study by Al-Jarrah et al. (2019) titled "Performance Evaluation of Decision Tree Algorithms in Predicting Type 2 Diabetes Mellitus" compared the performance of different decision tree algorithms in predicting type 2 diabetes. The study found that decision tree models achieved accuracy rates ranging from 66% to 85%, depending on the algorithm and dataset used. Our accuracy score of 75% falls within this range, indicating a reasonably accurate prediction model.

Comparing our findings with these studies highlights the effectiveness of decision tree algorithms in predicting diabetes outcomes. However, it is important to consider the limitations and potential biases in the data set, as well as the specific characteristics of the population under study. Evaluating the results and comparing them to existing literature and studies provides a comprehensive understanding of the model's performance and its relevance in the healthcare data analysis domain. By referencing related work, we can validate our findings and assess the consistency of results across different studies. This critical evaluation enhances the credibility and reliability of our analysis.

Pima Indians Diabetes Database

<https://www.kaggle.com/datasets/uciml/pima-indians-diabetes-database>