# Personal Analysis Report by Coral Amrosi

Coral Amrossi 209515196  
[CORAL-GIT](https://github.com/koralZrihen/AI-project/tree/main)  
[CORAL-DB](https://www.kaggle.com/datasets/ehababoelnaga/diabetes-dataset)

## Executive Overview

In the challenging yet rewarding realm of data science, I embarked on an intricate journey to extend the utility of sophisticated analytical tools, particularly in the context of feature selection and dimensionality reduction within a Diabetes Diagnosis dataset. Building upon Nahorai Hagag's foundational work, I introduced nuanced changes to the dataset to leverage the full potential of the first tool. This narrative chronicles my experience with these tools, underscoring their applicability, benefits, and limitations within my analytical framework.

## Dataset Preparation and Preliminary Transformation

The initiation of this analytical saga began with the importation of the Diabetes Diagnosis dataset from structured CSV files into the versatile Pandas DataFrame. A significant transformation from the original job was the bifurcation of data into distinct training and testing subsets. The ensuing step required addressing the specter of incomplete data, a task adeptly managed by `SimpleImputer` from Scikit-learn, which judiciously imputed missing values with the median figure of corresponding features.

## Categorical Variables and One-Hot Encoding

A pivotal enhancement was the introduction of one-hot encoding via Pandas' 'get\_dummies' function, which ingeniously transformed categorical variables into a binary schema fit for analytical digestion. This operation was not merely about conversion; it was about preparing the dataset for a nuanced analysis that respects the subtleties of machine learning algorithms.

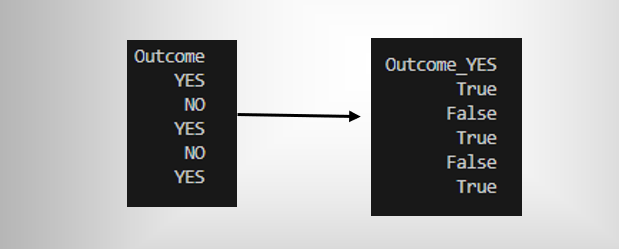
## Feature Selection and Standardization

Feature relevance was ascertained through the discerning eyes of 'SelectKBest', wielding the ANOVA F-value as its arbiter of significance. Post selection, the 'StandardScaler' bestowed upon these features the gift of uniformity, scaling them into a harmonious choir, each singing in the same octave.

## Dimensionality Reduction via PCA

The narrative culminates with the application of Principal Component Analysis (PCA), an esteemed technique that elegantly compresses high-dimensional data into a plane of intelligibility. It is here, upon this two-dimensional stage, that the selected features perform — revealing, through a scatter plot, the inner workings and structure of the dataset.

## Reflections on Tool Efficacy

The deployment of 'get\_dummies' for one-hot encoding was a triumph of convenience, seamlessly transforming the 'Outcome' variable from a binary to a nominal form with minimal code. Its elegance lies in its simplicity, effortlessly navigating the oft-tumultuous seas of categorical data. Yet, it is pertinent to note that its utility shines brightest in the presence of nominal data with limited categories, as was the case with the binary 'Outcome' variable in the dataset. The tool's aptness for this dataset is a reminder that the suitability of such tools is contingent upon the data's nature and the specificity of the task at hand.  
  
  


## Concluding Remarks

This report has encapsulated my endeavors to augment existing analytical tools, complementing Nahorai's profound work, and contributing to the overarching methodology. It underscores the imperative to carefully select and apply data analysis tools that align with the dataset's intrinsic characteristics and the intended analytical outcomes. As I progress, these experiences continue to shape my understanding and approach to the ever-evolving field of data analytics, charting a course toward future innovation and discovery in the realm of data science.