Predictive Model for Diabetes Diagnosis

The major goal of this project is to create a diagnostic model that could predict whether a patient has diabetes based on diagnostic parameters in the dataset.

Python emerged as the tool of choice in the pursuit of this goal. Python is well-known for its abilities in data analysis and machine learning, making it an ideal companion for this project. I started the adventure of processing and modelling the data using this powerful programming language.

Model Selection: For the project's predictive part, I used logistic regression, a commonly used technique for binary classification tasks. Using logistic regression, we created a model that can categorise patients as diabetic or non-diabetic depending on input features.

Addressing Class Imbalance: Addressing class imbalance within the dataset was one of the major issues encountered during the study. Recognising the potential biases that can result from imbalanced data, we corrected the class distribution using the Synthetic Minority Over-sampling Technique (SMOTE). SMOTE aided in the production of synthetic data points, ensuring that both classes were represented sufficiently in the dataset.

Threshold Adjustment: To fine-tune the model's performance, we made the critical decision to adjust the classification threshold. This change, which dropped the threshold to 0.4, was made to reduce the number of false positives. By doing so, we emphasised the necessity of controlling Type I mistakes, which is a critical factor in diagnostic applications.

The effective use of Python, logistic regression, and the intelligent integration of SMOTE to correct class imbalance demonstrate the dedication to solid data preprocessing and modelling. The delicate decision to change the classification threshold emphasises the project's Practicality and attention to real-world repercussions.