\*\*Presentation Script: Predictive Analysis of Medical Students' Susceptibility to Diabetes\*\*

Hello everyone! Today, I'm presenting my LHL Final Project, which is focused on predictive analysis concerning the likelihood of medical students developing diabetes.

Initially, I sourced a database from Kaggle, encompassing data of 200,000 students. This database meticulously documented the students' health conditions. However, a significant chunk of data was missing. From this foundation, I set two primary objectives:

1. To employ the existing data in training a predictive model, with an aim to supplement the diabetes section, which lacked around 20,000 entries.

2. To execute regression analysis on the supplemented data to identify health features impacting the onset of diabetes.

In pursuit of these objectives, I adopted the following steps:

1. Data cleaning and Exploratory Data Analysis (EDA).

2. Model training. Initially, I opted for Logistic Regression and Random Forest. However, due to suboptimal results, I transitioned to XGBoost, which predicted the missing data with an accuracy rate of 89%.

3. I embarked on regression analysis, seeking correlations between various features and diabetes.

4. Visualization of the data using Tableau.

The results unveiled that:

The logistic regression indicated a minimal overall correlation. However, some notable observations were:

1. The strongest negative correlation related to gender, suggesting females are less prone to diabetes.

2. Blood type O showed a negative correlation with diabetes.

3. Non-smokers were relatively less prone to the disease.

4. Features with a positive correlation included Blood type B, Blood pressure, Cholesterol, Heart rate, Blood type A, and Body temperature.

The visualization from Tableau further depicted:

1. Smoking, gender, and blood type had a minimal influence on diabetes susceptibility.

2. The data suggested individuals with lower body temperature, height, and blood pressure were more prone to diabetes. Conversely, those with higher weight, cholesterol, heart rate, age, and BMI faced increased risks.

Throughout this journey, I confronted several challenges:

1. A discrepancy between the database's description and its actual content, prompting a swift project objective recalibration.

2. Difficult-to-detect duplicates in the data, compelling me to restart my work from scratch.

3. A meager correlation between data features and the predictive target.

4. Abundant missing data complicated the model training process.

Looking ahead, potential enhancements could be:

1. Exploring samples beyond the confines of medical students.

2. Conducting duplication analysis on columns that might serve as primary keys.

3. Further refining and adjusting the prediction model.

4. Intensifying collaborations with medical experts to enhance model accuracy from a medical perspective.