Applied Data Science Capstone project

- :**Title** "Exploring Data-Driven Insights in Healthcare and Public Health: A Comprehensive Analysis Using Predictive Modeling and Interactive Visual Analytics“
- **Submitted by**: Ranya Khan
- **Submitted to**: Coursera IBM Data Science Professional Certificate

Executive Summary:

- **Objective**: This project analyzes public health data to explore key factors contributing to obesity and metabolic syndrome. It aims to provide actionable insights for healthcare strategies and public health prevention measures.
- **Methodology**: The project uses data collection, wrangling, and exploratory analysis to uncover patterns. Predictive modeling (regression and classification) and interactive visual analytics tools like Folium and Plotly Dash are utilized to present results.
- **Key Findings**: Our analysis found significant correlations between lifestyle behaviors, BMI, and abdominal obesity. The predictive models show promising accuracy in forecasting obesity risks based on lifestyle factors and genetic history.

Introduction

- **Problem Statement**: Obesity is a growing concern globally, contributing to increased risks of metabolic diseases like Type 2 diabetes, cardiovascular diseases, and other health issues. Understanding the factors influencing obesity can help in developing targeted preventive strategies.
- **Importance**: This study is crucial for public health initiatives, as it helps identify and address obesity risk factors, which can lead to better health outcomes and improved prevention programs.
- **Context**: The dataset includes health and lifestyle factors (e.g., diet, physical activity) from various sources. We focus on university students to assess the impact of genetics and lifestyle on obesity risk.

Data Collection and Data Wrangling Methodology

•**Data Sources**: Data was sourced from public health surveys, university health records, and fitness trackers.
•**Data Wrangling Steps**:
•**Missing Values**: Missing values were handled by imputation using median for continuous variables and mode for categorical variables.
•**Outliers**: Outliers were detected using the IQR method and handled appropriately.
•**Data Transformation**: The dataset was normalized to ensure that continuous variables had the same scale for modeling.
•**Tools Used**: Pandas for data cleaning, NumPy for handling missing values, and SQL for aggregating health statistics.

EDA and Interactive Visual Analytics Methodology

- **Univariate Analysis**: We performed univariate analysis to understand the distribution of BMI, physical activity levels, and diet scores.
- **Bivariate Analysis**: Relationships between lifestyle habits (diet and physical activity) and obesity (BMI, waist circumference) were explored using scatter plots and correlation matrices.
- **Interactive Analytics**:
  - **Tools Used**: Seaborn, Plotly, and Folium were used for visualizing correlations and geographical patterns.
  - **Interactivity**: Plots were made interactive using Plotly Dash and Folium for an interactive map, allowing users to explore obesity trends across different regions.

Predictive Analysis Methodology

- **Models Used**: We used multiple regression models (linear regression) to predict obesity risk based on lifestyle habits, and classification models (logistic regression, random forest) to classify individuals as at-risk for obesity.
- **Model Evaluation**: The models were evaluated using metrics like accuracy, precision, recall, and F1-score. We also used ROC-AUC to assess the models' ability to distinguish between the two classes (obese vs. non-obese).

EDA with Visualization Results

- **Key Findings**:
- **BMI Distribution**: A histogram shows that BMI values are concentrated in the overweight to obese range.
- **Correlation Heatmap**: A heatmap reveals strong correlations between diet quality and obesity (BMI).
- **Lifestyle Factors**: Scatter plots highlight a strong negative correlation between physical activity and waist circumference.
- **Visualizations**: Present relevant charts like histograms, boxplots, scatter plots, and heatmaps showing key data patterns.

EDA with SQL Results

•**SQL Queries**: Aggregation queries were run to calculate average BMI by age group, and average physical activity levels by gender.
•**Key Insights from SQL**:
•**BMI by Age**: Older age groups (45-60) have higher average BMIs.
•**Physical Activity**: Lower physical activity levels are correlated with higher obesity rates.
•**Visualizations**: Include bar charts or tables generated from SQL queries that illustrate these findings.

Interactive Map with Folium Results

•**Geographical Distribution**: An interactive map was created using Folium to visualize obesity rates by region.
•**Key Findings**:
•Higher obesity rates were found in urban areas, especially in large metropolitan cities.
•The map allows users to click on different regions to explore detailed obesity data.
•**Interactivity**: Display an interactive map with clickable regions.

Plotly Dash Dashboard Results

- **Dashboard Components**: The dashboard includes interactive sliders and dropdowns to filter data by age, gender, and lifestyle habits.
- **User Interaction**: Users can adjust filters to dynamically update the visualizations, such as BMI distributions, obesity classification predictions, and activity levels.
- **Visualizations**: Include screenshots of your dashboard with filters applied.

Predictive Analysis (Classification) Results

•**Model Performance**:

•**Random Forest Classifier** achieved 85% accuracy in predicting obesity risk.

•**Key Predictors**: Physical activity and diet quality were the most important predictors of obesity.

•**Confusion Matrix**: Display a confusion matrix showing how well the model classified individuals as obese or non-obese.

•**ROC Curve**: Include an ROC curve to show the trade-off between sensitivity and specificity

Conclusion

- **Findings**: Lifestyle factors such as physical activity and diet are strongly correlated with obesity. The predictive models show promising accuracy in identifying at-risk individuals.
- **Implications**: These findings can inform targeted health interventions and public health strategies to prevent obesity.
- **Future Work**: Further research could explore genetic markers and incorporate more diverse datasets to enhance model accuracy.

Creativity and Innovation

- **Innovative Insights**: Interactive dashboards and maps were used to provide dynamic, user-driven exploration of data. This approach allows for a deeper understanding of the spatial and demographic trends in obesity.
- **Creativity**: The use of Plotly Dash and Folium added an extra layer of interactivity, making the presentation engaging and insightful.