Bushra Haque
Data Management for Data Science Final project
07/2025
Github: https://github.com/BushraHaque78/DataManage-Project.git

Defining the Project:
This project explores the relationship between walking activity and blood glucose levels in individuals with Type 2 Diabetes. While walking is generally known to help lower blood sugar, the effect can vary significantly depending on factors such as time of day and recent meals.

In this project, I prepared datasets using pandas and established relationships between them through data cleaning and integration. I built a Streamlit web application that enables users to input their walking activity and obtain a predicted glucose level, incorporating adjustments for time of day and meal status. Additionally, I used R for exploratory data analysis, creating visualizations such as scatter plots, bar charts, and regression analyses to reveal patterns in the data and highlight trends between walking activity and glucose levels.

The ultimate goal was to provide an interactive, data-driven tool that can help individuals better understand and manage their glucose patterns, laying the foundation for more personalized, real-time recommendations for diabetes management.

Strategic Aspect Involved:
My approach demonstrates key data management and analysis concepts covered in class, applied to real-world health data. I began by cleaning and preparing the data using pandas, including handling missing values and transforming categorical variables like time_of_day through encoding, concepts emphasized in our data cleaning and preprocessing lectures. Integrating two datasets (steps activity and glucose readings) to create a time-synced dataset suitable for analysis and modeling.

In R, I applied visualization techniques such as scatter plots, boxplots, and regression analysis to explore trends and relationships, a direct application of our discussions on graphical representation of data. For modeling, I implemented a Random Forest Regressor in Python using scikit-learn, reflecting the machine learning applications covered in class, particularly regression techniques with numerical features.

Finally, I built and deployed an interactive Streamlit web application to make these insights accessible in a dynamic format, simulating a real-world health data tool. This project reflects key topics from class including data integration, feature engineering, encoding categorical variables, visualization techniques, and predictive modeling workflows.

Why is this project important and why did I chose to research this:
This project is important because it addresses a practical, real-world need for people living with Type 2 Diabetes, especially older adults who may struggle with complex health apps or tracking tools. By creating a simple, easy-to-use web application, this project empowers individuals to track their walking activity and better understand how it might influence their blood glucose levels. This tool isn't just valuable for individuals; it also has broader importance for researchers and healthcare providers. As wearable devices and continuous glucose monitors become more common, tools like this can help analyze and visualize personal health data, paving the way for more personalized, data-driven care.

I chose to research this topic because diabetes management is a growing challenge globally, particularly among older populations who could benefit from accessible digital tools. Even simple, intuitive apps like this can support better daily decision-making and improve quality of life. Additionally, the project demonstrates how combining data science techniques with user-friendly interfaces can translate raw health data into meaningful insights, a critical step for both individual self-care and population health research.

Related Works:
Before starting this project, I explored popular libraries for data analysis, visualization, and machine learning, including pandas, scikit-learn, Streamlit, and ggplot2 in R. While existing commercial tools track glucose data, many are too technical or costly for broad use. This project builds on accessible, open-source tools to create a simple, interactive solution that helps users understand how walking activity relates to glucose management.

Data Technique:
The datasets used in this project were sourced from publicly available health datasets on Kaggle. The glucose data is derived from the OhioT1DM dataset, which contains continuous glucose monitoring (CGM) records for individuals with diabetes, capturing glucose readings over time. The steps dataset includes timestamped walking activity data, allowing for analysis of how recent physical activity relates to blood glucose trends.

To prepare the third dataset, I merged glucose level data with step count data from two CSV files. It first loads the data into pandas DataFrames and creates a `timestamp` column for the glucose data by combining `date` and `time` columns or converting an existing `timestamp`. Both DataFrames were sorted chronologically, and a 30-minute window defined to calculate the total steps taken prior to each glucose measurement. This total is stored in a new column, `steps_last_30min`. Additionally, the code categorizes the time of day into morning, afternoon, evening, and night based on the hour of the `timestamp`. Finally saving the CSV file named `model_ready_dataset.csv`.

	date	time	level	timestamp	steps_last_30min	time_of_day
2	18-01-2022	00:01:00	179	2022-01-18 00:01:00	0.0	night
3	18-01-2022	00:06:00	183	2022-01-18 00:06:00	5139.0	night
4	18-01-2022	00:11:00	187	2022-01-18 00:11:00	5139.0	night
5	18-01-2022	00:16:00	191	2022-01-18 00:16:00	5139.0	night
6	18-01-2022	00:21:00	195	2022-01-18 00:21:00	12555.0	night
7	18-01-2022	00:26:00	199	2022-01-18 00:26:00	12555.0	night
8	18-01-2022	00:31:00	204	2022-01-18 00:31:00	12555.0	night
9	18-01-2022	00:36:00	209	2022-01-18 00:36:00	7416.0	night
10	18-01-2022	00:41:00	211	2022-01-18 00:41:00	7416.0	night

	date	time	level
2	18-01-2022	00:01:00	179
3	18-01-2022	00:06:00	183
4	18-01-2022	00:11:00	187
5	18-01-2022	00:16:00	191
6	18-01-2022	00:21:00	195
7	18-01-2022	00:26:00	199
8	18-01-2022	00:31:00	204

```
glucose_df = pd.read_csv('glucose.csv')
steps_df = pd.read_csv('steps_with_timestamps.csv')

if 'date' in glucose_df.columns and 'time' in glucose_df.columns:
    glucose_df['timestamp'] = pd.to_datetime(glucose_df['date'] + ' ' + glucose_df['time'], dayfirst=True)
else:
    glucose_df['timestamp'] = pd.to_datetime(glucose_df['timestamp'], dayfirst=True)

steps_df['timestamp'] = pd.to_datetime(steps_df['timestamp'])

glucose_df = glucose_df.sort_values('timestamp')
steps_df = steps_df.sort_values('timestamp')

window = pd.Timedelta(minutes=30)
total_steps_list = []
```

Visual work:

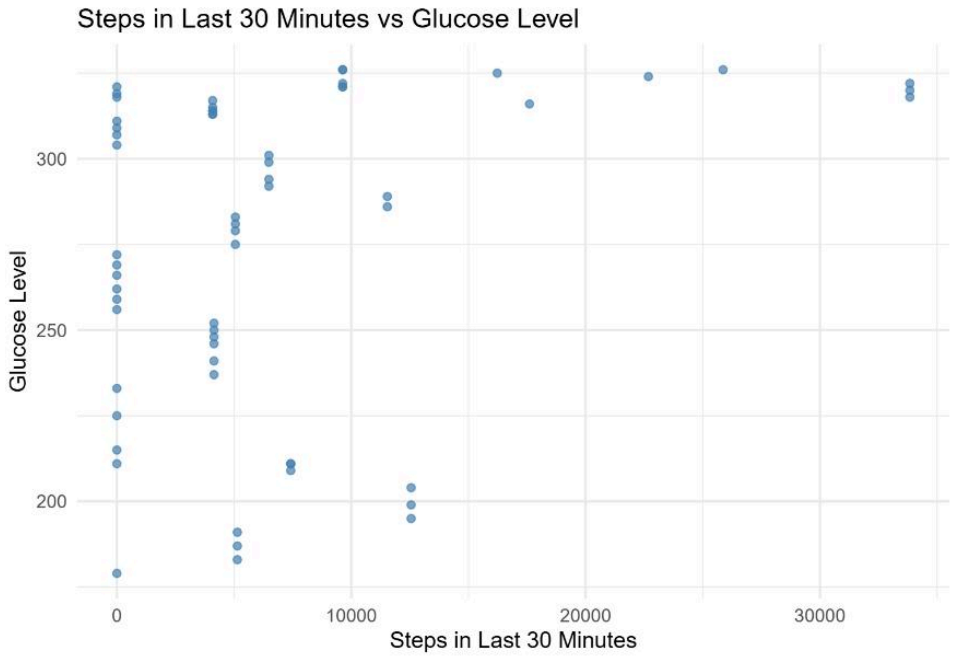
By providing visual analysis, individuals can better understand the potential impact of their physical activity on their health, particularly in relation to glucose management. thereby encouraging people to consider their habits and make informed decisions about their health and wellness.

Scatter plot:

The code utilizes data visualization techniques, specifically employing the ggplot2 library in R to create a scatter plot. This plot illustrates the relationship between the number of steps taken in the last 30 minutes (steps_last_30min) and glucose levels (level) from the dataset model_ready_dataset.csv. By using geom_point(), the code generates points on the graph, with customizations for color and transparency, and includes labels for the axes and a title to enhance clarity. The theme_minimal() function is applied to give the plot a clean and modern appearance.

Overall, I used this to aim at exploring and visualizing the correlation between physical activity (steps) and glucose levels.

```
ggplot(df, aes(x = steps_last_30min, y = level)) +
  geom_point(color = "steelblue", alpha = 0.7) +
  labs(
    title = "Steps in Last 30 Minutes vs Glucose Level",
    x = "Steps in Last 30 Minutes",
    y = "Glucose Level"
  ) +
  theme_minimal()
```



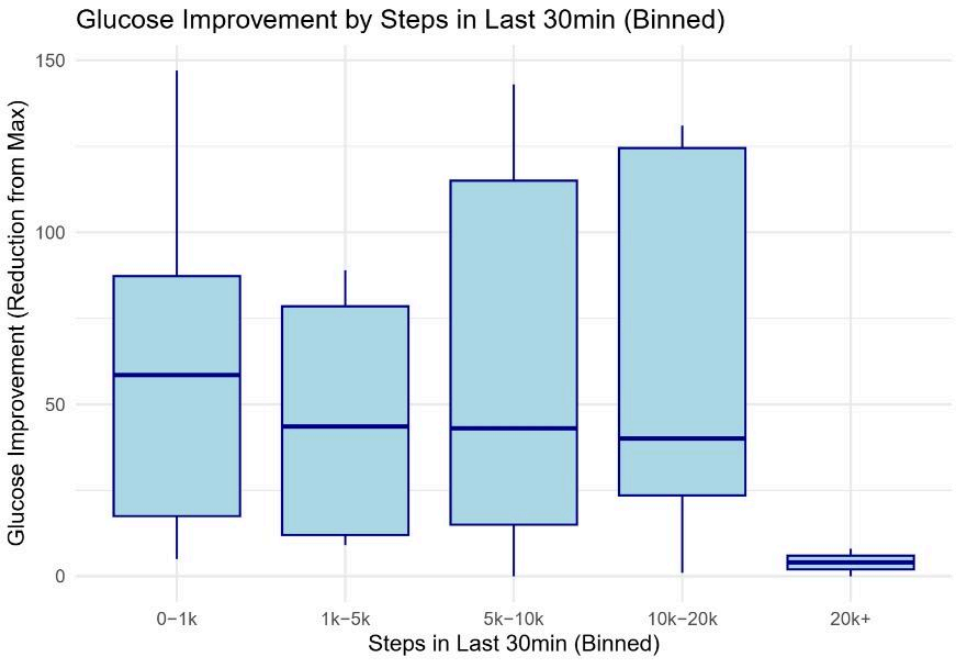
As you can see the glucose levels are ranging from different step levels. As 30 minutes is not enough time for a significant change, you can still see how it makes a gradual change in the plot. We also keep in mind that the absence of a strong short-term correlation may reflect the complexity of individual physiology and time lags between activity and measurable glucose changes, not a lack of benefit.

Box plot:

I then created a box plot using the ggplot2 library in R to visualize the relationship between binned step counts taken in the last 30 minutes and glucose improvement. The code reads the dataset model_ready_dataset.csv and calculates a new variable, glucose_improvement, which represents the reduction in glucose levels from the maximum recorded level. The step counts are then categorized into bins (e.g., "0-1k", "1k-5k", etc.) using the cut() function.

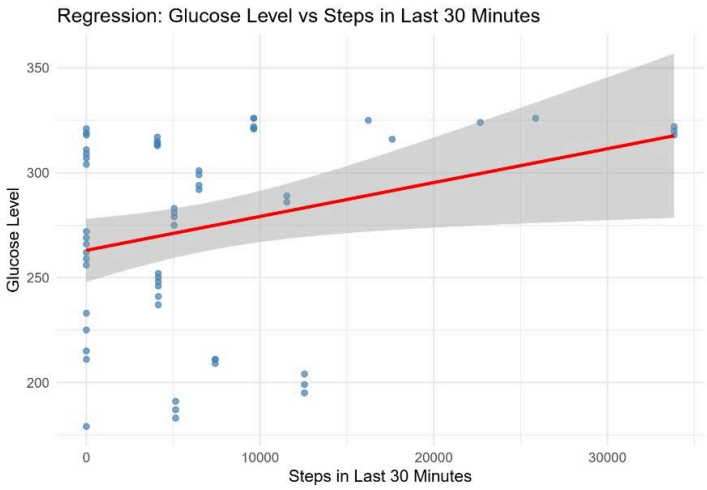
The box plot is generated with geom_boxplot(), displaying the distribution of glucose improvement for each step count bin. The plot is customized with labels for the title and axes, and a minimal theme is applied for clarity.

```r
df <- df %>%
  mutate(glucose_improvement = max(level, na.rm = TRUE) - level)

df <- df %>%
  mutate(steps_bin = cut(steps_last_30min,
                         breaks = c(-Inf, 1000, 5000, 10000, 20000, Inf),
                         labels = c("0-1k", "1k-5k", "5k-10k", "10k-20k", "20k+")))

ggplot(df, aes(x = steps_bin, y = glucose_improvement)) +
  geom_boxplot(fill = "lightblue", color = "darkblue") +
  labs(
    title = "Glucose Improvement by Steps in Last 30min (Binned)",
    x = "Steps in Last 30min (Binned)",
    y = "Glucose Improvement (Reduction from Max)"
  ) +
  theme_minimal()
```



The boxplot summarizes how glucose improvement varies across different levels of walking activity. While there is variability within each group, it suggests that moderate walking (1k–10k steps) tends to be associated with higher reductions in glucose from baseline.

Regression analysis:
Although walking can play a part in decreasing glucose levels, it all ties back to lifestyle and other habits one builds.



This regression suggests other complexity of glucose regulation: factors like food intake, insulin administration, and physiological lags between exercise and glucose response can obscure this short-term relationship.

Random Forest Regressor:
I applied a Random Forest regression model to predict glucose level. The model consisted of 500 trees, with one variable considered at each split. The root mean squared error (RMSE) is approximately 41 mg/dL, indicating that the typical prediction error was around 41 mg/dL. The model explained approximately 11.6% of the variance in glucose levels, suggesting that while walking activity and time of day provide some predictive signal, most of the variation in glucose levels remains unexplained by these two variables alone.

This relatively low explanatory power highlights the complexity of glucose regulation, which depends on multiple factors such as meal intake, medication timing, stress, and individual physiology factors not captured in this dataset. However, this model provided a useful foundation for understanding how walking and time of day contribute to short-term glucose variation.

```r
df$time_of_day <- as.factor(df$time_of_day)

rf_model <- randomForest(level ~ steps_last_30min + time_of_day, data = df, ntree = 500)


print(rf_model)
```

jistic-Regression.pdf

```
## No. of variables tried at each split: 1
##
##           Mean of squared residuals: 1883.001
##                     % Var explained: 11.07
```



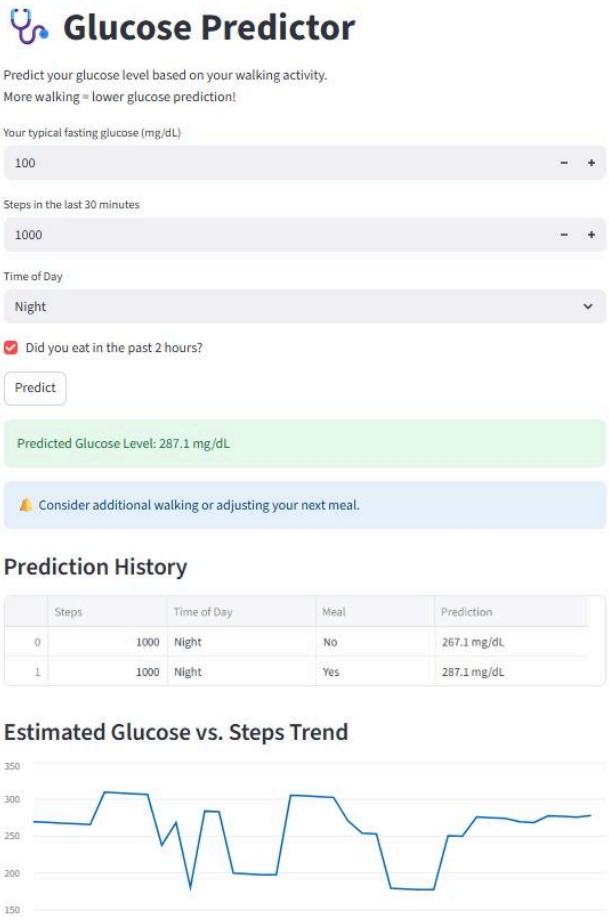
```r
predicted_levels <- predict(rf_model, df)

rmse <- sqrt(mean((predicted_levels - df$level)^2))
print(paste("RMSE:", round(rmse, 2)))
```

```
## [1] "RMSE: 41.18"
```

Streamlit web showcase:

Lastly I created a web application using Streamlit that allows users to predict their glucose levels based on their walking activity. I imported essential libraries such as streamlit, pandas, numpy, and sklearn, with the RandomForestRegressor model being utilized for making predictions. The application reads the dataset (model_ready_dataset.csv), to process the time_of_day column into categorical codes, and define the features and target variable for the model. After training the Random Forest model on the data, the application is configured with a title and introductory text to inform users about its purpose.

This will allow users to input their typical fasting glucose level, the number of steps taken in the last 30 minutes, select the time of day, and indicate if they have eaten recently. Upon clicking the "Predict" button, it will generate a prediction based on the user inputs, making adjustments for the number of steps and recent meals. The adjusted prediction is displayed along with health-related messages. Finally, the application maintains a history of predictions and visualizes the relationship between steps and predicted glucose levels through the datasets. This interactive tool effectively combines machine learning with user engagement to promote better health management.



In conclusion, this project effectively explores the relationship between walking activity and blood glucose levels in individuals with Diabetes, providing a comprehensive and interactive tool for better health management. By leveraging data cleaning, integration, and visualization techniques, the project highlights how physical activity can influence glucose levels while considering factors such as time of day and recent meals. The development of a user-friendly Streamlit web application makes these insights accessible, empowering individuals to track their walking habits and understand their impact on glucose management.