Contenido

[**Background Information** 2](#_Toc204544196)

[**Objective:** 2](#_Toc204544197)

[**Methodology** 3](#_Toc204544198)

[Who 3](#_Toc204544199)

[Where 3](#_Toc204544200)

[When The exact date of data collection is not specified. However, it can be reasonably assumed that the data originates from around the year 2024, based on the timing of the competition in which it was provided. 3](#_Toc204544201)

[How 3](#_Toc204544202)

[**Data Description** 4](#_Toc204544203)

[**Data Exploration** 5](#_Toc204544204)

[**Missing Data:** 5](#_Toc204544205)

[Nulls by columns 5](#_Toc204544206)

[Nulls by user 7](#_Toc204544207)

[**Current situation dataset** 8](#_Toc204544208)

[Sliding Window (ML/Forecasting) – Concept 8](#_Toc204544209)

[Input 8](#_Toc204544210)

[Target 8](#_Toc204544211)

[Time Series 9](#_Toc204544212)

[Overall 9](#_Toc204544213)

[**Suggest how to manage it?** 9](#_Toc204544214)

**Background Information**

**Objective:**

The purpose of this dataset is to provide comprehensive information on several key health indicators. It includes data on:

* Blood glucose levels
* Insulin levels
* Heart rate readings
* Carbohydrate intake
* Step counts recorded at various times
* Calorie burns during specific periods
* Indicator of blood glucose level in future for train dataset

Overall physical activity levels

Additionally, the dataset aims to support the prediction of future blood glucose trends based on the combined analysis of these variables, the dataset does not specify clearly how to predict the blood glucose indicator, it’s disposed to be used as a source to predict the indicator using the techniques of data analysis

# **Methodology**

## Who

The dataset contains information from nine young adults living with type 1 diabetes.

## Where

The data was collected in the United Kingdom, although no specific city is mentioned.

**When**  
The exact date of data collection is not specified. However, it can be reasonably assumed that the data originates from around the year 2024, based on the timing of the competition in which it was provided.

## How

The dataset was collected as part of the BrisT1D Blood Glucose Prediction Competition, which ran from September to November 2024. It features real-world data from nine young adults in the UK living with type 1 diabetes. Participants wore continuous glucose monitors, used insulin pumps, and were provided with smartwatches to track physical activity. The goal of the study was to gather diverse, time-stamped health metrics to support the development of algorithms that predict blood glucose levels one hour into the future using the previous six hours of data.

This data collection effort aimed to address the complexity of managing type 1 diabetes, where factors presumably like carbohydrates, physical activity in steps and heart rates, and levels of insulin. Although the dataset does not prescribe a specific modeling method, it was designed to encourage data scientists to explore predictive techniques using real-world conditions.

# **Data Description**

train.csv

id - row id consisting of participant number and a count for that participant

p\_num - participant number

time - time of day in the format HH:MM:SS

bg-X:XX - blood glucose reading in mmol/L, X:XX(H:MM) time in the past (e.g. bg-2:35, would be the blood glucose reading from 2 hours and 35 minutes before the time value for that row), recorded by the continuous glucose monitor

insulin-X:XX - total insulin dose received in units in the last 5 minutes, X:XX(H:MM) time in the past (e.g. insulin-2:35, would be the total insulin dose received between 2 hours and 40 minutes and 2 hours and 35 minutes before the time value for that row), recorded by the insulin pump

carbs-X:XX - total carbohydrate value consumed in grammes in the last 5 minutes, X:XX(H:MM) time in the past (e.g. carbs-2:35, would be the total carbohydrate value consumed between 2 hours and 40 minutes and 2 hours and 35 minutes before the time value for that row), recorded by the participant

hr-X:XX - mean heart rate in beats per minute in the last 5 minutes, X:XX(H:MM) time in the past (e.g. hr-2:35, would be the mean heart rate between 2 hours and 40 minutes and 2 hours and 35 minutes before the time value for that row), recorded by the smartwatch

steps-X:XX - total steps walked in the last 5 minutes, X:XX(H:MM) time in the past (e.g. steps-2:35, would be the total steps walked between 2 hours and 40 minutes and 2 hours and 35 minutes before the time value for that row), recorded by the smartwatch

cals-X:XX - total calories burnt in the last 5 minutes, X:XX(H:MM) time in the past (e.g. cals-2:35, would be the total calories burned between 2 hours and 40 minutes and 2 hours and 35 minutes before the time value for that row), calculated by the smartwatch

activity-X:XX - self-declared activity performed in the last 5 minutes, X:XX(H:MM) time in the past (e.g. activity-2:35, would show a string name of the activity performed between 2 hours and 40 minutes and 2 hours and 35 minutes before the time value for that row), set on the smartwatch

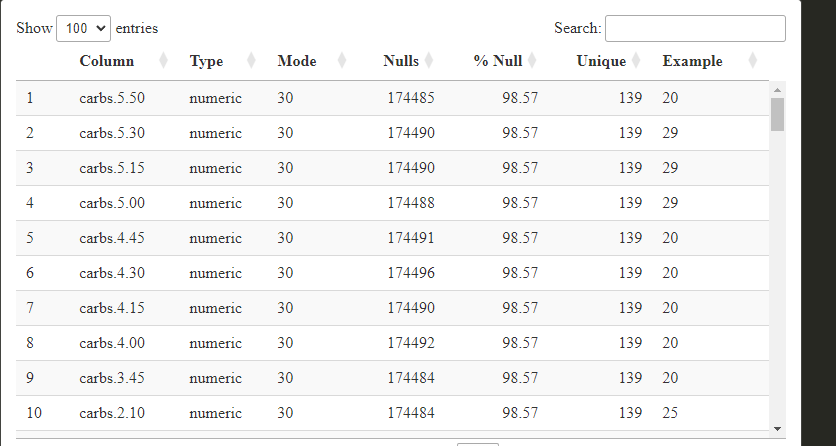
bg+1:00 - blood glucose reading in mmol/L an hour in the future, this is the value you will be predicting (not provided in test.csv)

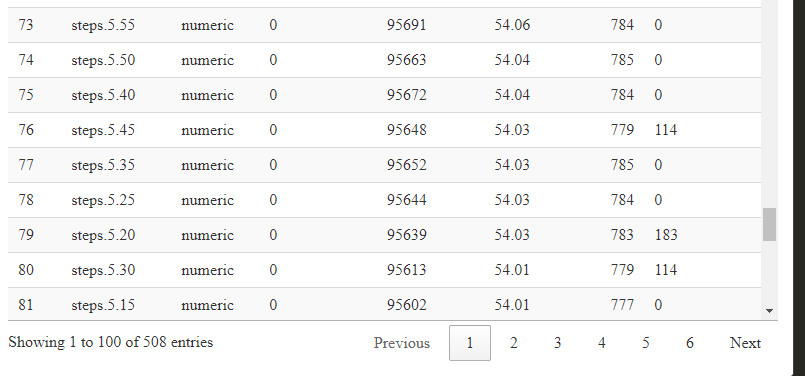
# **Data Exploration**

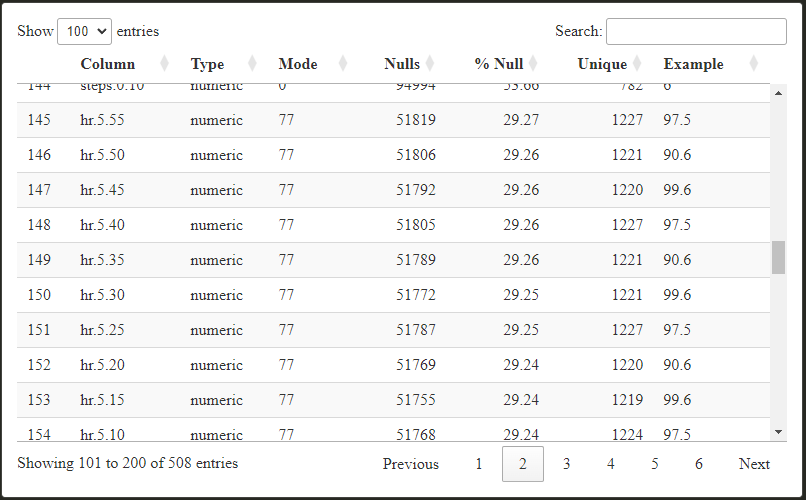
Accuracy  
Pending

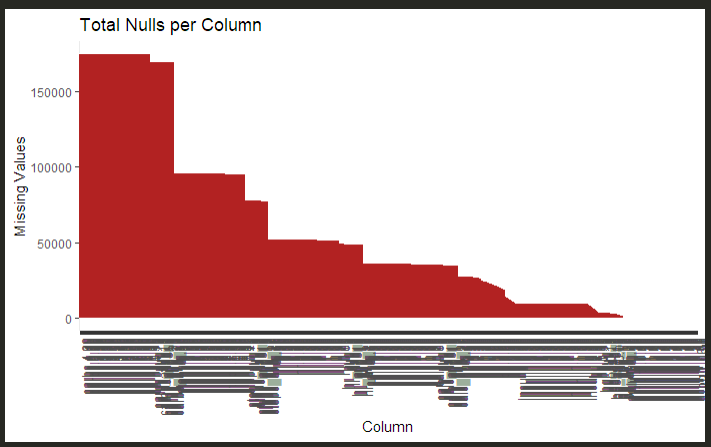
# **Missing Data:**

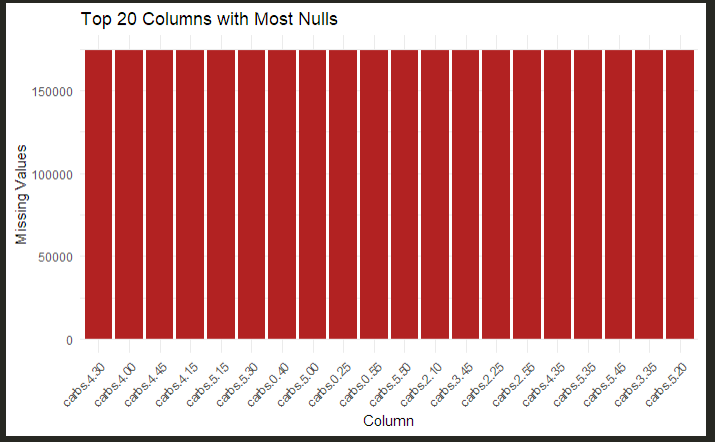
## Nulls by columns



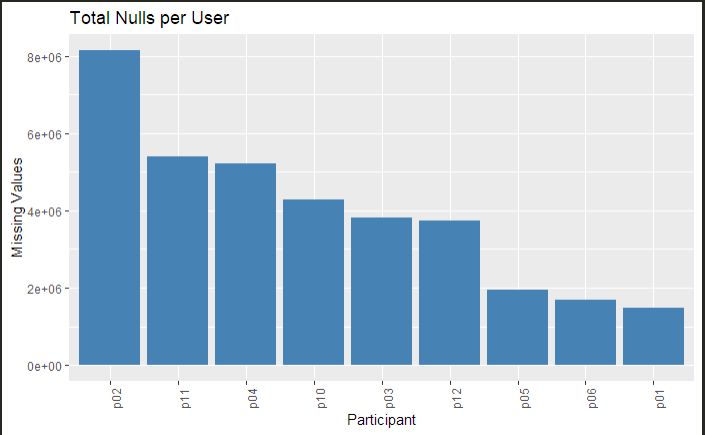








## Nulls by user



# **Current situation dataset**

## Sliding Window (ML/Forecasting) – Concept

A sliding window is a method used in time series or forecasting tasks where a fixed number of past data points (rows) are taken as input to predict a future value.

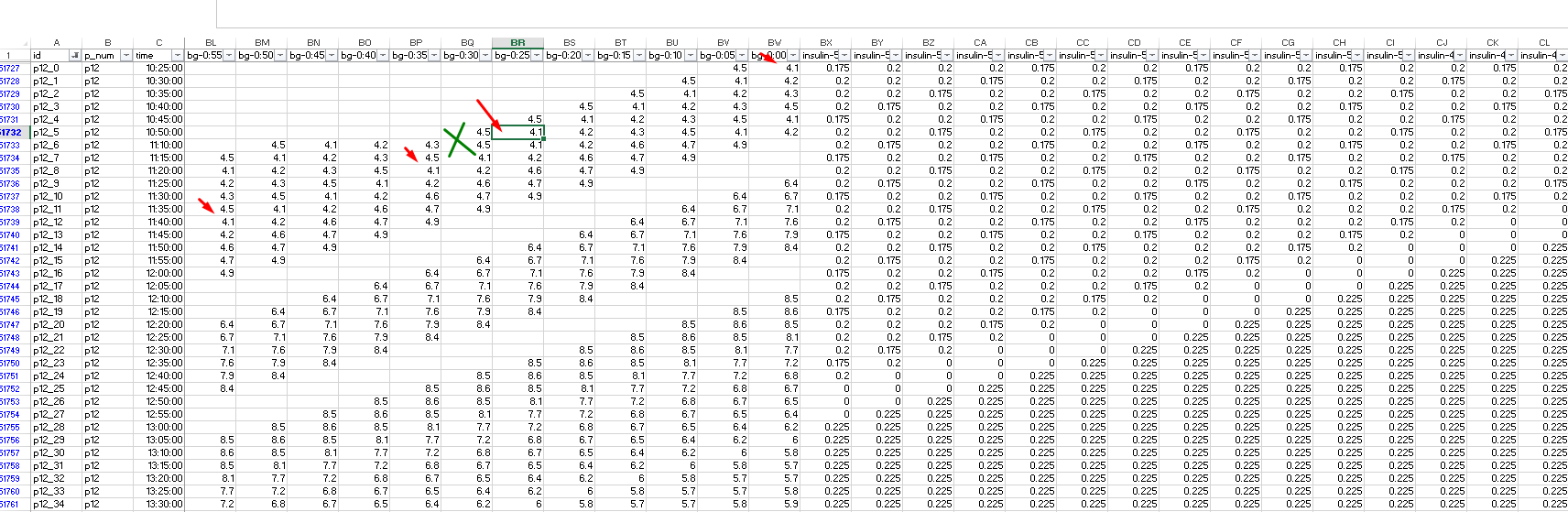
* Window length: Number of past rows to use (e.g., 4 rows = 1 hour).
* Stride: How far the window moves forward each time (e.g., 1 row = 15 mins).
* Columns each 72 columns represent 1h

## Input

Combined values of blood glucose, insulin, and meals over the window.

## Target

Future blood glucose value (e.g., in 15 or 60 minutes).



## Time Series

A **time series** is a sequence of data points collected or recorded **over time** at **regular intervals** (e.g., every 5 minutes, daily, monthly).

**In this scenario we have on columns a row in different distributions**

## Overall

The majority of the dataset consists of **numeric variables** representing **measured values** (e.g., blood glucose levels, insulin doses, or meal counts). These variables are continuous or discrete and are suitable for statistical analysis and modeling.

# **Suggest how to manage it?**