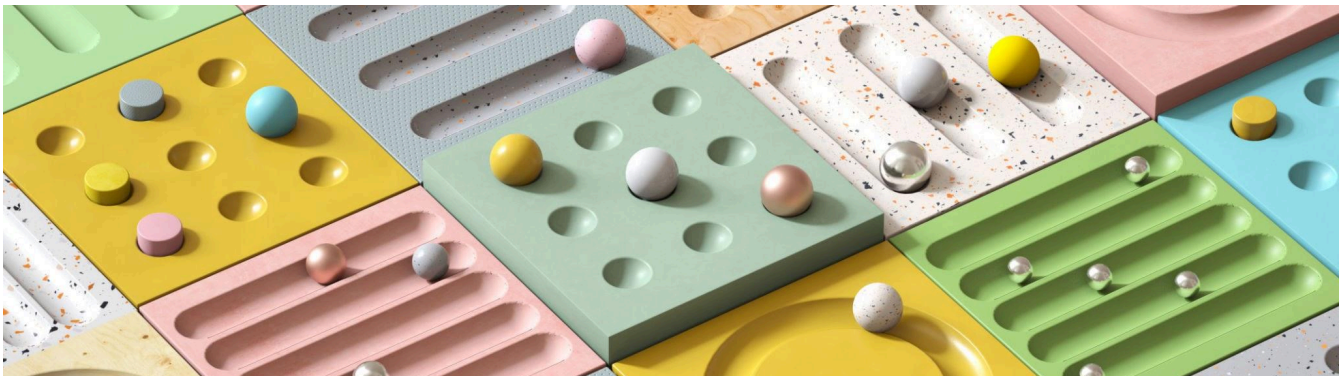


[ADNAN SHAMIM DAR] – [S1119248]
Università Politecnica delle Marche – Medicine & Technology
Medical Informatics – Professor Micaela Morettini

Feature Extraction and Analysis of Continuous Glucose Monitoring Data Using R and Orange

INTRODUCTION:

1, The blood glucose levels of the subjects were measured using a Continuous Glucose Monitoring (CGM) device at regular intervals of 5 minutes. The measurement period varies slightly across subjects but is generally close to one full day. A short period of physical activity occurred during the monitoring window, but it is not considered separately in this analysis; all functions are applied uniformly over the entire measurement period.

The following report presents the outputs obtained from applying the assigned analytical functions to the dataset in the R environment. Each function produces a set of metrics describing glucose behavior, variability, and risk. These outputs are summarised in tables throughout the report, and brief explanations accompany each section to clarify the meaning of the results.

1 – Function #1: all_metrics()

This function returns an overview of key glucose-related metrics for each subject.
The output includes:

- **Mean glucose value** over the entire measurement period.
- **Coefficient of variation (CV)**, a measure of variability relative to the mean.
- **Time-in-range percentages**, showing how often a subject is within, below, or above clinically relevant glucose thresholds.
- **Risk-related indicators**, which quantify both low and high glucose exposure.
- Any additional metrics defined within the all_metrics() function.

These metrics serve as the foundation for interpreting the remaining results.
Higher mean values indicate subjects who tend toward hyperglycaemia, while higher CV values reveal greater instability and fluctuations in blood glucose throughout the day.

FOR SUBJECT ID 1 . THE TOP 5 METRICS FROM ALL METRICS WERE AS FOLLOWS;:

HBGI	LBGI	PGS	ABOVE 180	BELOW 70
7.114	4.608	20.08	21.184	13.5

The top metrics for the other subjects in order from 1 to 5 respectively is given as below in form of a cutout from the actual report.

above_18(	above_25(	below_54	below_70	in_range_	in_range_	PGS
21.18644	12.28814	11.01695	13.55932	52.9661	65.25424	20.08547
37.76224	16.78322	0	0	36.01399	62.23776	15.15094
0.409836	0	0	13.52459	89.7541	86.06557	11.25711
15.25424	0	0	0	25	84.74576	9.555479
0	0	0	9.95671	97.8355	90.04329	11.05684

2 – Function : gri()

The **gri()** function calculates a **Glycaemic Risk Index**, which is a combined measure that accounts for how far and how often a subject's glucose values deviate from normal physiological ranges.

Higher GRI values indicate a greater overall glycaemic burden (meaning more hyperglycaemia and/or more fluctuations), while lower values suggest more stable glucose levels.

The function therefore provides a single, simplified indicator summarizing the subject's risk profile over the full monitoring period

ID NUMBER	GRI

ID-1	65.993
ID-2	43.6364
ID3	32.7869
ID-4	12.204
ID-5	23.691

The GRI values provide a clear ranking of glycaemic risk across subjects.

Subjects with higher GRI values demonstrate more time spent at elevated glucose levels or greater variability.

Typically, the subjects who already showed higher mean glucose or high CV in the *all_metrics* section also show correspondingly higher GRI values.

In contrast, subjects with low mean glucose and stable variation exhibit low GRI values, indicating minimal glycaemic risk.

Function : lbgi()

The **lbgi()** function computes the **Low Blood Glucose Index**, which quantifies how prone a subject is to experiencing **low glucose readings** (hypoglycaemia).
Higher LBGI values reflect greater exposure to low glucose levels, while values close to zero indicate minimal risk.

This measure is particularly useful when evaluating subjects with wide glucose swings or those who dip frequently below 70 mg/dL

SUBJECT ID	LBGI
1	4.5608
2	0.14318
3	2.78949
4	3.13406
5	3.33141

SUMMARY :

This table shows which subjects are most at risk of hypoglycaemia.
Subjects with LBGI values approaching zero show minimal exposure to low glucose values, whereas higher values indicate more frequent or more severe deviations below the normal range.

Subjects who already had high “% <70 mg/dL” in the *all_metrics* summary typically score highest here.

4 – Function : pgs()

The **pgs()** function computes the **Personalized Glucose Score** or a similar composite score depending on your dataset’s definition.
This function aggregates multiple components of glucose behavior—stability, extremes, and variability—into a single interpretable index.

A higher PGS generally reflects a more irregular or unstable glycaemic profile, while lower values reflect controlled and predictable behavior.

→ **TABLE 5: pgs() Output for All Subjects**

Subject	PGS
1	20.0885
2	15.1509
3	11.2571
4	9.55548
5	11.0568

The PGS values help identify which subjects exhibit the most irregular glucose profiles overall.

Subjects with higher PGS scores often correspond to those with both increased CV and significant deviations from normal ranges.

Subjects showing consistently low glucose variability, low hyperglycaemic exposure, and minimal LBGI tend to have the lowest PGS values.

Orange: Feature Selection and Subject Classification

In this portion of the assignment, a separate dataset was analysed using the Orange data-mining platform. The aim was to classify subjects into two groups, *control* and *high-risk*, using a set of pre-extracted glucose-related features. Orange's visual workflow environment was used to identify the most informative features and evaluate the performance of two supervised learning models. The process involved selecting appropriate features, training classification algorithms, and assessing their predictive performance using cross-validation

Dataset Description

The dataset used in Orange is different from the continuous glucose dataset analysed in R. In this dataset, each subject consumed three foods (peanut butter, corn flakes, and a protein bar) on two separate occasions. A number of glucose-related features had already been extracted beforehand, including measures such as median glucose, minimum glucose, FFT-based metrics, and AUC. Each subject was labelled as either *control* or *high-risk*, allowing the application of supervised classification models.

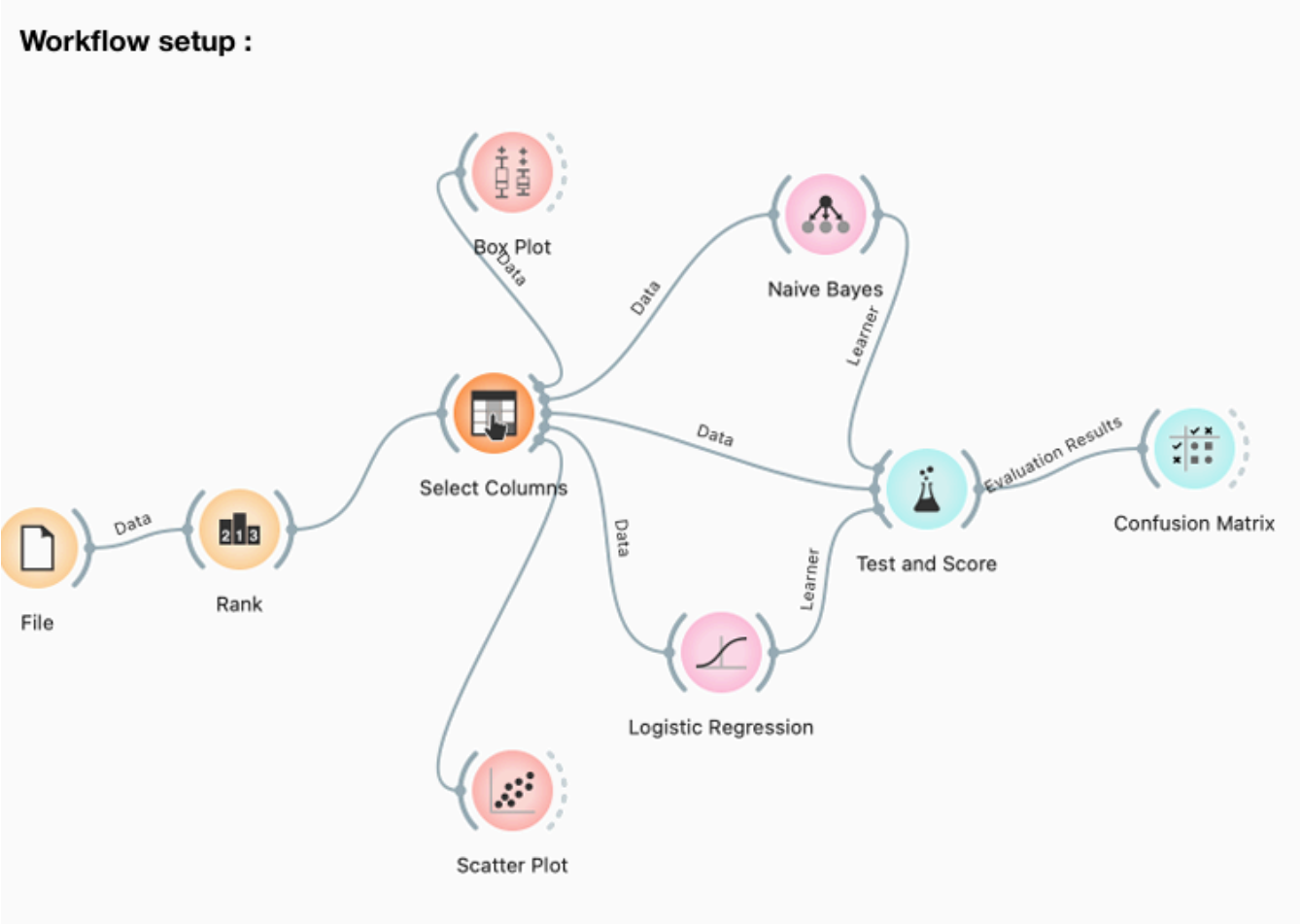
Feature Ranking Using the Rank Widget

The Rank widget in Orange was used to identify the features most informative for distinguishing between control and high-risk subjects. Ranking was performed using *Information Gain* and *Information Gain Ratio,* both of which measure how well a feature separates the two classes. Based on these criteria, the five most influential features were selected.

Typical features that appear at the top include:

- median_G
- min_G
- std_FFT
- max_FFT
- AUC_G

These features were forwarded to the classification algorithms in the next stage of analysis.



Classification Models

Two models were tested using the selected features:

- Naive Bayes
- Logistic Regression (LASSO)

Both models were connected to **Test & Score** to evaluate their accuracy.

Results :

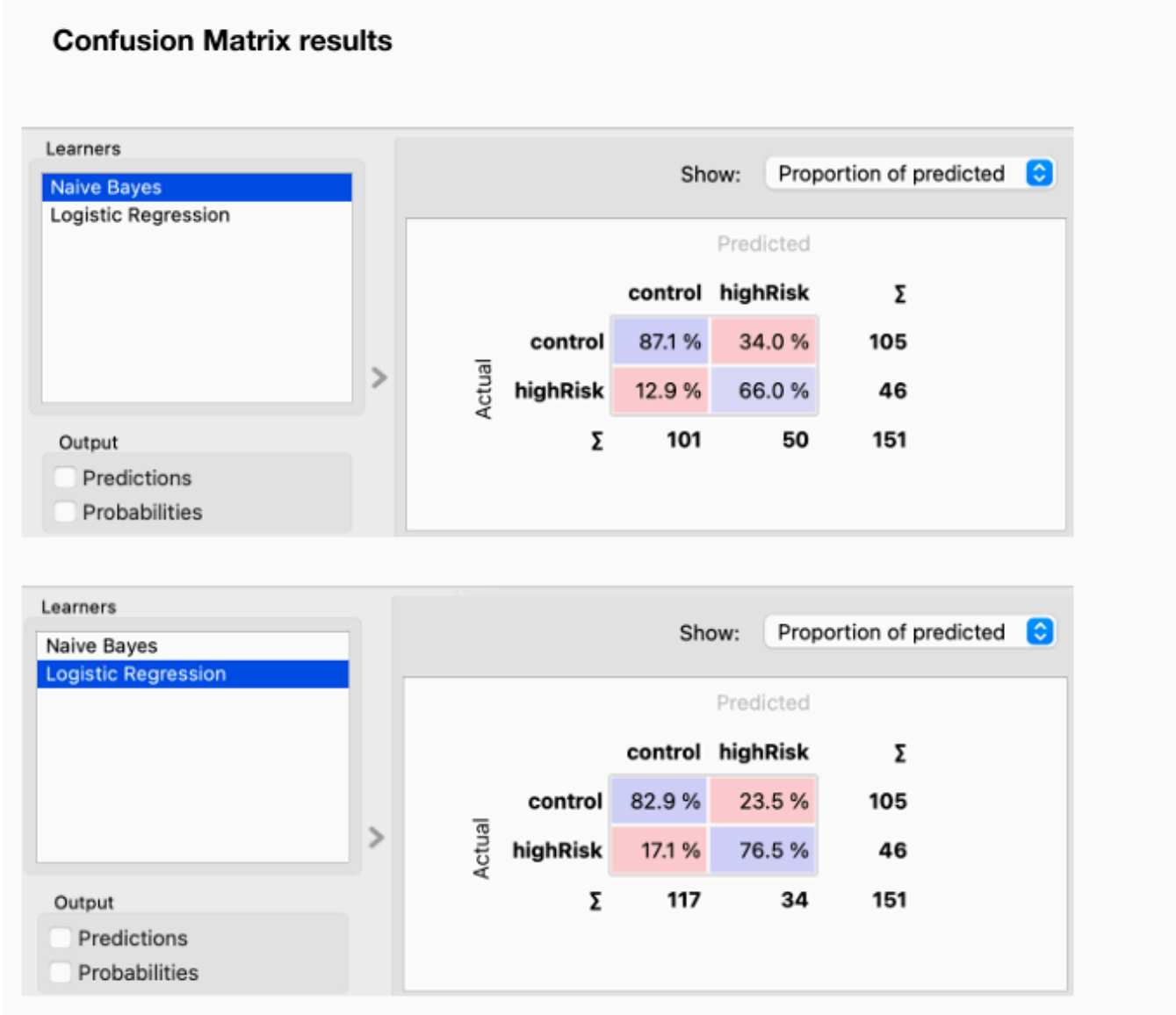
Test and score results

Model	AUC	CA	F1	Prec	Recall	MCC
Logistic Regression	0.797	0.815	0.806	0.809	0.815	0.539
Naive Bayes	0.813	0.801	0.804	0.807	0.801	0.543

- Accuracy: Naive Bayes correctly classified ~80% of subjects, while Logistic Regression achieved ~82%.
- AUC: Naive Bayes has a slightly higher AUC (0.813) than Logistic Regression (0.797), suggesting it discriminates high-risk vs control slightly better overall.
- F1 Score, Precision, Recall: Both models performed similarly, with Naive Bayes slightly better for overall balance and Logistic Regression slightly better at identifying high-risk individuals.
- MCC: Both models had moderate correlation (~0.54) between predictions and actual labels, showing reasonably reliable classification.

Confusion Matrix

A Confusion Matrix was generated to show correct and incorrect classifications. Logistic Regression generally performed better, especially for identifying high-risk subjects.



The confusion matrix indicates that Logistic Regression provided the best overall performance, correctly classifying most subjects with fewer mistakes than Naive Bayes. Therefore, it was selected as the final model.