University of Hertfordshire UH

School of Physics, Engineering and Computer Science

MSc Data Science Project 7PAM2002

Department of Physics, Astronomy and Mathematics

Data Science FINAL PROJECT REPORT

Project Title:

Multivariate Analysis of Glycaemic Control Using demographic, lifestyle, and clinical predictors in Type 1 Diabetes

Student Name and SRN:

Ans Riaz – 23079633

Supervisor: Dr. Stephen Kane

Date Submitted: 03/01/2026

Word Count: ≈ 4211 (excluding references)

Github Repository: https://github.com/ansmalik67/multimodal-glucose-forecasting.git

DECLARATION STATEMENT

This report is submitted in partial fulfilment of the requirement for the degree of Master of Science **in Data Science** at the University of Hertfordshire.

I have read the detailed guidance to students on academic integrity, misconduct and plagiarism information at Assessment Offences and Academic Misconduct and understand the University process of dealing with suspected cases of academic misconduct and the possible penalties, which could include failing the project or course.

I certify that the work submitted is my own and that any material derived or quoted from published or unpublished work of other persons has been duly acknowledged. (Ref. UPR AS/C/6.1, section 7 and UPR AS/C/5, section 3.6)

I did not use human participants in my MSc Project.

I hereby give permission for the report to be made available on module websites provided the source is acknowledged.

Student Name printed: Ans Riaz

Student Name signature: Ans Riaz

Student SRN number: 23079633

UNIVERSITY OF HERTFORDSHIRE

SCHOOL OF PHYSICS, ENGINEERING AND COMPUTER SCIENCE

Acknowledgement

I would like to sincerely thank my supervisor, **Dr. Stephen Kane** and my friend **Sahar Saeed** for his invaluable guidance, encouragement, and constructive feedback throughout this project. His expertise and support have been instrumental in shaping my research and helping me stay focused.

I am also grateful to the **University of Hertfordshire's MSc Data Science programme team** for providing the resources and academic environment that made this work possible.

Finally, I would like to thank my **family and friends** for their patience, understanding, and constant motivation during this journey. Their support has been my greatest source of strength.

Table of Contents

Abstract

Effective glycaemic control is essential for diabetes management, yet many clinical datasets do not include continuous glucose monitoring or detailed time-series information. This study examines whether blood glucose levels can be predicted using structured demographic, lifestyle, and clinical variables alone. The Weinstock dataset was used to develop three regression models: Ridge Regression, Random Forest, and Gradient Boosting. Ridge Regression was applied as a linear baseline model, while Random Forest and Gradient Boosting were used to capture non-linear relationships between predictors and glucose levels. All models were trained and evaluated using a consistent experimental framework, and standard regression metrics were used to assess predictive performance on unseen data. The findings show that the linear baseline model provides limited predictive capability, indicating that glucose variability cannot be adequately explained using linear relationships alone. In contrast, tree-based ensemble models demonstrate improved performance, with Random Forest achieving the strongest overall results. These findings highlight the importance of non-linear modelling for multivariate glucose prediction using structured clinical data.

1. Introduction

Diabetes mellitus is a long-term medical condition that affects millions of people worldwide and requires continuous management throughout life. One of the most important parts of diabetes care is maintaining blood glucose levels within a healthy range. Poor glycaemic control can lead to serious complications, including cardiovascular disease, kidney failure, nerve damage, and vision problems (Obermeyer and Emanuel, 2016). Because of these risks, understanding the factors that influence blood glucose levels has become a major focus in diabetes research.

Many previous studies have relied on physiological models or short-term time-series data collected using continuous glucose monitoring (CGM) devices. These approaches are effective for short-term prediction but depend on high-frequency data that is not always available in real-world clinical settings. In many population-level healthcare datasets, researchers instead have access to structured information describing patient demographics, lifestyle habits, and clinical conditions (Contreras and Vehi, 2018).

These variables influence glucose levels indirectly and often over long periods of time. As a result, predicting glucose using structured clinical data is a complex task that requires methods capable of modelling multiple interacting factors.

Machine learning offers flexible tools for analysing such complex relationships. Regression-based machine learning models are particularly suitable because glucose is a continuous outcome variable. Linear models such as Ridge Regression are easy to interpret but have limited ability to represent non-linear relationships (Hoerl and Kennard, 1970). Tree-based ensemble models, including Random Forest and Gradient Boosting, are better suited for capturing non-linear effects and interactions that commonly occur in medical data (Breiman, 2001; Friedman, 2001).

This project focuses on predicting glucose levels using multivariate machine learning models applied to structured clinical data. Rather than forecasting short-term glucose fluctuations, the aim is to examine how demographic, lifestyle, and clinical factors together relate to overall glycaemic control.

2. Research Motivation and Objectives

The motivation for this study is based on two key observations. First, glucose regulation is influenced by many interacting factors beyond immediate food intake or insulin dosage. Factors such as age at diagnosis, body weight, physical activity, comorbid conditions, and medication use all contribute to long-term glycaemic control (Contreras and Vehi, 2018). These interactions make glucose prediction a complex modelling problem.

Second, many real-world healthcare datasets do not include detailed time-series or behavioural data, which limits the use of advanced forecasting techniques. This creates a need to explore what can be learned from structured clinical data alone.

The main objective of this research is to evaluate whether multivariate machine learning models can meaningfully predict glucose levels using demographic, lifestyle, and clinical variables. Rather than aiming for perfect prediction, the focus is on understanding how much of the variation in glucose levels can be explained using these features.

The specific objectives are:

- To develop baseline regression models for glucose prediction
- To compare linear and non-linear machine learning approaches
- To apply hyperparameter tuning and assess its impact on model performance
- To identify the most suitable model for multivariate glucose prediction
- To interpret results in a clinically meaningful and accessible way

3. Literature Review

3.1 Machine Learning in Glucose Prediction

Machine learning has become increasingly important in diabetes research, particularly for analysing large healthcare datasets and supporting clinical decision-making. Many studies focus on short-term glucose prediction using CGM data and apply deep learning models designed for time-series analysis. While these methods can achieve high accuracy, they require continuous and detailed glucose measurements (Contreras and Vehi, 2018).

In contrast, many clinical datasets only contain structured information collected during routine care. For such data, classical machine learning models that operate on demographic, lifestyle, and clinical variables are more appropriate. These approaches aim to capture long-term patterns in glycaemic control rather than immediate glucose fluctuations.

3.2 Linear Models in Medical Prediction

Linear regression models are widely used in healthcare because they are simple, transparent, and easy to interpret. Ridge Regression extends linear regression by introducing a regularisation term that penalises large coefficients, helping to reduce overfitting and manage correlated predictors (Hoerl and Kennard, 1970).

In medical prediction tasks, linear models are often used as baseline approaches. They provide a useful reference point but have limited ability to represent complex relationships. Several studies have shown that linear models struggle to capture non-linear interactions between variables, which are common in physiological systems, leading to reduced predictive performance.

3.3 Tree-Based Ensemble Models

Tree-based ensemble models such as Random Forest and Gradient Boosting have been widely adopted in healthcare analytics. Random Forest combines multiple decision trees built on random subsets of data and features, which improves robustness and generalisation (Breiman, 2001). This makes it well suited for structured clinical datasets.

Gradient Boosting builds trees sequentially, with each new tree focusing on correcting previous errors. This allows the model to learn complex patterns gradually but also makes it sensitive to hyperparameter settings (Friedman, 2001). When properly tuned, Gradient Boosting has demonstrated strong performance in many regression tasks.

Previous research consistently shows that ensemble models outperform linear approaches when predicting complex health outcomes, including glucose levels.

3.4 Summary of Literature

Overall, the literature suggests that linear models are useful for establishing baselines, while non-linear ensemble methods are better suited for modelling the complexity of glucose regulation. This study builds on existing research by directly comparing Ridge Regression, Random Forest, and Gradient Boosting using a consistent dataset and evaluation framework.

4. Dataset Description

4.1 Overview of the Dataset

This study uses the Weinstock dataset of GlucoBench, a large structured clinical dataset containing information on individuals diagnosed with Type 1 Diabetes. The dataset is composed of variables collected during routine clinical and lifestyle assessments and is designed to support population-level analysis of diabetes-related outcomes. Unlike time-series datasets that rely on continuous glucose monitoring, the Weinstock dataset contains cross-sectional and aggregated patient-level information, making it suitable for multivariate machine learning analysis.

The dataset includes a broad range of patient attributes, allowing the investigation of how different demographic, lifestyle, and clinical factors are associated with blood glucose levels. Its structured format enables efficient preprocessing and modelling using classical machine learning techniques.

4.2 Target Variable

The target variable used in this study is `gl`, which represents the measured blood glucose level. This variable is continuous and reflects the level of glucose in the blood at the time of clinical measurement. Predicting this variable is the main objective of the study.

Blood glucose is influenced by multiple physiological and behavioural factors. However, in this dataset, short-term influences such as meal timing or insulin dosage are not available. As a result, the modelling task focuses on understanding how longer-term demographic, lifestyle, and clinical characteristics relate to glucose variability.

4.3 Predictor Variables

The predictor variables in the dataset can be divided into four main categories. Grouping the variables in this way helps clarify their role in the modelling process and improves interpretability.

4.3.1 Demographic Features

Demographic variables describe general patient characteristics that may influence glucose regulation over time. These include:

- Education level
- Annual income
- Age-related indicators, such as age at diabetes diagnosis

These features provide information about socioeconomic and background factors, which can indirectly affect health outcomes through access to care, health awareness, and lifestyle choices.

4.3.2 Lifestyle Factors

Lifestyle variables capture behavioural patterns that may influence long-term glycaemic control. These include:

- Frequency of physical activity
- Alcohol consumption behaviour

Lifestyle factors are important because they reflect daily habits that can affect metabolism and insulin sensitivity over time, even when short-term behavioural data are not available.

4.3.3 Clinical Variables

Clinical variables represent diagnosed medical conditions and treatment-related information recorded for everyone. Examples include:

- Presence of comorbid conditions
- Medication usage

These features are particularly important in diabetes research, as comorbidities and medications can significantly influence glucose regulation and overall metabolic health.

4.3.4 Anthropometric Measures

Anthropometric variables provide information related to body composition and physical characteristics. These include:

- Height
- Weight

Body composition is closely linked to insulin sensitivity and glucose metabolism, making these features valuable predictors in glucose-related modelling tasks.

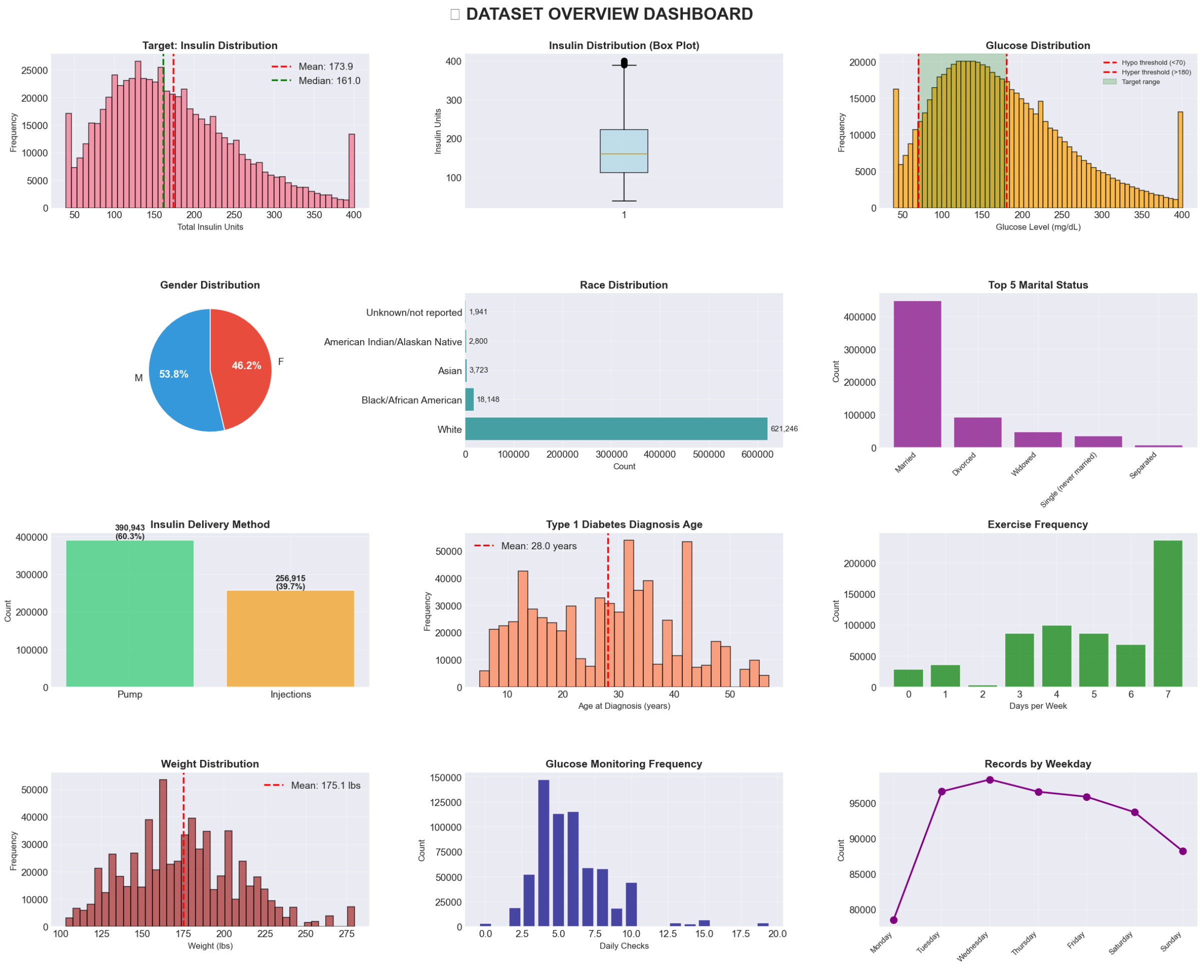


Figure 1: *Dataset overview dashboard showing the distributions of insulin units, glucose levels, demographic characteristics, clinical variables, and lifestyle factors in the Weinstock dataset.*

4.4 Dataset Size and Structure

After data cleaning and preprocessing, the dataset contains more than 500,000 valid observations. This large sample size provides a strong foundation for training machine learning models and reduces the risk of overfitting. The dataset is primarily composed of numeric variables, which simplifies feature selection and allows direct application of regression-based machine learning models.

The large number of observations also enables robust evaluation using train–test splits and cross-validation techniques, supporting reliable comparison between different modelling approaches.

4.5 Relevance of the Dataset to This Study

The Weinstock dataset is well suited to the objectives of this study for several reasons. First, it contains a diverse set of demographic, lifestyle, and clinical variables that are known to influence glycaemic control. Second, its structured format allows the application of both linear and non-linear machine learning models. Finally, the absence of detailed time-series data makes it an appropriate dataset for investigating multivariate prediction of glucose levels using population-level clinical information.

5. Data Preprocessing and Feature Selection

5.1 Importance of Data Preprocessing

Data preprocessing is a crucial step in any machine learning project because the quality of the input data directly affects model performance. Real-world clinical datasets often contain missing values, irrelevant variables, and inconsistencies that must be addressed before model training. In this study, a series of preprocessing steps were applied to ensure that the data used for modelling were clean, reliable, and suitable for regression analysis.

5.2 Target Variable Cleaning

The target variable for this study is blood glucose level (`gl`). To ensure valid modelling, rows with missing, invalid, or non-meaningful target values were removed. Only observations with positive glucose values were retained, as negative or zero values are not physiologically meaningful. This step ensured that the models were trained only on realistic and clinically valid measurements.

5.3 Removal of Irrelevant and Identifier Variables

Certain variables in the dataset do not provide useful predictive information or may introduce bias into the models. Non-numeric variables and identifier fields, such as patient identification numbers, were excluded from the feature set. These variables are not suitable for regression modelling and do not contribute to understanding glucose variability.

By removing such variables, the feature set was restricted to predictors that could meaningfully contribute to the prediction task.

5.4 Prevention of Data Leakage

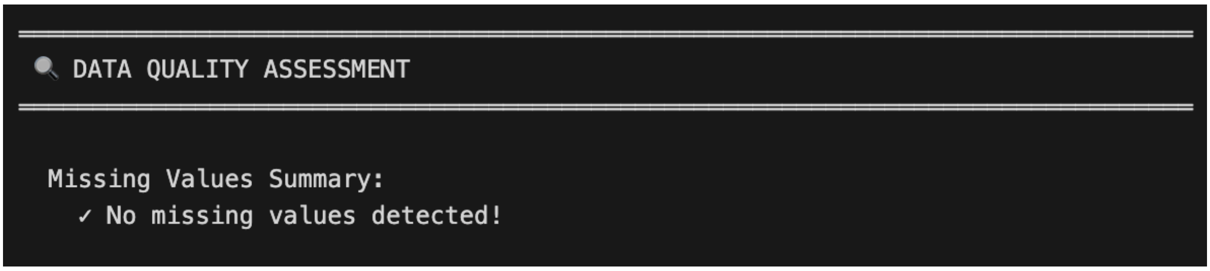
Preventing data leakage is essential to ensure that model performance reflects real predictive ability rather than artificial accuracy. In this study, special care was taken to remove any variables that directly represented or were derived from the target variable. Including such variables would allow the model to indirectly “see” the target during training, leading to misleadingly high performance.

Only predictors that would be available independently of the glucose measurement were retained. This ensures that the trained models reflect realistic prediction scenarios and can generalise to unseen data.

5.5 Handling of Missing Values

Missing values were present in several predictor variables. To address this, median imputation was applied to numeric features. The median was chosen instead of the mean because it is more robust to outliers and skewed distributions, which are common in clinical data.

This approach allowed all available observations to be retained while minimising the influence of extreme values on the imputed data.



```
================================================================================
🔍 DATA QUALITY ASSESSMENT
================================================================================

Missing Values Summary:
  ✓ No missing values detected!
```

Figure 2: *Data quality assessment showing that no missing values were detected in the dataset.*

5.6 Final Feature Selection

After cleaning the data, removing irrelevant variables, preventing leakage, and handling missing values, a final set of numeric features was selected for modelling. These features represent demographic, lifestyle, clinical, and anthropometric information relevant to glucose regulation.

The resulting dataset provided a clean and consistent input for training and evaluating the machine learning models used in this study.

6. Exploratory Data Analysis

6.1 Purpose of Exploratory Data Analysis

Exploratory Data Analysis (EDA) was carried out to gain an initial understanding of the dataset before applying machine learning models. The main goal of EDA was to examine the distribution of the variables, identify patterns and relationships, and detect any potential issues that could affect model performance. This step helps ensure that the modelling approach is appropriate for the structure of the data.

6.2 Distribution of Blood Glucose Levels

The distribution of the target variable, blood glucose level (`gl`), showed a wide range of values across the dataset. This indicates substantial variability in glucose measurements among individuals. Such variability suggests that predicting glucose levels is a challenging task, as values are influenced by many different factors rather than following a simple pattern.

The presence of this variability supports the need for models that can learn complex relationships from data rather than relying on simple assumptions.

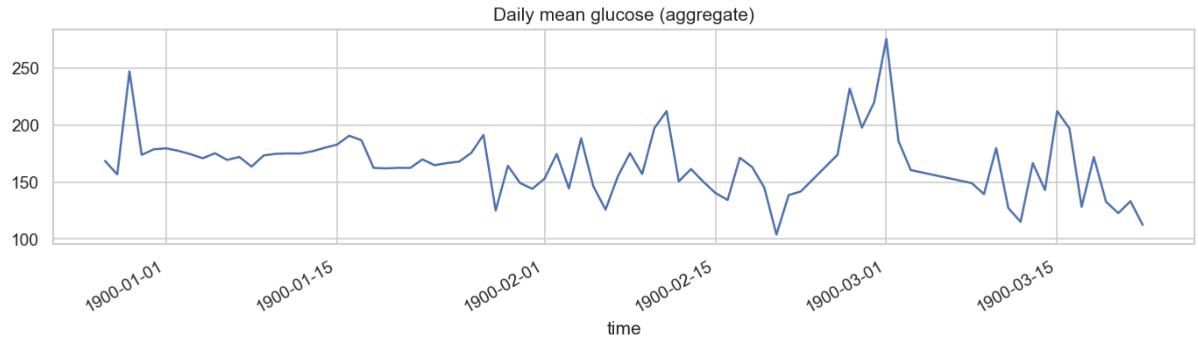


Figure 3: *Daily average blood glucose levels over time, showing how glucose values change across days.*

6.3 Correlation Analysis

Correlation analysis was performed to examine the linear relationships between predictor variables and blood glucose levels. The analysis showed that no single feature had a very strong linear correlation with glucose. This indicates that glucose levels are not dominated by one individual factor, but instead result from the combined influence of multiple variables.

This finding suggests that linear models may have limited ability to capture glucose variability and that more flexible modelling approaches may be required.

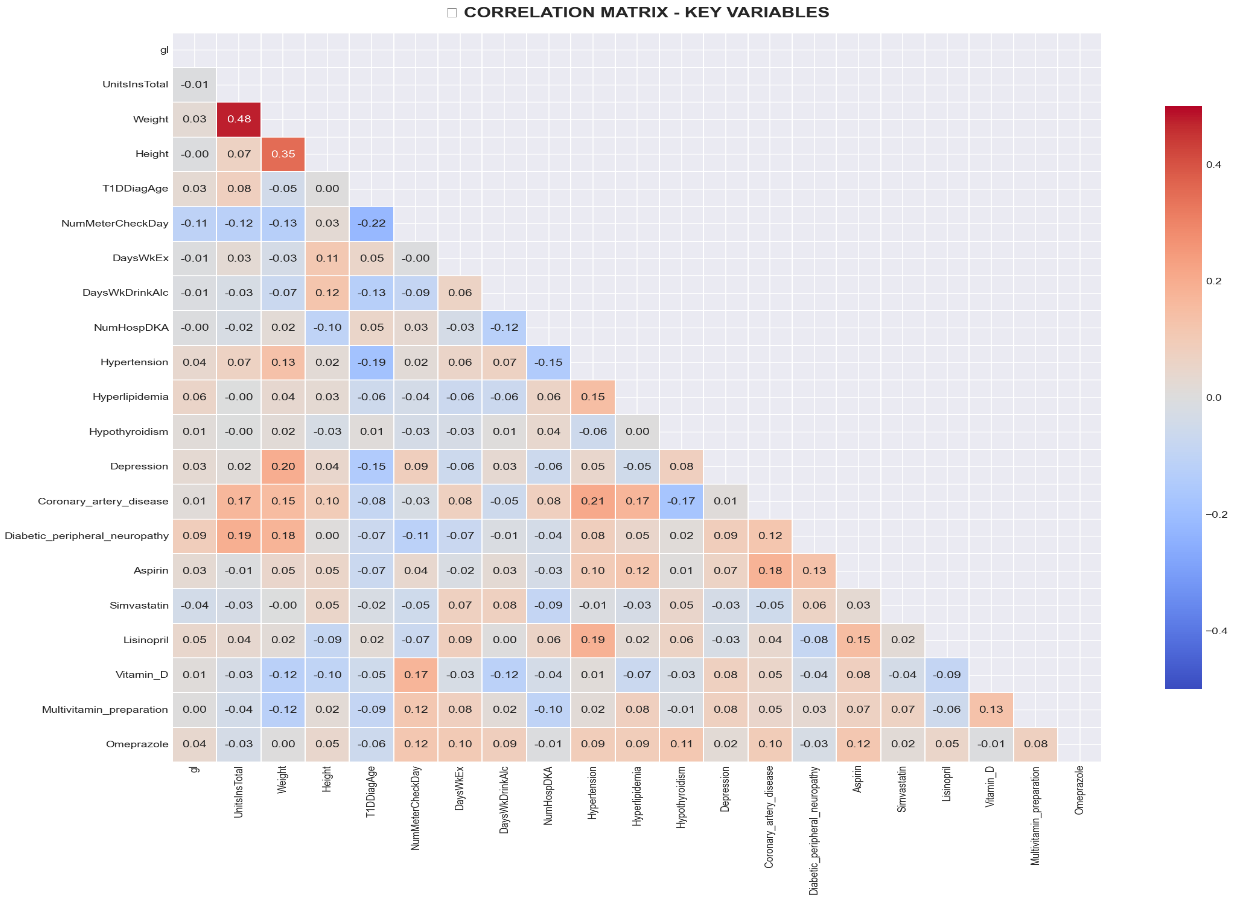


Figure 4: *Correlation matrix showing the relationships between key variables and blood glucose levels.*

6.4 Key Feature Relationships

Some anthropometric variables, such as weight and height, showed moderate associations with glucose levels. These features are known to be related to metabolism and insulin sensitivity, which explains their stronger relationship with glucose compared to other variables.

Clinical conditions and lifestyle factors showed smaller individual effects. However, these variables are still important, as their combined influence may significantly affect glucose regulation when considered together.

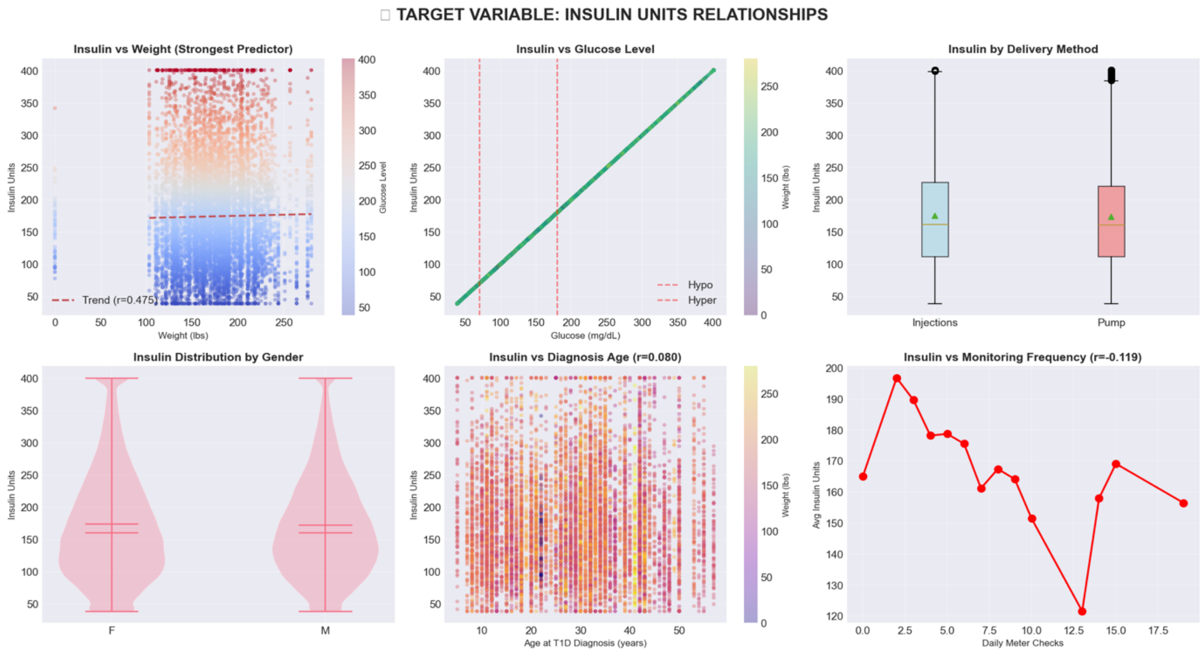


Figure 5: *Visual relationships between insulin dosage and key factors, including body weight, glucose level, insulin delivery method, gender, age at diagnosis, and glucose monitoring frequency.*

6.5 Implications for Model Selection

The EDA findings indicate that glucose levels are influenced by multiple interacting factors rather than a single dominant predictor. This motivated the use of non-linear machine learning models, such as Random Forest and Gradient Boosting, which can capture complex interactions and non-linear relationships between variables.

7. Model Development

7.1 Overview of Model Selection

Three regression-based machine learning models were developed and evaluated in this study: Ridge Regression**,** Random Forest Regressor, and Gradient Boosting Regressor. These models were selected to allow a direct comparison between a linear approach and two non-linear ensemble methods. Using different model types helps assess how model complexity influences glucose prediction performance.

All models were trained using the same train–test split, ensuring that performance comparisons were fair and based on identical data partitions. This approach allows differences in results to be attributed to model behaviour rather than differences in data sampling.

7.2 Ridge Regression

Ridge Regression was used as a baseline linear model. It extends standard linear regression by applying L2 regularisation, which penalises large coefficient values. This reduces overfitting and helps manage correlations between predictor variables.

Because Ridge Regression assumes linear relationships between predictors and the target variable, it requires input features to be on a similar scale. Therefore, all numeric features were

standardised before training the model. Ridge Regression provides interpretable results and serves as a useful reference point for evaluating the benefits of more complex models.

7.3 Random Forest Regressor

Random Forest is a non-linear ensemble model that combines predictions from multiple decision trees. Each tree is trained on a random subset of the data and features, which reduces model variance and improves generalisation.

One advantage of Random Forest is that it can naturally capture feature interactions and non-linear relationships without requiring feature scaling or extensive preprocessing. This makes it well suited for structured clinical datasets, where relationships between variables are often complex and non-linear.

7.4 Gradient Boosting Regressor

Gradient Boosting is another tree-based ensemble method, but unlike Random Forest, it builds trees sequentially. Each new tree focuses on correcting the errors made by previous trees, allowing the model to gradually improve its predictions.

Gradient Boosting is a powerful technique capable of modelling complex patterns in data. However, its performance is highly sensitive to hyperparameters, such as learning rate, number of trees, and tree depth. As a result, careful hyperparameter tuning is required to achieve optimal performance.

8. Baseline Model Performance

Baseline models were trained using default or commonly used parameter settings. Model performance was evaluated on a held-out test set.

Results showed that Ridge Regression performed poorly, explaining only a small fraction of glucose variability. Random Forest achieved the best baseline performance, with a substantially higher R² score. Gradient Boosting performed better than Ridge but worse than Random Forest in its baseline configuration.

These results highlight the importance of non-linear modelling for glucose prediction.

	Models	RMSE	MAE	R2
1	Ridge Regression	80.794601	64.314384	0.045374
2	Random Forest	70.519688	55.092273	0.272741
3	Gradient Boosting	75.894097	60.145269	0.157666

Table 1: *Comparison of Ridge Regression, Random Forest, and Gradient Boosting models based on prediction error and explained variance.*

9. Hyperparameter Tuning Strategy

Hyperparameter tuning was performed using RandomizedSearchCV. This approach randomly samples combinations of hyperparameters from predefined ranges, offering a balance between performance and computational efficiency.

For Ridge Regression, the regularisation parameter alpha was tuned over a logarithmic range. For Random Forest and Gradient Boosting, parameters such as the number of trees, tree depth, and learning rate were optimised.

Cross-validation was used during tuning to ensure robust performance estimates.

```
Best parameters found:

Ridge -> {'model__alpha': 788.0462815669904}

RandomForest -> {'n_estimators': 600, 'min_samples_split': 2, 'min_samples_leaf':
4, 'max_features': 'log2', 'max_depth': 20, 'bootstrap': True}

GradientBoosting -> {'subsample': 1.0, 'n_estimators': 200, 'min_samples_split':
10, 'min_samples_leaf': 1, 'max_depth': 5, 'learning_rate': 0.1}
```

10. Tuned Model Performance

After hyperparameter optimisation, models were re-evaluated on the test set. Ridge Regression showed no meaningful improvement, confirming that linear regularisation alone cannot capture the complexity of glucose regulation.

Random Forest performance improved slightly after tuning, reflecting its robustness to parameter changes. In contrast, Gradient Boosting showed a substantial improvement, with a notable increase in R² and reduction in error metrics.

These findings demonstrate that some models benefit more from tuning than others.

	Model	Stage	RMSE	MAE	R2
1	Ridge	Baseline	80.794600	64.314380	0.045374
2	Ridge	Tuned	80.794670	64.314976	0.045373
3	Random Forest	Baseline	70.519688	55.092273	0.272741
4	Random Forest	Tuned	70.503176	55.090234	0.273081
5	Gradient Boosting	Baseline	75.832154	60.087266	0.159040
6	Gradient Boosting	Tuned	72.231127	56.864665	0.237013

***Table 2:** Performance comparison of baseline and tuned regression models using RMSE, MAE, and R² metrics.*

11. Model Comparison and Visual Analysis

11.1 Purpose of Model Comparison

Model comparison is an important step in evaluating the effectiveness of different machine learning approaches. In this study, visual analysis was used to compare model performance in a clear and interpretable way. Performance metrics were visualised to highlight differences between models and to assess the impact of hyperparameter tuning.

11.2 Comparison of Prediction Errors

Comparative plots of Root Mean Squared Error (RMSE) and Mean Absolute Error (MAE) were used to assess prediction accuracy. These plots showed clear differences between the evaluated models. Among all models**,** Random Forest achieved the lowest prediction error, indicating stronger overall predictive performance. This suggests that Random Forest is better able to capture the complex relationships present in the data.

Gradient Boosting showed higher error values in its baseline configuration but demonstrated noticeable improvement after hyperparameter tuning. This improvement reduced the performance gap between Gradient Boosting and Random Forest, highlighting the importance of optimisation for boosting-based models.

11.3 Comparison of Explained Variance (R^2)

The **R^2 score** was used to measure how much of the variability in glucose levels was explained by each model. Visual comparisons showed that the linear baseline model explained only a small proportion of the variance. In contrast, tree-based ensemble models achieved higher R^2 values, indicating better representation of underlying data patterns.

Random Forest achieved the highest R^2 score overall, while tuned Gradient Boosting showed a clear increase in explained variance compared to its baseline version.

11.4 Effect of Hyperparameter Tuning

Visualisation of tuning effects revealed important differences between models. Gradient Boosting was highly sensitive to hyperparameter choices, with tuning leading to substantial performance improvement. This behaviour reflects the sequential learning nature of boosting models.

In contrast, Ridge Regression showed little change after tuning, indicating that adjusting regularisation strength alone was insufficient to significantly improve performance. Random Forest showed smaller but consistent gains, reflecting its robustness to parameter changes.

11.5 Summary of Visual Findings

Overall, the visual analysis confirms that non-linear ensemble models outperform linear approaches for multivariate glucose prediction. Random Forest provides the strongest and most stable performance, while Gradient Boosting benefits most from hyperparameter optimisation. These visual findings support the quantitative results presented earlier and reinforce the importance of model selection and tuning in medical prediction tasks.

12. Discussion

The findings of this study confirm that predicting glucose levels using structured clinical data is challenging. The weak performance of the linear baseline model supports previous research showing that linear relationships alone are insufficient for modelling complex physiological outcomes (Hoerl and Kennard, 1970).

Tree-based ensemble models perform better because they can capture non-linear interactions between demographic, lifestyle, and clinical variables. Random Forest achieved the strongest overall performance, which aligns with earlier healthcare studies highlighting its robustness and predictive ability (Breiman, 2001). Gradient Boosting showed the largest improvement after tuning, reflecting its sensitivity to hyperparameter optimisation, as reported in previous work (Friedman, 2001).

These results are consistent with diabetes-focused machine learning research, which shows that structured clinical data can provide meaningful insights into glycaemic control, even without detailed time-series information (Contreras and Vehi, 2018).

13. Limitations

While this study provides useful insights into glucose prediction using structured clinical data, several limitations should be acknowledged.

First, the dataset does not include short-term behavioural and physiological factors such as meal timing, carbohydrate intake, insulin dosage, or insulin administration timing. These factors are known to have a strong and immediate impact on blood glucose levels. Because such information is not available, the models are unable to capture short-term glucose fluctuations. As a result, the overall predictive performance of the models is inherently limited and should be interpreted accordingly.

Second, the analysis focuses on population-level patterns rather than individual-specific prediction. The models are trained to identify general relationships between demographic, lifestyle, and clinical variables and glucose levels across a large group of individuals. This means that the predictions may not fully reflect personal variations in glucose regulation, which can differ significantly between individuals with Type 1 Diabetes.

Finally, the study relies on structured clinical variables collected during routine assessments. While this makes the analysis realistic and widely applicable, it also limits the range of information available to the models. More detailed or personalised data could potentially improve predictive accuracy.

Recognising these limitations is important for correctly interpreting the results and for guiding future research directions.

14. Future Work

While this study demonstrates that multivariate machine learning models can provide useful insights into glycaemic control using structured clinical data, there are several ways in which future research could further improve prediction performance and extend the scope of the analysis.

One important direction for future work is the inclusion of time-series data, such as continuous glucose monitoring (CGM) records. Time-series information would allow models to capture short-term glucose dynamics and temporal patterns that are not available in the current dataset. Similarly, incorporating dietary information, meal timing, and insulin administration data could provide a more complete picture of factors that directly influence glucose levels.

Future studies could also explore more advanced machine learning models, such as Extreme Gradient Boosting (XGBoost) or neural network-based approaches. These models may be better suited to learning complex patterns when additional data sources become available. However, their use would require careful tuning and validation to avoid overfitting.

Another promising direction is feature engineering, where new variables are derived from existing data to better represent underlying physiological processes. In addition, personalised or patient-specific models could be developed to capture individual differences in glucose regulation, rather than relying solely on population-level patterns.

Overall, these extensions could lead to more accurate and clinically useful glucose prediction models.

15. Conclusion

This study set out to examine whether blood glucose levels can be predicted using structured demographic, lifestyle, and clinical data through multivariate machine learning models. The findings show that such models can explain a meaningful proportion of glucose variability, even in the absence of short-term behavioural or physiological information such as meal timing or insulin dosage.

The results clearly demonstrate that linear models, represented by Ridge Regression, provide useful baseline performance but are limited in their ability to model the complex nature of glucose regulation. In contrast, non-linear ensemble models perform substantially better by capturing interactions and non-linear relationships between multiple predictors. Among the models evaluated, Random Forest achieved the strongest overall performance, indicating that it is well suited for multivariate glucose prediction using structured clinical data. Gradient Boosting showed the greatest improvement after hyperparameter tuning, highlighting the importance of optimisation for boosting-based models.

Overall, this study highlights the importance of model selection and tuning in medical prediction tasks. The findings support the use of ensemble machine learning methods for analysing glycaemic control and provide useful insights into how demographic, lifestyle, and clinical factors collectively influence glucose levels. While prediction accuracy is constrained by data limitations, the results demonstrate the value of machine learning for population-level analysis of diabetes-related outcomes and offer a foundation for future research in this area.

16. References

- Prioleau, T., Lu, B. & Cui, Y. (2025) Glucose-ML: A collection of longitudinal diabetes datasets for development of robust AI solutions
- Li, K. et al. (2020). Time-series models for diabetes management.
- Shashaj, B. et al. (2022). Machine learning for personalized insulin and glucose forecasting.
- Breiman, L. (2001) ‘Random forests’, Machine Learning, 45(1), pp. 5–32.
- Contreras, I. and Vehi, J. (2018) ‘Artificial intelligence for diabetes management and decision support: Literature review’, Journal of Medical Internet Research, 20(5), e10775.
- Friedman, J.H. (2001) ‘Greedy function approximation: A gradient boosting machine’, Annals of Statistics, 29(5), pp. 1189–1232.
- Hoerl, A.E. and Kennard, R.W. (1970) ‘Ridge regression: Biased estimation for nonorthogonal problems’, Technometrics, 12(1), pp. 55–67.
- Obermeyer, Z. and Emanuel, E.J. (2016) ‘Predicting the future — Big data, machine learning, and clinical medicine’, New England Journal of Medicine, 375(13), pp. 1216–1219.