# Project Milestone - STATS 315B
# Using Machine Learning to Predict HbA1c and Analyze Glucose Management Strategies in Pediatric Type 1 Diabetes

**Nicolas Fertout, Thomas Bougnon**
Department of Statistics, Department of Aeronautics and Astronautics
Stanford University
nfertout@stanford.edu, thomasbn@stanford.edu

## Abstract

We compared basic regression models, including OLS and Lasso, with State-of-the-art machine learning techniques, like Random Forest (RF) to predict HbA1c levels using CGM and patient data. Of the traditional regressions, Lasso achieved the best performance, with an average error (MAE) of 0.34 percent points. Using more complex ML algorithms further improved results. The RF surpassed all other models in terms of performance, with an MAE of 0.31, 98.64% of predictions within 1 percent point, and 81.90% within 0.5 percent points of the true HbA1c value. These findings demonstrate the effectiveness of ML methods in predicting HbA1c measurements using CGM and patient data.

## 1 Project Description

This project takes place in the context of a study conducted in the Stanford Diabetes Research Center, and specifically related to type-1 diabetes. As part of this study, patients with type-1 diabetes wore Continuous Glucose Monitoring (CGM) devices collecting real-time glucose data every 5 minutes for one year. Simultaneously, multiple hemoglobin A1c (HbA1c) measurements were taken to assess average blood sugar levels. The primary objective of this ML project is to leverage supervised and unsupervised learning techniques to predict HbA1c levels using CGM data and patient-specific information. One of the key objectives of this milestone is to replicate a similar study (available at this link) to explore the associations between HbA1c and CGM data, as well as to demonstrate the benefits of employing ML models in this context.

## 2 Methodology

### 2.1 Dataset

To conduct our analysis, we required a comprehensive dataset with ample HbA1c measurements, CGM data, and patient characteristics such as weight and ethnicity. To meet these criteria, we selected publicly available data from the study Continuous glucose monitoring and intensive treatment of type 1 diabetes. This dataset included five HbA1c measurements per patient over a one-year monitoring period, taken at the beginning of monitoring and every 13 weeks thereafter. We created one row per HbA1c measurement, along with the glucose data collected for a period of two months prior to the measurement. As some readings did not have enough associated data, or other patients stopped the experiment earlier than expected, we ended up with a dataset consisting of 866 HbA1c measurements from 230 patients with type-1 diabetes. We first structured our data as followed:

Submitted to 37th Conference on Neural Information Processing Systems (NeurIPS 2023). Do not distribute.

Table 1: Dataset structure with Time Series.

| Patient ID | Patient Characteristics | Glucose Time Series | HbA1c Measurements |
|---|---|---|---|
| 1 | sex, age, weight... | 150, 180, 210... | 8.3, 6.3, 9.4, 8.8, 8.4 |
| 2 | sex, age, weight... | 250, 80, 140... | 8.1, 8.4, 9.1, 9.0, 8.8 |
| ... | ... | ... | ... |

### 2.2 Feature Engineering

Now, we aimed to both replicate and enhance the CGM statistics used in the original paper. We replicated the following CGM statistics: Mean, Standard Deviation (SD), Coefficient of Variation (CV), and the percentage of time spent in various glycemic ranges: Hypoglycemia [<54 mg/dL], Clinical Hypoglycemia [54–69], Target Range [70–180], Conservative Target Range [70–140], Above Target Range [181–250], and Far Above Target Range [>250].

In addition to the replicated statistics, we introduced new CGM statistics: Minimum, Maximum, Median, Mean Absolute Deviation (MAD). Furthermore, we computed statistics related to the Rate of Change (ROC) of CGM data, from which we derived the following metrics: Mean ROC, SD ROC, Minimum ROC, Maximum ROC, Median ROC, MAD ROC.

We end up with the following dataset for our prediction task (note that each patient ID is associated with several HbA1c measures):

Table 2: Dataset structure with CGM Statistics.

| Patient ID | Patient Characteristics | Glucose Characteristics | HbA1c |
|---|---|---|---|
| 1 | sex, age, weight... | CGM statistics (mean, SD, ...) | 8.3 |
| 1 | sex, age, weight... | CGM statistics (mean, SD, ...) | 6.3 |
| 1 | sex, age, weight... | CGM statistics (mean, SD, ...) | 9.4 |
| 1 | sex, age, weight... | CGM statistics (mean, SD, ...) | 8.8 |
| 2 | sex, age, weight... | CGM statistics (mean, SD, ...) | 8.7 |
| .. | ... | ... | ... |

In the Results section, we analyzed two datasets: one with only replicated statistics and another with all CGM statistics (resp. "Replicated" and "Full"). This allowed us to assess the performance of our models using different data configurations, and to see whether our new features led to improvements.

### 2.3 Prediction Models: Training and Evaluation

We implemented and evaluated four classes of regression models to predict HbA1c levels: linear regression (OLS), providing a baseline model, Ridge & Lasso regressions, introducing regularization, decision tree (DT), giving us interpretable insights and random forest (RF), to capture complex relationships between features.

We split the dataset using a 75:25 ratio, leading to a training set with 665 instances, and a test set with 221 instances. We used to training set to fine-tune hyperparameters like regularization strength ($\lambda$) and tree parameters like `max_depth` and `min_sample_leaf` (min. number of samples per leaf).

We evaluated the models using multiple metrics, including Mean Squared Error (MSE) and Mean Absolute Error (MAE). Additionally, and in line with the paper we intended to reproduce, we assessed the proportion of predictions within 1 percent point and within 0.5 percent point of the true HbA1c values. These metrics provided insights into the accuracy and precision of the models' predictions. We then compared the performance of the different models based on the evaluation metrics. The models with lower MSE and MAE, as well as higher proportions of predictions within 1 percent point and 0.5 percent point, were considered superior.

# 3 Results

Table 3: Performance metrics on the test set for different models.

| Model | Dataset | MSE | MAE | Within 1% Accuracy | Within 0.5% Accuracy |
|---|---|---|---|---|---|
| Null | NA | 0.663 | 0.613 | 82.81% | 51.58% |
| OLS | Replicated | 0.208 | 0.366 | 97.29% | 70.14% |
| OLS | Full | 0.203 | 0.355 | 97.29% | 71.95% |
| Ridge | Replicated | 0.192 | 0.350 | 97.74% | 74.21% |
| Ridge | Full | 0.187 | 0.345 | 98.19% | 75.11% |
| Lasso: Optimal Lambda | Replicated | 0.195 | 0.353 | 97.74% | 74.66% |
| Lasso: Optimal Lambda | Full | 0.186 | 0.343 | 97.74% | 75.57% |
| Lasso: One-SE Lambda | Replicated | 0.195 | 0.351 | 97.74% | 74.66% |
| Lasso: One-SE Lambda | Full | 0.206 | 0.359 | 97.74% | 73.76% |
| Decision Tree | Full | 0.225 | 0.375 | 96.83% | 69.68% |
| **Random Forest** | **Full** | **0.159** | **0.306** | **98.64%** | **81.90%** |

# 4 Discussion

While analysing the results, it is important to keep in mind the metrics from the Null model, which assumes a constant prediction for all samples, and gives a reference of what is achievable with minimum effort on our dataset.

The OLS models, both improved the prediction performance significantly. In particular, they achieved a high Within 1% Accuracy (around 97%), nearly as good as most other models. One significant limitation is found in the Within 0.5% Accuracy. It got slightly improved from around 70% to 72% using our additional features, but remains below most other models.

Adding regularization with Ridge and Lasso further improved results. The best performing Lasso model achieved an MSE of 0.186 and MAE of 0.343, with high accuracy. It is followed closely by the Full Ridge model (which outperformed Lasso on the Within 1% Accuracy). It is worth noticing that Optimal Lasso used 24/30 features (Full), and 18/20 (Replicated), while a $\lambda$ one SE away from the optimal $\lambda$ gave close results using only 4 variables: Mean Glucose, Proportion in Target, Proportion Far Above Target, Height/Median (for Replicated/Full resp.).

Although the DT model clearly under-performed, the associated tree provided valuable insights regarding the most useful features for splitting patients into heterogeneous HbA1c level groups: Proportion in Target, Mean Glucose, SD of Glucose, Height, Age. These were all part of the original statistics we replicated, demonstrating their suitability for this task.

Finally, as anticipated from its complexity and performance in the original paper, the RF outperformed all other models. It achieved an average error of only 0.306, with remarkably high Within 1% and 0.5% Accuracies (98.64% and 81.90% resp.) – substantial improvement over the other models.

In conclusion, this milestone demonstrated the effectiveness of machine learning techniques in predicting HbA1c levels in pediatric type-1 diabetes. Incorporating additional statistics and fine-tuning models resulted in significant improvements. The superior performance of RF highlights its potential as a valuable tool for accurate and reliable predictions of HbA1c levels, improving diabetes management and patient outcomes.

# 5 Next Steps

In the next phase, we will incorporate advanced models like RNN, LSTM, TCN, or Transformers, leveraging directly the time aspect of the data to improve HbA1c prediction by understanding the structure of CGM time series. Additionally, we aim to explore ensemble methods for combining multiple models to enhance predictive accuracy. Validating our models on external datasets will provide insights into their generalizability and robustness.

Finally, we will conduct an analysis to identify and compare the glucose management differences between patients using insulin pumps and those utilizing closed-loop systems.

# 6 References


Overview Study: A New Technology-Enabled Care Model for Pediatric Type 1 Diabetes

Paper we reproduced: Improved individual and population-level HbA1c estimation using CGM data and patient characteristics

Study that provided the dataset: Continuous glucose monitoring and intensive treatment of type 1 diabetes

Additional paper for potential next steps: A Glycemia Risk Index (GRI) of Hypoglycemia and Hyperglycemia for Continuous Glucose Monitoring Validated by Clinician Ratings