Blood Glucose Prediction Project

Mohammed El mendili *February 8, 2021*

Diabetes is spreading around the world. In order to control blood sugar levels and prevent hypoglycemia, frequent blood glucose (BG) monitoring is needed by diabetes patients and their healthcare professionals [1]. However, BG measurement is often complicated especially because it is required very often and hence not very compatible with a normal daily life.

Application of our model: We develop a Machine Learning based model that predicts Blood Glucose levels (mg/dl) for patients with Insulin Dependent Diabetes Mellitus (IDDM) at specific times/periods and after some specific injection have been made. This tool could be used to predict the BG measurmeent for patients using minimum information (e.g. collected from their smartphones). Our model could be, for example, integrated in a smartphone application that could give real help to patients and healthcare professionals.

Ethics: The model is based on patient's personal and sensitive data. Thus, such a tool shouldn't be used without his explicit agreement. If one wants to retrain the model using new data, then all data privacy laws should be applied.

Dataset: Our dataset is constructed by files for 70 different patients. Diabetes files consist of four fields per record. Each field is separated by a tab and each record is separated by a newline.

File Names and format:
(1) Date in MM-DD-YYYY format
(2) Time in XX:YY format
(3) Code (Regular insulin dose, NPH insulin dose, UltraLente insulin dose, etc)
(4) Value (i.e. Blood glucose level in mg/dl) - Target value

1 Cleaning of the Dataset

In order to clean our dataset, we proceeded using the following steps:

- We begin by extracting, merging all the records and dropping lines that contain NaN values. We found **66** files in this case and we dropped them.
- We drop files that contains incompatible Code Values as they are surely misleading. We dropped **121** files like this.
- Since the variable Code is indeed Categorical rather than numerical, we create a one-hot encoding columns for this variables and we drop the main one. We also reset the index of our dataset.
- At this stage, we also noticed that some lines are duplicated due the merge between different patient files. We remove these duplications. We dropped **1804** files here.
- To handle the Time related features. We begin by transforming their type from object to datetime and then we define some functions that extracts Year, Month, Day, Hour and minutes. Lines with errors are dropped afterwards (we found **12** of them at this stage).
- We then ended up with a dataset of **25427 rows and 26 columns**.

2 Modeling

We tested two main models in our dataset. We used the LightGBM model (Using LightGBM Python Package, to run this one has to launch in terminal the command : **pip install lightgbm**). LightGbm is a Microsoft-developed implementation for **The Gradient Boosting Machine**. We also use the **Random Forest** model. We divided our dataset into Train/Test sets. The test set is 30% of the dataset.

The following table reports our results on the test set. The LighGBM model is the best, and hence it is the one we keep.

Figure 2 plots the predicted values versus the actual Blood Glucose values.

```
LGBMRegressor(boosting_type='gbdt', class_weight=None, colsample_bytree=1.0,
              importance_type='split', learning_rate=0.1, max_depth=-1,
              min_child_samples=20, min_child_weight=0.001, min_split_gain=0.0,
              n_estimators=100, n_jobs=-1, num_leaves=31, objective=None,
              random_state=None, reg_alpha=0.0, reg_lambda=0.0, silent=True,
              subsample=1.0, subsample_for_bin=200000, subsample_freq=0)
RandomForestRegressor(bootstrap=True, ccp_alpha=0.0, criterion='mse',
                      max_depth=None, max_features='auto', max_leaf_nodes=None,
                      max_samples=None, min_impurity_decrease=0.0,
                      min_impurity_split=None, min_samples_leaf=1,
                      min_samples_split=2, min_weight_fraction_leaf=0.0,
                      n_estimators=100, n_jobs=None, oob_score=False,
                      random_state=None, verbose=0, warm_start=False)
```

Figure 1: Our selected two models to solve the Blood Glucose prediction task

Model	RMSE	MSE	MAE	R squared
Random Forest	3071	55	31	0.64
LightGBM	2713	52	30	0.68

Table 1: Performance metrics for our models

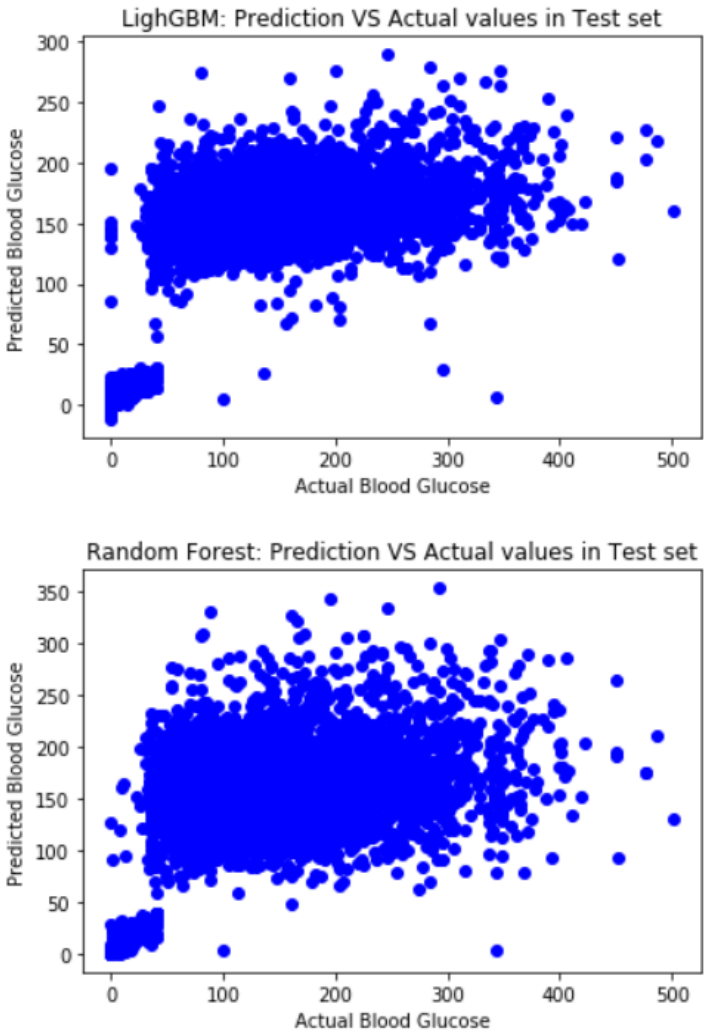

Figure 2: Predictions VS actual values for out models

References

[1] Juan Li, Chandima Fernando., *Smartphone-based personalized blood glucose prediction. ICT Express, Volume 2, Issue 4, 2016, Pages 150-154..*

[2] Li, Juan and Jun Kong., *Cell phone-based diabetes self-management and social networking system for American Indians. 2016 IEEE 18th International Conference on e-Health Networking, Applications and Services (Healthcom) (2016): 1-6.*