**CMPE541 - PROJECT #2**

**Due Week 12**

The objective of this project is:

* Have Experience in Data Engineering.
* Extract Sensible Features from the corpus.
* Prepare Data for Machine Learning Tasks.
* Have Experience in Correlation Analysis and Feature Selection.

If you have finished the first part of the project successfully you would have a Data Collection that consist of 4 files in your disposal. Within these 4 files, you should have all information about all of the patients combined and sorted with respect to time (being sorted with respect to time is critical here, as you will see in a couple of minutes.)

If you have not finished the project #1 by the time of week 8, or are not confident about the correctness/integrity of your data collection (which you should not be), you can ask one of your colleagues to share their collection and proceed with the second part using their collection. However in that case it is imperative that you notify this in the second project's report, providing the appropriate honor and citation to your colleague.

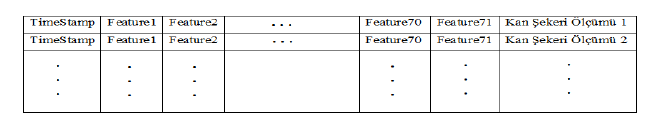
In the second part of the project you will have two tasks in your hand:

* First, you will extract several features which would relate to patients' diabetes better.
* Since we will be extracting these features by hand, it is possible that some of these features are redundant/duplicated. Most classification algorithms respond very poorly to both of these situations and thus we need to eliminate redundancies and duplications in our dataset. Thus we will implement our own feature selection algorithms.

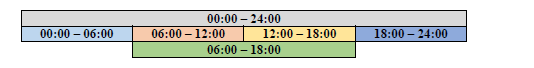
**Task 1: Feature Engineering**

At the end of Task 1, you will strip all non-blood sugar measurements from the collection (The collection will contain only blood sugar measurements in the end). Let's call the resulting data collection a "corpus". Ideally, you would want a single file in the corpus containing one line for each of the blood sugar measurement. Each line will contain the user ID, time of the measurement, a set of features (which will be described below), and the numeric value for the blood sugar measurement. The entries at each line will be stored in a comma separated form.

The resulting file can be illustrated in the following figure:



Time Zones: Since the blood sugar levels change significantly through day, we would like to divide day into time zones. The figure below represents different time zones. Note that work hours are represented as a separate time zone since daily metabolism is controlled strongly with one's habits (and work is an undeniable habit). You may want to divide time zone features described below into two: bedtime, morning, evening, and night; and work hours, non-work hours. Two groups of time zones mean two groups of time zone related features in the descriptions below.



Features:

For each of the blood sugar measurements in the data collection you should extract the following features:

* General Features
  + Average of all Blood sugar levels up to that date.+
  + Average of Unspecified Blood sugar levels up to that date.+
  + Average of all Blood sugar levels in the same category (pre- post- or unspecified, whichever group the measurement belongs to) up to that date+
  + Average of all Blood sugar levels up to that date within the same time zone.
  + Average of all Blood sugar levels in the same category up to that date within the same time zone.
  + Average of all Blood sugar levels up to that date both in the same measurement group (pre- post- unspecified) and in the same time zone

Note that especially for time zone features, you may need more than one feature (since time zone groups also overlap.

* Blood Sugar Level Features
  + Average of all pre-breakfast Blood sugar levels up to that date.
  + Average of all pre-lunch Blood sugar levels up to that date.
  + Average of all pre-supper Blood sugar levels up to that date.
  + Average of all pre-snack Blood sugar levels up to that date.
  + Average of all pre-all (only remove unspecified measurements) Blood sugar levels up to that date.
  + Average of all post-breakfast Blood sugar levels up to that date.
  + Average of all post-lunch Blood sugar levels up to that date.
  + Average of all post-supper Blood sugar levels up to that date.
  + Average of all post-all (only remove unspecified measurements) Blood sugar levels up to that date.
* Time Related Features
  + Average of all 00-06 Blood sugar levels up to that date.
  + Average of all 06-12 Blood sugar levels up to that date.
  + Average of all 12-18 Blood sugar levels up to that date.
  + Average of all 18-24 Blood sugar levels up to that date.
  + Average of all 00-06 Food Intake levels up to that date.
  + Average of all 06-12 Food Intake levels up to that date.
  + Average of all 12-18 Food Intake levels up to that date.
  + Average of all 18-24 Food Intake levels up to that date.
  + Average of all 00-06 Activity levels up to that date.
  + Average of all 06-12 Activity levels up to that date.
  + Average of all 12-18 Activity levels up to that date.
  + Average of all 18-24 Activity levels up to that date.
  + Average of all Blood Sugar levels in the same time zone.
  + Average of all Food Intake levels in the same time zone.
  + Average of all Activity levels in the same time zone.
* Food Intake Features
  + Average of Food Intake in the last 4 hours.
  + Average of Food Intake in the last 6 hours.
  + Average of Food Intake in the last 1.5 hours.
  + Average of Food Intake in the last 24 hours.
  + Average of Food Intake in the last 2 hours.
  + Average of Food Intake in the last 1 hours.
  + Time Elapsed since last meal (in hours, rounded down)
  + Time Elapsed since last meal (in minutes, rounded down)
* Insulin Medication Features
  + Average of Regular Insulin Level in the last 6 hours.
  + Average of NPH Insulin Level in the last 14 hours.
  + Average of Ultralente Insulin Level in the last 24 hours.
  + Maximum of Regular insulin in the last 6 hours.
  + Maximum of NPH insulin in the last 14 hours.
  + Maximum of Ultralente insulin in the last 24 hours.
* Activity Features
  + Total Activity in the last 2 hours.
  + Total Activity in the last 4 hours.
  + Total Activity in the last 6 hours.
  + Total Activity in the last 12 hours.
  + Total Activity in the last 24 hours.
  + Average Activity in the last 2 hours.
  + Average Activity in the last 4 hours.
  + Average Activity in the last 6 hours.
  + Average Activity in the last 12 hours.
  + Average Activity in the last 24 hours.
* Unspecified Features
  + Average of unspecified Blood sugar levels up to that date.

It is possible that for some you may struggle (implementation-wise or in terms of understanding) Notify the ones that you have not implemented.

At the end of the task, you should have a single corpus file, containing as many lines as there are blood sugar measurements in the original project documents.

**Task 2: Feature Selection**

Using the **Chi-Square, Pearson Measurement and Covariance Analysis Methods**, select 10%, 25%, 50%, 75%, and 90% of the above features. The features you select should be the least correlated ones. Keep the results for the Next Chapter of the Project, report the technique you developed in the report and turn it in.

At the end of the task, you may want to have many many corpus files, for different feature selection strategies you have implemented. As a suggestion, I would design and try multiple feature selection methods, and keep all of them separately as the result of the project 3 depends very heavily on the feature selection method you have adopted.