# Data mining applied to artificial pancreas time series data

Juan de Dios Huarancca Cahuana
*Arizona State University*
*Ira A. Fulton Schools of Engineering*
*jhuarancca@gmail.com*

### *Abstract*

Data mining is a set of techniques to explore and analyze data to discover data patterns; we are going to analyze three projects regard to data mining; to address these initiatives we are going to approach using the CRISP-DM guide.

The first project is related to extracting metrics from an artificial pancreas time series data, the second project is about training and testing a machine model, and the third one is regarded to evaluate clustering models using measures such as SSE, purity, and entropy.

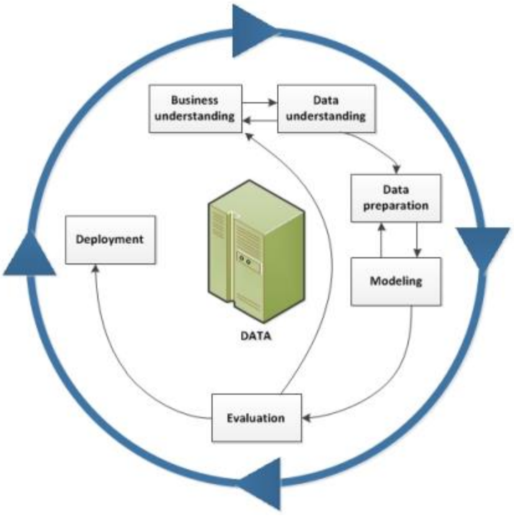
### *Keywords*

*Clustering, DBSCAN, K-means, SSE, Entropy, Purity, CRIPS-DM, SVM*

## I. INTRODUCTION

Based on the CRISP-DM model and in figure 1, a guide to approach data mining projects, to manage a data mining initiative we need to follow up these steps: understand the business use case, understand the data, and prepare data, modeling, evaluation, and deployment [1].

Figure 1: Data mining life cycle



Regarding business understanding, the company is testing the Medtronic 670G system for an artificial pancreas, this system have a continuous glucose monitor (CGM) and a sensor, which is used for the collection of blood glucose data. So, the business use cases were solved with data mining techniques like preparing, modeling, and evaluating data mining algorithms to get insights from this continuous glucose dataset.

For data understanding tasks, data collection techniques were applied to the continuous glucose sensor and the insulin pump, both in CSV file format needed to import using Pandas libraries. Then, using data explorations and descriptions scripts the dataset was analyzed and cleaned up.

The CGM sensor data have a date time stamp, a glucose measure, and the auto mode exit fields; this data was gathered every 5 minutes.

Extract glucose level features from an artificial pancreas data was the goal of the first project; this was mainly focused on data preparation steps from the CRIPS-DM guide, data was cleaned up and metric was extracted to use as input for a machine model. In order to have a robust machine model it is recommend to input data features such as bias or variance of the data, instead of raw data, especially in time series data.

The objective of the second initiative was training and testing a machine model; is going to focus on modeling steps from the CRISP-DM guide, this initiative was focused on data preparation and then on modeling a machine learning algorithm.

Finally, the third project was focused on the evaluation and deployment steps of the CRISP-DM guide; because, two clustering techniques were evaluated using metrics to measure their performance.

## II. DATA MINING SOLUTIONS

### *A. Extracting metrics from time series data*

Extract glucose levels properties from an artificial pancreas had the main scope of data preparation, the main inputs for this initiative were the continuous glucose sensor and insulin pump data.

Scripts for data cleanup processes were executed to delete null and noisy values from the dataset, data construct, and extract new data scripts considering manual and automated data generation were worked also; the new dataset was six metrics based on CGM (Continuous Glucose Measure) such as hyperglycemia, hypoglycemia, and percentage of CGM in a specific range, these metrics for three intervals: daytime, overnight and the whole day.

Finally, meal and no-meal data were merged into a dataset as a part of integrating data steps. This data selection, data cleanup, data preparation, and data integration scripts were developed using Python and Pandas libraries.

### B. Training and testing a machine learning model

For the train and model initiative, a machine model was built, trained, and tested with the purpose to predict if a given dataset is meal and no meal time series data.

The data source for this machine model was a glucose measurement dataset from two patients, data preparation scripts were executed to extract meal, no-meal data, and some pattern metrics from time series data; to train and test the machine learning model the data was split for train and testing purpose, 70% for train and 30% for test the model.

One of the most important feature that were extracted from time series data was the Discrete Fourier Transform, the Fourier Transform get data patterns as frequency of data from time series data; for instance, considering the frequency of the glucose in an specific period of time a machine model could predict if this frequency is for a meal or no meal data. .

For select modeling techniques we analyzed the data available, the mining goal was to recognize if a dataset is a meal or no-meal data; and this modeling algorithm does not need any special data to build the model. With this consideration, we chose the SVM algorithm to implement the model.

The SVM is part of the supervised machine learning algorithms and is implemented mainly for classification data mining use cases; to implement an SVM technique in this project labeled data as an input was prepared from time series glucose data extracted from artificial pancreas.

To calibrate the model and measure its performance, precision, accuracy, recall, and F1 score metrics were calculated. Finally, the machine model was deployed using Pickle to predict if a dataset is a meal or no-meal data.

This machine learning model was implemented using Python, Pandas, Numpy, Scikit-learn, and Pickle libraries.

### C. Clustering algorithms validation

The initiative regard to clustering validation had the objective to validate and choose the clustering algorithm with better performance, clustering is a technique to group similar types of data; CGMData.csv and InsulinData.csv dataset were the input for clustering, and some metrics from time series data was extracted and has been prepared as a new dataset for Kmeans and DBSCAN algorithms.

The Kmeans and DBSCAN techniques are the two most majority used clustering algorithms which group data based on different criteria [3], DBSCAN is a cluster algorithm based on density and Kmeans uses a partition clustering approach, for these projects we are going to implement these two clustering algorithms.

After the clustering models were calibrated, SSE (Sum of the Squared Error), Entropy, and Purity were calculated to measure, assess, and compare these two clustering models. This initiative was developed and deployed using Python, Pandas, Scikit-learn, and Python Pickle.

## III. RESULT ANALYSIS

### A. Extracting metrics from time series data

One of the main and more difficult steps for machine learning development is the data preparation process; It is recommended to do not to input raw data for a machine learning model, so, some data patterns such as variation, frequency, or mean were extracted. These data patterns will be better input for a machine learning algorithm to get patterns and insight from data.

The final result, after selecting, cleaning up, and constructing new data, and integrating it; was data metrics that are ready to input in a machine model for training and testing.

The metrics extracted from time series data were six features from continuous glucose measure; we got artificial pancreas metrics for three intervals like daytime, midnight, and the whole day.

### B. Training and testing a machine learning model

The mining goal for this new machine model was to predict if a given dataset is a meal or no meal data; data preparation steps was one of the important phase to have a robust machine; So, script for cleanup , transform and extract time series data features were executed as a part of the project.

According to the objective of the mining, we needed to use a classification machine learning model; so, we chose the SVM technique.

After train and testing, we got a machine model that helped us to evaluate and predict a given dataset, this model will predict if the data is a meal or no meal dataset.

### C. Clustering algorithms validation

The goal of the third project was to get some metrics such as SSE, purity, and entropy to evaluate the quality and precision of Kmeans and DBSCAN clustering models; machine models that have better metrics is going to have better result to classify new dataset.

The SSE is, the sum of the squared error, which measures the clustering performance calculating the error of each data point to the centroid, the entropy express the degree of objects of a single class belong to each cluster, and, the purity measure the cluster fraction that consists of objects in a specified class [2].

As it is shown in table 1, according to these metrics, the Kmeans clustering model is better because it had a minor SSE, lower entropy, and high purity. From the result analysis it is recommend to apply Kmeans algorithm because it fit better for this data set and will have more precision to cluster new dataset.

**Table 1: SSE (Error Sum of Squares), entropy, and purity results**

| SSE | | Entropy | | Purity | |
|---|---|---|---|---|---|
| Kmeans | DBSCAN | Kmeans | DBSCAN | Kmeans | DBSCAN |
| 30.54 | 562.20 | 0.52 | 1.47 | 0.70 | 0.32 |

## IV. CONTRIBUTIONS TO THE PROJECT

The CRISP-DM guide [1] was used to approach and address these three data mining initiatives, this guide recommends following these steps: understand the business, understand the data, data preparation, modeling, evaluation, and deployment.

Playing a data analyst role, the business challenges were analyzed and a technical solution was defined according to the tools and theories that I got during the data mining course; for instance, in the first initiative, explore, extract and transform the raw data into a pattern data were executed using python libraries to delete data with nulls values, extract bias, variance and Fourier values; and this new data helped us to train and test machine models.

According to the mining goal for each project, I had decided how to approach the solution deciding which tools, algorithms, or models we needed to implement. Data cleaning techniques were executed for the first initiative, a classification learning model like a Support Vector Machine was implemented for the second project, and Kmeans and DBSCAN clustering algorithms were implemented for the third project.

One of the main challenge faced on this projects was how to deploy ad consume a machine model; data Engineer skills was learnt and used to design and implement the machine learning architecture for these data mining projects, Python Pickle library was used for this implementation.

I had to learn, develop and improve my machine-learning engineering skills during the planning, and execution of these data mining projects; for example, according to the mining goal I needed to decide which machine model would have a better result.

Finally, two clustering models were evaluated, I need to use skills to read, interpret, and asses these two clustering models using metrics like SSE, entropy, and purity, and decided which model is going to have a better result for the business.

## V. LESSON LEARNED

### *A. Extracting metrics from time series data*

A better understanding of the CRIPS-DM guide to approaching data mining projects, mainly the data preparation steps; because I needed to apply scripts to clean up, explore, generate and transform data.

In the first project related to extracting time series properties from a dataset, the challenge was to split the data from a specific time interval, for this I needed to implement a python code to iterate in the data and split data to have better insight from the data.

It is very important to understand the business because this will give us a context about the data that we are working on; for instance, I needed to understand that there was data generated manual and automatically, each patient has a dataset in one file, for each patient we have data about the insulin measure and data when patient ingest some meal.

### *B. Training and testing a machine learning model*

For the second project about train and testing a machine learning model data preparation steps needed to run for the clustering model, extract meal and no meal data, and also got eight data patterns for this datasets was one of the goals for this initiative.

My knowledge and experience regarding to features for times series data was improved, my skill to extract data features like bias, variance and Fourier Transform tools and prepare this data for a machine model was improved; Fourier Transform is a powerful tool to extract frequencies from a time series data

For training purposes we got 70% of the data and 30% of the data to test the model, So, I had to review extra information on how to implement an SVM algorithm with Python.

For deployment of the model, I have to investigate how to implement Pickle to publish and consume a machine model.

### *C. Clustering algorithms validation*

One of the challenges here is that we need to implement formulas in Python to calculate SSE, purity, and entropy for a clustering model, and, also understand how to read and interpret the metrics to assess these two clustering models.

Using the object serialization process we were able to translate data structures or object states into a format, so, this can be stored and reconstructed later; another challenge in the project was learning how to serialize a machine model and how to consume this; this was implemented using python Pickle library [4].

Finally, I can say that I developed some data analysis, data engineering, and machine learning engineering skills; because I needed to approach the business goal, plan, execute and deploy the data mining solutions.

Concerning tools and new technologies, I improved my knowledge about machine learning models such as Kmeans, and DBSCAN, I also increased my knowledge and experience with libraries like SVC, Python, and Pickle.

## REFERENCES

[1] IBM, 'IBM SPSS Modeler CRISP-DM Guide', 2022. [Online]. Available:https://www.ibm.com/docs/en/SS3RA7_18.4.0/pdf/ModelerCRISPDM.pdf. [Accessed: Oct-18- 2022].

[2] Pang-Ning Tan, Michael Steinbach, Anuj Karpatne, and Vipin Kumar, 'Introduction to Data Mining (Second Edition) Webpage for First Edition (2005)'. 2005. [Online].Available:https://www-users.cse.umn.edu/~kumar001/dmbook/index.php. [Accessed: Oct-17-2022]

[3] Sanjay Chakraborty, N.K. Nagwani, and Lopamudra Dey. ‘Performance Comparison of Incremental K-means and Incremental DBSCAN Algorithms’. 2011. [Online]. Available: https://arxiv.org/ftp/arxiv/papers/1406/1406.4751.pdf. [Accessed: Oct-16-2022]

[4] Abdou Rockkikz. ‘How to Use Pickle for Object Serialization in Python’. 2022. [Online]. Available: https://www.thepythoncode.com/article/object-serialization-saving-and-loading-objects-using-pickle-python. [Accessed: Oct-15-2022]