Machine Learning Documentation

When it comes to the machine learning part, the first challenging task was to find a suitable dataset for our project. The task wasn't easy and even to this day, we do not have the perfect dataset that we were thinking of when choosing our topic. However, we found a publicly available dataset which contains the following elements :

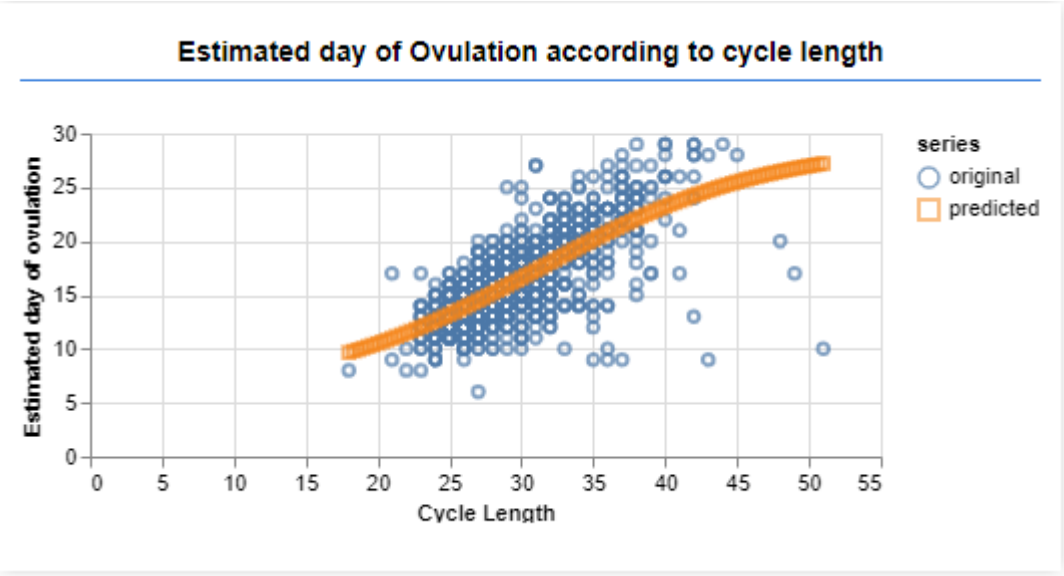
Essentially, we have data about different women's (one ID per individual) menstrual cycle.

We could only use the columns that we could ask to the users of our services or calculate/predict.

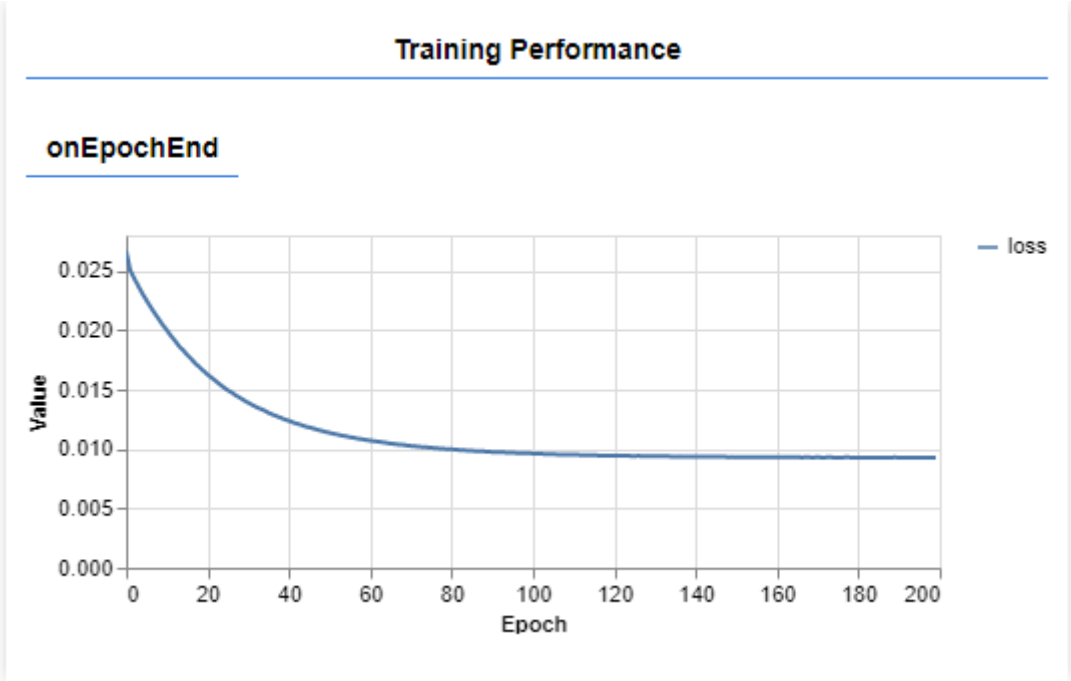
Therefore, we choose to only keep: ClientID, CycleNumber, LengthofCycle, MeanCycleLength, EstimatedDayofOvulation, LengthofMenses, MeanMensesLength, Age, NumberPregnancies and BMI.

To clean and modify the dataset, we used python on a Jupyter Notebook. The dataset contained rows with empty columns (missing values) and needed to be cleaned before we could use it for machine learning purposes.

First off, we learned how to use tensorflow.js and tried to realize a simple non linear regression model to predict the estimated day of ovulation of a cycle according to its length as a first practicing example. We started by plotting the relationship between the 2 and the predicted values of the model.

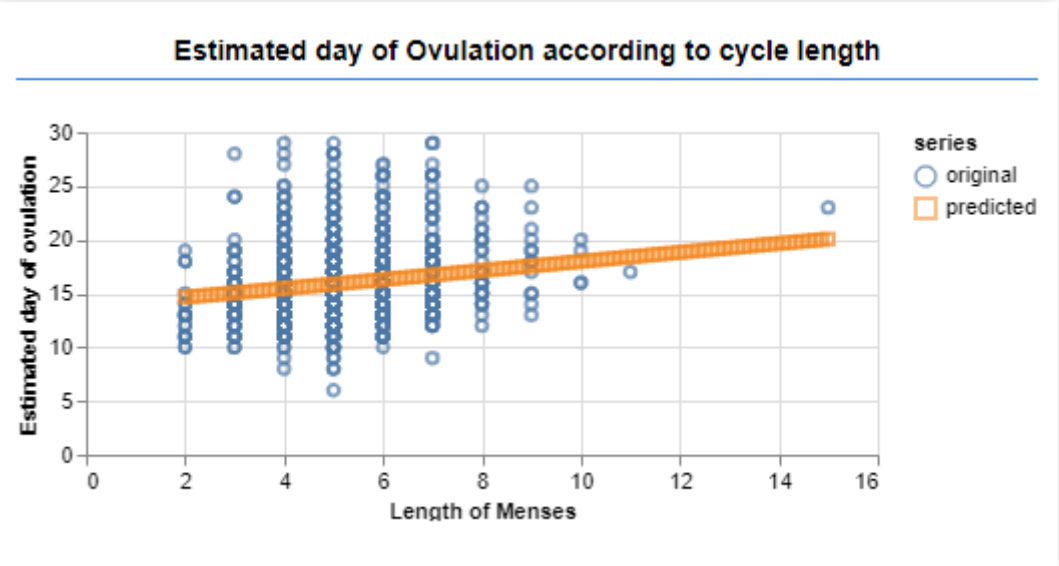


As we can see, it seems that the estimated day of ovulation is linked to the ovulation day. It looks like the ovulation day increases when the cycle length increases. The relationship between those 2 elements seems to be relevant.



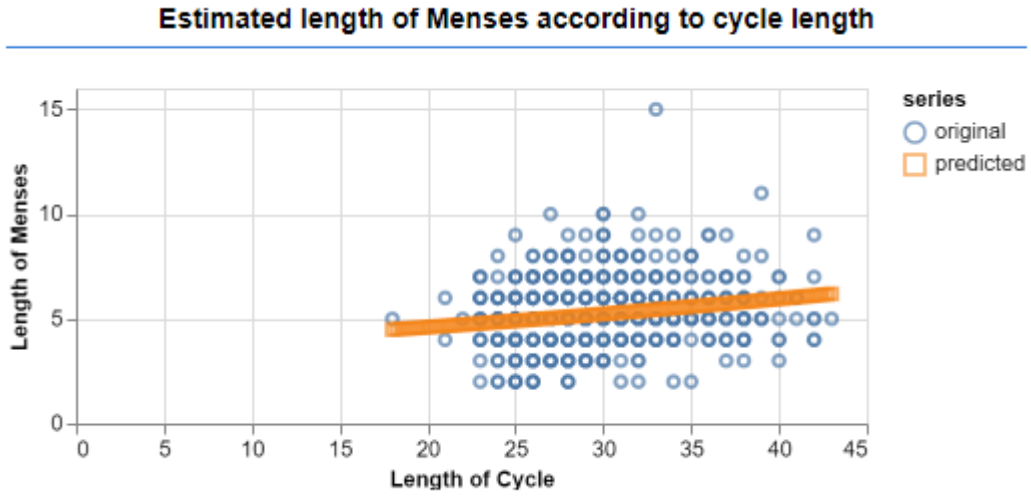
The loss function on this model shows us a significant decrease of the loss throughout the training an a final value of 0.0092 during training, and 0.0096 during testing. For example, with a cycle length of 28 the model predicts the ovulation day to be on the 15th day which matches what we can see on the graph generated previously.

Then we tried to do the same with other elements of the dataset to find a relationship between them. We first tried to predict the ovulation day according to the length of menses.

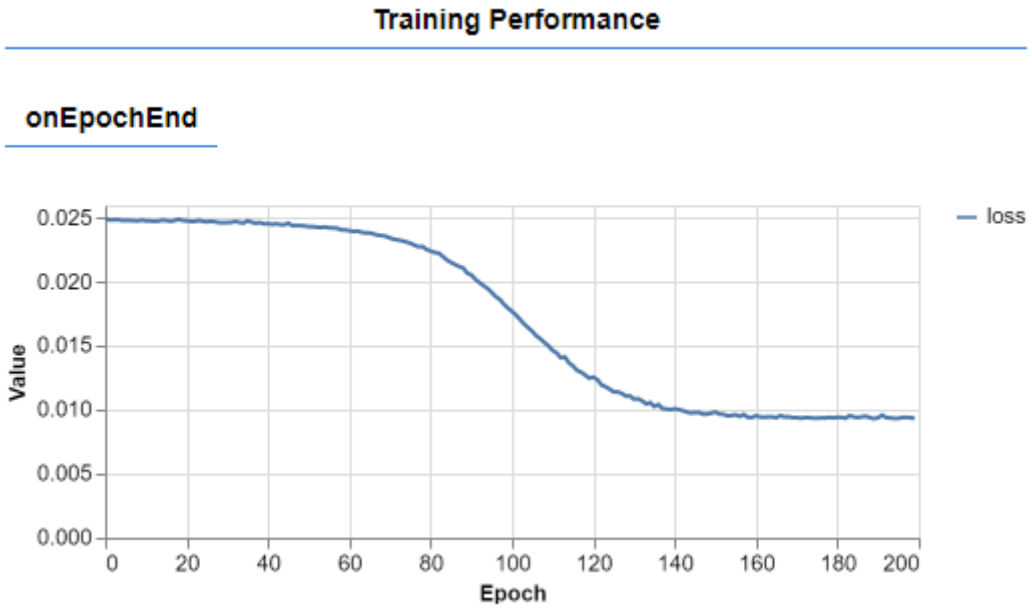


As we can see, we cannot observe a clear relationship between the 2. It looks like the ovulation day seems to be sooner with the lower values of the length of Menses. However, the range of value for the ovulation day is very broad for every data points of the Length of Menses variable.

We also tried to predict the length of Menses of a cycle based on its length but couldn't observe a clear relationship between the 2.



We obtained similar results when trying to use BMI, Age and number of pregnancies one at a time to predict the ovulation day Moreover, when using LengthofCycle, LengthOfMenses, BMI, Age and NumberOfPregnancies to predict the day of ovulation, we obtained results similar to when we only used the Length of Cycle.



As we can see on the graph, at the end of training we had a loss of 0.0093 and also a loss of 0.0093 on testing. This doesn't look like an improvement compared to the loss of 0.0092 obtained before. We can see that the loss value on the testing set was slightly lower and closed to the loss obtained on the training set when using those 5 features, but the difference too small to not be significant enough and draw conclusions.

At the end of this "training" and features research using a non linear regression model in order to learn how to clean a dataset, import it, create an AI model and train it on the dataset, we had to move to other AI models and use the dataset differently for multiple reasons.

First off, our goal isn't to wait for the end of a cycle to predict when was the ovulation day that already happened, that would be pointless as the user needs to know a prediction for the next cycle, and not a cycle that already ended.

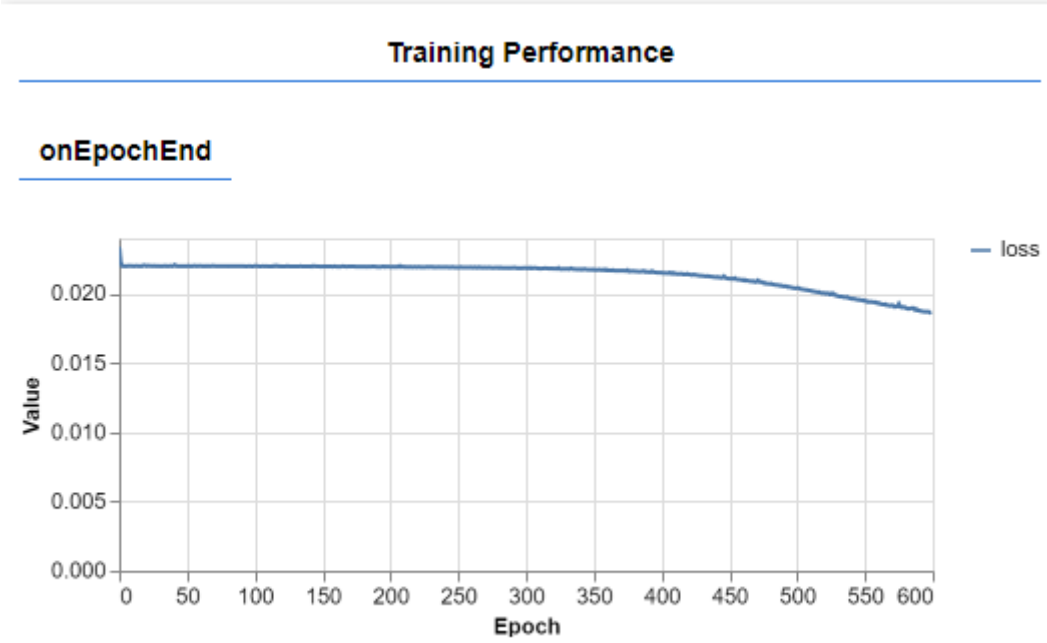
Our AI models need to predict the length of the next cycle, length of Menses and the estimated day of ovulation based on the data obtained about previous cycles.

Previous research papers that we read used the body basal temperature at their main feature to predict the day of ovulation, as it gives a clear and precise indication as when it happened. However, this is not something we can obtain with our service. The BBT (body basal temperature) needs to be measured daily, when waking up, with a special kind of thermometer that most people do not own. But what we could learn from this research paper is how they used CNN (Convolutional Neural Network) models and LSTM (Long Short Term Memory model, a type of Recurrent Neural Network or RNN) models to predict the day of ovulation and how effective those model were, especially the LSTM model.

As some of us where and are still working on an LSTM model using tensorflow.js, other members of the team worked on creating an artificial dataset that we could use both for an LSTM model and a regular NN (Neural Network) model. The reason to create an artificial dataset is to introduce more irregularity in the cycle length, length of Menses and day of ovulation of individuals following a pattern, to try to create a model that would be able to predict those 3 same features for next cycles to help people with irregular cycles.

To this day, this artificial dataset has been made and we are working on making it work with an LSTM model and a regular Neural Network model.

But before moving on to an LSTM or CNN model, we tried to predict the ovulation day using the mean cycle length and the length of menses of the current cycle. As the ovulation day happens a few days after the menses, we could use the mean value of previous cycles and the length of menses of the current cycle to try to predict the ovulation day. Sadly this method did not provide really good results. This could be explained by the fact that only the length of Menses varies from cycle to cycle of an individual, same as the ovulation day. However we've seen that the ovulation day does not show a clear relationship with the length of Menses but more with the length of Cycle, which here does not change between cycles of a same individual because we are using the same mean value every time.



We only managed to obtain a loss of 0.187 on the training set and 0.2 on the testing set, but after a huge number of epochs compared to what we had for previous model. Which could mean that we only overfitted the data and do not necessarily have a working model.

Appended :

Non linear regression model using 1 feature :

http://www.mediafire.com/file/g4xsgjklt8a0ovn/regression1feature.html/file

Non linear regression model using multiples features (here using 5 features):

http://www.mediafire.com/file/3vka6tbdgzb852d/tensorflow5features.html/file

Non linear regression model using mean length of Cycle, length of menses to predict ovulation:

http://www.mediafire.com/file/ayafw3b1mymrdfx/regressionusingmeancyclelength.html/file

Original Dataset :

http://www.mediafire.com/file/a1nrr7yluwqtgcg/FedCycleData.csv/file

Jupyter Notebook File for the Dataset cleaning :

http://www.mediafire.com/file/e4xdf3af0k0ct6e/AI_PROJECT.ipynb/file