Progress Summary Report: Milestone II

Zach Thomas and Shima Oruji

Note: First we would like to thank you for all your comments on our work and we believe that they are making this project much more meaningful. We would like to clarify one possible misunderstanding. In the initial proposal feedback, we received few comments and we were supposed to reply to that with the changes until September 16. We updated the proposal and pushed our changes on September 15, just one day before the deadline in the old repo here. You can verify the time by checking the commit time. As we discussed shortly at that time and received your verbal approval, we decided to perform additional work by implementing the models in more depth and adding a website and API to the project scope. Since this was not pushed into the repo you made for each project in the class github, you might not see this. Please accept this as this was an honest mistake.

For your feedback on the first milestone, as you suggested we decided to implement the models from the low level and not use the high level libraries in our model development. However, we still would like to bring the Pycaret library in our final presentation to let the class know about its benefits. It can give a quick idea on how the dataset looks like and what are the typical expected performances from different models. However, as you pointed out properly, for developing a practical classifier, we always need to dive deeper and perform additional analysis.

Updates on Milestone 2

We have reworked our entire work up to this point to learn more about the models we will be using for our final project and allow for more customization and alteration of these models. We have built neural networks, decision trees and random forests from scratch. This was accomplished using only numpy and pandas, which are two fundamental libraries that help with mathematical operations and reading datasets, respectively. These models will be used to predict diabetes onset and be displayed using our API.

The neural network was made conventionally, using feature vectors, weights, and activation functions to determine the characteristics of inputs that best predict diabetes onset. The network is fully connected, so every input neuron affects each output neuron. Our preliminary results show that the trained neural network model has an accuracy of 69% which is a bit lower than that (77%) of the best model obtained from the PyCaret library.

The decision tree we made uses entropy to sort out the qualities that separate diabetic patients from nondiabetic patients. We currently have the capability to train the decision tree on any given dataset, including the diabetes dataset. Our tree takes the dataset, filters out the outcomes column, then finds the best point to split the entries into two separate paths. This is done by calculating the entropy of the split and finding the split which produces the most entropy on both sides. Since our problem is a binary classification problem, the developed tree remains thin but has developed deep. Our random forest model uses bootstraps, or evenly sized randomized divisions of our dataset, across multiple instances of our tree model to attempt to predict diabetes onset from what it can amalgamate from the results of the trees.

For the next step, we are planning to dive deeper into the models and our preliminary results and determine the key factors affecting the final accuracy of the trained models. We are planning to train and use the best possible model in our website and API. We will also seek to finish our work on the website and API by milestone 3.