**CAPSTONE**

**DATS 6501**

**RENZO P. CASTAGNINO**

**PRE-PROPOSAL**

**AN INTERPRETABILITY MODEL FOR MACHINE LEARNING**

Machine Learning is a set of methods and techniques that computers use to learn and make predictions based on data. Its main goal is to allow computers to learn automatically and take decisions without human intervention. A Machine Learning algorithm learns a model by estimating parameters or learning structures. This is measured by a score or loss function that is usually minimized. Thus, one of the most important aspects to consider is the accuracy of the model. However, trusting the model performance could lead us to ignore why the model is making certain decisions. This means that a single metric like accuracy is not a complete description of the model. In certain cases, knowing why the model made a decision can help to understand deeper the problem, as well as the data and why a model might not perform better.

Interpretability arises from an incomplete way of evaluating a model, to understand how it came to the prediction. To facilitate learning, understand why predictions or behaviors are created by the machines, interpretability is crucial. Nowadays, Machine Learning is playing an important role in real-world tasks and explaining why a Machine Learning model made a decision could be crucial. For instance, if a model is getting biases from the training data, it could happen that your model predictions discriminate based on demographics. This problem can’t be solved by the loss function.

The main goal of this project is to train a model that produces predictions, to evaluate and interpret the results, and understand the reasons that lead to certain decisions. Specifically, A Convolutional Neural Network model will be used to apply the concept of interpretability. To achieve this, either SHAP or Class Activation Maps will be used depending on the application that will be work.