**[Github Repository](https://github.com/Mono-Blaine/CMPS-261-Project.git)**

**Group Name:**

Deeplearners

**Platform Used:**

We used *Google Colab* to train our models.

To avoid uploading the huge dataset every time we worked on our project, we uploaded it to Google Drive which we made automatically mount whenever we used *Colab*.

**Preprocessing:**

We tried 3 ways to preprocess our data, the first 2 of which failed as they lead to an extreme loss of data. The dataset contained non-float values such as “error”, “s”, and “”1.01””.

The first 2 approaches were selecting only the columns that were of type *float64*. This led to dropping 2 columns because they contained 3 non-float values. Obviously, this isn’t a good solution as it involved dropping 2 features.

Our third and final approach, which we achieved after lots of research, involved iterating over every cell in the dataset and modifying it in such a way that it becomes acceptable by our models. If a value contains quotation marks, they were removed to keep the float value inside them. Otherwise, if the value is a non-digit, it was replaced with 0.

**Models Tested:**

We used the following models in our attempt to solve the project:

* Logistic Regression: including scaling the data using the *StandardScaler* and *imputing* any missing values. We achieved an accuracy of 64.2%.
* XGBoost: using *XGBClassifier*. We achieved an accuracy of 73.8%.
* Neural Networks: using a dense Neural Network written in *TensorFlow*. We achieved an accuracy of .

**Validation:**

For logistic regression and *XGBoost*, we used 5-fold cross validation.

In full transparency, 5-fold cross validation failed on logistic regression due to a non-acceptable value (NaN) that we were unable to remove from our dataset.

As for neural networks, we did not validate our results.