# Report

After creating the dataset (about 7,400 rows) by merging data from SNDs and Laboratory Report Workbooks by matching them on id where it was pre-feature engineered, the dataset was then cleaned by removing rows with empty cells.

The algorithm chosen was LightGBM which is a gradient boosted tree which has the advantage of being the fastest gradient boosted tree algorithm as well as being able to balance the soil type classes which was heavily skewed with the “clay” soil type representing 68% of the dependent set with “Gravel” having only one instance.

The model’s hyperparameters were tuned with GridSearchCV to try and achieve highest possible performance. The model was then fitted and tested using the test set and this led to a precision score of 92%, recall of 88% and F-1 score of 89%. This means the model performed very well with a good ratio of True Positives and True Negatives to False Positives and False Negatives.

However, the model scored an accuracy of 97.7% which is a quite high and distant from the other classification scores. This led me to believe the model was overfitting on the dataset. I then tested and confirmed the overfit by cross-validating the accuracy of the model on 10 randomly sampled sets of the data and found the mean score to be 72%.

This means the model has the potential to do better with a lot more data points by using a larger dataset. This will mean a lot more SND and Workbook reports to be extracted as this would help improve performance and reduce overfitting.