**Mini Hackathon 2021**

**Preliminary Round**

**Report Submission**

**TEAM NAME: ENIGMA**

Team Leader’s Name : J.K. Kenneth Chamara

Team Members :

1. D.M. Janitha Prathapa

2. G. Yuvin Nimnaka Perera

3. G.L. Udari Kaushalya

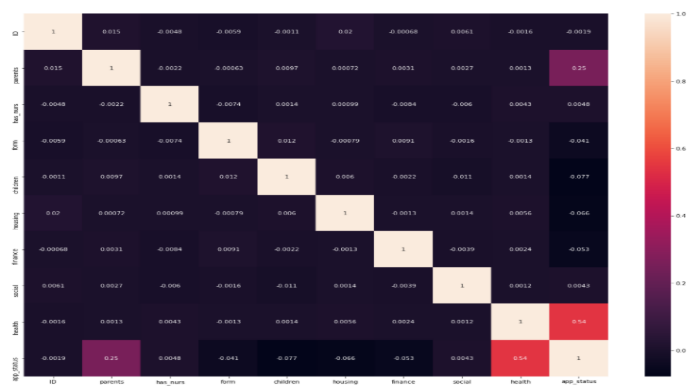
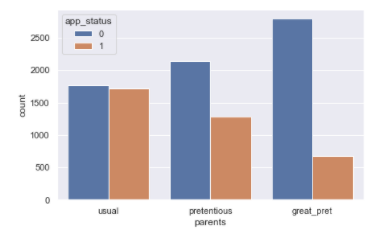
Kaggle User Name : enigmacrew

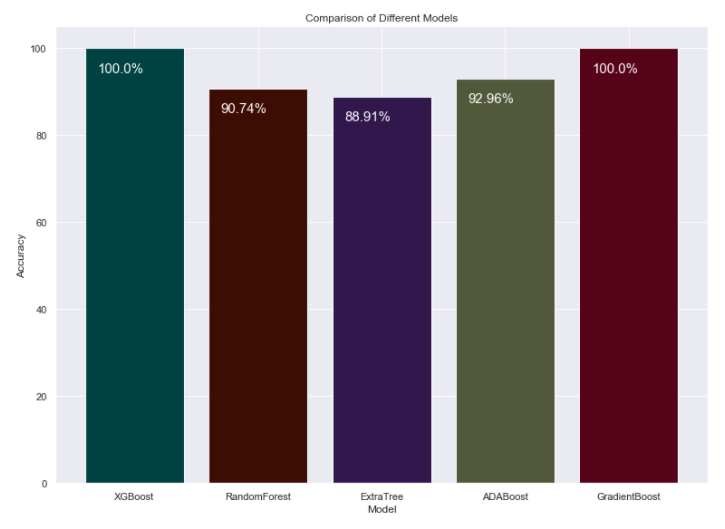
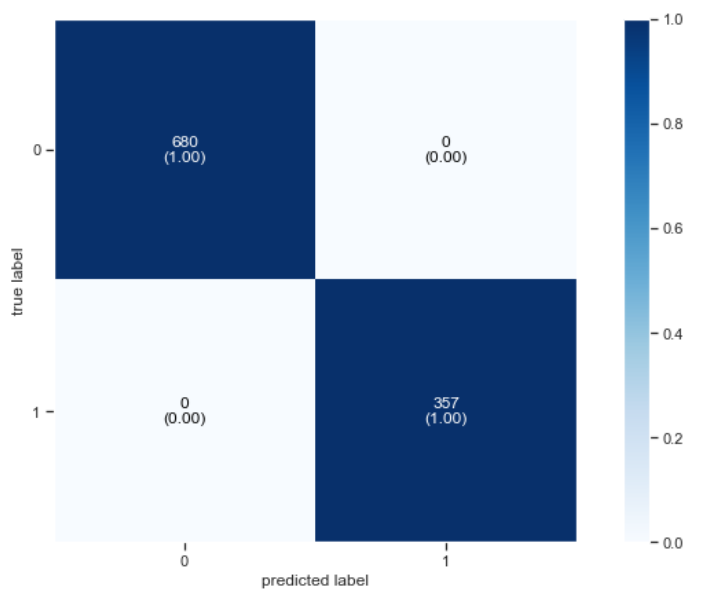
Kaggle Display Name : Enigma Crew

1. **Introduction**

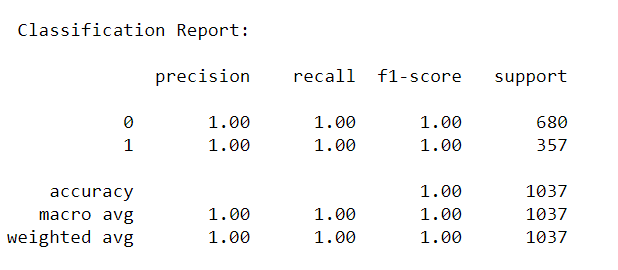
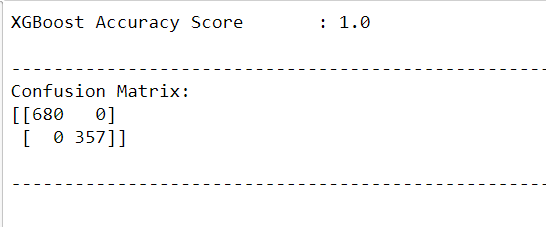
This is a binary classification problem regarding to a nursery application data set consisted with a response variable as 'app\_status' which predicts whether the applicant would be selected or not with another nine independent variables as 'parents', 'has\_nurs', 'form', 'children', 'housing', 'finance', 'social' and 'ID'. In this approach, few classifiers such as Random Forest, XGBoost, ADABoost, Extra Tree Classifier and Gradient Boost Classifier are tested with introducing class wise error rate, predicting model accuracy and confusion matrix. Implementation is done using Python and pandas, NumPy, sklearn, collections, matplotlib and seaborn are the main libraries used here. Finally, a stack model has created and outcomes of the tested models are used as the input of the prediction.

1. **Methodology**

After the data loading step using the given CSV files, data are preprocessed by checking null values and duplicate rows. Exploratory Data Analysis has done by using Correlation Heat map to identify which variables are correlated and Count plots are used to visualize the distribution behavior of data into different categories in different variables.

The Training data set has split into two main parts as train and test with 90% of train data set. Classifier Model class has implemented to use in different Machine Learning Algorithms and Random Forest, Extra Tree Classifier, AdaBoost Classifier and Gradient Boost Classifier Cross Validation has done and for all these mentioned classifiers, models have trained and calculated prediction. In the next step, Function for a stack model cross validation has implemented and the outputs given by the mentioned above classifiers are taken as the inputs of the Stack Model. Finally, the outcome of all the tested models are compared with the predicting accuracy and selected the most accurate model for the final prediction implementation. Model Comparison for the classifiers mentioned above has given 88.91% for Extra Tree, 90.74% for Random Forest, 92.96% for ADA Boost and 100% Accuracy for both XG Boost and Gradient Boost Classifiers. The outcomes of all other models except XGBoost has taken into a NumPy array and input for the XGBoost Classifier and predict the final outcome.

1. **Results**

As the result of using XGBoost classifier after the analyzing, it gave a 1.0 Accuracy in Kaggle and the Area Under the Curve (AUC) become 1. For all the classifiers, Class wise error was calculated in the testing steps and finally for the XGBoost the Error valued (Class wise Error) as 0.

1. **Conclusion**

According to the given binary classification, the behavior of the Data is identified by performing an exploratory analysis. Few classifier models have been trained and tested for prediction accuracy to select the best model. As one of the most accurate models, XGBoost Classifier is chosen and used for the final prediction (accuracy :1.0). Total Data set (training and testing) has the all possible combinations for these nine features as 12960 and since the accuracy of the prediction (test data: 2592) is 1.0 and the training accuracy (train data: 10368) is also 1.0, this can be concluded as the dataset will give the correct outcome for any given input combination.