***FUNDAMENTALS OF DATA ANALYTICS***

**Google Cloud & NCAA® March Madness Analytics**

|  |  |
| --- | --- |
| All Roll Numbers of your Team: | 18MIS7023  18MIS7061 |
| Student names in your team : | K. Hemanth Chowdary  A. Krishnapriya |
| Contribution by each team member in percentage to complete this project: (Try not to mention equal proportions) | 18MIS7023(Hemanth) ---51%  18MIS7061(Priya) ---49%  (Almost we Both have did equally) |
| Slot (E2/F1/F2) | E2 |
| Project Title | **Google Cloud & NCAA® March Madness Analytics – Basketball Tourney** |
| Did you complete the Project? If No, how much percentage you have done? | Yes Completed the full project |
| Is it a competition (YES/NO)? | Yes |
| If YES, competition link | <https://www.kaggle.com/c/march-madness-analytics-2020/overview> |
| Data set used and its link | <https://www.kaggle.com/c/march-madness-analytics-2020/data?select=MPlayByPlay_Stage2> |
| How many rows and columns in the data set(s) and what are its data types | Rows-11681180(whole data which are going to predict)  Columns-17  Data types- Numeric and String |
| Describe about your project (What, Why, Final Goal) | **What-** The project is about the basket ball game and the data of the basketball game played in past 5 years data is given. In the data they have given the id,season,daynum,WinningTeamID(WTeamID),LostTeamID(LTeamID),WinningFinalScore(WFinalScore), LostFinalScore(LFinalScore),WinningTeamCurrentScore(WCurrentScore),LostTeamCurrentScore(LCurrentScore) and Events types with elapsed times of every event occurs and event and playerid’s with X,(axis)Y(axis of distance),area.  **WHY:-** This is userfull for the present year or next year game to predict the results by seeing the data of previous . we are going to find who will win on the every round  **Final Goal:-** As per the competion we should predict the whether the team win or loose.But to directly we cannot predict the win or loose.  We can only predict with the WFinalScore and LFinalScore.  For our project Class Labels are **WFinalScore** and **LFinalScore.**  Finally by comparing predicted values win loose is predicted |
| Challenges Identified in your project | 1)To identify the Class Labels  2)How to reduce the data and reducing should not loose the valuable information needed for the class label  3)Applying ML algorithm and predicting 2 class labels and finally win or lost |
| Pre-processing techniques used | 1)un wanted columns are removed first  2)By the class Labels which columns are not impacting the class label are removed  3)Removing the columns --- As we are having the class labels as WFinalScore and LFinalScore .so just we need data that the different combination of WTeamID and LTeamID .So no need of duplicate combinations of them . So Just **duplicated rows** are removed (distinct method is used)  **\*\*\*11681180 rows to 11,000 rows has been changed after pre-processoing\*\*\*\***  --"already we have applied initially different method like removing outliers,unwanted events are removed but those pre processing is not worked(predicting properly)” |
| Did your Team spent more time on Pre-processing or for implementing ML algorithms? | **On Pre-Processing** (65% - 70%time)  ML Algorithm(30%-35% time) |
| What Machine Learning Algorithms Implemented and their accuracies | 1)**Multiple Regression**  Before Pre-processing, ML **Accuracies:** R2 value  For class label **WFinalScore** : 0.21  For ClassLabel **LFinalScore** :0.25  After Pre-processing, ML **Accuracies:** R2 value  For class Label **WFinalScore** : 0.42  For class Label **LFinalScore** : 0.8  But finally with the predicted scores I have calculated the win or lose I got the **95.7%**correct results got these are got by just comparing WFinalScore and the LFinalScore    2)**Decision Tree**  Before Pre-processing, ML **Accuracies:** Because of the high data almost more than 100lakhs of data is there so can’t able predict before applying pre-process  After Pre-processing, ML **Accuracies:** there are two class Labels and predicted for both the class labels separately with the same test and train data for both class labels . As mentioned in above just I have calculated the won or lost in the form of 1’s and 0’s  By finding the won or lost for the predicted final scores and we can check the accurate percentage by comparing the actual and predicted wins are same on the data and got the percentage is 96.26 % correctly predicted  3)Random Forest  After Pre-processing, ML  **Accuracies:** here won and lost predicted result is 97.92%correctly predicted wons and 2.1 % got wrong result of final result |
| Which machine learning algorithm you implemented gave highest accuracy and WHY?  (Explain this in very detail) | Random Forest ML gave the more predicted results of 97.92 % correctly.  While random forest is a collection of decision trees,and we have already applied the decision tree so the that tis ML Algorithm got the more accurate final results.Random forest can train the data as much as effective way and speed and predict the more accurate results |
| Highest accuracy identified from the Competition site existing submissions and what ML algorithm they have used and how they achieved high performance?  (Explain this in very detail)  (Mention the links of the python code of the earlier competition submissions of the same project from which you have taken inspiration) | Our project is taken from the Kaggle.com  Our Competition is not yet completed in the Kaggle website and there are no submissions are submitted till now. So there are no existing submissions on our project |
| Final Conclusion | In the project we have applied and tested in so many ways to predict the better way of results and finally executed the more accurately predicted methods and all we have tested and used method are in the R document what we have sent.And we have predicted for both class labels separetly and mentioned in the R file .we have predicted the values for the class labels and to know how much accuracy we have did some process after predicting **the both class label values and finally we have found the win or lost and checked the accuracy by comparing test actuals values and predicted values.**  We have removed lakhs of data because those rows will not impact on the class label any more because **for 1 row there are so many duplicatd rows so by this the prediction will be corrupt because for one combination got the wFinalScore team will win and same combination set is resulted in another row that they fail.**so for only one combination we are getting the won and lost status so **I have removed the all rows of duplicate** |