Final Project Report
Introduction to Data Analytics

Project Title:
To predict if a student would go to college.

Prepared by:
Karan Punjabi(N01514624)

ITE 5201 –Summer 2022
Humber College

1. Problem Statement

To predict if a student would attend college or not based upon various features like Interest, Average Grades, Residence, and Parent was in college.

2. Dataset Description

	type_school	school_accreditation	gender	interest	residence	parent_age	parent_salary	average_grades	parent_was_in_college	will_go_to_college
0	Academic	A	Male	Less Interested	Urban	56	6950000	84.09	False	1
1	Academic	A	Male	Less Interested	Urban	57	4410000	86.91	False	1
2	Academic	B	Female	Very Interested	Urban	50	6500000	87.43	False	1
3	Vocational	B	Male	Very Interested	Rural	49	6600000	82.12	True	1
4	Academic	A	Female	Very Interested	Urban	57	5250000	86.79	False	0

This dataset contains various columns including Type_School, school_accreditation, Interest, Residence, Parents_Age, Parents_Salary, Average Grades, and Parent was in college.The dataset has 1000 records. Parents_salary is in Indonesian Rupee. While most of data has just two classes.

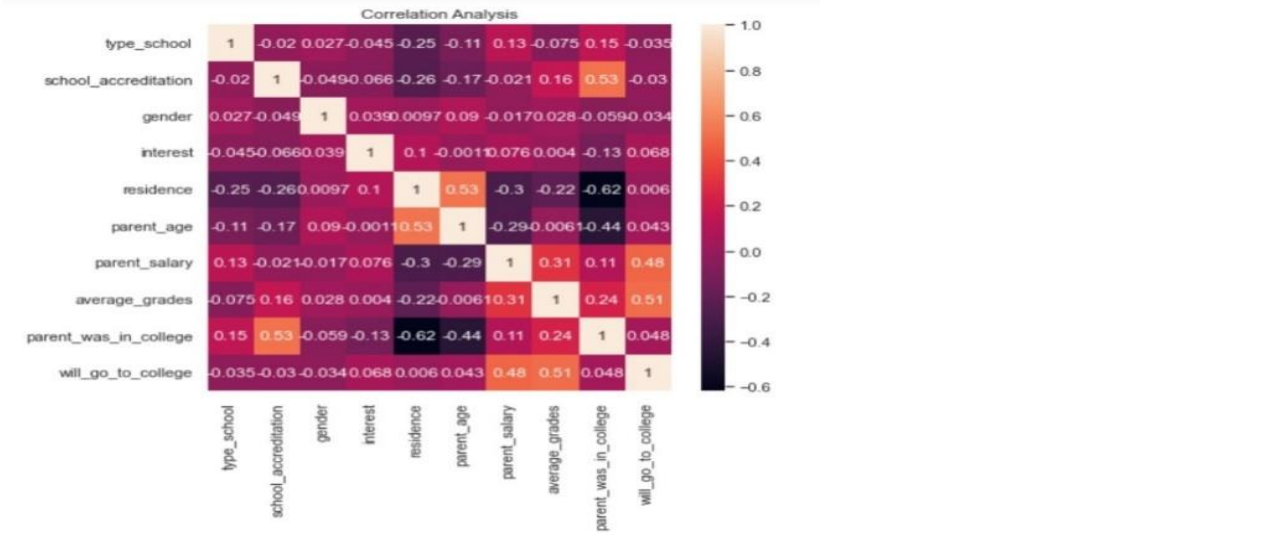
3. Dataset Analysis and Observations

	type_school	school_accreditation	gender	interest	residence	parent_age	parent_salary	average_grades	parent_was_in_college	will_go_to_college
0	0	0	1	1	1	56	6950000	84.09	0	1
1	0	0	1	1	1	57	4410000	86.91	0	1
2	0	1	0	4	1	50	6500000	87.43	0	1
3	1	1	1	4	0	49	6600000	82.12	1	1
4	0	0	0	4	1	57	5250000	86.79	0	0

For this dataset most of the data has two classes so I did one hot encoding and converted data into 1 and 0. While I didn't change parent_age, parent_salary, and Average_grades.

While Interest has 5 classes which are interested, not interested, very interested, uncertain, and less interested and mapped them from 0 to 4.

HeatMap plot showing Correlation Analysis between different columns.

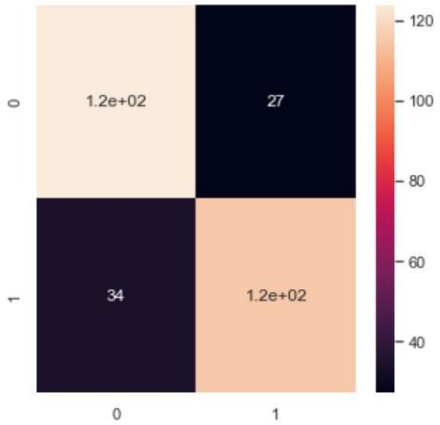


4. Proposed Analytical/Prediction Model

For this project I used two prediction model Logistic Regression and Random Forest Classifier.

Logistic Regression:

Confusion Matrix



Classification Report:

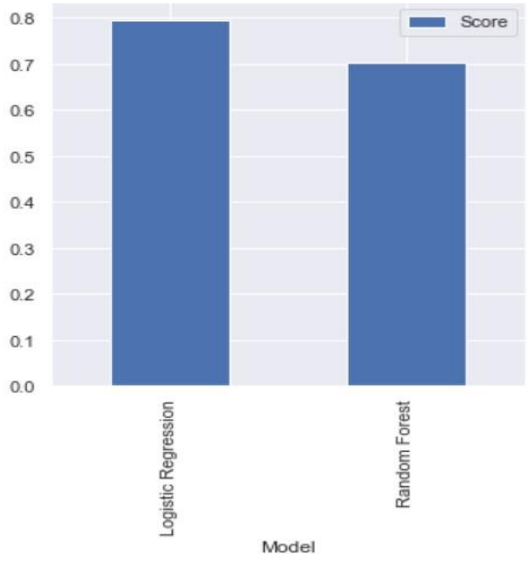
	precision	recall	f1-score	support
0	0.78	0.82	0.80	151
1	0.81	0.77	0.79	149
accuracy			0.80	300
macro avg	0.80	0.80	0.80	300
weighted avg	0.80	0.80	0.80	300

Random Forest Classifier:

Classification Report:

	precision	recall	f1-score	support
0	0.71	0.69	0.70	151
1	0.69	0.71	0.70	149
accuracy			0.70	300
macro avg	0.70	0.70	0.70	300
weighted avg	0.70	0.70	0.70	300

5. Results and Discussions



Results: So, the accuracy of Logistic Regression, to predict if a student would attend college or not is 80% while on the other hand, accuracy of Random Forest Classifier is 70%.

Comparing both the models: So, it can be said that accuracy to predict, if a student would attend

College or not, of Logistic Regression model is morewhen compared to Random Forest Classifier.