Department of Computer Science
FAST National University of Computer and Emerging Sciences
Karachi Campus

Predictive Analytics on the Academic Record of NUCES

FYP-1 REPORT

Submitted by :

Obaid Ur Rehman
17k-3848

Areeka Ajaz
17k-3913

Tooba Shahid
17k-3731

Submitted On :

19th January 2021

Supervisor :

Dr Jawwad Ahmed Shamsi

ABSTRACT

Predictive Analytics is the process of using past data to make future predictions . The past data is used to capture important trends with the help of a mathematical model and the model is then used to make predictions on current data . Our aim is to perform predictive analytics on the academic record of NUCES . The basis of any model is the features using which the model is built. In FYP-I we targeted those features which we will be using , in the second phase of our project , to build the predictive model. We used the past academic record of NUCES to make insights and find out correlations between different attributes. Finally a dashboard was built to display the insights and the analytics.

INTRODUCTION

Each year a number of students take admission in FAST NUCES . The students taking admission in FAST NUCES are from different educational backgrounds and different regions . Their academic performance throughout their university life is a reflection of different factors , not only but including their educational background , their previous academic records , the region/district from where they belong etc . The first phase of our project aimed to answer numerous questions about how these factors are related to the performance of students at FAST throughout their educational period.

PROPOSED WORK

Our proposed work for FYP-1 was divided into two parts, the first one being cleaning , transforming and EDA of the data followed by feature selection and the second part was integrating our EDA and feature selection with a dashboard.

Through our feature selection we aimed to answer the following :

1. Does the previous educational background (Intermediate / A levels) affect the performance at FAST?
2. Does the previous educational background (Matriculation / O levels) affect the performance at FAST?
3. What is the correlation between matriculation / equivalence grade and the performance at FAST?
4. What is the correlation between intermediate / equivalence grade and the performance at FAST?
5. Does there exist any correlation between the city/district (a person belongs to) and their academic performance?
6. Does the performance in initial CS courses affect the performance in the later ones?
7. Does academic performance vary campus wise ?
8. What role does gender play in academic performance ? Do girls tend to perform better than boys or vice versa ?
9. What is the correlation of a school with academic performance at FAST ?
10. What is the correlation of a college with academic performance at FAST ?
11. What is the correlation of year of admission with CGPA ?
12. What is the correlation of the year of graduation with CGPA ?
13. Does a degree program affect CGPA?

Feature selection was to be done using statistical methods such as Pearson correlation , ANOVA to find out correlation between cgpa and different attributes.

EXPLORATORY DATA ANALYSIS

Exploratory Data Analysis is one of the crucial steps in data analytics. Before we jump to learning and modeling the data , EDA was to be performed. In our case , the EDA was first performed for all the data as a whole (all FIVE campuses together) and then separately for each campus .

SUMMARY OF EXPLORATORY DATA ANALYSIS

Attributes	Exploratory Data Analysis					
	Overall	Faisalabad	Islamabad	Karachi	Lahore	Peshawar
No. of Students	39217	3468	11305	8459	12756	3229
Degree Programs	11	5	9	6	7	6
Gender	M: 31866	M: 3001	M: 8611	M: 7255	M: 10011	M: 2988
	F: 7351	F: 467	F: 2694	F: 1204	F: 2745	F: 241
City	151	69	119	78	80	109
Secondary Education	SSC: 30952	SSC: 2961	SSC: 8346	SSC: 6936	SSC: 9769	SSC: 2940
	Olevel: 7070	Olevel: 411	Olevel: 2554	Olevel: 1407	Olevel: 2517	Olevel: 181
Higher Secondary	HSSC: 29634	HSSC: 2930	HSSC: 7793	HSSC: 5900	HSSC: 10041	HSSC: 2970
	Alevel: 5102	Alevel: 281	Alevel: 2146	Alevel: 659	Alevel: 1900	Alevel: 116

RESULTS AND DISCUSSIONS

Now when the data is pre processed , transformed and cleaned , we move forward to the feature selection . The results here are also explored in two ways: first for the whole data and then separately for each campus.

For attributes where the data set was imbalanced down sampling is done to balance the dataset before finding out correlations.

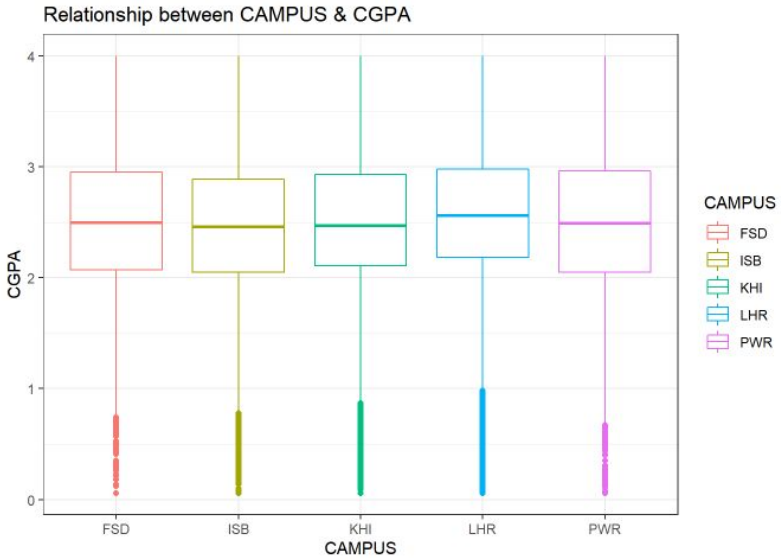
Working on the whole data :

Null Hypothesis 1: Campus affect CGPA
Alternative Hypothesis 1: Campus doesn't affect CGPA .

```
##               Df Sum Sq Mean Sq F value Pr(>F)
## CAMPUS         4    2.0  0.4921   2.472 0.0426 *
## Residuals   2495  496.6  0.1991
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

With p value less than significance value it can be seen that the null hypothesis is clearly rejected.

This can also be shown from the box plot below.



It can clearly be seen that the CGPA for each campus is almost the same , so campus doesn't really matter in terms of student performance.

Null Hypothesis 2: Degree Program affect CGPA
Alternative Hypothesis 2: Degree Program doesn't affect CGPA .

```
##               Df Sum Sq Mean Sq F value   Pr(>F)
## PROG_CODE      5   5.44  1.0876   5.436 5.94e-05 ***
## Residuals   1194 238.87  0.2001
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

This null hypothesis is also rejected with p value less than significance value.

Null Hypothesis 3: Gender affect CGPA
Alternative Hypothesis 3: Gender doesn't affect CGPA .

```
##               Df Sum Sq Mean Sq F value   Pr(>F)
## GENDER         1   9.85   9.846    49.4 3.49e-12 ***
## Residuals   1198 238.77   0.199
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

And clearly CGPA doesn't relate to gender.

Null Hypothesis 4: City affect CGPA
Alternative Hypothesis 4: City doesn't affect CGPA .

```
##               Df Sum Sq Mean Sq F value  Pr(>F)
## CITY          16   6.24  0.3897   2.103 0.00667 **
## Residuals   1003 185.84  0.1853
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

Comparing the p value and significance value , we accept the null hypothesis .

Null Hypothesis 5: Secondary Education (SSC/O Level) affect CGPA
Alternative Hypothesis 5: Secondary Education (SSC/O Level) doesn't affect CGPA .

```
##              Df Sum Sq Mean Sq F value   Pr(>F)
## SECONDARY     1   3.67   3.674   17.11 3.81e-05 ***
## Residuals   998 214.21   0.215
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

With p level almost equal to significance value we accept the null hypothesis.

Null Hypothesis 6: School affect CGPA
Alternative Hypothesis 6: School doesn't affect CGPA .

```
##              Df Sum Sq Mean Sq F value   Pr(>F)
## SCHOOL       32  13.42  0.4193   2.237 0.000142 ***
## Residuals   627 117.54  0.1875
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

As can be seen that p value is greater than significance level , clearly accept the null hypothesis.

Null Hypothesis 7: Higher Secondary Education (HSSC / A Level) affect CGPA
Alternative Hypothesis 7: Higher Secondary Education (HSSC / A Level) doesn't affect CGPA .

```
##                   Df Sum Sq Mean Sq F value   Pr(>F)
## HIGHER_SECONDARY   1   6.74   6.739   35.11 4.28e-09 ***
## Residuals        998 191.53   0.192
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

With a p value way too small than the significance value , we reject the null hypothesis.

Null Hypothesis 8: College affect CGPA
Alternative Hypothesis 8: College doesn't affect CGPA .

```
##              Df Sum Sq Mean Sq F value   Pr(>F)
## COLLEGE      51  18.95  0.3716   1.878 0.000254 ***
## Residuals   988 195.48  0.1978
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

Since p value is greater than significance value , the null hypothesis is accepted.

Null Hypothesis 9: Admission Year affect CGPA
Alternative Hypothesis 9: Admission Year doesn't affect CGPA .

```
##              Df Sum Sq Mean Sq F value Pr(>F)
## FIRST_SEM    12   3.09  0.2571   1.188  0.288
## Residuals   637 137.85  0.2164
```

With a greater p value , there is no evidence against the null hypothesis hence accepted.

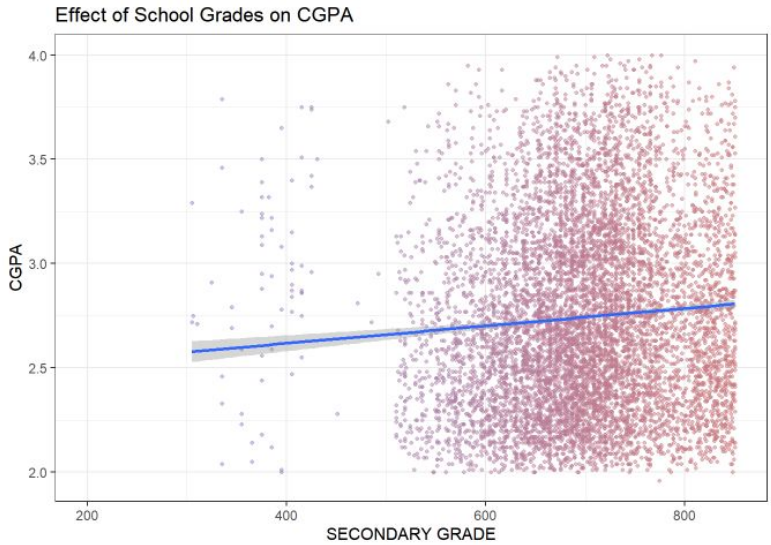
Null Hypothesis 10: Graduation Year affect CGPA
Alternative Hypothesis 10: Graduation Year doesn't affect CGPA .

```
##               Df Sum Sq Mean Sq F value Pr(>F)
## LAST_SEM      30  73.69  2.4564   18.42 <2e-16 ***
## Residuals   1519 202.54  0.1333
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

Since the p value is less than the significance value , we will go with the alternative hypothesis.

For the remaining attributes we calculate correlation coefficient .

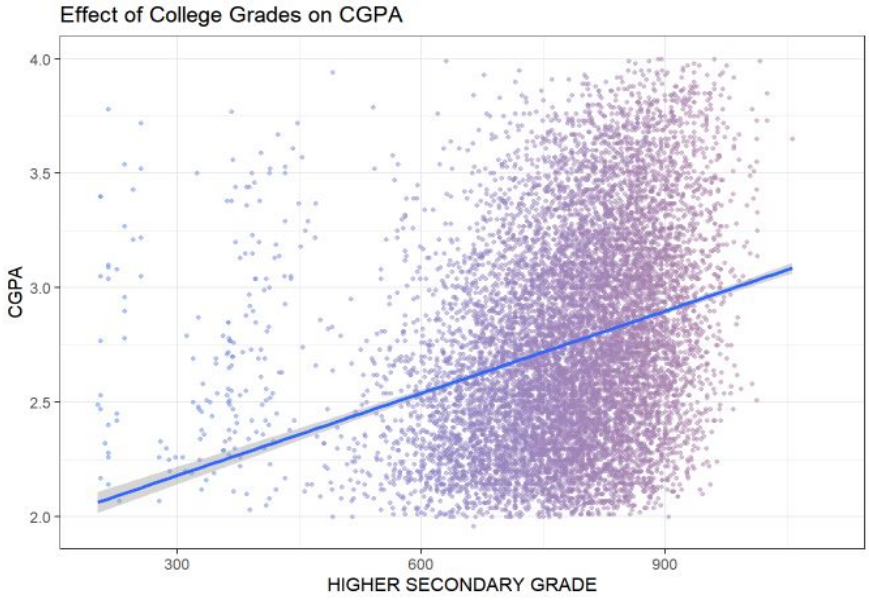
Effect of School Grades on CGPA:




```
##
##  Pearson's product-moment correlation
##
## data:  data$SEC_GRADE and data$CGPA
## t = 9.459, df = 10741, p-value < 2.2e-16
## alternative hypothesis: true correlation is not equal to 0
## 95 percent confidence interval:
##  0.07210533 0.10961323
## sample estimates:
##       cor
## 0.09089151
```

The value of 0.09 shows that there is no correlation between Secondary Grade and CGPA.

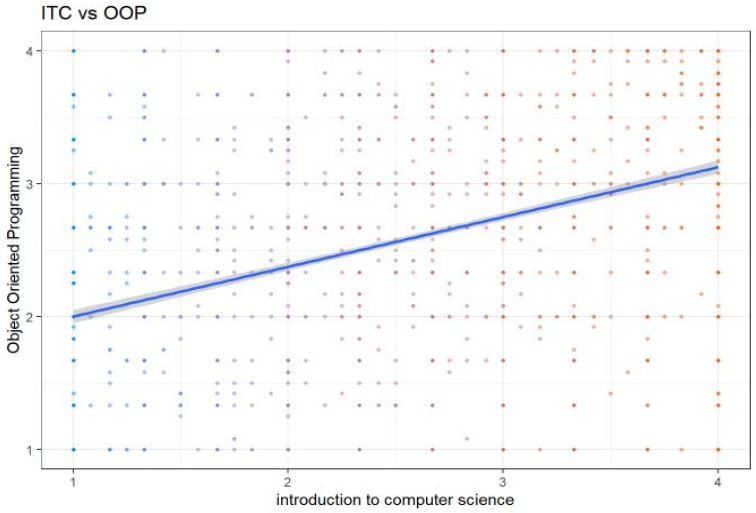
Effect of College Grades on CGPA:




```
##
##  Pearson's product-moment correlation
##
## data:  data$HIG_SEC_GRADE and data$CGPA
## t = 27.917, df = 10741, p-value < 2.2e-16
## alternative hypothesis: true correlation is not equal to 0
## 95 percent confidence interval:
##  0.2423769 0.2776395
## sample estimates:
##       cor
## 0.2600949
```

Higher Secondary Grade and CGPA show a very weak correlation.

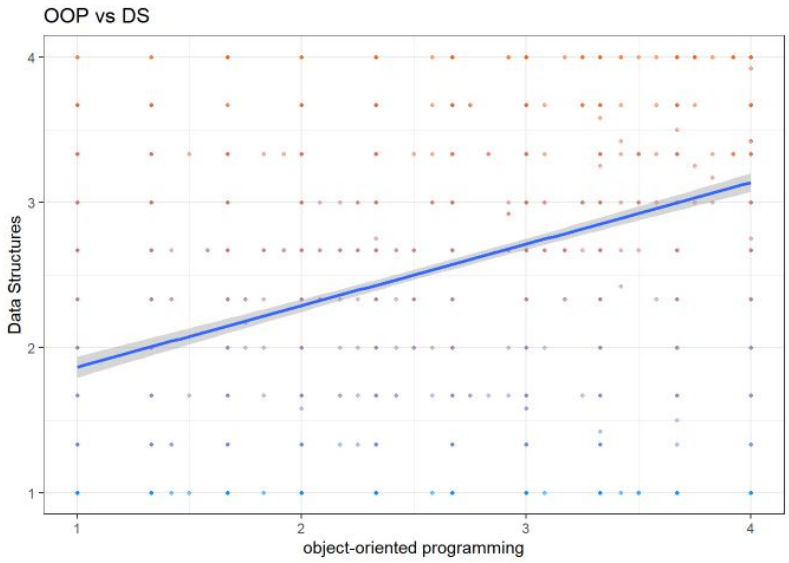
ITC vs OOP:



```
##
##  Pearson's product-moment correlation
##
## data:  CS_courses$`object-oriented programming` and CS_courses$`introduction to computer science`
## t = 25.652, df = 3679, p-value < 2.2e-16
## alternative hypothesis: true correlation is not equal to 0
## 95 percent confidence interval:
##  0.3617646 0.4165830
## sample estimates:
##       cor
## 0.3895187
```

The value 0.38 of correlation coefficient shows that somehow performance of Introduction to Computing and Object Oriented Programming is related.

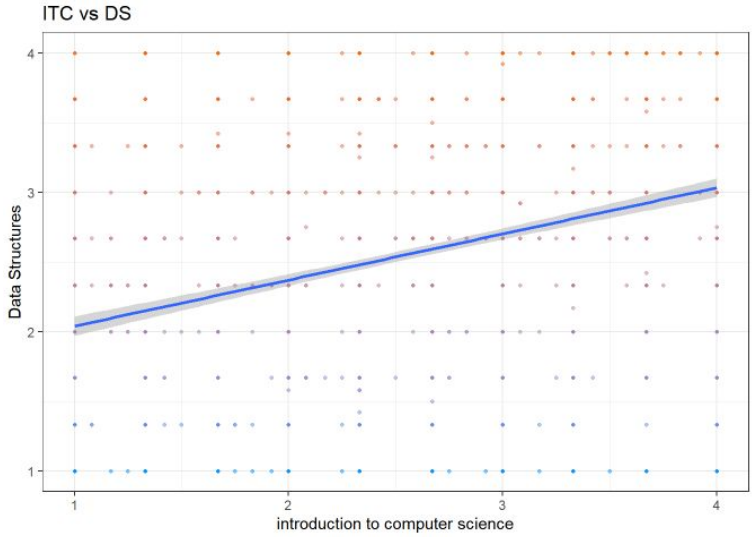
OOP vs DB:




```
##
##  Pearson's product-moment correlation
##
## data:  CS_courses$`object-oriented programming` and CS_courses$`data structures`
## t = 21.18, df = 2114, p-value < 2.2e-16
## alternative hypothesis: true correlation is not equal to 0
## 95 percent confidence interval:
##  0.3826000 0.4529283
## sample estimates:
##       cor
## 0.4183911
```

The courses Object Oriented Programming and Data Structures also have a positive relationship.

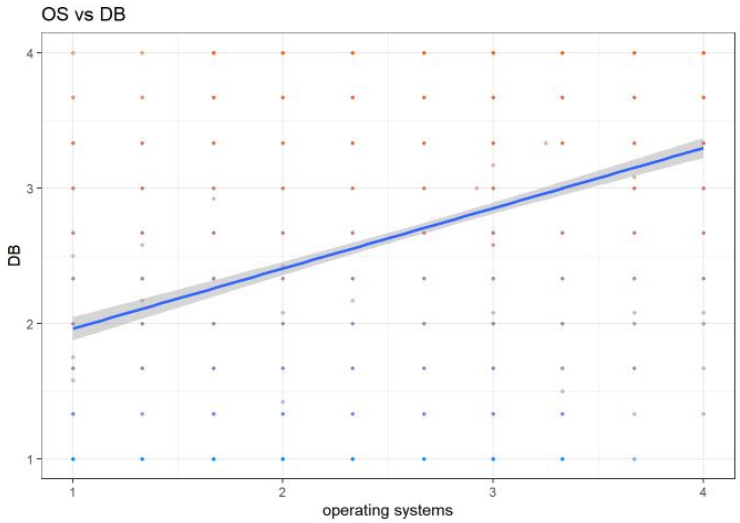
ITC vs DS:




```
##
##  Pearson's product-moment correlation
##
## data:  CS_courses$`introduction to computer science` and CS_courses$`data structures`
## t = 16.653, df = 2114, p-value < 2.2e-16
## alternative hypothesis: true correlation is not equal to 0
## 95 percent confidence interval:
##  0.3023191 0.3776760
## sample estimates:
##       cor
## 0.3405443
```

As compared to OOP , Introduction to Computing affects the grade in Data Structures less than it does in OOP.

OS vs DB:



```
##
##  Pearson's product-moment correlation
##
## data:  CS_courses$`operating systems` and CS_courses$`database systems`
## t = 18.952, df = 1424, p-value < 2.2e-16
## alternative hypothesis: true correlation is not equal to 0
## 95 percent confidence interval:
##  0.4063716 0.4893245
## sample estimates:
##       cor
## 0.4488144
```

With the correlation coefficient of 0.44 Operating Systems and Database Systems show a relation.

CONCLUSION AND FUTURE WORK

Student performance is dependent upon different factors and they may slightly vary for different data of students , but more or less there are few factors which stay the same.
The summary of our feature selection is shown in the table below :

Features	Evaluation Metric					
	Overall	Faisalabad	Islamabad	Karachi	Lahore	Peshawar
Degree	No	Yes	Yes	No	Yes	Yes
Gender	No	Yes	No	No	No	Yes
City	Yes	Yes	Yes	Yes	Yes	Yes
Secondary Education	Yes	Yes	No	Yes	No	Yes
School	Yes	Yes	Yes	Yes	Yes	Yes
Higher Secondary	No	No	No	Yes	No	Yes
College	Yes	Yes	Yes	Yes	Yes	Yes
Admission Year	Yes	Yes	Yes	Yes	Yes	No
Graduation Year	No	No	No	No	No	No
School Grade	No	No	No	No	No	No
College Grade	No	No	No	No	No	No

So we selected degree , city , secondary education , school , college and admission year as important factors that affect student performance.

Since our FYP limited our scope to finding correlation between CS courses only , future work can be done on finding correlations between courses of other domains also for example EE , SS , MT etc.

The work will be further carried to build a predictive model using the selected features . Different models will be implemented including Logistic regression , Decision trees , Random Forest , Support Vector Machine , KNN etc. The performance for each model will be compared in terms of accuracy and other performance metrics. Finally , the whole work will be integrated with a fully functional website which will support data visualization and query processing features.