BIRMINGHAM CITY
University

EDM FOR STUDENT'S UNIVERSITY SELECTION IN UNITED STATES

Darshan Tamboli (STUDENT ID: 21171488)

CMP7206 Data Mining

1/7/22

CONTENTS

1. DOMAIN DESCRIPTION

Nowadays, education had already produced an enormous amount of relevant data, which can be used to generate a range of research values. But nowadays, most universities and colleges experience enrolment difficulties. For most people, higher education is the desired part of preparing for professional life.

Admission to the university for higher education has thus become a critical and difficult topic. As a result, University enrolment projection systems are required to assist students in enrolling in the correct university as well as assist institutes to allocate students.

However, this decision-making process is extremely complicated, as so many students have to go on to college each year. Because this procedure is based on more than just student academic scores, it is also influenced by individuals' backgrounds.

DM is an effective technique that is best characterized as the computerized mode of deriving relevant data and understandings such as trends and relationships (U. M. Fayyad, 1996).

The findings obtained by DM tools can benefit universities and colleges in a variety of ways, including not only making correct judgments, having better extensive planning in instructing students, projecting individual actions with more precision, and allowing the organization to more efficiently utilize resources and manpower. It has the effect of increasing the process's efficacy and proficiency (Yacef, 2009) (A. AL-Malaise, 2014)].

This study makes an attempt to tackle higher learning data requirements; give indications of educational monitoring and assessment that are consistent, reliable, full, and meaningful; and publish meaningful, great information to data to academic authorities, education policymakers, students, educational institutes.

2. Problem Definition

This research aims solutions for Educational Data Mining problems by applying descriptive and predictive techniques such as hierarchical clustering, k-means clustering, Decision Trees, and regression analysis. This research will assist in presenting effective algorithms for students, universities, and government authorities in order to simplify the admission process and increase university enrolment numbers

Due to the rapid increase in the cost of living on campus, the admission yield of the certain university is dropped, and due to the availability of a large number of universities students are facing issues while selecting universities as per their requirements. Government officials can work on certain aid to boost enrolment in such universities and students can target the desired university.

3. Dataset Description

University Admissions (Qian, 2018)**:**

It is a publicly available dataset on university admissions in the United States. The collection includes information including over 1517 universities it has more than 100 columns. Its characteristics allow you to obtain insights into admission data for each university from various perspectives: from the number of applicants, how many students offered admissions, how many actually enrolled, the total price for in-state or out-of-state students to live on campus, total part-time enrolment as well as full-time enrolment, grant aid or a loan aid, etc.

Table 1: Sample of University Admissions Dataset

Name	Applicants	Admission	Enrolled to	Percent ad	Admission	Total price	Total price	State abbr	Sector of i	Control of	Historically	Degree of	Carnegie C	Total enro	Full-time e	Part-time e	U
Alabama A	6142	5521	1104	90	20	21849	27441	Alabama	Public, 4-y	Public	Yes	City: Midsi	Master's C	5020	4439	581	
University	5689	4934	1773	87	36	22495	31687	Alabama	Public, 4-y	Public	No	City: Midsi	Research U	18568	11961	6607	
University	2054	1656	651	81	39	23466	35780	Alabama	Public, 4-y	Public	No	City: Midsi	Research U	7376	4802	2574	
Alabama S	10245	5251	1479	51	28	18286	25222	Alabama	Public, 4-y	Public	Yes	City: Midsi	Master's C	6075	5182	893	
The Univer	30975	17515	6454	57	37	27000	41500	Alabama	Alabama	Public, 4-y	Public	No	City: Small	Research U	34752	29498	

This report is meant to divide the universities into different groups to understand the common factors and challenges.

3.1 Data set Pre-processing

Before analysis of any data set, it should first be pre-processed. This implies that perhaps the information should be outlined with a particular goal in mind in order for a conclusion to be obtained.

i. Columns like “Level of the institution”, “Tribal college” were removed because all record has common data so it’s not going to help for any kind of discovery.
ii. Derived one column to store a Boolean flag to identify a university or a college that is historically black.
iii. Some columns have less relevance to targeted knowledge discovery so removed
iv. “FIPS state code” column has same data as “State abbreviation” column so removed this column in data pre-processing
v. Some universities do not have any value in the required column so removed these entries

4. EXPERIMENTS

This chapter explains the various methodologies being used to evaluate datasets. In particular, the justification for picking a particular methodology is detailed in the following sections of the study.

4.1 CLUSTERING

Clustering is the process of organizing a series of data objects as clusters so that elements inside a cluster seem to be more "identical" to one another than elements in those other clusters (G. Nizar, 2005). The theory of likeness could be stated in a number of ways, depending on the study's objective, context hypotheses, and earlier information about the topic.

Each k cluster as formed should meet the two requirements listed below [(Rezende, 2010):

1. Every cluster has to include at least one item.
2. Any item must always be associated with just a single cluster.

Clustering is often referred to as unsupervised learning since it is generally executed when no metadata about the association of data objects is provided

Clustering will help to group Universities based on the number of students who get enrolled from admission offered by the university and the number of students who applied for university.

4.2 DECISION TREE

The decision tree is a hierarchical arrangement of decision nodes that are connected by branches (Witten, 2005). The very last nodes are known as leaf nodes, so they are connected with a decision variable's subgroups in the course of the training process. In data mining, decision trees are the most often used stratification technique (Wilhelmiina Hämäläinen, 2010).

The decision tree-based prediction approach often needs a huge quantity of data, as well as in addition to getting excellent outcomes, each phase (i.e., tree creation, training, and testing) should have its individual data.

The decision tree can help students during the selection of a suitable university for their studies as per their requirements.

4.3 REGRESSION

Regression is supervised learning. The goal of regression is to define the connection between a specific numerical dependent variable (value to predict) and single or multiple numeric predictors (Lantz, Machine Learning with R - Second Edition, 2015).

The dependent variable, as the name indicates, is determined by the effect of the independent variable or variables (Lantz, Machine Learning with R Second edition, 2015).

We always have this type of mental link; When we predict any child's age depending upon how tall she is, we believe that taller will be the older.

The regression analysis has been widely used to model complicated connections between data variables, estimate the influence of a procedure on a result, and extrapolate in to next.

Regression can help to understand the factor affecting the enrolments in universities and with the help of these factors, we can predict the enrolments.

5 Analysis and Results

5.1 Clustering

The primary function in hierarchical clustering is just to merge the two adjacent clusters together into a bigger cluster. It organizes the items together into a hierarchical forming a tree-like graph known as a dendrogram. A distance between splits or joins (referred to as height) is depicted upon that y-axis on the dendrogram graph.

When calculating distance, this is a wise idea to rescale these numbers, and as such the mean equals one, as well as standard deviation equals zero. This conversion minimizes the influence of outliers and permits a single observation to be compared to the mean.

Table 2: Sample data for clustering

	Applicants.total	Admissions.total	Enrolled.total
1	6142	5521	1104
2	5689	4934	1773
3	2054	1656	651
4	10245	5251	1479

Table 3: Scaled Data for clustering

	Applicants.total	Admissions.total	Enrolled.total
1	-0.047267609	0.432993450	0.03036695
2	-0.098239492	0.297951134	0.54594307
3	-0.507252283	-0.456169327	-0.31874513
4	0.414404875	0.370878585	0.31936701

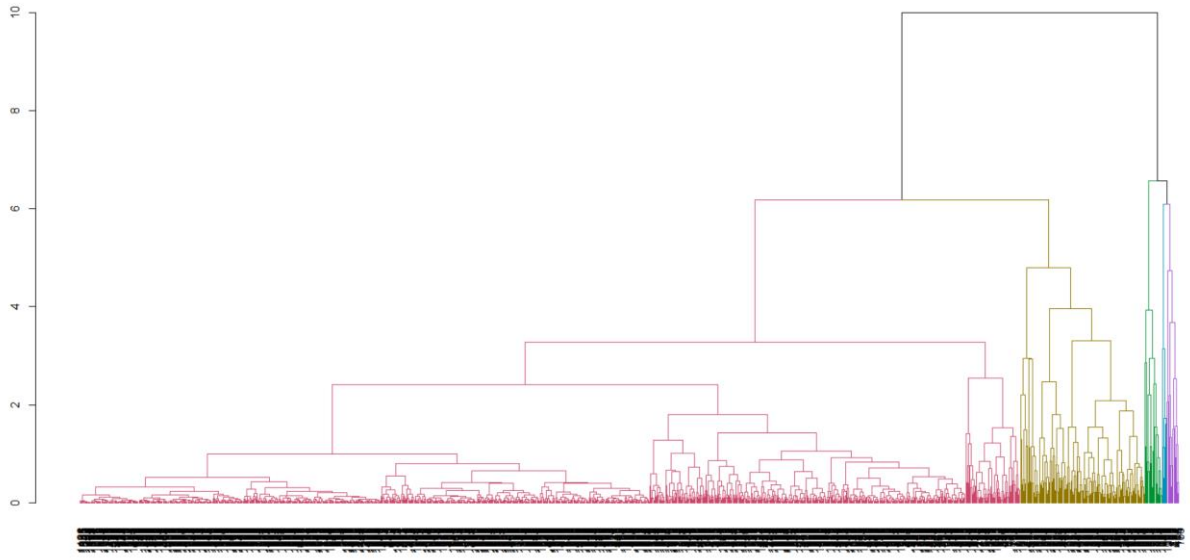

Fig 1: hierarchical Clustering

There are several basic differences seen between hierarchical clustering versus k-means clustering methods, while either can outperform the other under certain situations. Figure 2 indicates the data set is very complex as huge information is stored in this dataset. It's not easily readable with hierarchical clustering. Let's try k-mean clustering to present the same data into clusters.

Due to its simplicity, efficiency, as well as flexibility, K-means clustering is really a popular technique (ERIN, 2020). So, the k-means technique is used here to divide data items into different clusters.

When employing k-means clustering, an initial step is to specify the number of clusters (k) that would be formed mostly in the correct output. The Elbow technique is used to estimate the appropriate cluster centers.

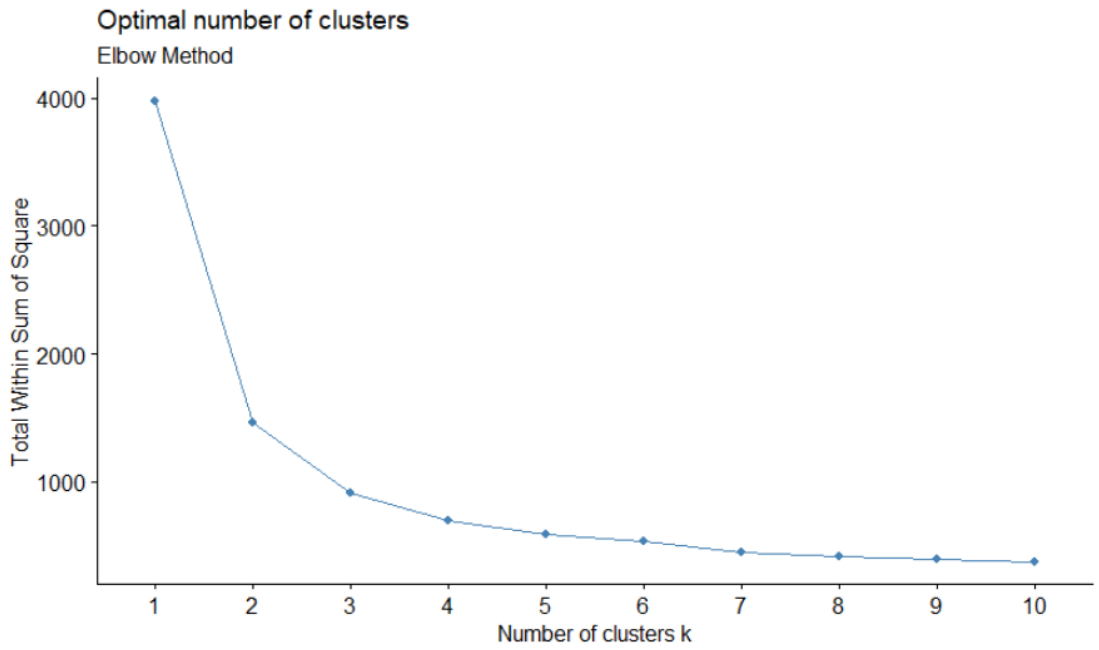

Fig 2: Elbow to find out optimal clusters

Figure 2 illustrates that the ideal k is four when the curve begins to have a declining return.

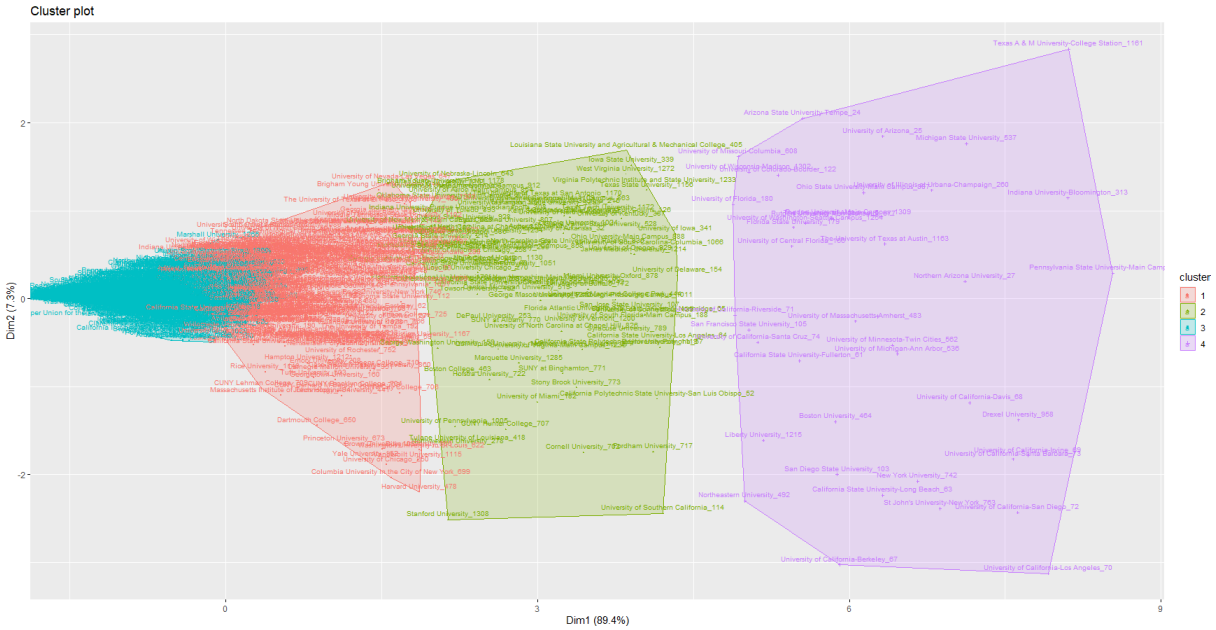

Fig 3: k-means Clustering

Figure 3 represents the universities into 4 clusters depending on the number of student applicants, offered admission by universities and the actually enrolled students. Businesses or students can target similar types of universities as per required actions. This clustering enables the identification of different universities having similar characteristics in terms of the number of applications, admission offers, and actual enrolment.

5.2 Decision Tree

A decision tree is a tree-shaped diagram that represents options and associated outcomes. The graph's nodes play the part of an incident or a decision, while the graph's edges reflect underlying decision rules or situations. It is widely used in R for Machine Learning and DM applications.

In a decision tree, first, we have to select required set columns to consider while creating a tree then, split the cleaned data set into two: one should be used to create a training module i.e., to build an appropriate decision tree, and another to put our model here to test

Table 4: Sample data for decision tree

	Admissions.yield...total	Total.price.for.s	Control.of.institution	Historically.Black.Boolean
1	20	27441	Public	1
2	36	31687	Public	0
3	39	35780	Public	0
4	28	25222	Public	1
5	37	41500	Public	0

Table 4 represent the sample data of columns considered for creating a decision tree to help the student who wants to choose a university from “Public” or a “Private not-for-profit” on the basis of admission yield in history, the total price for out of state students living on campus, and historically black university or a college

Any student or a consultant who is assisting a student can predict the possibilities of admission into the university which is “Public” or “Private not-for-profit” with different factors in mind such as admission yield, the total price for students living on campus, and historically black university or a college.

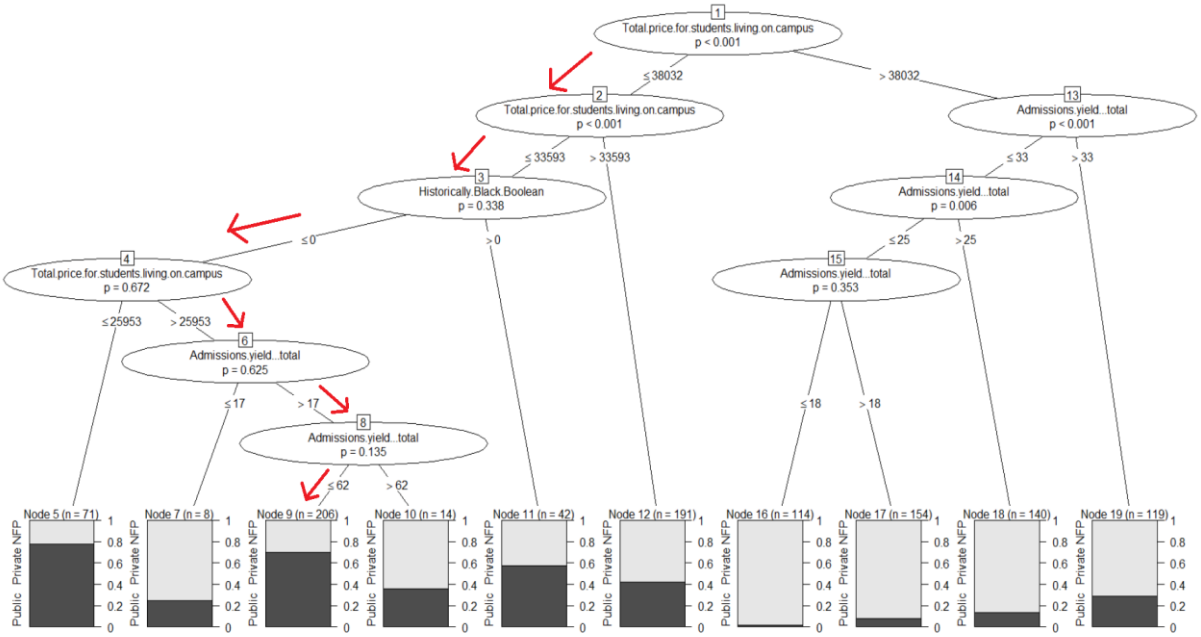

Fig 4: Decision Tree to choose “Public” or a “Private not-for-profit” university

As shown in figure 4, A Decision tree can help to predict higher chances of admissions and will assist to make the decision of choosing the type of university easily. For example, if a student who is having a budget of around 30000 dollars but wants a Historically Black University and wants to try for a university who is having an admission yield of around 50% then this decision tree will guide him to choose the type of university. Then the system will guide him to apply for a public university it has a higher probability of around 60%

```
> predict_unseen <-predict(tree, train)
> table_mat <- table(train$Control.of.institution, predict_unseen)
> table_mat
             predict_unseen
              Private NFP Public
  Private NFP         586     96
  Public              154    223
> accuracy_Test <- sum(diag(table_mat)) / sum(table_mat)
> print(paste('Accuracy for test', accuracy_Test))
[1] "Accuracy for test 0.763928234183192"
```

Fig 5: Misclassification error of train data

To assess classification efficiency, using a confusion matrix seems to be a preferable option. Inside a confusion matrix, the row shows an actual target, and each denotes a projected target. As calculated in figure 5, the training dataset has a score slightly above 76%.

5.3 Regression Analysis

Regression analysis is a statistical technique that examines the connection among a dependent factor and one or many factors, as well as the interactions between them.

The goal of linear regression is just to create a statistical equation that describes a dependent variable Y as a response of single or multiple X variables. As a result, when just the X seems available, this regression model may be used to estimate the Y.

```
Call:
lm(formula = Enrolled.total ~ Applicants.total + Admissions.total +
    DMData$Total.price.for.students.living.on.campus + DMData$Percent.of.freshmen.receiving.institutional.grant.aid,
    data = DMData)

Residuals:
    Min      1Q  Median      3Q     Max
-5510.9  -184.6     0.1   133.9  4565.5

Coefficients:
                                                              Estimate Std. Error t value Pr(>|t|)
(Intercept)                                                 1.088e+03  6.653e+01  16.357  < 2e-16 ***
Applicants.total                                            1.842e-02  3.539e-03   5.206 2.24e-07 ***
Admissions.total                                            2.217e-01  6.786e-03  32.666  < 2e-16 ***
DMData$Total.price.for.students.living.on.campus           -1.115e-02  1.676e-03  -6.657 4.08e-11 ***
DMData$Percent.of.freshmen.receiving.institutional.grant.aid -6.958e+00  6.375e-01 -10.914  < 2e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 550.3 on 1321 degrees of freedom
Multiple R-squared:  0.8207,    Adjusted R-squared:  0.8201
F-statistic:  1511 on 4 and 1321 DF,  p-value: < 2.2e-16
```

Fig 6: Summary of Regression Model

Figure 6 shows that the model's p-value is 2.2e-16, with a scientific value of 0.00000000000000022, signifying that it really is extremely close to zero as well as below than 0.05. The adjusted R-square is for this model is 0.8201 which is good. Let’s check for any improvement scope.

```
Call:
lm(formula = Enrolled.total ~ Applicants.total + Admissions.total +
    DMData$Total.price.for.students.living.on.campus + DMData$Percent.of.freshmen.receiving.institutional.grant.aid +
    DMData$Percent.of.freshmen.receiving.any.financial.aid +
    DMData$Percent.of.freshmen.receiving.other.federal.grant.aid,
    data = DMData)

Residuals:
    Min      1Q  Median      3Q     Max
-5402.8  -200.7    -7.3   150.4  4585.0

Coefficients:
                                                          Estimate Std. Error t value Pr(>|t|)
(Intercept)                                             1.504e+03  1.944e+02   7.740 1.96e-14 ***
Applicants.total                                        1.715e-02  3.643e-03   4.706 2.79e-06 ***
Admissions.total                                        2.201e-01  6.846e-03  32.154  < 2e-16 ***
DMData$Total.price.for.students.living.on.campus       -1.379e-02  1.875e-03  -7.355 3.33e-13 ***
DMData$Percent.of.freshmen.receiving.institutional.grant.aid -5.242e+00  8.701e-01  -6.024 2.20e-09 ***
DMData$Percent.of.freshmen.receiving.any.financial.aid -3.751e+00  2.088e+00  -1.796   0.0727 .
DMData$Percent.of.freshmen.receiving.other.federal.grant.aid -5.519e+00  1.219e+00  -4.527 6.52e-06 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 545.7 on 1319 degrees of freedom
Multiple R-squared:  0.824,     Adjusted R-squared:  0.8232
F-statistic:  1029 on 6 and 1319 DF,  p-value: < 2.2e-16
```

Fig 7: Summary of Regression Model with additional variables

Figure 7 indicates minimal improvement in adjusted R-square now it’s 82.32% which is quite better. The model indicates these are the factor that mostly affects enrolment.

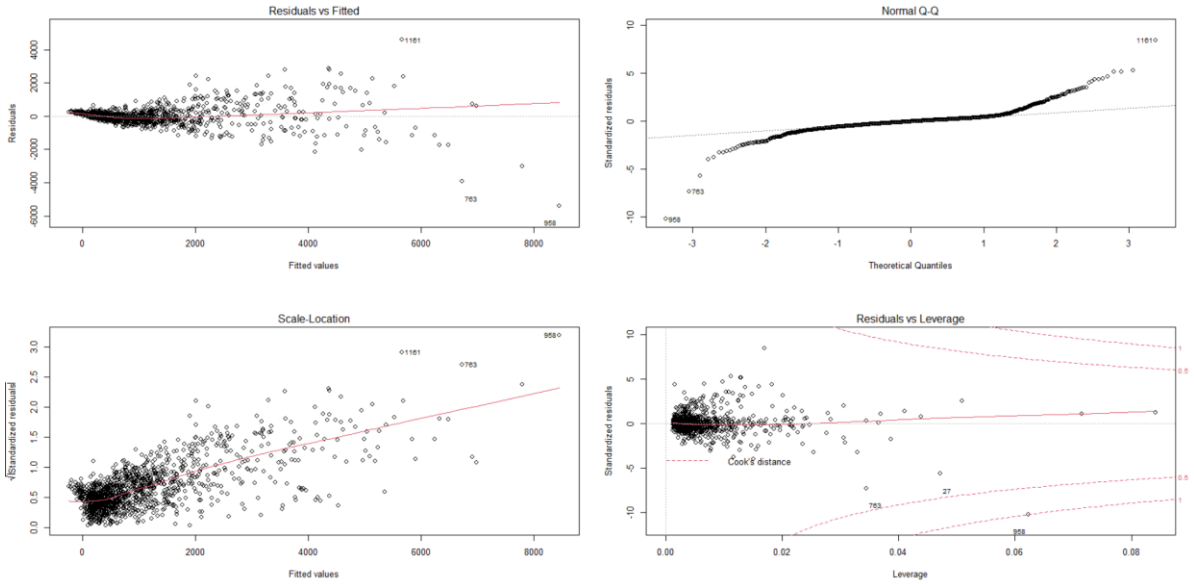

Fig 8: Summary of Regression Model with additional variables

As shown in figure 8, The unexplained variation is referred to as residuals. Those are not like model errors, but they have been estimated from that, as a result, observing a prejudice inside the residuals also would imply a prejudice inside the error.

In the upper right corner of there is a Normal Q-Q plot, The observed residuals under this model form a nearly one-to-one connection also with theoretical residuals.

6 CONCLUSION

The research findings addressed every one of the concerns expressed at the outset of the problem description using a total of four data mining approaches, which include two descriptive and two predictive techniques.

Clustering will help students to identify universities having similar characteristics in terms of how many people apply for particular universities, how many of them get admission offers, and how many students actually get to enrol in their respective universities. If government official wants to boost enrolment in universities that get the least enrolments, then this clustering will also help these government officials to target a set of universities.

If any student wants to get an idea before applying for any university based on the type of university (i.e., Private or Public) and having a certain budget in mind, the Decision tree can assist him to apply a particular type where he will be having higher probabilities of getting an admission. The regression model can help universities or government officials to identify factors affecting the enrolment of students in the universities.

This research may be extended in the coming time by applying different techniques.

7 REFERENCES

1. U. M. Fayyad, G. Piatetsky-Shapiro, P. Smyth and R. Uthurusamy, "Advances in knowledge discovery and data mining," 1996.
   Available at: https://ojs.aaai.org/index.php/aimagazine/article/view/1230/1131
   (Accessed: January 2022)

2. R. S. J. D. Baker and K. Yacef, "The State of Educational Data Mining in 2009 : A Review and Future Visions," Journal of Educational Data Mining, vol. 1, no. 1, pp. 3-16, 2009.
   Available at: https://jedm.educationaldatamining.org/index.php/JEDM/article/view/8
   (Accessed: January 2022)

3. A. AL-Malaise, A. Malibari and M. Alkhozae, "STUDENTS' PERFORMANCE PREDICTION SYSTEM USING MULTI AGENT DATA MINING TECHNIQUE," International Journal of Data Mining & Knowledge Management Process (IJDKP) , vol. 4, 2014
   Available at: https://aircconline.com/ijdkp/V4N5/4514ijdkp01.pdf
   (Accessed: January 2022)

4. G. Nizar , C. Michel , B. Nozha, "Unsupervised and Semi-supervised Clustering: a Brief Survey ,"INRIA Rocquencourt , B.P 105, France , pp. 1-12, August 15,2005
   Available at: http://cedric.cnam.fr/~crucianm/src/BriefSurveyClustering.pdf
   (Accessed: January 2022)

5. Rezende , H. R, Esmin , A.A.A ,"Proposed Application of Data Mining Techniques for Clustering Software Project," INFOCOMP – Special Edition, Brazil, pp. 43-48, Jul.2010.
   Available at: https://flosshub.org/sites/flosshub.org/files/art06.pdf
   (Accessed: January 2022)

6. ERIN HODGSON JAN 29, 2020 "CLUSTERING ALGORITHMS: WHICH ONE IS RIGHT FOR YOUR BUSINESS?"
   Available at:
   https://www.dotactiv.com/blog/clustering-algorithms (Accessed: January 2022)

7. Witten, I.H., Frank, E. (2005). Data Mining: practical machine learning tools and techniques. Morgan Kaufmann series in data management system
   Available at:
   http://www.academia.dk/BiologiskAntropologi/Epidemiologi/DataMining/Witten_and_Frank_DataMining_Weka_2nd_Ed_2005.pdf
   (Accessed: January 2022)

8. W. H¨am¨al¨ainen and M. Vinni, "Classifiers for educational data mining," Handbook of Educational Data Mining, 2010

Available at: https://www.researchgate.net/publication/48336339_Classifiers_for_educational_technology (Accessed: January 2022)

9. Lantz, Brett. Machine Learning with R - Second Edition, Packt Publishing, Limited, 2015. Available at: https://ebookcentral.proquest.com/lib/bcu/reader.action?docID=2122139 (Accessed: January 2022)

10. Samson Qian, "College Admissions Admission/Class Demographics by University",2018 Available at: https://www.kaggle.com/samsonqian/college-admissions (Accessed: January 2022)

Appendix: R Code

- **Clustering**

```
#Fetch Data Set
> DMData <- read.csv("DMDataSet_Cleaned.csv")
> View(DMData)
#Remove any blanks
> DMDataCleaned <- DMData[complete.cases(DMData),]
> View(DMDataCleaned)
# DMData's Labels
> DMDataCleaned.labels=DMDataCleaned$Name
> table(DMDataCleaned.labels)
> DMDataCleaned_Data <- select(DMDataCleaned, 2:4)
> View(DMDataCleaned_Data)
#Scale the Data
> DMDataCleaned_Data_Scale <- scale(DMDataCleaned_Data)
> View(DMDataCleaned_Data_Scale)
#Calculate the distance
> DMDataCleaned_Data <- dist(DMDataCleaned_Data_Scale)
#hierarchical algorithm
hc.out_DMData <- hclust(DMDataCleaned_Data, method="complete")
hc.out_DMData
#Dendodram
plot(hc.out_DMData)
#clusters
rect.hclust(hc.out_DMData,k=4, border = 2:5)
abline(h = 6, col = 'red')
avg_dend_obj <- as.dendrogram(hc.out_DMData)
install.packages("dendextend")
library(dendextend)
avg_col_dend <- color_branches(avg_dend_obj, h = 6)
plot(avg_col_dend)
#Calculate the how many clusters needed
> fviz_nbclust(DMDataCleaned_Data_Scale, kmeans, method = "wss") + labs(subtitle =
"Elbow Method")
#k-means
> km.out <- kmeans(DMDataCleaned_Data_Scale, centers=3, nstart=100)
> print(km.out)
#Visualizing the clusters
> km.clusters <- km.out$cluster
> rownames(DMDataCleaned_Data_Scale) <- paste(DMDataCleaned$Name,
1:dim(DMDataCleaned)[1], sep="_")
> fviz_cluster(list(data=DMDataCleaned_Data_Scale, cluster = km.clusters), labelsize = "8",
pointsize = .1)
```

❖ **Decision Tree**

```
set.seed(555)
#Select Features
DMData_Data<- select(DMData,6,8,11,13)
View(DMData_Data)
#Change to Factor
DMData_Data$Control.of.institution <- factor(DMData_Data$Control.of.institution)
#Partition the data into 2 sets
pd <- sample(2,nrow(DMData_Data),replace= TRUE, prob = c(0.8,0.2))
train <- DMData_Data[pd==1,]
validate <- DMData_Data[pd==2,]
View(validate)
#Decision tree creation
tree <- ctree(Control.of.institution~., data=train,controls = ctree_control(mincriterion = 0.99,minsplit=100))
#Visualization of decision tree
plot(tree)
#Misclassification of errors
p1 <- predict(tree,train)
tab1 <- table(Predicted = p1, Actual = train$Control.of.institution)
tab1
1 - sum(diag(tab1))/sum(tab1)
p2 <- predict(tree,validate)
tab2 <- table(Predicted = p2, Actual = validate$Control.of.institution)
tab2
1 - sum(diag(tab2))/sum(tab2)
#Accuracy check of the model
predict_unseen <-predict(tree, train)
table_mat <- table(train$Control.of.institution, predict_unseen)
table_mat
accuracy_Test <- sum(diag(table_mat)) / sum(table_mat)
print(paste('Accuracy for test', accuracy_Test))
```

❖ **Regression Analysis**

```
DMData <- read.csv("DMDataSet_Cleaned.csv" , stringsAsFactors = TRUE)
str(DMData)
model <- lm(Enrolled.total ~ Applicants.total + Admissions.total +
DMData$Total.price.for.students.living.on.campus +
DMData$Percent.of.freshmen.receiving.institutional.grant.aid +
DMData$Percent.of.freshmen.receiving.any.financial.aid +
DMData$Percent.of.freshmen.receiving.other.federal.grant.aid, data=DMData)
summary(model)
par(mfrow=c(2,2))
plot(model)
```