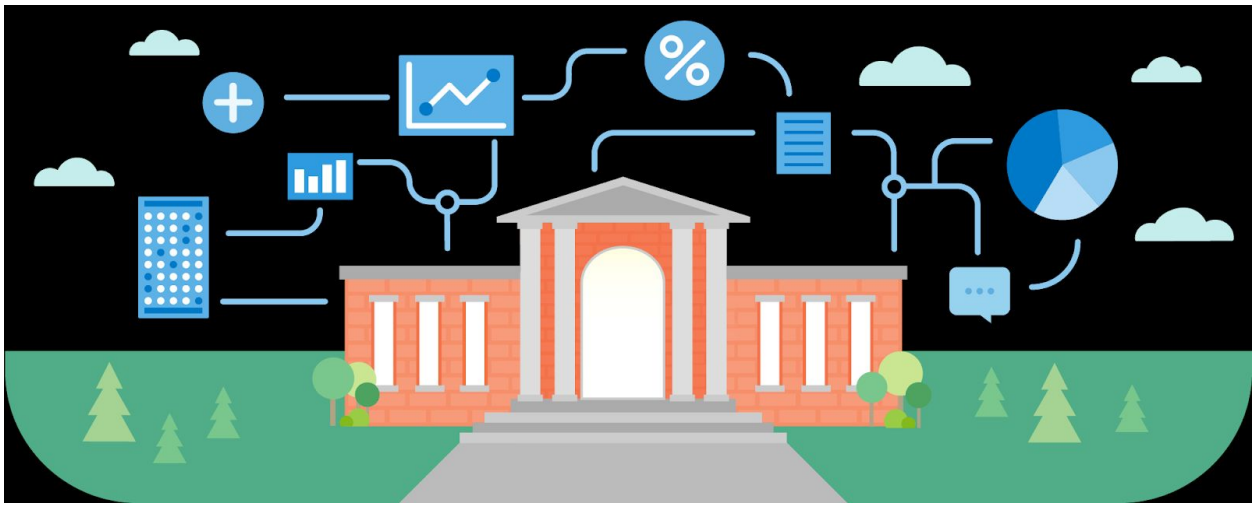
DATA SCIENCE WITH MACHINE LEARNING PROJECT

TITLE : COLLEGE ADMISSION PREDICTION USING MULTILINEAR REGRESSION AND NATURAL LANGUAGE PROCESSING



GROUP DETAILS :

GROUP NUMBER : 16

CAMPUS : RR CAMPUS

TEAM DETAILS :

1. Name : Anish D
   SRN : PES1201801334

2. Name : Natasha Bonala
   SRN : PES1201801653

3. Name : Ann Freida F
   SRN : PES1201801729

4. Name : Krishna PC
   SRN : PES1201801162

CONTENTS **PG NO**

INTRODUCTION ON MULTILINEAR REGRESSION

Thus education preparation students often have multiple questions about universities in which they can get admission and scholarship and accommodation. One of the main concerns is getting admitted to their dream university.The data attributes used to predict student's performance can include many features, such as student grades in some materials which were studied previously,demographic information such as sex, age and address, and social information such as parents cohabitation status, mother's and father's job, family size and so on.

Regression is a supervised machine learning technique that shares the same concept as classification in using a training dataset to make a prediction.It attempts to determine the strength and character of the relationship between one dependent variable (usually denoted by Y) and a series of other variables (known as independent variables).
The general form of each type of regression is:

- **Simple linear regression:** $Y = a + bX + u$
- **Multiple linear regression:** $Y = a + b_1X_1 + b_2X_2 + b_3X_3 + ... + b_tX_t + u$

Where:

- Y = the variable that you are trying to predict (dependent variable).
- X = the variable that you are using to predict Y (independent variable).
- a = the intercept.
- b = the slope.
- u = the regression residual.

There are multiple benefits of using regression analysis. They are as follows:

1. It indicates the **significant relationships** between dependent variable and independent variable.
2. It indicates the **strength of impact** of multiple independent variables on a dependent variable.

PROBLEM STATEMENT :

Prediction is the task of estimating the value of an unknown output variable based on the values of input variables. In education the output variables that refer to students' performance can be in the form of marks, numeric values or decisions categorical values.

To predict the "Chance of Admit" based on the different parameters that are provided in the dataset for a given university

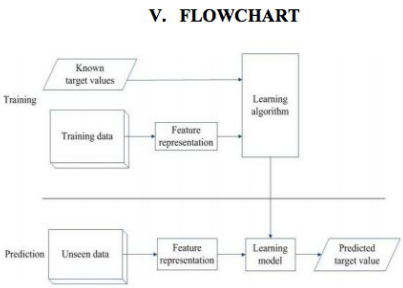
.

LITERATURE SURVEY :

Literature survey 1

Bayesian Networks Algorithms have been used to create a decision support network for evaluating the application submitted by foreign students of the university[5].

This model was developed to forecast the progress of prospective students by comparing the score of students currently studying at university. The model thus predicted whether the aspiring student should be admitted to university on the basis of



various scores of students. Since the comparisons are made only with students who got admission into the universities but not with students who got their admission rejected so this method will not be that much accurate.

Literature survey 2

The focus here is on the students who want to do their Masters in America. Students have to write GRE and TOEFL/IELTS. Once they have attended the exams they have to prepare their SOP and LOR, which are crucial factors that their admission depends upon. develop a model which will tell the students their chance of admission into a respective university. This model should consider all these crucial factors which play a vital role in the student admission process and should have high accuracy. The model name is UAP. To access this model a simple user interface must be developed.

IEEE paper	Our project
Uses a classifier model	Uses a regression model
Uses the k means clustering classifier	Uses multilinear regression(multilinear has the best results when compared to polynomial ,decision tree etc)
Provides binary output .The output is whether the student will get admitted or not	Provides a probabilistic output, which is more accurate
Sum of squared errors for 129 data entries is 669.0. This is because the input data provide contained response variables to be in the binary format(Yes-1 No-0).Hence did not provide adequate information to the classifier	Sum of squared errors of 200 data entries 0.0022139 Our project the dataset has response variables which were probabilistic that provided more information to the regressor model and Hence our model appears to be more successful

Literature survey 3

Data Analysis : Prior to multivariate analysis, we check the skewness and the kurtosis coefficient.Skewness ranges from -1 to +1.Kurtosis coefficients range beyond -1 to +1. This shows that the dataset is symmetric but it has outliers which makes the kurtosis coefficients rage beyond +/-1

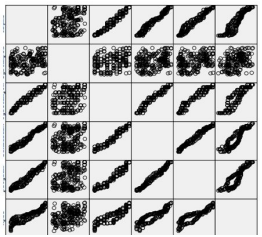
The scatter plot below shows the pairwise relation plots between the variables:

Almost all plots are seen to be elliptical.

So as to see if there were relations/dependencies , correlations, variance increase factors and condition indices were examined.The correlations values are not higher than 8.

All VIF values are less than 10.This points out that there are no multiple relations between variables.

Now to check whether or not the five independent variables in the model predict the KPSS value significantly correct, a multiple regression model gives the values shown below.The



model's degree of predicting the dependent variable is found to be R=0.932 The model's degree of explaining the variance is seen to be R^2 = 0.87.This shows that all five variables are significant in predicting the dependent variable KPSS score.This suggests that the model can significantly predict the dependent variable very well.

KPSS = 9.811 + 1.157 Measurement + 0.090 Educational Psychology - 0.339 Teaching Methods – 0.195 Guidance + 0.078 Curriculum Development

Literature survey 4

Citation : A Study on Multiple Linear Regression Analysis

End-Term-Scores received from

1.measurement and evaluation

2.educational psychology

3.curriculum development

4.guidance`

5.teaching methods

Variables	B	SH_B	β	T	p	Paired-r	Partial
Constant	9.811	2.261	-	4.340	.000	-	-
Measure.	1.157	.114	1.421	10.136	.000	.924	.554
Edu.Psyc.	.090	.028	.082	3.230	.001	.348	.207
Ins.Meth.	-.339	.103	-.404	-3.291	.001	.884	-.211
Guidance	-.195	.165	-.221	-1.181	.239	.894	-.077
Curri.De.	.078	.130	.093	.604	.547	.891	.040

In this project, I referred to the 2003 IEEE 58th Vehicular technology conference. We propose a flexible hierarchical framework for admission control, based on this architecture which aims for the prediction based admission scheme. The proposed system consists of two cascaded hybrid recommenders working together with the help of college predictor for achieving high performance. The college predictor algorithm uses the college's students' GPA and similar data for predicting most probable college. A prototype system has been implemented and tested with live data available. In addition to the high prediction accuracy rate, it also provides flexibility which is an advantage as the system can predict suitable colleges that match the students' profiles and the suitable track channels through which the students are advised to enter. The system is highly adaptive, since it can be turned up with other decision makers attributes performing trusted needed tasks faster and fairly

Literature survey for natural language processing

Literature survey 5

Sentiment Analysis for Social Media: A Survey

This problem for sentiment classification they proposed a model called dual sentiment analysis (DSA). They also extended the DSA framework from polarity (positive-negative) classification to 3-class (positive -negative-neutral) classification, by taking the neutral feelings into consideration.concentrated on engineering students' Twitter posts to understand problems and glitches in their educational experiences.

Literature survey 6

Challenges faced in this paper

A. Incremental Approach

B. Parallel Computing For Massive Data

C. Credibility/Behavior/Homophily

D. Sarcasm

E. Grammatically Incorrect Words

F. Review Author Segmentation

H. Handling Noise and Dynamism

Literature survey 7

Masses of users share their feelings on social media, making it a valuable platform for tracking and exploring public sentiment. Social media is one of the biggest platforms where massive instant messages are published every day which makes it an ideal source for capturing the opinions towards various curious topics, such as products, goods or celebrities, etc. The main goal of this paper is to give an overview of latest updates in sentiment analysis and classification methods.

Literature survey 8

The limitations of today's natural language processing technology are as follows:

1. Current systems have limited discourse capabilities that are almost exclusively handcrafted.
2. Domains must be narrow enough so that the constraints on the relevant semantic concepts and relations can be expressed using current knowledge presentation techniques, i.e., primarily in terms of types and sorts.
3. Handcrafting is necessary, particularly in the grammatical components of systems (the component technology that exhibits least dependence on the application domain).
4. The user must still adapt to the machine, but as the products testify, the user can do so Effectively.

ANALYSIS :

A Variance Inflation Factor(VIF) detects multicollinearity in regression analysis. Multicollinearity is when there's correlation between predictors (i.e. independent variables) in a model; it's presence can adversely affect your regression results.This gives you the R-squared values, which can then be plugged into the VIF formula. "i" is the predictor you're looking at (e.g. x1 or x2):

$$\mathrm{VIF} = \frac{1}{1 - R_i^2}$$

Variance inflation factors range from 1 and upwards. The numerical value for VIF tells you (in decimal form) what percentage the variance (i.e. the standard error squared) is inflated for each coefficient. For example, a VIF of 1.9 tells you that the variance of a particular coefficient is 90% bigger than what you would expect if there was no multicollinearity —

	feature	VIF
0	GRE Score	1436.417266
1	TOEFL Score	1343.763370
2	University Rating	21.499440
3	SOP	31.870699
4	CGPA	1035.520076
5	Research	2.857031

if there was no correlation with other predictors.

First, we determine the dependencies between dependent variables, we perform VIF calculations.
Above are the VIF values for each of the independent variables.
Here, we find that most of the VIF values are above 10 ,which means that the variables are heavily interdependent. Therefore, multilinear regression is a good approach.

The relation between the dependent and independent variables can be obtained as:
Therefore , the " chance of permit " in equation form can be given as :

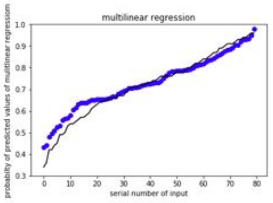
```
[[ 0.00173741  0.00291958  0.00571666 -0.00330517  0.02235313  0.11893945
  0.02452511]]
```

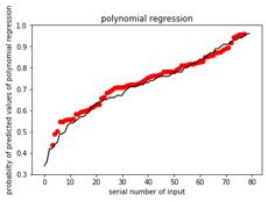
Chance of Admit = 0.001737 GRE Score + 0.00291 TOEFL Score + 0.0057 University Rating – 0.0033 SOP + 0.0223 LOR + 0.1189 CGPA + 0.0245 Research

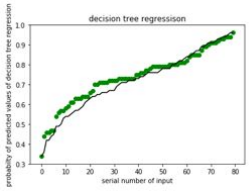
We also find the correlation between the variables:
Testing using different methods of regression :
Here, we use multilinear regression, polynomial regression and decision tree regression. The plots are as follows:

	Serial No.	GRE Score	TOEFL Score	University Rating	SOP	LOR	CGPA	Research	Chance of Admit
Serial No.	1.000000	-0.097526	-0.147932	-0.169948	-0.166932	-0.088221	-0.045608	-0.063138	0.042336
GRE Score	-0.097526	1.000000	0.835977	0.668976	0.612831	0.557555	0.833060	0.580391	0.802610
TOEFL Score	-0.147932	0.835977	1.000000	0.695590	0.657981	0.567721	0.828417	0.489858	0.791594
University Rating	-0.169948	0.668976	0.695590	1.000000	0.734523	0.660123	0.746479	0.447783	0.711250
SOP	-0.166932	0.612831	0.657981	0.734523	1.000000	0.729593	0.718144	0.444029	0.675732
LOR	-0.088221	0.557555	0.567721	0.660123	0.729593	1.000000	0.670211	0.396859	0.669889
CGPA	-0.045608	0.833060	0.828417	0.746479	0.718144	0.670211	1.000000	0.521654	0.873289
Research	-0.063138	0.580391	0.489858	0.447783	0.444029	0.396859	0.521654	1.000000	0.553202
Chance of Admit	0.042336	0.802610	0.791594	0.711250	0.675732	0.669889	0.873289	0.553202	1.000000

1.Multilinear regression 2.Polynomial regression 3.Decision Tree regression







Multilinear and polynomial regression seems to give a better fit than decision tree regression.
We then compare the error functions of the three methods of regression.

Regression Models	Error
Multilinear	0.00222
Polynomial	0.0179
Decision tree	1.6493

We see that the multilinear regression model has the least error, therefore giving the best fit.

CONCLUSION :

Determining the factors that affect the student's performance in academic institutions is a very interesting task since it will help educators to enhance their learning and teaching process. In this context, we have proposed a methodology that well examines the students attributes and selects among them the most important to build a prediction model.Our methodology consists in applying different methods for the most important variables and then using the selected variables to build different linear regression models and the best approach is seen to be the Multilinear Regression Method.

NATURAL LANGUAGE PROCESSING ON COLLEGE REVIEWS

INTRODUCTION TO NLP :

Natural language processing (**NLP**) is concerned with the interactions between computers and human language, in particular how to program computers to process and analyze large amounts of natural language data. The result is a computer capable of "understanding" the contents of documents, including the contextual nuances of the language within them. The technology can then accurately extract information and insights contained in the documents as well as categorize and organize the documents themselves.**Sentiment analysis** (or **opinion mining**) is a technique used to determine whether data is positive, negative or neutral.Sentiment analysis models focus on polarity (*positive, negative, neutral*) but also on feelings and emotions (*angry, happy, sad*, etc), urgency (*urgent, not urgent*) and even intentions (*interested v. not interested*).

HOW IT WORKS?

NLP deals with applying algorithms that extract the rules of a natural language and convert it so a computer can understand

Many different techniques are used for this process, including:

- **Lemmatization:** grouping inflected forms of a word into a single form
- **Word segmentation:** separating a large piece of text into units
- **Parsing:** analyzing the grammar of a sentence
- **Word sense disambiguation:** determine meaning to word based on contex

PROBLEM STATEMENT :

Classifying college reviews using Natural Language Processing methods.

METHODOLOGY :

College reviews are a great source of information for students. From the universities' point of view, online reviews can be used to gauge the students' feedback on the services provided to the students and the faculty.However, since these reviews are quite often overwhelming in terms of numbers and information, an intelligent system, capable of finding key insights (topics) from these reviews, will be of great help for the students. This system will serve two purposes:

1. Enable students to quickly extract the key topics covered by the reviews without having to go through all of them
2. Help the university get student feedback in the form of topics.

ANALYSIS AND RESULTS :

1. Basic Feature Extraction

We can use text data to extract a number of features even if we don't have sufficient knowledge of NLP.

1.1 Number of Words

One of the most basic features we can extract is the number of words in each review. The basic intuition behind this is that generally, the negative sentiments contain a lesser amount of words than the positive ones.

	Review	word_count
0	everyone's doing research! whether you are a s...	22
1	if the grading could be a little more lenient,...	29
2	great.	1
3	the lack of field trips, participation, and mo...	21
4	college is good.	3

1.2 Number of characters

This feature is also based on the previous feature intuition. Here, we calculate the number of characters in each review. This is done by calculating the length of the review.

	Review	char_count
0	everyone's doing research! whether you are a s...	137
1	if the grading could be a little more lenient,...	168
2	great.	6
3	the lack of field trips, participation, and mo...	123
4	college is good.	16

1.3 Average Word Length

We will also extract another feature which will calculate the average word length of each review. This can also potentially help us in improving our model.Here, we simply take the sum of the length of all the words and divide it by the total length of the review:

	Review	avg_word
0	everyone's doing research! whether you are a s...	5.272727
1	if the grading could be a little more lenient,...	4.827586
2	great.	6.000000
3	the lack of field trips, participation, and mo...	4.904762
4	college is good.	4.666667

1.4 Number of stopwords

Generally, while solving an NLP problem, the first thing we do is to remove the stopwords.Stop words" usually refers to the most common words in a language.They are usually removed.But sometimes calculating the number of stopwords can also give us some extra information which we might have been losing before.Here, we have imported stopwords from *NLTK*, which is a basic NLP library in python.

	Review	stopwords
0	everyone's doing research! whether you are a s...	9
1	if the grading could be a little more lenient,...	14
2	great.	0
3	the lack of field trips, participation, and mo...	11
4	college is good.	1

1.5 Number of Uppercase words

Anger or rage is quite often expressed by writing in UPPERCASE words which makes this a necessary operation to identify those words.

2. Basic Pre-processing

Data preprocessing and cleaning is an important step before any text mining task, in this step, we will remove the punctuations, stopwords and normalize the reviews as much as possible. After every preprocessing step, it is a good practice to check the most frequent words in the data. So far, we have learned how to extract basic features from text data. Before diving into text and feature extraction, our first step should be cleaning the data in order to obtain better features. We will achieve this by doing some of the basic pre-processing steps on our training data.

2.1 Lower case

The first pre-processing step which we will do is transform our reviews into lower case. This avoids having multiple copies of the same words. For example, while calculating the word count, ‘Analytics’ and ‘analytics’ will be taken as different words.

```
0    everyone's doing research! whether you are a s...
1    if the grading could be a little more lenient,...
2                                               great.
3    the lack of field trips, participation, and mo...
4                                     college is good.
Name: Review, dtype: object
```

2.2 Removing Punctuation

The next step is to remove punctuation, as it doesn’t add any extra information while treating text data. Therefore removing all instances of it will help us reduce the size of the training data.

```
0    everyones doing research whether you are a soc...
1    if the grading could be a little more lenient ...
2                                               great
3    the lack of field trips participation and moti...
4                                     college is good
Name: Review, dtype: object
```

2.3 Removal of Stop Words

As we discussed earlier, stop words (or commonly occurring words) should be removed from the text data. For this purpose, we can either create a list of stopwords ourselves or we can use predefined libraries.

```
0    everyones research whether social sciences ste...
1    grading could little lenient would however und...
2
3    lack field trips participation motivation enga...
4
Name: Review, dtype: object
```

2.4 Stemming

It is done as a pre-processing step to prepare the text for further advanced processing.Stemming refers to the removal of suffices, like “ing”, “ly”, “s”, etc. by a simple rule-based approach. This produces the root form of a word.

```
[0    everyon research whether social scienc stem st...
1    grade could littl lenient would howev understa...
2
3    lack field trip particip motiv engag primari r...
4
Name: Review, dtype: object]
```

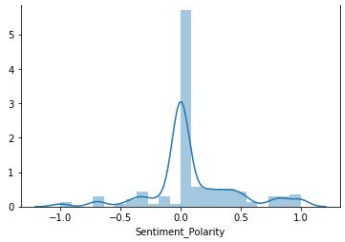
3. Advance Text Processing Using Sentimental Analysis

Up to this point, we have done all the basic pre-processing steps in order to clean our data. Now, we can finally move on to extracting features using NLP techniques.We use TextBlob for Sentimental Analysis.

We can see that the analysis returns a tuple representing polarity and subjectivity of each tweet. Here, we only extract polarity as it indicates the sentiment as value nearer to 1 means a positive sentiment and values nearer to -1 means a negative sentiment.Then a value of sentiment is associated for each review.

```
0    (0.011111111111111112, 0.10555555555555556)
1                  (0.0945, 0.4213333333333333)
2                                   (0.0, 0.0)
3                (0.125, 0.44999999999999996)
4                                   (0.0, 0.0)
```

Let’s see the distribution of sentiment polarity of the review



	Review	sentiment
0	everyones research whether social sciences ste...	0.011111
1	grading could little lenient would however und...	0.094500
2		0.000000
3	lack field trips participation motivation enga...	0.125000
4		0.000000

Here, we see that most of the reviews are positive and have a polarity between 0 - 0.5

CONCLUSION :

We find that our method can improve general text classification by using sentimental analysis for the unique reviews given by different individuals.

Sentiment Analysis leads to development of better and good business management, reviews.

REFERENCES :

1. A Study on Multiple Linear Regression Analysis
   Gü lden Kaya Uyanık and Neşe Gü ler / Procedia - Social and Behavioral Sciences 106 (2013) 234 – 240
2. 2003 IEEE 58th Vehicular Technology Conference VTC 2003 Fall (IEEE) Cat. No. 03CH37484
3. Prediction of the admission lines of college entrance examination based on machine learning 14-17 Oct. 2016 2nd IEEE International Conference on Computer and Communications (ICCC)
4. Prediction for University Admission using Machine Learning (IJRTE)
5. Natural Language Processing: State of The Art, Current Trends and Challenges.- Diksha Khurana, Aditya Koli, Kiran Khatter and Sukhdev Singh Department of Computer Science and Engineering