## Dr. AMBEDKAR INSTITUTE OF TECHNOLOGY

(An Autonomous Institute, Affiliated to Visvesvaraya Technological University, Belagavi, Accredited by NAAC, with ‘A’ Grade)

**Near Jnana Bharathi Campus, Bengaluru – 560056**



**DEPARTMENT OF INFORMATION SCIENCE AND ENGINEERING**

**Project Report on**

**“**Admission prediction in college**”**

**Submitted in partial fulfillment of the requirement for the award of the Degree of**

**Bachelor of Engineering In Information Science and Engineering**

Submitted by

**M P NEHA(1DA19IS054)**

**VIJAYLAKSHMI S KALASA(1DA19IS054)**

**For the academic year 2022-23.**

**Under the Guidance of**

**HOD**

**Dr. M V VIJAYA KUMAR**

**Professor & Head Dept. of ISE**

**Dept. of Information Science and Engineering**

**Dr. Ambedkar Institute of Technology**

**Bengaluru – 56.**

**Dr. AMBEDKAR INSTITUTE OF TECHNOLOGY**

Near Jnana Bharathi Campus, Bengaluru-560056

**(An Autonomous Institution, Aided by Government of Karnataka)**



**CERTIFICATE**

This is to certify that the project work titled **“Admission prediction in college”** is carried out by M P NEHA **[1DA19IS017]**, VIJAYLAKSHMI S KALASA **[DA19IS054]** is a bonafide student of **Dr. Ambedkar Institute of Technology, Bengaluru**, in partial fulfillment for the award of **Degree in Bachelor of Engineering** in **Information Science and Engineering** during the academic year 2021-2022. It is certified that all the corrections/suggestions indicated during Internal Assessment have been incorporated in the Project report deposited in the department. The Project report has been approved as it satisfies the academic requirements in respect of Project work prescribed for the said Degree

**ACKNOWLEDGEMENT**

The sense of jubilation that accompanies the successful completion of this Project would be incomplete without mentioning and thanking all the people who played a vital role in the completion of this project by providing endless encouragement and support.

I would like to profoundly thank the **Management - PVPWT** of Dr. Ambedkar Institute of Technology, fr providing such a healthy environment to learn and implement new technologies.

I wish to express my sincere thanks to **Dr. MEENAKSHI M,** Principal, Dr.AIT, who has always been a great source of inspiration while carrying out this Project work.

I am extremely grateful to **Dr. M V Vijaya Kumar**, Professor and Head, Department of ISE, Dr.AIT for providing us constant encouragement and permitting us to utilize the required laboratory facilities and a congenial working environment for the successful completion of this Project work.

I am highly indebted to my to **Dr. M V Vijaya Kumar**, Professor, Department of ISE, Dr.AIT for constant guidance and support, as well as for providing necessary information regarding the Project work.

I would also like to thank all the teaching and non-teaching staff members of the Department of Information Science and Engineering Department for their support during the course of this Project implementation.

Last but not the least, I would like to thank my parents and friends whose constant encouragement and support was crucial in execution and completion of this Project.

**M P NEHA**

**VIJAYLAKSHMI S KALASA**

### ABSTRACT

In today’s era we see a lot of students pursuing their education away from their home countries. The main country targeted by these international students is The United States of America. Majority of the international students in the United States of America are from India and China. In the past decade the number of Indian students pursuing post graduate education from the other countries has rapidly increased. With the increase in the number of international students studying in the foreign countries, each applicant has to face a tough competition to get admission in their dream university. Generally as the students don’t have much idea about the procedures, requirements and details of the universities in the other countries they seek help from the education consultancy firms to help them successfully secure admission in the universities which are best suitable for their profile, for this they have to invest huge amount of money as consultancy fees.

Apart from these the education consultancy firms there are few websites and blogs that guide the students on the admission procedures. The drawback of the currently available resources is that they are very limited and also they are not truly dependable taking into consideration of their accuracy and reliability. The aim of this research is to develop a system using machine learning algorithms, we will name it as Student Admission Predictor (SAP). It will help the students to identify the chances of their application to an university being accepted.

# TABLE OF CONTENTS

**CONTENTS Pg No**

[CHAPTER 1 : INTRODUCTION 5](#_TOC_250026)

* 1. [Overview](#_TOC_250025) 5
  2. [Objectives](#_TOC_250024) 7

[CHAPTER 2 : LITERATURE SURVEY](#_TOC_250023) 8

CHAPTER 3 : SYSTEM REQUIREMENT 11

SPECIFICATION

* 1. [Hardware requirements 1](#_TOC_250022)1
  2. [Software Requirements 1](#_TOC_250019)1

3.3 [Software Description 1](#_TOC_250017)2

[CHAPTER 4 : SYSTEM DESIGN 2](#_TOC_250016)6

* 1. [Architectural Design](#_TOC_250015) 16
  2. [Methodology](#_TOC_250014) 18
  3. [Flowchart 2](#_TOC_250013)3

4.4Algorithm 23

4.5Existing system 25

4.6 Proposed system 25

CHAPTER 5 : IMPLEMENTATION & DEVELOPMENT 27

* 1. Data set 27
  2. [Data set extraction and transformation](#_TOC_250012) 27
  3. Evaluation 28
  4. [Data set features](#_TOC_250011) 29
  5. [Research significance](#_TOC_250010) 30
  6. Applications 31

[CHAPTER 6 : OUTPUT](#_TOC_250004) 32

[CHAPTER 7 : CONCLUSION](#_TOC_250003) 40

7.1 References 42

# CHAPTER 1 : INTRODUCTION

# 1.1 OVERVIEW

# Specific preparation plays a crucial part in your life. Thus education preparation students often have multiple questions about universities which they can get admission and scholarship and accommodation. One of the main concerns is getting admitted to their dream university. It's seen that students still choose to obtain their education from universities that are known internationally. And when it comes to international graduates, the United States of America is the first preference of the majority of them. With most world-renowned colleges, Wide variety of courses available in each discipline, highly accredited education and teaching programs, student scholarships, are available for international students.

# According to estimates, there are more than 10 million international students enrolled in over 4200 universities and colleges including both private and public across the United States Most number of students studying in America are from Asian countries like India, Pakistan, Srilanka, Japan and China. They are choosing not only America but also UK, Germany, Italy, Australia and Canada. The number of people pursuing higher studies in these countries are rapidly increasing.

# The background reason for the students going to abroad universities for Masters is the no. of job opportunities present are low and number of people for those jobs are very high in their respective countries. This inspires many students in their profession to pursue postgraduate studies. It is seen that there is quite a large number of students from universities in the USA pursuing Masters in the field of computer science, the emphasis of this research will be on these students.

# Many colleges in the U.S. follow similar requirements for student admission. Colleges take different factors into account, such as the ranking on aptitude assessment and academic record review. The command over the English

# 

# language is calculated on the basis of their performance in the English skills test, such as TOEFL and IELTS. The admission committee of universities takes the decision to approve or reject a specific candidate on the basis of the overall profile of the applicant application.

# The majority of international students studying in the USA are from India and China. In the past decade, India has seen a huge increase in the number of students opting to pursue their education from foreign universities in countries like The USA, Ireland, Australia, Germany, etc. Although there are significant universities and colleges in India, students are finding it difficult to get admission in the highly ranked colleges and also getting a job is a challenge as the ratio of number students to the number work opportunities available is quite high.

# India is one of the leading counties in the number of software engineers produced each year; it becomes tough for the students to find jobs in elite companies due to high competition. This motivates a good number of students to pursue post-graduation in their field. It is seen that the number of students pursuing Masters in Computer Science field from universities in the USA is quite high; the focus of this research will be on these students.

**OBJECTIVES**

The idea behind creating this project is to build a machine learning model that could detect emotions from the speech humans have with each other all the time. Nowadays personalization is something that is needed in all the things that is experienced everyday. So why not have an emotion detector that will gauge human emotions and in the future recommend different things based on human mood. This can be used by multiple industries to offer different services like marketing company suggesting customers to buy products based on human emotions, automotive industry can detect the person’s emotions and adjust the speed of autonomous cars as required to avoid any collisions etc.

The primary objective of this research is to develop a system to solve the problems the international students are facing while applying for universities in the many countries. We will be developing a Student Admission Predictor (SAP) system which will help the students to predict the chances of their application being selected for a particular university for which they wish to apply based on their profile. Also, the system will provide a recommendation of universities to the student to which the student has a high possibility of getting admission. Multiple machine learning classification algorithms were evaluated to develop the system

• To understand the basic concept of Machine learning graduation admission model and its applications.

• To design an efficient prediction model which will be able to support different types of modules.

• To increase the accuracy to predict chances of admission in university.

• To design and implement a model in order to achieve an efficiency.

• To develop a model which will be precise in its operations.

# 

# CHAPTER 2 : LITERATURE SURVEY

In this paper the author [1] had described that Predicting students’ academic performance has been a research area of interest in recent years, with many institutions focusing on improving the students’ performance and the education quality. The analysis and prediction of students’ performance can be achieved using various data mining techniques. Moreover, such techniques allow instructors to determine possible factors that may affect the students’ final marks. To that end, this work analyzes two different undergraduate datasets at two different universities. Furthermore, this work aims to predict the students’ performance at two stages of course delivery (20% and 50% respectively).

This analysis allows for properly choosing the appropriate machine learning algorithms to use as well as optimize the algorithms’ parameters. Furthermore, this work adopts a systematic multi-split approach based on Gini index and p-value. This is done by optimizing a suitable bagging ensemble learner that is built from any combination of six potential base machine learning algorithms. It is shown through experimental results that the posited bagging ensemble models achieve high accuracy for the target group for both dataset

In this paper the author [2] had described that A plethora of research has been done in the past focusing on predicting student’s performance in order to support their development. Many institutions are focused on improving the performance and the education quality; and this can be achieved by utilizing data mining techniques to analyze and predict students’ performance and to determine possible factors that may affect their final marks.

To address this issue, this work starts by thoroughly exploring and analyzing two different datasets at two separate stages of course delivery (20% and 50% respectively) using multiple graphical, statistical, and quantitative techniques. The feature analysis provides insights into the nature of the different features considered and helps in the choice of the machine learning algorithms and their parameters.

Furthermore, this work proposes a systematic approach based on Gini index and p-value to

select a suitable ensemble learner from a combination of six potential machine learning algorithms

In this paper the author [3] had described that Regression analysis is a statistical technique for estimating the relationship among variables which have reason and result relation. Main focus of univariate regression is analyzing the relationship between a dependent variable and one independent variable and formulates the linear relation equation between dependent and independent variable.

Regression models with one dependent variable and more than one independent variable are called multilinear regression. In this study, data for multilinear regression analysis is occur from Sakarya University Education Faculty student's lesson (measurement and evaluation, educational psychology, program development, counseling and instructional techniques) scores and their 2012- KPSS score. Assumptions of multilinear regression analysis- normality, linearity, no extreme values- and missing value analysis were examined.

The data that verify the assumptions were analyzed with multiple regression and lessons measurement and evaluation, instructional techniques, counseling, program development and educational psychology were estimated

In this paper the author [4] had described that Predictive modelling has found its place in this century for providing an in-depth view and in helping humans in their day-to-day activity. In this paper, I have analyzed and predicted the possibility of a person getting an admit for graduate courses in the United States based on a supervised machine learning algorithm using Python and its various libraries on a dataset.

After implementing immense research on the dataset, explored the relationship between each factor which contribute in one or the other way to get an admit. Finally, using linear regression, allowed the program to predict the data from the use

In this paper the author [5] had described that Machine learning has been leveraged to increase the effectiveness of intrusion detection systems (IDSs). The focus of this approach, however, has largely be on detecting known attack patterns based on outdated datasets.

In this paper, we propose an ensemble feature selection method along with an anomaly detection method that combines unsupervised and supervised machine learning techniques to classify network traffic to identify previously unseen attack patterns. To that end, three different feature selection techniques are used as part of an ensemble model that selects 8 common features.

Moreover, k- Means clustering is used to first partition the training instances into k clusters using the Manhattan distance. A classification model is then built based on the resulting clusters, which represent a density region of normal or anomaly instances.

This in turn helps determine the effectiveness of the clustering in detecting unknown attack patterns within the data. The performance of our classifier is evaluated using the Kyoto dataset, which was collected between 2006 and 2015

# CHAPTER 3 : SYSTEM REQUIREMENT SPECIFICATION

### 3.1 HARDWARE REQUIREMENTS

Hardware required:

 Processors: Intel Atom® processor or Intel® Core™ i3 processor

 Disk space: 1 GB or more

 Ram: 4GB or more

 Operating systems: Windows 7 or later, macOS, and Linux

**3.2 SOFTWARE REQUIREMENTS**

* + - * Python versions: 3.7
      * Numpy
      * Tensorflow-gpu 1.8
      * Keras
      * Pandas
      * matplotlib

**3.3 SOFTWARE DESCRIPTION**:

 The release you are looking at is **Python 3.7.0**, the initial **feature release** for the legacy **3.7** series which is now in the **security fix** phase of its life cycle.

Among the major new features in Python 3.7 are:

* [PEP 539](http://www.python.org/dev/peps/pep-0539), new C API for thread-local storage
* [PEP 545](http://www.python.org/dev/peps/pep-0545), Python documentation translations
* New documentation translations: [Japanese](https://docs.python.org/ja/), [French](https://docs.python.org/fr/), and [Korean](https://docs.python.org/ko/).
* [PEP 552](http://www.python.org/dev/peps/pep-0552), Deterministic pyc files
* [PEP 553](http://www.python.org/dev/peps/pep-0553), Built-in breakpoint()
* [PEP 557](http://www.python.org/dev/peps/pep-0557), Data Classes
* [PEP 560](http://www.python.org/dev/peps/pep-0560), Core support for typing module and generic types
* [PEP 562](http://www.python.org/dev/peps/pep-0562), Customization of access to module attributes
* [PEP 563](http://www.python.org/dev/peps/pep-0563), Postponed evaluation of annotations

**NUMPY**

NumPy 1.22.0 is a big release featuring the work of 153 contributors spread over 609 pull requests. There have been many improvements, highlights are:

* Annotations of the main namespace are essentially complete. Upstream is a moving target, so there will likely be further improvements, but the major work is done. This is probably the most user visible enhancement in this release.
* A preliminary version of the proposed Array-API is provided. This is a step in creating a standard collection of functions that can be used across applications such as CuPy and JAX.
* NumPy now has a DLPack backend. DLPack provides a common interchange format for array (tensor) data.
  + - New methods for quantile, percentile, and related functions. The new methods provide a complete set of the methods commonly found in the literature.
* The universal functions have been refactored to implement most of [NEP 43](https://numpy.org/neps/nep-0043-extensible-ufuncs.html#nep43). This also unlocks the ability to experiment with the future DType API.
* A new configurable allocator for use by downstream projects.

**PANDAS**

**pandas** is a Python package that provides fast, flexible, and expressive data structures designed to make working with "relational" or "labeled" data both easy and intuitive. It aims to be the fundamental high-level building block for doing practical, **real world** data analysis in Python. Additionally, it has the broader goal of becoming **the most powerful and flexible open source data analysis / manipulation tool available in any language**. It is already well on its way towards this goal.

Here are just a few of the things that pandas does well:

* Easy handling of [**missing data**](https://pandas.pydata.org/pandas-docs/stable/user_guide/missing_data.html) (represented as NaN, NA, or NaT) in floating point as well as non-floating point data
* Size mutability: columns can be [**inserted and deleted**](https://pandas.pydata.org/pandas-docs/stable/user_guide/dsintro.html#column-selection-addition-deletion) from DataFrame and higher dimensional objects
* Automatic and explicit [**data alignment**](https://pandas.pydata.org/pandas-docs/stable/user_guide/dsintro.html?highlight=alignment#intro-to-data-structures): objects can be explicitly aligned to a set of labels, or the user can simply ignore the labels and let Series, Data Frame, etc. automatically align the data for you in computations
* Powerful, flexible [**group by**](https://pandas.pydata.org/pandas-docs/stable/user_guide/groupby.html#group-by-split-apply-combine) functionality to perform split-apply-combine operations on data sets, for both aggregating and transforming data
  + - * Make it [**easy to convert**](https://pandas.pydata.org/pandas-docs/stable/user_guide/dsintro.html#dataframe) ragged, differently-indexed data in other Python and NumPy data structures into Data Frame objects
* Intelligent label-based [**slicing**](https://pandas.pydata.org/pandas-docs/stable/user_guide/indexing.html#slicing-ranges), [**fancy indexing**](https://pandas.pydata.org/pandas-docs/stable/user_guide/advanced.html#advanced), and [**subsetting**](https://pandas.pydata.org/pandas-docs/stable/user_guide/indexing.html#boolean-indexing) of large data sets
* Intuitive [**merging**](https://pandas.pydata.org/pandas-docs/stable/user_guide/merging.html#database-style-dataframe-or-named-series-joining-merging) and [**joining**](https://pandas.pydata.org/pandas-docs/stable/user_guide/merging.html#joining-on-index) data sets
* Flexible [**reshaping**](https://pandas.pydata.org/pandas-docs/stable/user_guide/reshaping.html) and [**pivoting**](https://pandas.pydata.org/pandas-docs/stable/user_guide/reshaping.html) of data sets

# Matplotlib

Matplotlib is a comprehensive library for creating static, animated, and interactive visualizations in Python. Matplotlib makes easy things easy and hard things possible.

* Create [publication quality plots](https://ieeexplore.ieee.org/document/4160265/citations?tabFilter=papers).
* Make [interactive figures](https://mybinder.org/v2/gh/matplotlib/mpl-brochure-binder/main?labpath=MatplotlibExample.ipynb) that can zoom, pan, update.
* Customize [visual style](https://matplotlib.org/stable/gallery/style_sheets/style_sheets_reference.html) and [layout](https://matplotlib.org/stable/tutorials/provisional/mosaic.html).
* Export to [many file formats](https://matplotlib.org/stable/api/figure_api.html#matplotlib.figure.Figure.savefig) .
* Embed in [JupyterLab and Graphical User Interfaces](https://matplotlib.org/stable/gallery/#embedding-matplotlib-in-graphical-user-interfaces).
* Use a rich array of [third-party packages](https://matplotlib.org/mpl-third-party/) built on Matplotlib.

# Keras

**Keras** is an [open-source](https://en.wikipedia.org/wiki/Open-source_software) [software](https://en.wikipedia.org/wiki/AI_software) library that provides a [Python](https://en.wikipedia.org/wiki/Python_(programming_language)) interface for [artificial neural networks](https://en.wikipedia.org/wiki/Artificial_neural_network). Keras acts as an interface for the [TensorFlow](https://en.wikipedia.org/wiki/TensorFlow) library.

Up until version 2.3, Keras supported multiple backends, including [TensorFlow](https://en.wikipedia.org/wiki/TensorFlow), [Microsoft Cognitive Toolkit](https://en.wikipedia.org/wiki/Microsoft_Cognitive_Toolkit), [Theano](https://en.wikipedia.org/wiki/Theano_(software)),

 As of version 2.4, only [TensorFlow](https://en.wikipedia.org/wiki/TensorFlow) is supported. Designed to enable fast experimentation with [deep neural networks](https://en.wikipedia.org/wiki/Deep_learning), it focuses on being user-friendly, modular, and extensible. It was developed as part of the research effort of project ONEIROS (Open-ended Neuro-Electronic Intelligent Robot Operating System),[[5]](https://en.wikipedia.org/wiki/Keras#cite_note-5) and its primary author and maintainer is [François Chollet](https://en.wikipedia.org/wiki/Fran%C3%A7ois_Chollet), a [Google](https://en.wikipedia.org/wiki/Google) engineer. Chollet is also the author of the [Xception](https://en.wikipedia.org/wiki/Xception) deep neural network model.[[6]](https://en.wikipedia.org/wiki/Keras#cite_note-6)



# CHAPTER 4 : SYSTEM DESIGN

### 4.1 ARCHITECTURAL DESIGN

### In this section, we will describe the architecture of the Student Admission Predictor system. The figure below explains the flow of the system:

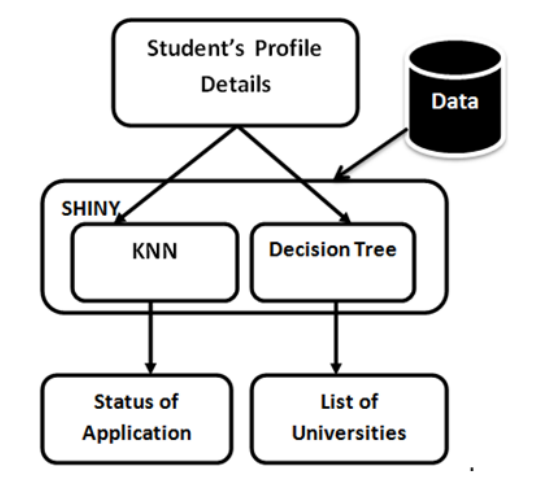
### • The student will enter his/her profile details using the user interface developed in shiny.

### • The user interface code will interact with the KNN and Decision Tree models to provide the users with the required result.

### • The KNN algorithm will be used to determine the chance of the student of securing admission in a particular university based on his/her profile.

### • The Decision Tree algorithm will be used to determine the rank of college to which is most suitable for the student based on his/her profile and provide the student with the list of universities which fall in that rank.

### • Once the models have been executed the result will be provided to the student as the output on the user interface



#### 4.2 METHEDOLOGY

Problem Understanding: Initially first we have to spend some time on what are the problems or concerns students having during their pre admission period and we should set the solutions to those problems as objectives of this research.

Data Understanding: Data should be collected from multiple sources like yocket and also consider all the factors including which will play a tiny role in student admission process.

Data Preparation: Data should be cleaned that is removing the noise in the data and filling the missing values or extreme values and finalising the attributes/factors which will have crucial importance in student admission process

Building Models: several ML models have to be developed using various machine learning algorithms for admission to a particular university and the user interface has to be developed to access those models

Evaluation: Developed models are evaluated according to their accuracy scores. Once the model is finalized that model will be merged with node red for final deployment

Data Cleaning and Analysis: • Inspecting feature values that help identify what needs to be done to clean or pre-process until you see the range or distribution of values typical of each attribute.

• You may find missing or noisy data, or anomalies such as the incorrect data form used for a column, incorrect measuring units for a particular column[7], or that there are not enough examples of a specific class.

• You can know that without machine learning, the problem is actually solvable.

The data cleaning process has several key benefits to it:

1. This eliminates major errors and inconsistencies which are unavoidable when dragging multiple data sources into one dataset.

2. Having data cleaning software will make everyone more effective as they will be able to get easily from the data what they need.

3. Fewer mistakes mean happy clients, and less unhappy workers.

4. The ability to chart the various functions, and what your data is supposed to do and where it comes from your data.

Data Visualization:

• After analying the data, we will be able to know what the features and labels are, so from the above data, the label we have to consider is Chance of Admission[8] and then we have to consider the parameters that influence or play a major role in Chance of Admission

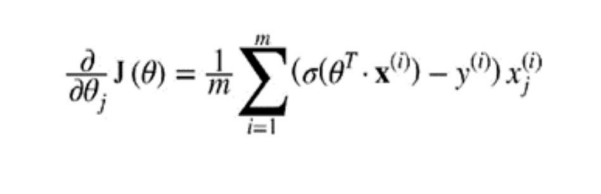
• We can get to know certain features that are more affected by the visualization (or) analysis or the use of feature importance method in decision tree

• Once the data visualization is done, we have to do predictive modelling for this purpose first we divide the data into train part and test part. • we will develop model using machine learning algorithms on the train data and test

model accuracy on the test data part.

• we will see which algorithms giving highest accuracy according to what parameters and take that for final consideration.

Linear Regression Model Linear Regression is a supervised learning machine learning algorithm and one of the most well-known algorithms in machine learning and statistics. Linear Regression is an attractive model for researchers because its representation is simple, and it works well for many problems. Learning algorithms are used to estimate the coefficients of the LR model. The objective of Linear regression model is to figure out the relationship between two variables by fitting a linear model to the training data. It is a predictive algorithm that provides a Linear relationship between Prediction (Call it ‘Y’) and Input (Call is ‘X’). The simplest form of the regression model with one feature and one target variable is defined by the formula

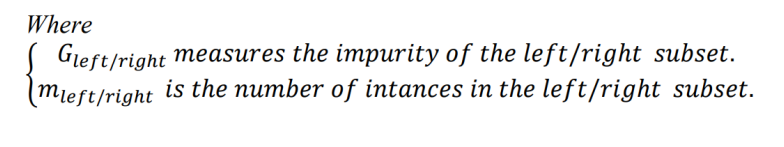


y = c + b\*x, Where y is the target variable value, c is a y-intercept, b is the slope , and x is the value of the feature variable [5]. To train a Linear Regression model, you need to find the value of θ that minimizes the RMSE by the Equation MSE cost function for a Linear Regression model: MSE (X, hθ) = 𝟏𝟏 𝒎𝒎 ∑ ( 𝒎𝒎 𝒊𝒊=𝟏𝟏 θT .x(i) – y(i) )2 Where hθ =is the hypothesis function using the model parameters θ, m= number of samples in dataset, θT = is the transpose of θ , x = is the instance’s feature vector , θT.x(i) = is the dot product of θT and x(i) , and y = expected value.

Decision Tree Model

A decision tree is a machine learning model that uses a tree-like graph of decisions. Decision trees are widely used algorithms in data mining and machine learning because of their simplicity and the ease of their interpretation. Decision Tree algorithms can model non-linear relationships. The cost function that the algorithm tries to minimize is given by the following formula:

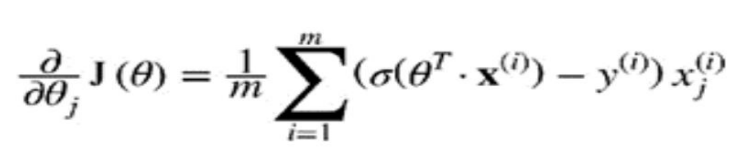




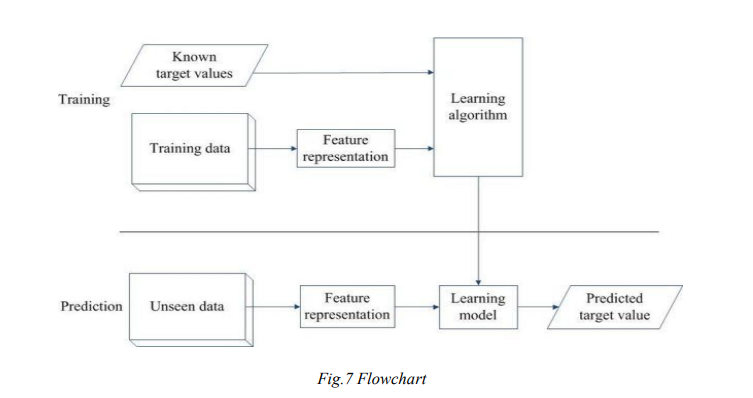
The idea is really quite simple: the algorithm first splits the training set in two subsets using a single feature k and a threshold tk (e.g., petal length ≤ 2.45 cm). It searches for the pair (k, tk) that produces the purest subsets (weighted by their size) [6]. To train a Linear Regression model, you need to find the value of θ that minimizes the RMSE by the Equation MSE cost function for a Linear Regression model: MSE (X, hθ) = 𝟏𝟏 𝒎𝒎 ∑ ( 𝒎𝒎 𝒊𝒊=𝟏𝟏 θT .x(i) – y(i) )2 Where hθ =is the hypothesis function using the model parameters θ, m= number of samples in dataset, θT = is the transpose of θ , x = is the instance’s feature vector , θT.x(i) = is the dot product of θT and x(i) , and y = expected value.

Logistic Regression Model

Logistic Regression (also called Logit Regression) is another technique originates from the field of statistics. It is widely used for binary classification problems (problems with two class values) and commonly used to estimate the probability that a data sample belongs to a particular class. Below is an example logistic regression formula: y = e^ (b0 + b1\*x) / (1 + e^ (b0 + b1\*x)) Logistic regression is often used with regularization techniques to prevent overfitting. To use MSE with the partial derivatives of the cost function with regards to the jth model parameter θj is given by Equation :



**4.3 FLOW CHART**



**4.4 ALGORITHM**

For this work, several machine learning algorithms have been used, K- Nearest Neighbor and Linear Regression, Ridge Regression, Random Forest are used to predict students ' likelihood of university admission

K-Nearest Neighbors:

KNN algorithm is the most commonly used algorithm for classification and regression purpose. KNN stands for k nearest neighbor, here k indicates a integer value which will tell that with how manyneighbors comparisons should be made.

It can be used for both classification and regression purpose. Suppose if it is classification and the k value is 5 it will compare with nearest 5 neighbors and gives the mode value, if it is regression and the k value is 6 it will take the nearest six values and return its mean value.

**Linear Regression:**

It is an algorithm based on supervised learning of computers. It does the role of regression. Regression models a predictive goal value based on the independent variables. Mostly it is used to figure out the relation between variables and forecasting. Different regression models vary on the basis–the form of relationship between dependent and independent variables, are considered, and the number of independent variables used.

**Random Forest**:

Random Forest is a machine learning algorithm which is a combined effect of classification and regression and other tasks which operate by erection of decision trees at training time and outputs the class that is the mode of the classes or mean value of individual trees

**Regression**:

Ridge regression is a regression method that is quite similar to unadorned minus squares linear regression: simply adding a 22 penalty on parameters β to the linear regression objective function gives the ridge regression objective function. Ridge regression is an example of a shrinkage method: it shrinks the parameter estimates in the hopes of reducing uncertainty, increasing prediction accuracy, and aiding interpretation relative

**4.5 EXISTING SYSTEM**

* + - * A great number of researches and studies have been done on graduation admission datasets using different types of machine learning algorithms.
      * Bayesian Networks were used to create a decision support system for evaluating the application submitted by international students in the university.
      * This model was designed to predict the performance of the aspiring students by comparing them with the performance of students currently studying in the university and had similar profile during their application.
      * In this way based on the current students profile the model predicted whether the aspiring student should be granted admission to the university. Since the comparisons were made only with the students who were already admitted in the university and the data of the students who were denied admission were not included in the research this model proved to be less efficient due to the problem of class imbalance
    - Emotion recognition systems based on digitized speech is comprised of three fundamental components: signal preprocessing, feature extraction, and classification.
    - Acoustic preprocessing such as denoising, as well as segmentation, is carried out to determine meaningful units of the signal. Feature extraction is utilized to identify the relevant features available in the signal.
    - Lastly, the mapping of extracted feature vectors to relevant emotions is carried out by classifiers. In this section, a detailed discussion of speech signal processing, feature extraction, and classification is provided.

**4.6 Proposed System**

This system will include the following features, and based on these features

the model predicts the result.

• GRE(out of 340): The graduate record examination (GRE) is a standardized exam for measuring student's aptitude for abstract thinking. The score will be out of 340 points.

• Test of English as a Foreigner Language (TOEFL) score, which will be out of 120 points.

• University Ranking that indicates the ranking of higher education Institutes. The score will be out of 5 points.

• Letter of Recommendation Strength (LOR) : from a professional who has taught a student, out of 5 points.

• Undergraduate GPA (CGPA): CGPA is the average grades obtained by a student in all the semesters out of 10 points.

• Statement of purpose (SOP): A Statement of Purpose is an essay of purpose of applying to a specific course in a particular university, out of 5 points.

• Research Experience: "Research Experience" means any academic research activity.

• Chance of Admit(ranging from 0 to 1): Is the likelihood of a student's chance to be admitted to a university. 1 means great possibility, 0 means no chance of admission

# CHAPTER 5 : IMPLEMENTATION AND DEVELOPMENT

# The model was implemented as per the requirement of predicting the admission of students. The algorithm was developed with the use of programming languages and then the algorithm was tested against standard set of values and obtained the results by passing the different data as input. The model was a learning model which was then trained accordingly to perform the functions as instructed by the algorithms.

**5.1 Data-set**

This section describes, in brief, the data that has been used for the research. Data from multiple sources was used in this project, the major amount of data was extracted from public website Yocket(Yocket), data regarding the rankings, fees and enrolment in colleges was obtained from a leading educational consultancy firm The Mentors Circle in India. Data from both the sources was integrated together to form a staging data-set. For predicting the chance of a student getting shortlisted in universities the final data-set was divided into multiple data-sets each representing a particular university. For predicting the list of universities suitable for students based on their profile data of all the students the staging data-set was updated only to have records of students who had successfully secured admission in the universities. Below table shows the different features of the data-sets.

**5.2 Data-set Extraction and Transformation**

Data related to the college ranking was collected in .csv format, the data related

to students profile was extracted from (Yocket (2017)) using data extraction tool provided by (Mozenda (n.d.))in .csv files. Data being from public portal had multiple records with missing and irrelevant values; data cleaning was performed in Microsoft Excel by deleting the records having unwanted and missing values. Unwanted columns were removed from the data-set. Once the data-set was cleaned data was transformed to be suitable for the model. The original data-set had TOEFL or IELTS score as a representation of language, to have a consistent metrics for the language score of all the records were converted

to IELTS scale using the conversion table. (Prescholar (2017)). Similarly, the UGPA score of the students was represented in terms of percentage and CGPA; all the records of percentage were converted to CGPA by multiplying percentage score by 9.5. The values of Intake, Status and Rank fields were changed as shown below to have numerical values for the data to achieve better results for KNN

**5.3 Evaluation**

K-nearest neighbour and Multivariate logistic regression algorithms were used to create a model that can be used to predict the likelihood of success of a students application to the university based on his/her profile. Both algorithms were tested and their performance was evaluated based on different factors like Accuracy, Sensitivity, Specificity and Kappa value. As can be seen the figure given below model created using K-Nearest Neighbour outperformed the model created using Logistic Regression on all the performance measures. Also by looking at the variance in the values of the data KNN seemed to be the best-fit algorithm to create the Student Admission Predictor System.

Accuracy was considered to be main metric to evaluate the performance of the models, as the data used for creating the models was balanced. Also, prediction of the true positive and true negative scenarios was equally equivalent. The KNN model performed well with an overall average accuracy of 76%.

The decision tree model which was created to predict the rank of the universities suitable for the student provided the result with an accuracy of 80%.

**5.4 DATA SET FEATURES:**

1. GRE Score (Out of 340)

2. TOFEL Score (Out of 120)

3. University Rating (Out of 5)

4. Statement Of Purpose and Letter of Recommendation Strength (Out of 5)

5. Undergraduate GPA (Out of 10)

6. Research Experience (Either 0 or 1)

7. Chance Of Admission (Ranging from 0 to 100)

**Process**

1.Problem Understanding

2.Data Understanding

3.Data Preparation

4.Building Models

5.Evaluation

**Training:**

Initially in the training process(fig.7),

1. The Training data is given as input.

2. The representation of features and functionality are defined

3. The algorithm is developed as per the requirement using some programming platforms.

4. The standard or target values are also given so that the algorithm can predict admission of

5. Students against the standard value

**PREDICTION:**

The prediction is done after training

1. The unseen data or some random data is given as input

2. The representation of features and functionality are already defined

3. The learning model is inserted with the algorithm so as to function in the same way.

4. The obtained results are analyzed against the standard values or targets

**5.5 RESEARCH SIGNIFICANCE**

The principal objective of the research is to help the students who are aspiring to pursue their education in the other countries. The SAP system will help them to evaluate the chances of success in a particular university without being dependent on any education consultancy firm. It will help them in saving a huge amount of time and money spent in the application process. Also, it will help them to limit the number of applications made by the students by suggesting them the best universities where they have high chances of securing admission thereby by saving the amount of money spent by the students by applying in universities where they have less chance to secure admit based on their profile

**5.6 APPLICATION**

Student admission problem is very important in educational institutions. This project addresses machine learning models to predict the chance of a student to be admitted. This will assist students to know in advance if they have a chance to get accepted.

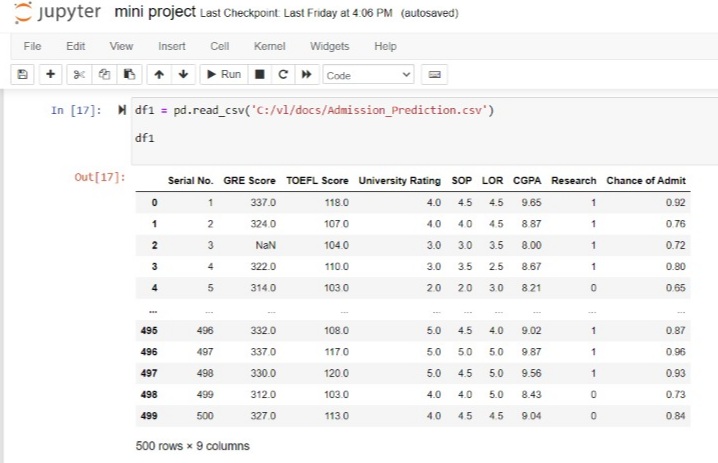
Machine learning models were performed to predict the opportunity of a student to get admitted to a master’s program. The machine learning models included are multiple linear regression, random forest, Multiple Linear Regression with Backward Elimination and random forest regression with backward elimination.

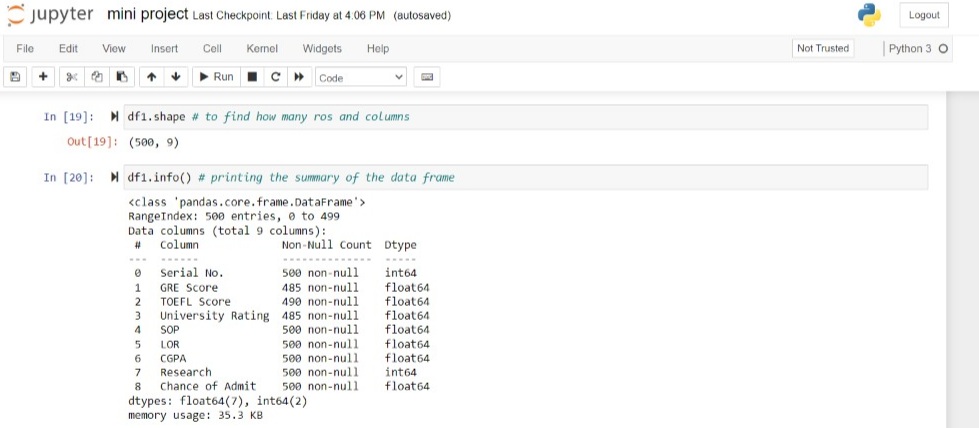
Experiments show that the Linear Regression model surpasses other models. Our aim would be to predict the “Chance of Admit” based on the different parameters that are provided in the dataset.

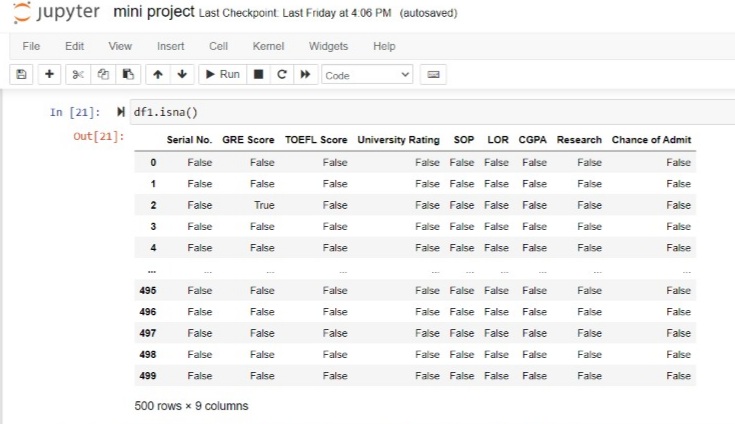
We will achieve this aim by using the Linear Regression model. Based on the data that we have, we will split out data into training and testing sets. The Training set will have features and labels on which our model would be trained. The label here is the “Chance of Admit”. Basically label is the output that we want and features are the parameters that drive us towards the output.

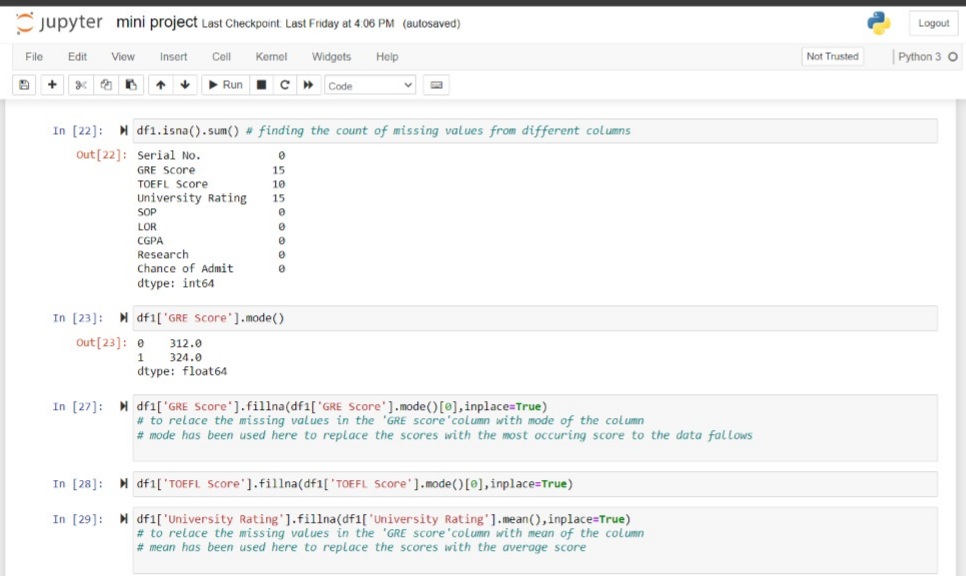
Once our model is trained, we will use the trained model and run it on the test set and predict the output. Then we will compare the predicted results with the actual results that we have to see how our model performs.

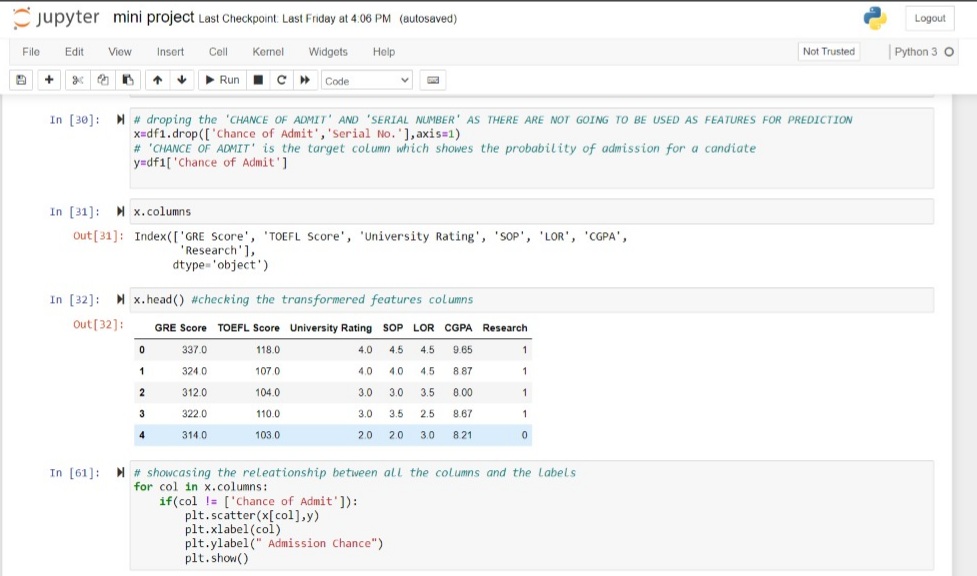
** CHAPTER 6: OUTPUT**

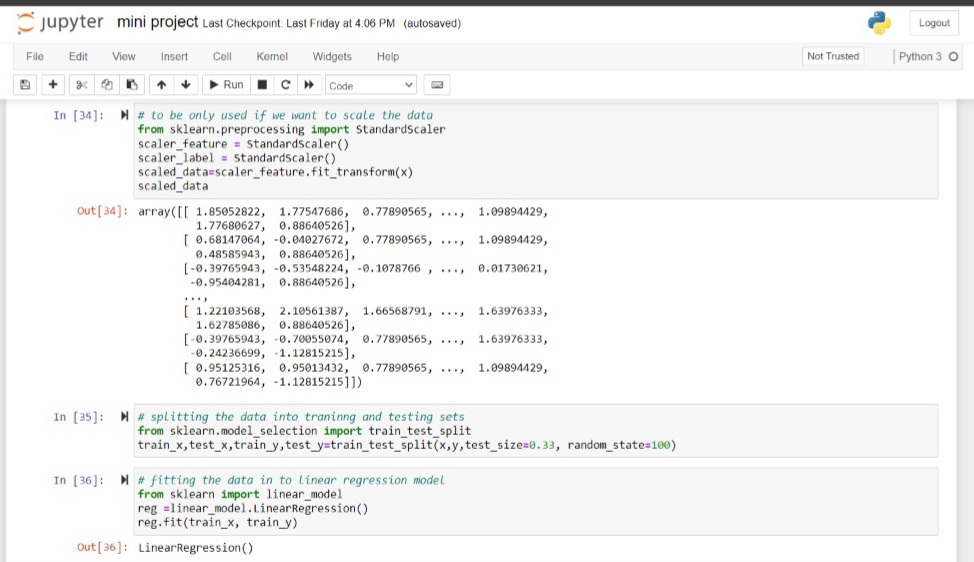


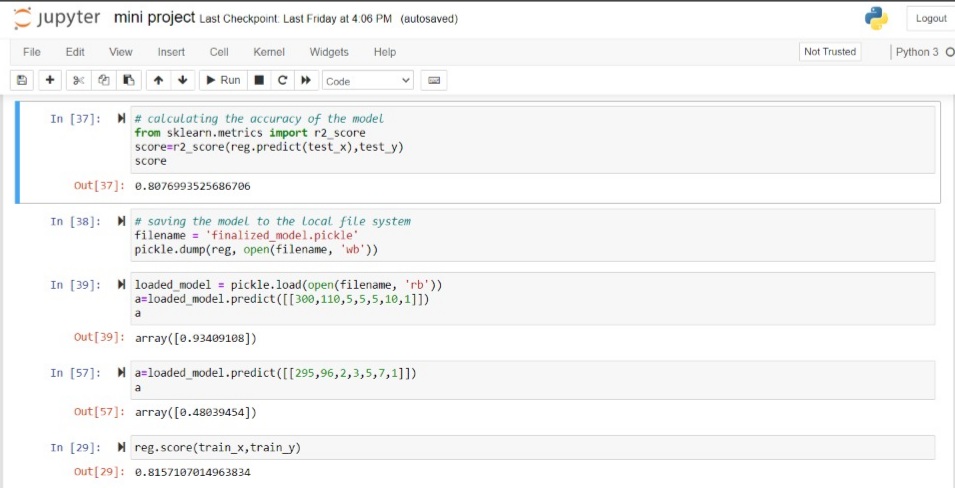












**CHAPTER 7: CONCLUSION**

The main goal of this work is to create a Machine Learning model which could be used by students who want to pursue their education in the US. Many machine learning algorithms were utilized for this research. Linear Regression model compared to other ones. Students can use the model to assess their chances of getting admission into a particular university. with an average accuracy of 79 percent.

A GUI was developed to make the program, from a non technical perspective, usable and user-friendly. Using node-red the user interface was developed. The ultimate goal of research will be accomplished successfully, as the system allows students to save the lot of time and money that they would spend on educational mentors and application fees for colleges where they have less chances of getting admissions.

The main limitation of this research is we developed models based solely on data from Indian Students studying Masters in Computer Science in the United States, we considered only few universities with different rankings. More information relating to new colleges and courses can be added to the curriculum in the future. The system may also be modified toa web-based application by making node-red modifications.

To solve the problem, it is possible to test other classification algorithms if they have high accuracy score than the current algorithm, the framework can be easily modified to support the new algorithm by changing the server code in the Node Red. Finally, students can have an open-source machine Learning model which will help the students to know their chance of admission into a particular university with high accuracy.

Multiple machine learning algorithms were developed and used for this research. KNN proved to best-fit for development of the system when compared with the Logistic regression model.

The model can be used by the students for evaluating their chances of getting shortlisted in a particular university with an average accuracy of 75%. Decision Tree algorithm was used to predict the universities which were best suitable for a student based on their profile. The decision tree algorithm proved to be 80% accurate.

A simple user interface was developed to make the application interactive and easy to use for the users from the non-technical background.

the limitation of the research we have created the models based only on the data of Indian Students studying Masters in Computer Science in the USA, we have considered only ten universities with different rankings.

In future, more data related to additional universities and courses can be added to the system. A great amount of variance in the criterion variables remains unexplained by the predictors in this study.

There may be room and a need to conduct more research to study the unexplained portion of variance in graduate performance. In this study, graduate performance was examined only in the form of GGPA.

As suggested by Kuncel et al. (2010), multiple aspects of student performance should be considered so as to have a more comprehensive picture about students’ performance.

The criteria may include the information like faculty ratings, degree attainment, degree completion, and research productivity.

Moreover, as to the admission criteria, committees should admit students based on not only cognitive abilities but also some noncognitive characteristics of the applications, such as motivation, interest, personality, and some other characteristics.

**7.1 REFERENCES**

[1] "Admissions Requirements," Berkeley Graduate Division, [Online]. Available: https://grad.berkeley.edu/admissions/requirements/. [Accessed 09 november 2019].

[2] I. Hatzilygeroudis, "PASS: An Expert System with Certainty Factors for Predicting Student Success," ResearchGate.https://www.researchgate.net/publication/221017874\_PASS\_An\_Expert\_Syste m\_with\_Certainty\_Factors\_for\_Predicting\_ Student\_Success.

[3] E. Roberts, "using machine learning and predictive modeling to assess admission policies and standards,"http://sites.tntech.edu/weberle/wpcontent/uploads/sites/87/2018/06/NSSR\_2013- 1.pdf.

[4] M. M. P. K. P. S. J. Heena Sabnani, "Prediction of Student Enrolment Using Data Mining Techniques," <https://www.irjet.net/archives/V5/i4/IRJET-V5I4408.pdf>.

[5] J. Brownlee, "Linear Regression for Machine Learning," <https://machinelearningmastery.com/linear-regression-formachine-learning/>.

[6] A. Géron, Hands on Machine Learning with Scikit Learn and Tensorflow, O'Reilly Media, 2017.

[7] R. S. Brid, "Decision Trees," Greyatom, https://medium.com/greyatom/decisiontrees-asimple-way-to-visualize-a-decision-dc506a403aeb.

[8] mohansacharya,"Graduate Admissions 2," kaggle, Available: <https://www.kaggle.com/mohansacharya/graduate-admissions>.

[9] "Data Correlation can make or break your Machine Learning Project," Medium, 14 Oct 2018. [Online]. Available: https://towardsdatascience.com/data-correlation-can-makeorbreak-your-machine-learning-project-82ee11039cc9.