Contents

[Introduction 1](#_Toc148793993)

[Data Collection 1](#_Toc148793994)

[Data Preparation 2](#_Toc148793995)

[Prediction Algorithm 3](#_Toc148793996)

[Deployment 6](#_Toc148793997)

[Conclusion and Recommendations 6](#_Toc148793998)

[Figure 1Sample Admission Data 2017/2018 2](#_Toc148794021)

[Figure 2 Cleaned data sample 2](#_Toc148794022)

[Figure 3 Modules Import 3](#_Toc148794023)

# Introduction

The ability to predict accurately the number of admissions an institution will receive is important for them to allocate resources correctly. This prediction cannot be done accurately by the use of traditional algorithms and requires the use of some new technologies. Machine Learning is one such technology that can be employed in order to create fully functional predictive models that are able to determine with lots of precision numbers of expected entries each year.

This form of prediction algorithm rely on data that has to be gathered, prepared and used to train the intelligent models that will then be used to determine with high accuracy the likely the numbers likely to be admitted to a certain program in the university.

The models rely on certain patterns that can be derived based on the year, programme cost and in extension the programme cost, cutoff points, total KCSE performance entries and the university rankings.

With more data points the accuracy of the predictions can be increased. In this demonstrational model due to limited data available the programme code and year as well as distribution between placement gender wise will be used. For the interface to allow for interaction between the users and the software streamlit was employed.

The entire program was build using python a language known for its suburb capabilities for use in ML and AI using packages such as Sklearn, Numpy for data cleaning and matplotlib for data visualization

# Data Collection

For the program to be able to predict the admission figures it had to be fed with data that was not pre-existing. Data for use with it was scarce and hard to find due to the lack of data banks dedicated to education institutions. Moi university Programme records were available from 2015 to 2022. This data showed the Programme Name, Year, Course Name and the Course Code.

This data was available in separate excel sheets files and a sample is as below.

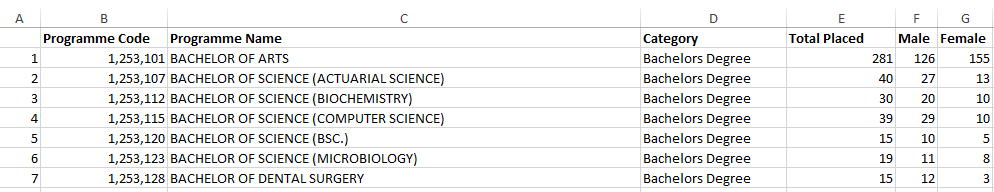


Figure 1Sample Admission Data 2017/2018

The data was distributed in different excel sheets but was standardized thus uniform in nature. It however could not be used in the original format and therefore required some preparation. The process of preparation is explained in the next chapter.

# Data Preparation

In order to use the data for training the model it had to be first consolidated into a single comma separated values (csv) sheet. This CSV sheet is what would then be used to train the model that is then used to predict the admission figures, During the preparation process excel was used to merge the different sheets into a single sheet. Then Programme Name field was removed to allow for uniform data and the Year Field was added. This is because the year is an important variable that influences the admission figures. The data head and first 5 rows is as follows.

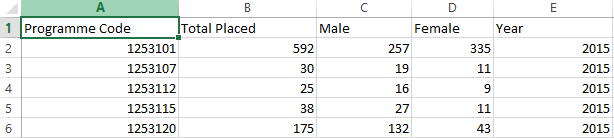
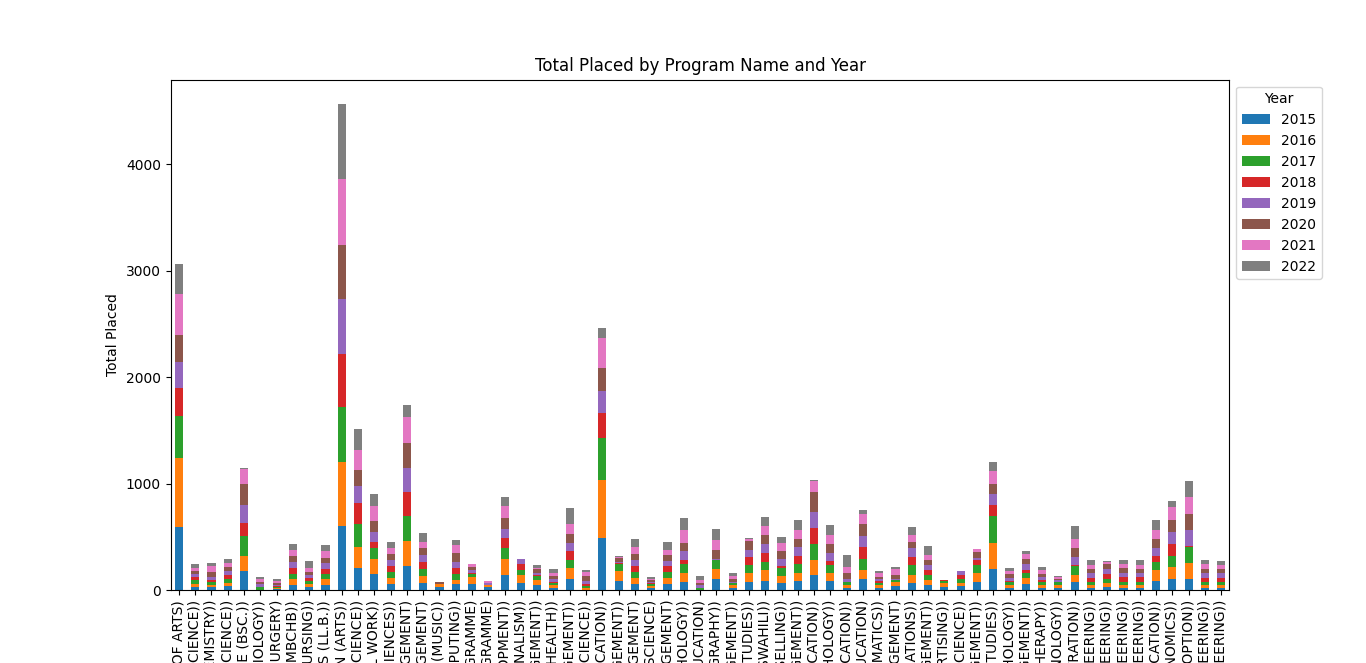


Figure 2 Cleaned data sample

After the cleaning process over 500 data-points were obtained this data points ranged over 23 different courses and 8 years. This data was later used to train the algorithms.

The data is visualized below using the matplotlib library.



The code used to generate the file is available with the visualization.py file.

# Prediction Algorithm

The prediction contains different aspects including, model training, user training and actual prediction. To allow for a comprehensive understanding of the code each block will be explained in detail.

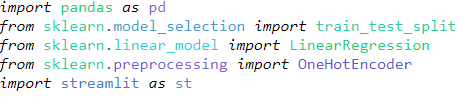


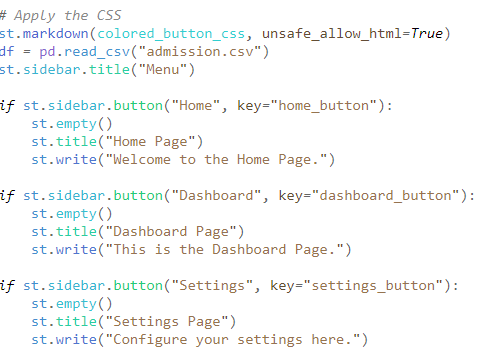
Figure 3 Modules Import

The first cell of code is for importing the necessary modules pandas is for data manipulation, sklearn is what is used for data training. Specifically, the model we will be using is Linear Regression which works by attaching weights to each field in our case the Year and Programme code.

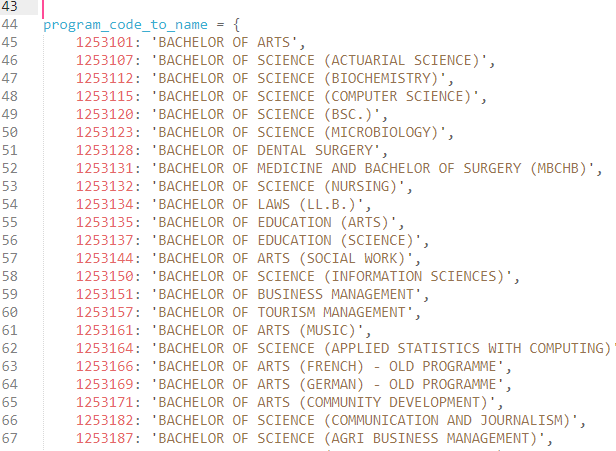
Streamlit is what is used to build the interface the user will use to interact with the platform



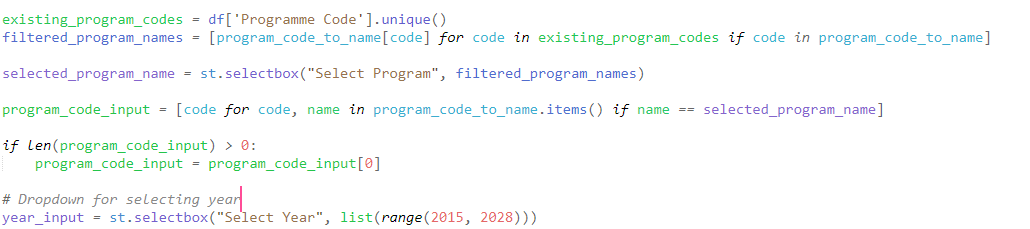
Some styling is done to customize how the buttons will look like and a menu is added.



The styling is used and the custom build using Streamlit is added. The data is loaded using pandas from the admissions csv file.



Line 44 to 110 involves maping the programme code to names by using a dictionary. This allows for the training to occur since algorithms are more customized to numbers but users programme names which are friendlier.



Using the earlier created dictionary the interface is created. The interface is a select button and uses the programme names.

The year interface is also created another select input field from the streamlit library.



Line 123 to 146 is the key part of the code. This listens to the click on the execute button being clicked. When the button is clicked the value for programme code and the year are fetched.

Line 129, 130 uses the HotOneEncoder to separate the data to allow each of the programme codes to be trained separately. This is because each programme has a different profile and therefore has to be trained uniquely with different weights.

In Line 132 and 133 is when the actual training occurs its split 0.8 to 0.2 with 80% of the data being the training data and 20% being the test data.

For the model in use Linear Regression is used.

In line 139 to 144 result of the training is used to perform a prediction and streamlit use to output the Predicted total placement for the year.

# Deployment

The programme is deployed vie streamlit to allow for public access. This starts with creating a github account and pushing the code there. After the creation of the github repository a streamlit account is required. The project is then deployed by creating a new site on the streamlit account.

# Conclusion and Recommendations

The project has potential to be used in many education institutions to influence important decisions in regard to first year admission. However, there is still room for improvement. The most crucial improvements start with adding more data points i.e the cutoff for each programme for each, the fees for each programme, the KCSE performance and the university ranking for that year. This is because these factors affect the distribution of students across the courses.