# python packages

NumPy is a fundamental package for scientific computing with Python, providing support for large, multi-dimensional arrays and matrices, along with a wide range of mathematical functions to operate on these arrays.

Pandas is a high-performance, easy-to-use data structures and data analysis toolkit. It provides powerful data manipulation and analysis capabilities, such as indexing, reshaping, merging, and joining of data sets.

Scikit-Learn is a popular machine learning library for Python, providing implementations of a wide range of algorithms for classification, regression, clustering, and more. It also includes many useful tools for model evaluation and selection, as well as for preprocessing and feature extraction.

Matplotlib is a popular and widely used plotting library for Python. It provides a high-level interface for creating attractive and informative statistical graphics in Python. With Matplotlib, you can create line plots, scatter plots, histograms, bar charts, and many other types of plots using a simple, intuitive API. Additionally, Matplotlib integrates well with other popular Python libraries, such as Pandas and NumPy, making it easy to use with large, real-world datasets.

# ML algorithms chosen for creating model

Logistic regression is a simple and widely used algorithm that is often used for binary classification tasks. It can be used to predict whether a student is likely to be accepted or rejected based on their scores and college ranking.

Random forests are an ensemble learning method that combines multiple decision trees to make more accurate predictions. They can be used in the same way as decision trees to predict a student's likelihood of acceptance based on their scores and college ranking.

Gaussian naive Bayes is a classification algorithm that is based on the idea of using Bayes' theorem to make predictions. It assumes that the features in the data are mutually independent and follows a Gaussian distribution. This allows the algorithm to make predictions based on the probabilities of different classes given the input data.

Support vector classifiers (SVCs) are a type of algorithm that can be used for classification tasks. They work by finding the hyperplane in the feature space that maximally separates the different classes in the data. This allows the algorithm to make predictions based on the distance of new data points to the hyperplane.support vector classifier could potentially be a useful algorithm to consider. It could be trained on the data to predict a student's likelihood of acceptance based on their scores and college ranking

K-nearest neighbors (KNN) is a classification algorithm that is based on the idea of using the "k" closest data points in the feature space to make predictions for new data. It works by calculating the distances between the new data point and the "k" nearest points in the training set, and then using these distances to determine the class of the new data point. KNN classifier could potentially be a useful algorithm to consider. It could be trained on the data to predict a student's likelihood of acceptance based on their scores and college ranking

* filling missing values with most frequent values using simpile imputer
* Normalizing and standardizing the data to ensure that all features are on a similar scale
* reduce the dimensionality of the data
* renamed some column names

Table

Description automatically generated

Graphical user interface, application

Description automatically generated

Table

Description automatically generated