Table of Contents

1. Introduction

The project requires us to analyse three real datasets of our choice from an open-source domain. These datasets will be used to demonstrate how to use machine learning, a tool we use to focus on data and algorithms to enable AI to imitate the way that humans learn, gradually improving its accuracy. These algorithms are based on dataset models so we can analyse any data we train on. This project will consist of completing the following three tasks:

1. **Regression**: where the problem consists of continuous target variable(s).
2. **Unsupervised Learning**: where the problem consists of identifying homogeneous population groups or dimension reduction techniques, which can then be used in the context of the empirical application
3. **Classification**: where the problem consists of categorical target variable(s).

We have sourced 3 datasets, all from Kaggle, for each of the mentioned tasks and will be using machine learning to build the algorithm to train the model to understand each task. We can train the dataset model to predict the output variables based on the inputs given on the dataset or train the model to identify and classify patterns based on the unlabeled data given from the dataset.

2. Price of Surveillance Cameras Dataset (Regression Learning)

2.1 Substantive Issue

Security is one of the most important aspects we all consider when we tackle our lives. It has become a factor we all can relate to, whether we need to add security to our homes, offices, rooms, and even our vehicles. Being able to see our prized possessions is another issue as we would like to have the assurance to be able to view our property from anywhere we are just to know whether they are safe or not. Security cameras make up the bulk of anyone's security needs but their prices of them can vary due to brand, type of camera, or even its features. For this dataset, we will be using regression learning to estimate the prices of security cameras based on other models and data from it. This will be based on understanding the pattern between each column and finally estimating the amount. This will be a great model for homeowners who intend on purchasing a security camera and want and estimated cost.

2.2 Methodology

Regression Learning will include loading the dataset and using exploratory data analysis(EDA) to focus on the top five brand names, encoding of category variables, and reducing the dimensions. We will be using three regression models, such as Linear Regression, Random Forest Regressor, and K-Nearest Neighbors to train and evaluate the metrics so that we can have a performing analysis and be able to predict the prices based on the model we used.

2.3 Dataset & Variables

The original dataset consists of 2934 rows with 6 columns. This dataset contains information about the prices of the security cameras and their features. Each row represents a different camera, and the columns include information such as its number, brand, category and even the price we are after.

Variables	Data type	Description
EAN	Object	The camera's identification number
item_description	Object	Description of Camera
brand_name	Object	The brand that sells the Camera
category	Object	Security & surveillance systems
currency	Object	The currency that the camera is selling on
Price	Object	Price of camera

Table 3: Price of Surveillance Cameras Dataset Description

2.4 Analysis

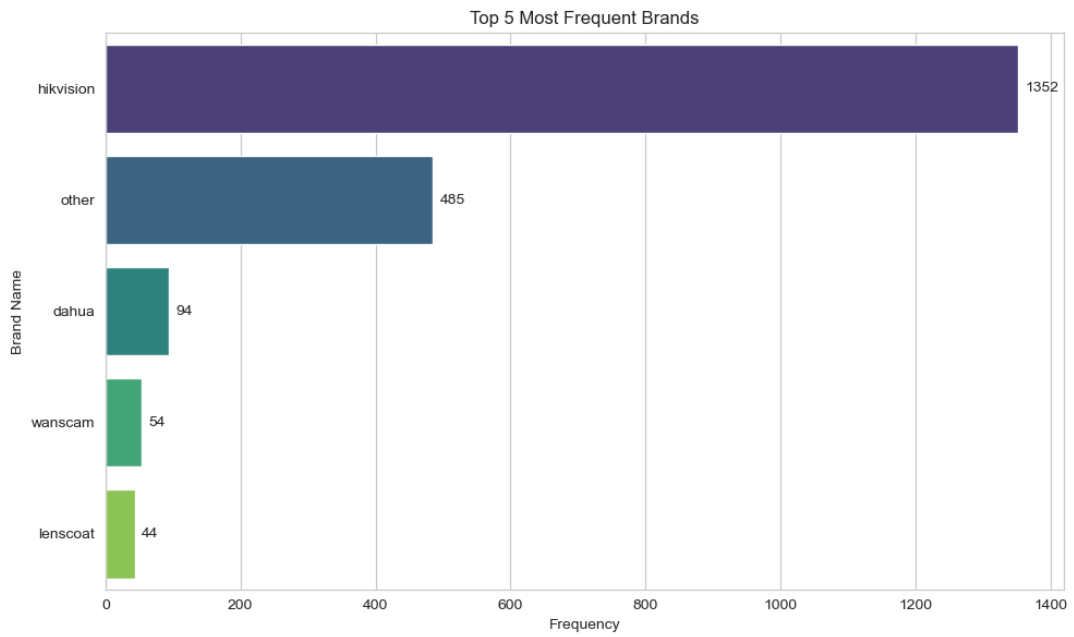


Figure 1: Distribution plot of Top 5 Most Frequent Brands

The bar chart displays the top 5 most frequent brands within a given dataset, showcasing the brand 'hikvision' as the most prevalent with a frequency of 1352. The 'other' category, which likely aggregates all brands not in the top four, has a significant presence with 485 occurrences. The brands 'dahua', 'wanscam', and 'lenscoat' follow, with noticeably lower frequencies of 94, 54, and 44 respectively. This visualization highlights 'hikvision' as the dominant brand among the surveyed group, substantially surpassing the others in frequency.

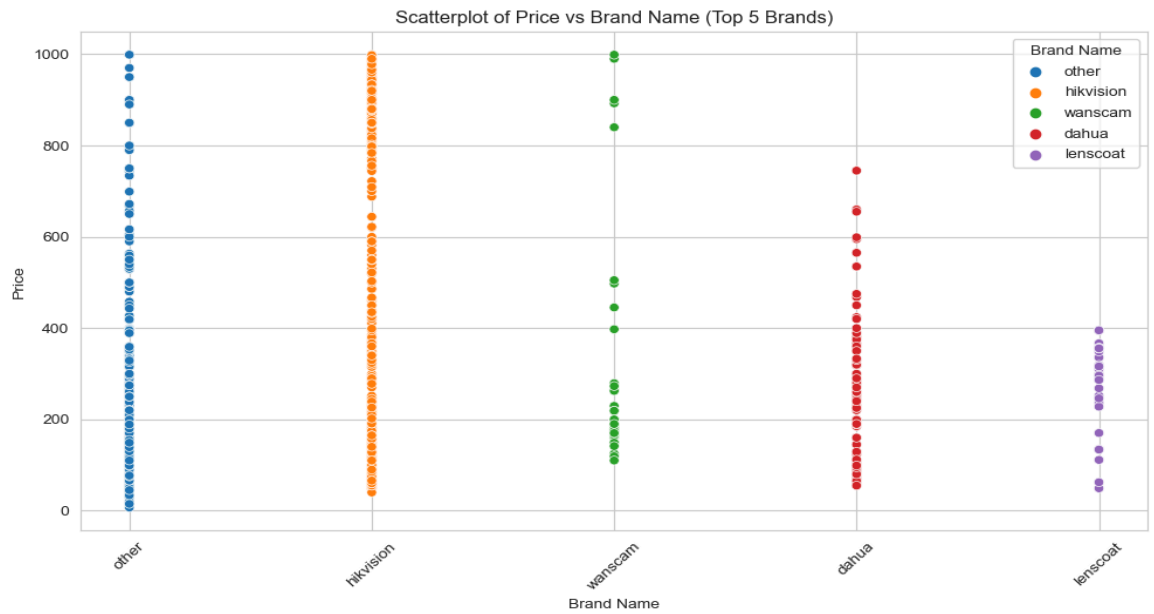


Figure 2: Scatterplot of Price vs Brand Name (Top 5 Brands)

The scatter plot represents the price distribution across the top 5 brands. The 'other' category exhibits a wide range of prices, suggesting it encompasses a diverse collection of brands. 'Hikvision' products show a tight price clustering, indicating consistent pricing. In contrast, 'dahua' displays a broad price range, suggesting a varied product line. 'Wanscam' and 'lenscoat' have narrower price ranges, which may reflect a more specialized or limited product selection.

2.5 Results

Linear Regression Model: Predicts the price based on linear relationships with features.

Random Forest Regressor Model: Utilizes an ensemble of decision trees to predict the price.

K Neighbors Regressor Model: Predicts the price based on the average of the prices of the k nearest neighbours in the feature space.

Plotting for Linear Model

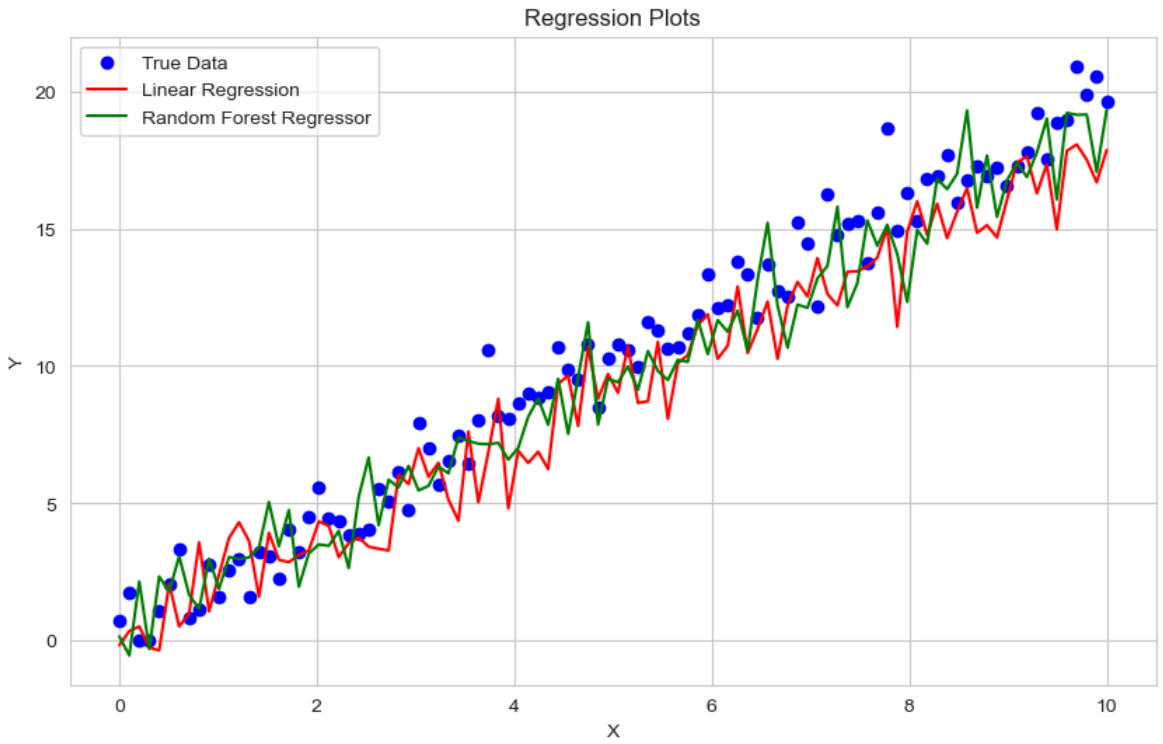


Figure 3: Linear & Random Forest Regression Plots

The Linear & Random Forest Regression plots will compare the performance of the two models against the true data which are indicated by the blue dots. This will show a non-linear trend with the green line which represents the Linear Regression model to capture with a straight line. This shows the limitations of the model for complex patterns. The Random Forest model however follows the true data trend closely which hints at a better and more suitable fit for the dataset based on its ability to model non-linear relationships.

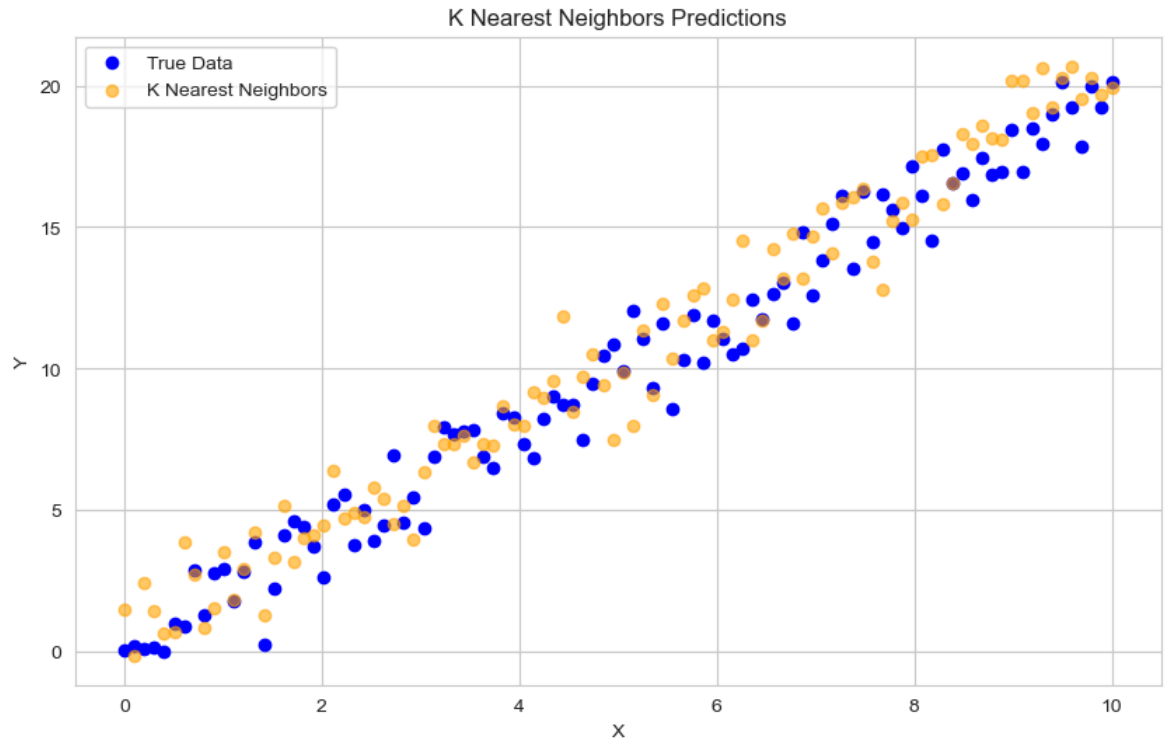


Figure 4: K Nearest Neighbors Regression Plot

This plot shows the K Nearest Neighbors algorithm's predictions while also comparing the true data. Marked with blue dots, the true data shows the actual values while the predictions are shown as orange dots. The model shows to be following the true data trend with some variation shown in regions where the points have a wider spread. This may show the localised averaging based on the KNN approach.

Regression Models Comparison

Ranked Comparisons of Regression Models		
Model	**RMSE**	**Adjusted-R2**
Linear Regression	246.157629	0.166968
Random Forest Regressor	246.091883	0.167413
K Nearest Neighbors	289.916567	-0.155530

Figure 5: Table of regression models

The models were evaluated based on their adjusted R^2 values and root mean square error (RMSE) metrics. The linear regression and random forest regressor models achieved similar adjusted R^2 values of approximately 0.17, indicating that they explain around 17% of the variance in the target variable. However, the random forest regressor slightly outperformed the linear regression model in terms of RMSE, with a value of approximately 246.09 compared to 246.16. On the other hand, the k-nearest neighbours (KNN) regressor exhibited a negative adjusted R^2 value, suggesting that the model performed worse than a horizontal line. Additionally, it had the highest RMSE among the three models, approximately 289.92, indicating larger prediction errors compared to the other models.

3. College Dataset (Unsupervised Learning)

3.1 Substantive Issue

Higher educations play an important role and step for all students who are seeking the next step in their education. It is mainly due to this that millions all over the world spend their time and effort to research and choose which college will be their next educational destination. For this project, we will simplify this by using unsupervised learning to cluster the data provided within the dataset to provide insights for students to choose which colleges they intend to apply to. This dataset will be used to identify key features and important data on each college to help understand the variables provided to cluster the data. This will help students classify and visualise which college they should focus on and conduct further research instead of blind research.

3.2 Methodology

We will be using K-Means clustering, Agglomerative Hierarchical Clustering, and DBSCAN, leveraging Principal Component Analysis (PCA) for dimensionality reduction. We will also use EDA to cluster the data found in the dataset to cluster its variables based on the algorithm used. For this model, PCA will be used to reduce the dimensions to help with visualising the dataset clusters. The model will separate the collages based on the features and show important data on the selection process. This will help provide suitable and informed decision-making based on the groups of collages that are categorized on their similarities and help students make the best decision.

3.3 Dataset & Variables

The original dataset consists of 777 rows and 19 columns and contains important information on college data and their main features and overall data. This dataset helps in providing the overall view of each college and can help give better insights into how students can choose and pivot to the right group of colleges for further research.

Variables	Data type	Description
Unnamed: 0	Object	Names of Colleges
Private	Object	Is the college private or not?
Apps	Integer	Number of applications
Accept	Integer	Number
Enroll	Integer	Number of new students enrolled.
Top10perc	Integer	Percentage of new students from top 10% of their high school class.
Top25perc	Integer	Percentage of new students from top 25% of their high school class.
F.Undergrad	Integer	Number of full-time undergraduates.
P.Undergrad	Integer	Number of part-time undergraduates
Outstate	Integer	Out-of-state tuition.
Room.Board	Integer	Room.Board
Books	Integer	Estimated book costs.
Personal	Integer	Estimated personal spending.
PhD	Integer	Percentage of faculty with Ph.D.'s
Terminal	Integer	Percentage of faculty with terminal degree.
S.F.Ratio	Float	Student/faculty ratio.
perc.alumni	Integer	Percentage of alumni who donate.
Expend	Integer	Instructional expenditure per student.
Grad.Rate	Integer	Graduation rate

Table 2: Credit Card Customer Data Description

3.4 Analysis

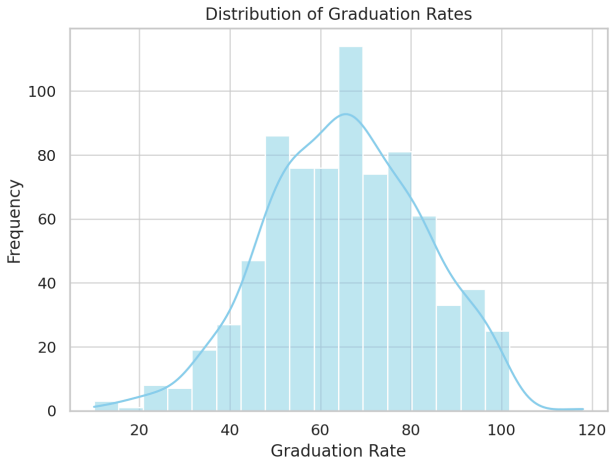


Figure 1: Distribution of Graduation Rates

This histogram shows the distribution of graduation rates among the colleges. The shape of the distribution shows that the higher graduation rates are not as common with a KDE estimation at the tail towards the higher rates. This suggests that not many colleges achieve higher graduation successes. Something interesting that is in the chart is the presence of a 100% graduation rate which should be further investigated.

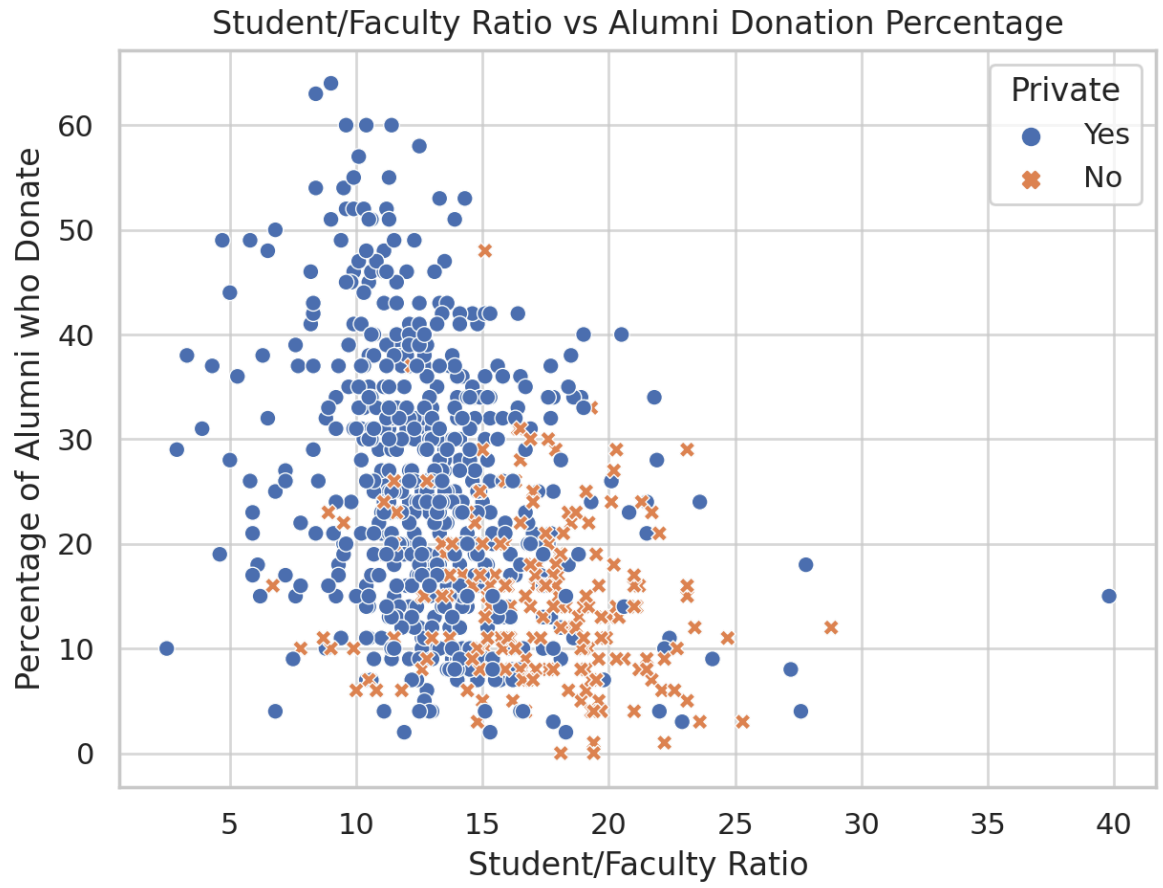
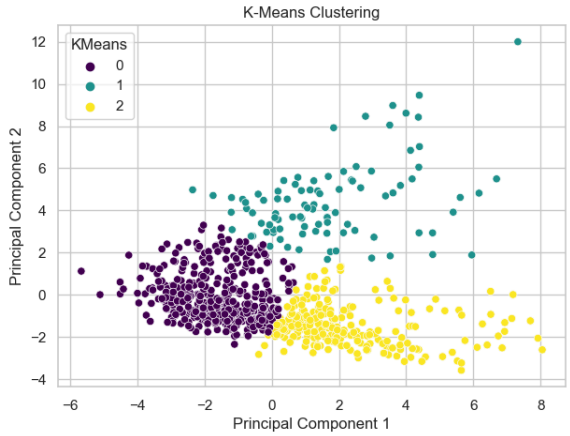


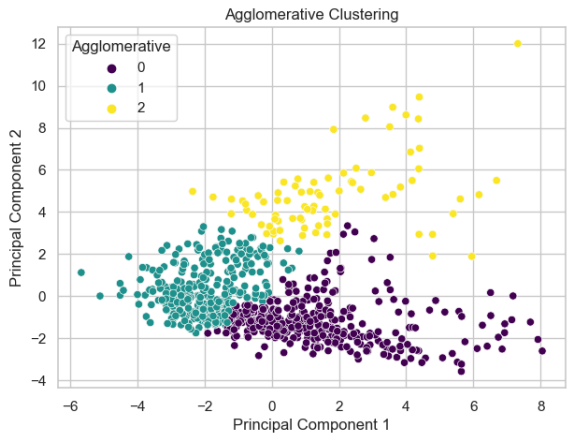
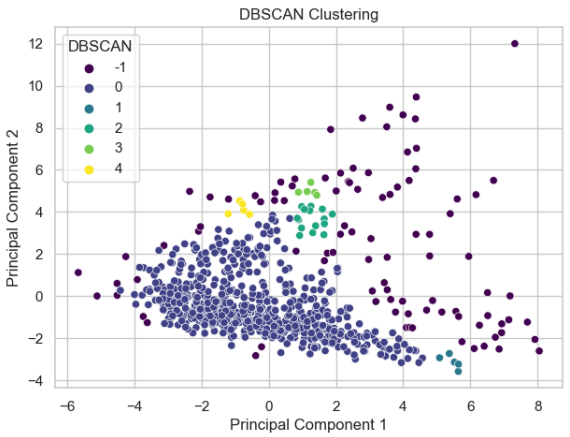
Figure 2: Correlation between features

This scatter plot shows the difference between the student and faculty ratio. This includes a percentage of alumni who donate to the colleges. Each plot is different, with private colleges as blue and non-private colleges as orange. We can infer that a high amount of private colleges have a lower ratio and a high percentage of donations. Other than that, there is no real trend to be donut between the student/faculty ratio and alumni donation percentages. Additionally, non-private institutions display a wider spread in student/faculty ratios, indicating more variability in these types of institutions.

3.5 Results

- **K-Means:** A popular centroid-based clustering algorithm.
- **Agglomerative Hierarchical Clustering:** A method that builds nested clusters by merging or splitting them successively.
- **DBSCAN (Density-Based Spatial Clustering of Applications with Noise):** A density-based clustering algorithm that groups together points that are closely packed together, marking as outliers points that lie alone in low-density regions.

Model Clustering using K-Means, Agglomerative Clustering, DBSCAN	
 *Figure 3: PCA Plot for K-Means Clustering*	The plot shows the partition of data points which have turned into three different clusters. This is done after reducing the dimensions using principal components. They are shown as different colours with Cluster 0, shown in purple, densely packed and centralized around the origin, suggesting a strong group similarity. Cluster 1 in yellow and Cluster 2 in green are more dispersed, with Cluster 2 extending further along the second principal component, possibly reflecting a variance in the underlying data characteristics that K-Means has detected.

PCA Plot for Agglomerative Clustering and DBSCAN	
 *Figure 4: PCA Plot for Agglomerative Clustering*	 *Figure 5: PCA Plot for DBSCAN Clustering*
This plot shows data segregated into three clusters. This hierarchical clustering technique uses the first two principal components for visualization. Distinct colours represent the clusters: purple for Cluster 0, cyan for Cluster 1, and yellow for Cluster 2. The varying tightness of the clusters reveals their diversity. Cluster 0's tight grouping shows that there is a high similarity.	The DBSCAN plot shows the data clustering based on density. The points are labelled as outliers (in yellow) and four distinct clusters are differentiated by colour. The largest cluster, shown in blue, shows density.

4. Admission Prediction (Classification)

4.1 Substantive Issue

Colleges receive over one million admissions from students annually and the numbers will not stop coming. Students submit multiple applications to enter college since it is the next step in their education journey. This calls for a need to automate and use the data we collect from student admissions and refine our way of admitting a student or not. We will be using the supervised learning technique from machine learning to add classification in the model to classify which student will be admitted or not based on his/her scores

4.2 Methodology

Based on this dataset, we will use EDA charts with he distributions of CGPA and GRE scores, alongside the balance of the target variable, admitted. This will be used to create and visualise a heatmap with a correlation between the features found in the dataset. We will apply supervised learning models to classify the admissions such as Logistic Regression, SVM with probability estimation enabled, and Random Forest. These models will evaluate and compare the variables and models based on the accuracy, precision, recall, and F1 score, alongside their ROC curves which will overall give us insights on the prediction capabilities and how effective the model is to classify the dataset.

4.3 Dataset & Variables

The dataset has up to 400 entries and 4 columns. This dataset includes all scores a student has accumulated across his academic years from GRE to GPA. This dataset has also labelled our target variable on whether a student will be admitted or not based on his scores. This will help us classify the factors that lead to admission or not based on the student's eligibility.

Variables	Data type	Description
gre	Integer	GRE scores (out of 340)
sop	Float	Statement of Purpose strength (out of 5)
cgpa	Float	Undergraduate GPA (out of 10)
admitted	Integer	Admission decision (1 for admitted, 0 for not admitted)

Table 1: Admission Prediction Description

4.4 Analysis

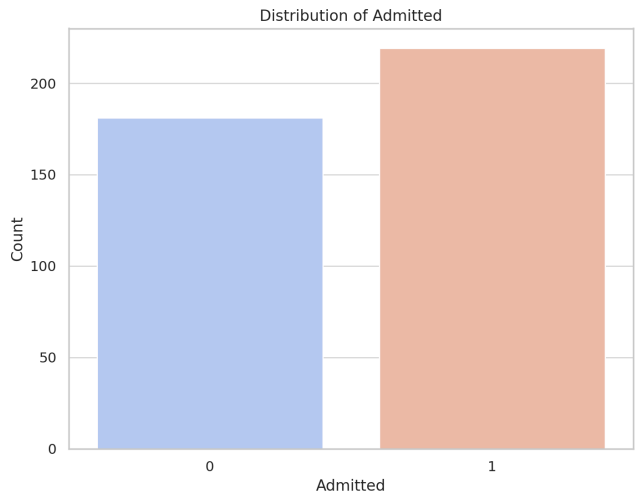


Figure 1: Distribution of Target Variable 'Admitted' Column

The distribution of the target variable admitted is shown above. It appears that the dataset is relatively balanced concerning the admission outcome, with approximately 54.75% of the students being admitted (label 1) and 45.25% not admitted (label 0). This balance suggests that the dataset has a good mix of both outcomes, which is beneficial for training supervised learning models without a strong bias towards one class.

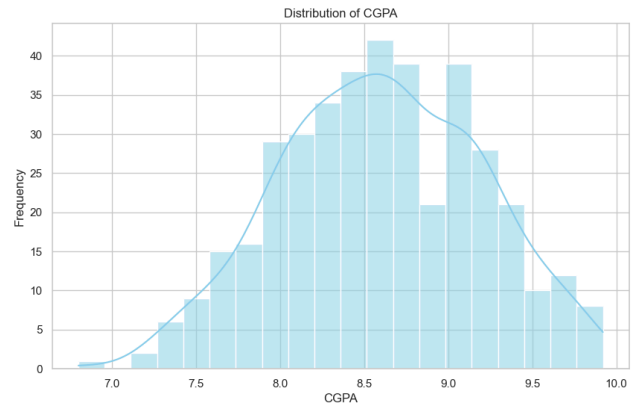


Figure 2: Distribution of CGPA

Visualization 1: Distribution of CGPA: This chart shows the distribution of CGPA scores among students. The CGPA scores are mostly distributed between 8 and 10, with the highest frequency around 8 to 9.

Distribution of GRE Scores: This histogram displays the distribution of GRE scores. GRE scores in this dataset range broadly from around 290 to 340, with a concentration of scores between 310 and 330.

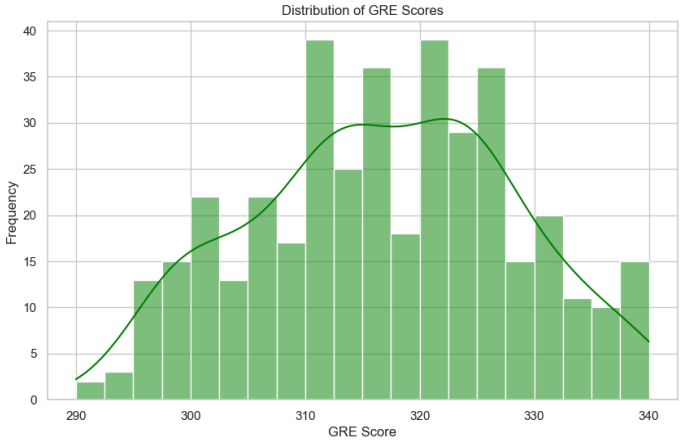


Figure 3: Distribution of GRE Scores

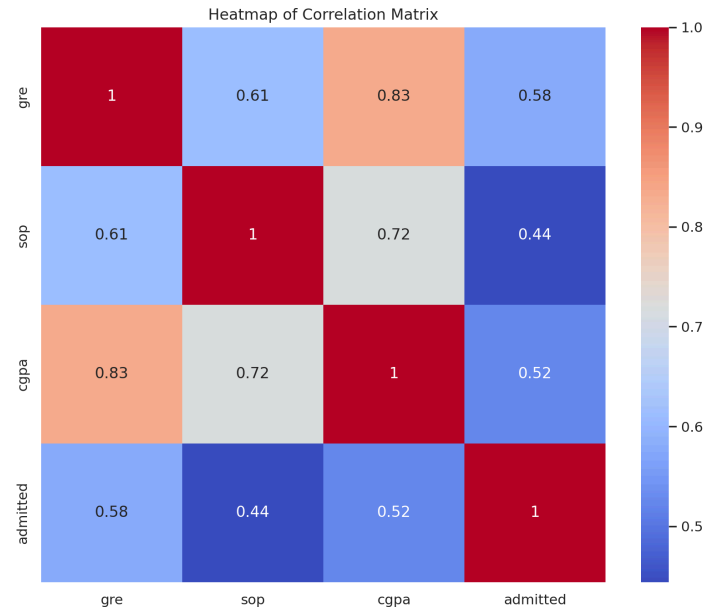


Figure 3: Heatmap of Correlation Matrix

Heatmap of Correlation Matrix: The heatmap visualizes the correlation between different variables in the dataset. The correlation values range from -1 to 1, where 1 indicates a perfect positive correlation, -1 indicates a perfect negative correlation, and 0 indicates no correlation. From the heatmap, we can see how each variable correlates with the others, including their relationship with the admission decision (admitted).

4.5 Results

Model Comparison

Figure 4: Model Comparison between three models

The SVM model has shown to perform better than other models on all metrics which makes it the most efficient and effective model for the dataset. Based on the results. The random forest model is still useful but still lags behind in overall performance. Based on the three models, they are able to capture linear separations (like SVM and Logistic Regression) and are more effective than ensemble methods like Random Forest. However, the choice of model can vary depending on the specific needs for precision, recall, or another metric based on the application's requirements.

Model	Accuracy	Precision	Recall	F1 Score
Logistic Regression	0.8000	0.8780	0.7660	0.8182
SVM	0.8125	0.9000	0.7660	0.8276
Random Forest	0.7125	0.7727	0.7234	0.7473

Accuracy: The SVM model leads with 81.25% accuracy, with Logistic Regression close behind at 80%; and Random Forest trails at 71.25%.

Precision: SVM boasts the highest precision at 90%, making it most reliable for predicting admissions, with Logistic Regression also high at 87.8%.

Recall: Both Logistic Regression and SVM show similar recall around 76.6%, outperforming Random Forest's 72.34% in identifying true positives.

F1 Score: SVM has the top F1 score at 0.8276, indicating the best balance of precision and recall, followed by Logistic Regression and then Random Forest.

ROC Curves

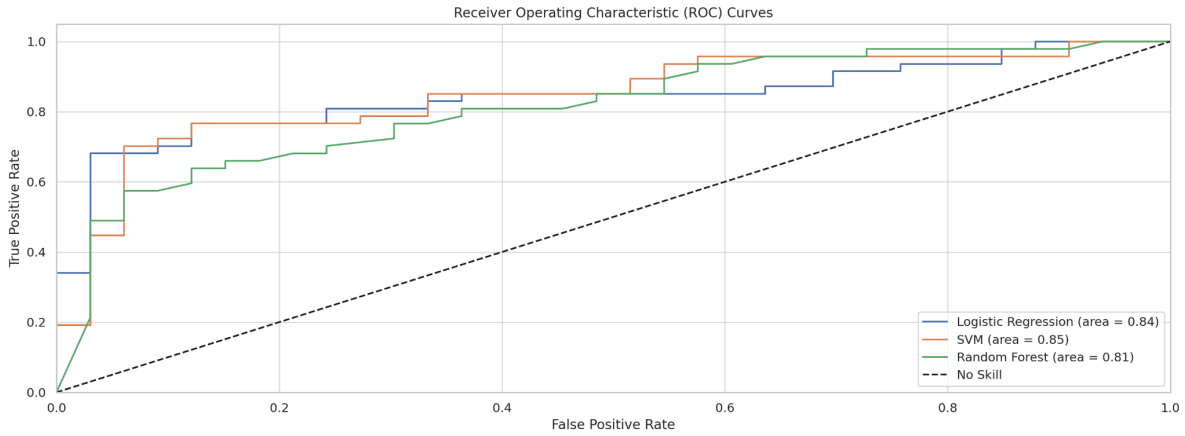


Figure 5: ROC Curves for Models

The curves show the ability of the model to classify based on their discrimination threshold but it can be shown in variations. The area under the curve shows the model's ability to identify between the classes, where a value of 1.0 represents a perfect model, and a value of 0.5 represents a model with no discriminative ability. In actuality, all models show a good ability to classify properly. This shows that they classify between admitted and not admitted students effectively. The SVM model, with the highest AUC, shows the best performance among the three.

5. References

Aljuaid, M. (2017, November 9). *Surveillance Camera*. Kaggle. Retrieved March 28, 2024, from https://kaggle.com/datasets/mansouraljuaid/surveillance-camera

Kumar, N. (2017, November 9). *College Dataset (unsupervised learning)*. Kaggle. Retrieved March 28, 2024, from https://www.kaggle.com/datasets/nishantpatyal/college-dataset-unsupervised-learnig

scikit-learn · PyPI. (n.d.). PyPI. Retrieved March 28, 2024, from https://pypi.org/project/scikit-learn/

Singh, S. (2017, November 9). *Admission Predict*. Kaggle. Retrieved March 28, 2024, from https://www.kaggle.com/datasets/sakshisaku3000/admission-predict

Waskom, M. (n.d.). *Installing and getting started — seaborn 0.13.2 documentation*. Seaborn. Retrieved March 27, 2024, from https://seaborn.pydata.org/installing.html