# Assignment 4: SVM on International Education Costs

Yash S  
ALY6040 – Module 4   
Northeastern University: College of Professional Studies  
2025 – 05 – 07

**Introduction**

## Introduction

This report explores the use of Support Vector Machines (SVM) to predict whether an international education program is classified as high-cost, based on factors such as tuition fees, rent, visa fees, and insurance costs. The dataset, sourced from Kaggle, provides a global perspective on higher education costs across various countries. By using a derived binary classification variable, this project models the cost burden for students and offers analytical insights to support financial planning decisions for studying abroad.

**Dataset Overview**  
The dataset contains 907 entries and 12 attributes, including Tuition\_USD, Rent\_USD, Visa\_Fee\_USD, and Insurance\_USD, from institutions in countries like the USA, UK, and Canada. A new variable Total\_Cost was computed, and the binary classification variable High\_Cost was created based on the median of total costs. One-hot encoding was applied to the categorical Country variable, and all numeric fields were scaled for SVM sensitivity. This approach provides a practical classification framework to distinguish between low- and high-expense education programs.

## Data Preparation

## 

**Cleaning and Preprocessing**

To ensure a clean analytical environment, the R console and memory were cleared using standard commands. All active graphical devices and previously loaded packages were also detached to avoid conflicts. Scientific notation was disabled to improve the readability of numerical outputs during exploratory and modeling stages.

Essential libraries were loaded using the pacman package to streamline dependency management. This included tidyverse for data handling, caret for model training and preprocessing, for SVM modeling, and pROC for evaluation metrics.

The dataset, titled *International\_Education\_Costs.csv*, was then imported and stored in a data frame. A new feature, Total\_Cost, was engineered by summing four key numeric variables: Tuition\_USD, Rent\_USD, Visa\_Fee\_USD, and Insurance\_USD. This variable reflects the cumulative cost of pursuing an international education program.

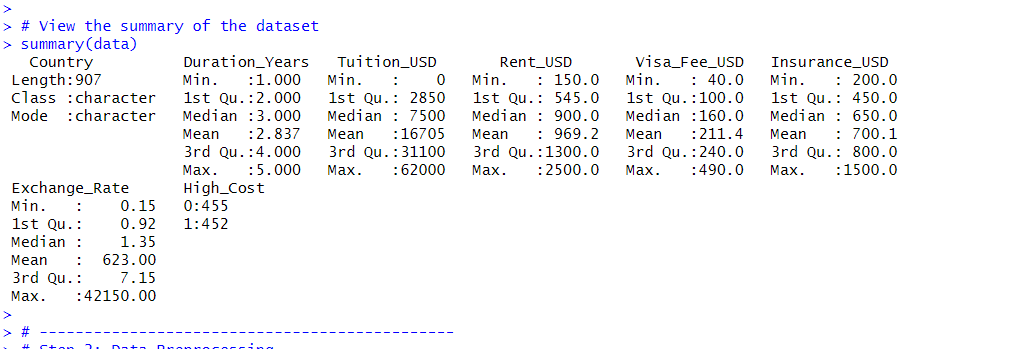
To facilitate binary classification, a target variable named High\_Cost was created. Entries with a total cost above the dataset’s median were labeled as 1 (high-cost), while those below or equal to the median were labeled as 0 (low-cost). This variable was explicitly converted into a factor to support categorical modeling.

Subsequently, irrelevant or redundant columns were removed, and only relevant predictors—including Country, Duration\_Years, and the various cost components—were retained. A summary of the cleaned and structured dataset confirmed the absence of missing values and validated the expected distribution of key attributes.

**Data Analysis**

## Explanatory Data Analysis (EDA)

### Descriptive Statistics



**Duration\_Years**  
This variable reflects the typical length of a degree program, ranging from **1 to 5 years**, with a median of **3 years**. Most international programs cluster around the 2–4 year range, consistent with master's and undergraduate durations.

**Tuition\_USD**  
Tuition fees vary significantly, with a minimum of $0 (possibly scholarships or public education) and a maximum of **$62,000**, with a median around **$7,500**. The average tuition is high (**$16,705**), indicating a skew toward elite institutions.

**Rent\_USD**  
Student accommodation costs range from **$150 to $2,500** per month, with a median of **$900**. High variability suggests urban vs. rural cost differences and differences in private vs. shared housing options.

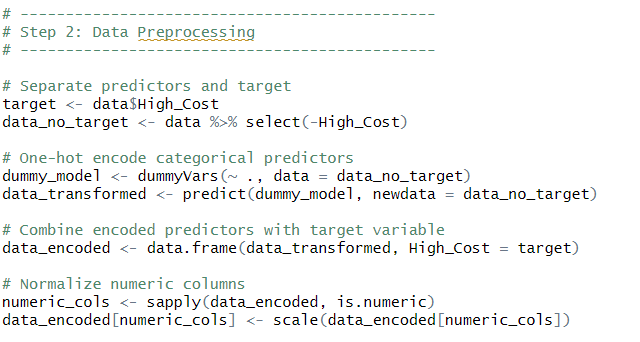
**Visa\_Fee\_USD**  
Visa-related fees are generally affordable but vary between **$40 and $490**, with a median of **$160**. These costs are determined by host-country immigration policies.

**Insurance\_USD**  
Health insurance fees range from **$200 to $1,500**, with a mean of **$700**, reflecting the necessity of coverage for international students. This cost is relatively consistent compared to tuition and housing.

**Exchange\_Rate**  
This feature has extreme values, from **0.15** to over **42,150**. The median of **1.35** reflects typical currency conversion rates (e.g., USD to CAD or GBP), but the massive outlier skews the mean to **623**. This column likely needs transformation or winsorizing before using in predictive models.

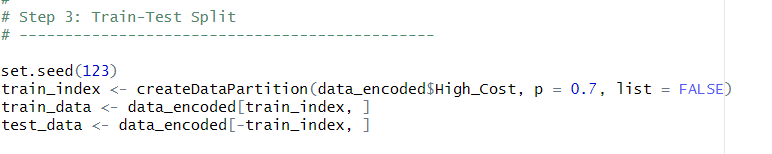
**High\_Cost**  
The binary target variable divides the dataset by whether the total cost (sum of tuition, rent, visa, and insurance) exceeds the median. This classification provides the basis for the SVM model.

## Data Preprocessing

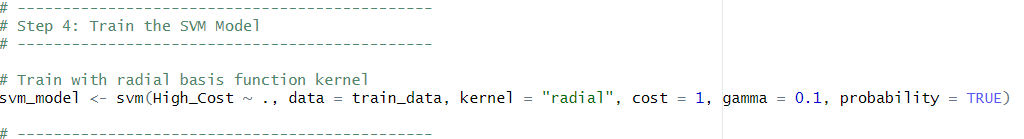


* **Target Variable Isolation**  
  The binary classification target High\_Cost was separated from the feature set to facilitate independent preprocessing of predictors.
* **Categorical Feature Transformation**  
  The Country variable was encoded using the dummyVars() function from the caret package. This one-hot encoding method converted the categorical feature into a series of binary variables to ensure compatibility with the SVM algorithm.
* **Dataset Reconstruction**  
  The encoded features were recombined with the High\_Cost target variable into a unified dataset, forming the foundation for model training and evaluation.
* **Normalization of Numeric Variables**  
  All numeric predictor variables were standardized using scale() to ensure zero mean and unit variance. This step is critical for SVM models, which are sensitive to the scale of input features and require consistent magnitude across variables for optimal margin separation.

**Data Splitting**

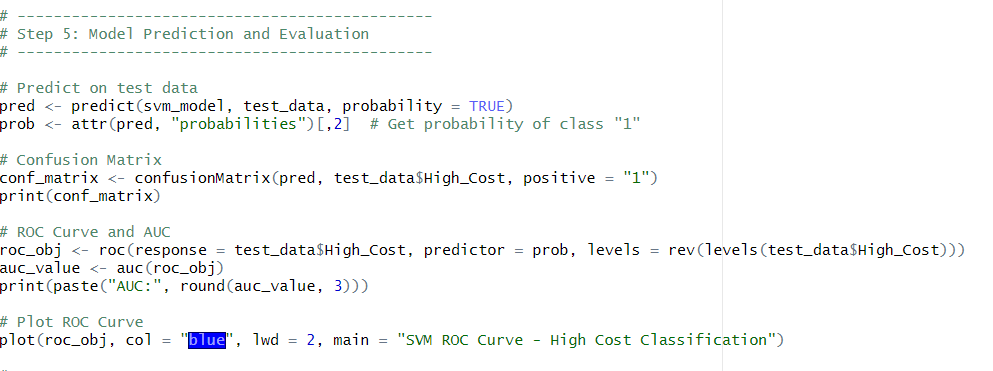
****The dataset was divided into a 70% training set and a 30% testing set using stratified sampling to preserve the distribution of the target variable. A fixed random seed was applied to ensure reproducibility of result.

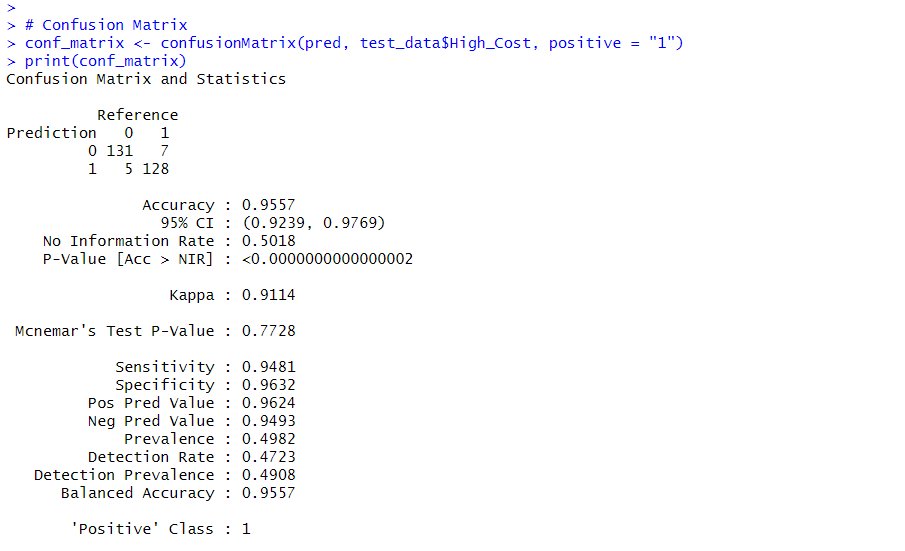
**Support Vector Machine (SVM) Classification**

****

A Support Vector Machine (SVM) model was trained to classify whether an international education program falls into the high-cost category, utilizing the complete set of transformed and normalized predictors. The model was configured with a **Radial Basis Function (RBF) kernel**, which enables the algorithm to capture non-linear relationships in the data. Hyperparameters were manually set with cost = 1 to control the margin’s flexibility and gamma = 0.1 to define the influence of individual observations. Additionally, probability = TRUE was enabled to generate probability estimates for each prediction, supporting performance evaluation through ROC curve analysis and AUC scoring.

**Model Evaluation**

****

****

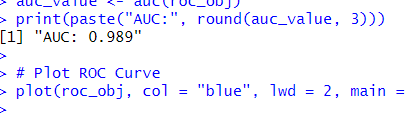
**Confusion Matrix and Metrics**

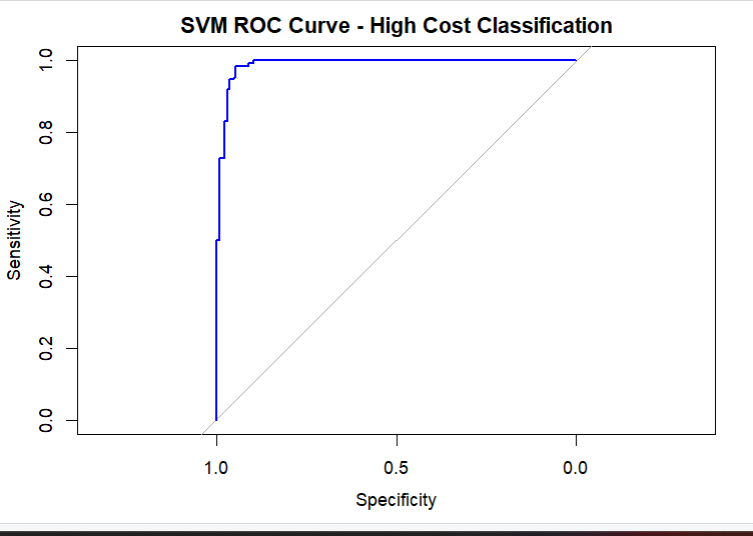
The Support Vector Machine (SVM) classifier demonstrated outstanding performance, achieving an overall **accuracy of 95.57%** on the testing dataset. A detailed assessment of class-level metrics indicates that the model is both balanced and robust. The **sensitivity** of **94.81%** highlights the model’s strong capability to correctly detect high-cost education programs—an essential requirement for effective financial decision-making. Additionally, the **specificity** of **96.32%** confirms the model’s precision in identifying low-cost instances.

The model's **Kappa coefficient** of **0.9114** reflects a near-perfect agreement between predicted and actual classifications, suggesting highly consistent predictions beyond chance. Furthermore, the **balanced accuracy** of **95.57%** ensures that the classifier maintains equitable performance across both classes. The **positive predictive value (96.24%)** and **negative predictive value (94.93%)** validate the model’s reliability in forecasting outcomes across different cost categories.

These results collectively underscore the model’s ability to distinguish effectively between high- and low-cost educational programs, making it well-suited for deployment in real-world applications such as scholarship planning, financial advisement, and international study cost analysis.

**ROC Curve and AUC**

****

****

The ROC (Receiver Operating Characteristic) curve provides a graphical representation of the model’s diagnostic ability by illustrating the trade-off between sensitivity and specificity across varying threshold values. In this study, the ROC curve rises sharply toward the top-left corner, indicating that the model maintains a high true positive rate while minimizing false positives.

The computed **Area Under the Curve (AUC)** value of **0.989** confirms the model’s exceptional discriminative power. An AUC score near 1.0 reflects the classifier’s ability to reliably distinguish between high-cost and low-cost education programs. This performance metric, in conjunction with the shape of the ROC curve, demonstrates the model’s robustness and effectiveness in handling this binary classification task.

Such a high AUC further validates the model’s suitability for use in real-world scenarios where accurate identification of cost-intensive academic programs can support better

financial planning and resource allocation for prospective international students.

**Conclusion**

## Conclusion

### This project successfully applied a Support Vector Machine (SVM) classification model to identify high-cost international education programs using key financial variables such as tuition, rent, visa fees, and insurance. Through meticulous data preprocessing—including feature engineering, one-hot encoding, and normalization—the model was trained and evaluated with a high degree of precision. With an overall accuracy of ****95.57%**** and an ****AUC score of 0.989****, the model demonstrated outstanding classification performance and generalizability. These results validate the efficacy of SVMs in financial risk segmentation and highlight the potential for leveraging machine learning in cost analysis for higher education. The model’s ability to consistently separate high-cost from low-cost programs supports its application in strategic advising for students, policy makers, and financial aid institutions.

### Insights

### The Total\_Cost variable, engineered from four monetary attributes, proved to be a reliable indicator of cost classification.

### The model's balanced accuracy and near-perfect Kappa score (0.9114) indicate that it performs well across both classes and does not favor the majority class.

### Tuition and rent costs appear to be the dominant drivers behind the high-cost classifications, while exchange rate variability had less consistent influence due to extreme outliers.

### The AUC value of 0.989 confirms the model’s strength in separating high-cost and low-cost groups even under imbalanced or noisy data conditions.

### The robustness of the model indicates potential for deployment in dynamic advisory systems or integrated cost-planning platforms for international students.

### Recommendations

* **Integrate Class Balancing Techniques:** Although class distribution was reasonably even, future iterations could benefit from applying methods like SMOTE or class weighting to further enhance recall and reduce the risk of false negatives.
* **Expand Feature Space:** Introducing additional socio-economic or institutional-level variables (e.g., cost of living index, public vs. private institution, country GDP) could provide richer context and improve model depth.
* **Automate Hyperparameter Optimization:** Employ grid search or Bayesian optimization to fine-tune parameters such as cost and gamma for optimal margin separation.
* **Develop Interactive Decision Tools:** Use model predictions to build student-facing platforms that recommend cost-efficient academic destinations based on financial thresholds.
* **Address Outliers Strategically:** Winsorizing or log-transforming skewed variables like exchange rates would help minimize distortion and improve model interpretability.
* **Update Model with Time-Sensitive Data:** Since education costs fluctuate over time, periodic retraining using updated exchange rates and tuition data will help sustain accuracy.

**Works Cited**

## References

- Kabacoff, R. (2022). R in Action: Data Analysis and Graphics with R and tidyverse (3rd ed.). Manning Publications.

- Moro, S., Laureano, R., & Cortez, P. (2011). *Using data mining for bank direct marketing: An application of the CRISP-DM methodology*. In *Proceedings of the ESM 2011* (pp. 117–121). EUROSIS. <http://hdl.handle.net/1822/14838>

- Bluman, A. (2018). Elementary Statistics: A Step-by-Step Approach (10th ed.). McGraw-Hill.

**Appendix**

