Interactive dashboard application:
Helping students analyze and compare the cost of International Education

Python for Engineering Data Analysis [EI04024]
Final Project

Authors:
Can Telli
ge63gar
03759727

Eylül Kırdak
ge89fev
03764297

Fuat Sekban
go56kub
03787300

Technical University of Munich
August, 2025

1. Introduction

The rising costs of higher education have become a critical factor for students when deciding what and where to study (Helen Li, 2013). In addition to paying for school, students must also pay for living expenses, insurance, and housing. The amount they pay for these things can vary greatly depending on the country. This makes it hard to know exactly how much international education costs. At the same time, universities and policymakers are under pressure to provide clear and easy-to-find information on these expenses so that people can make informed decisions. To address this challenge, our project focuses on developing an interactive dashboard that combines data visualization and machine learning techniques to explore the global costs of international education. We used information about tuition, rent, and other living costs in different countries to make a Streamlit-based application. It lets users see how education costs are set up, put countries into groups based on cost factors, and predict how much tuition will cost in individual countries and programs.

The primary research objectives of the project were threefold:

1. Exploring cost patterns across countries and fields of study using interactive visualizations.
2. Applying machine learning clustering methods to group countries with similar cost structures.
3. Developing a regression-based tuition prediction tool that estimates education costs under different scenarios.

By combining interactive visualization with predictive modeling, the dashboard aims to provide a practical and user-friendly tool to better understand and compare the financial demands of higher education worldwide.

2. Methodology

The analysis is based on the *Cost of International Education* dataset (Shamim, 2023), sourced from Kaggle, which contains 907 records from 71 countries. The data preparation involved standardizing column names, engineering a field_of_study category from raw program names, and creating a total_annual_cost metric. To ensure data quality for modeling, outliers were removed using the IQR method.

The dashboard was built using Streamlit for the multi-page structure and user interface. Data manipulation was handled by pandas and NumPy, while Plotly was used for all interactive visualizations. For the machine learning components, scikit-learn was used to implement the clustering (K-Means, Hierarchical, DBSCAN) and regression (Linear, Ridge, Random Forest) models.

3. The Interactive Dashboard: Tools & Insights

The core of this project is a multi-page dashboard designed to provide users with a suite of tools for exploring international education costs. The application is divided into three main analytical pages.

3.1 Global Cost Overview

The dashboard's first page provides tools for exploratory data analysis, organized into three tabs, each designed to answer a specific question.

The main view is a choropleth world map that provides an immediate overview of geographic cost distribution. To effectively visualize the skewed nature of global costs, the map includes a toggle for a logarithmic color scale, a standard technique for normalizing such data and revealing nuanced differences between lower-cost countries.

For more granular comparisons, the Country Ranker tab allows users to generate custom-filtered "Top-N" bar charts. The tool allows multi-faceted filtering, which enables a user to isolate and compare specific market segments (e.g., "Master's degrees in Data Science & AI").

Finally, to understand the *structure* of costs, we implemented a sunburst chart. This hierarchical plot is effective at visualizing the proportional breakdown of expenses (tuition, rent, etc.), allowing a user to quickly compare the financial composition of education across different countries.

3.2 Country Clustering

The second page of the dashboard provides a user-driven machine learning tool for grouping countries with similar cost profiles. The analytical merit of this page lies in its interactivity; rather than presenting a single, static model, it empowers the user to adjust model parameters dynamically. The tool allows users to choose from three distinct clustering algorithms for unique analytical perspectives: K-Means for clear segmentation, Hierarchical Clustering for understanding relationships, and DBSCAN for outlier detection. For rigorous analysis, the tool provides validation features such as the Elbow Method to justify the choice of k for K-Means, a Silhouette Score to measure cluster quality, and Principal Component Analysis for accurate 2D visualization.

This multi-model approach allows for a more robust analysis. For example, when clustering countries by average tuition and average rent, a K-Means model with k=3 achieves a strong Silhouette Score of 0.568, indicating that the clusters are dense and well-separated[1]. The model partitions countries into three distinct groups: a high-tuition/high-rent cluster (e.g., USA, Canada), a low-tuition/high-rent cluster (e.g., Switzerland, Denmark), and a large low-tuition/low-rent cluster representing the most affordable options.

The Hierarchical Clustering model visualizes the nested relationships between all countries through a dendrogram. This reveals not just *if* countries are in the same group, but *how closely related* they are. For instance, the dendrogram reveals nuanced relationships within the high-cost group[2]. It identifies a sub-cluster containing Singapore, Australia, and Hong Kong. Another distinct sub-group at a similar level of dissimilarity includes the UK, Canada, New Zealand, and the UAE. The USA is shown to be the most unique member of this tier, branching off from all others at the highest level.

Finally, running the same analysis with DBSCAN highlights a structural feature of the dataset and the importance of hyperparameter tuning. A large epsilon of 1.0 isolates only the most extreme global outliers (USA, Switzerland, Hong Kong), confirming the data is heavily skewed. However, a more sensitive epsilon of 0.5 provides a more granular analysis that answers a different question: "Where are the dense, common cost profiles?"

This model identifies a large, dense cluster of affordable countries (Cluster 0: avg. ~$3.9k/~$531) and two distinct tiers of high-cost countries[3]. Crucially, it also flags a number of countries like Netherlands and Israel as outliers. This contrasts with the K-Means result, which forced these "low-tuition/high-rent" countries into their own cluster. The DBSCAN result suggests that these countries are not a coherent "market segment" but rather a collection of individual, anomalous financial profiles. This demonstrates the power of the tool: a user can switch from K-Means to understand the broad market segments to DBSCAN to inspect whether some countries have truly unique cost structures.

[1] Figure 1
[2] Figure 2
[3] Figure 3

3.3 Tuition Predictor

The final page of the dashboard offers a predictive tool that estimates tuition costs based on program and location characteristics. Users can enter information such as the country, degree level, and field of study, as well as numeric factors including program duration and the living cost index, to generate tuition estimates[4]. For ease of use, these values are pre-filled with median values for the selected country but can be adjusted manually. After submission, the tool provides both a point estimate and an empirical 95% prediction interval, offering users a realistic range of potential outcomes.

The model was implemented as a Scikit-learn pipeline that imputes missing values, scales numeric features, and one-hot encodes categorical variables. Several regressors were tested, but a Random Forest Regressor configured with 300 estimators and no maximum depth restriction was selected for its superior performance in capturing nonlinear relationships. In this configuration, the selected features were program duration (in years), living cost index, country, degree level, and field of study.

Performance was evaluated through 5-fold cross-validation, reporting multiple metrics. The Random Forest achieved an RMSE of $3,738, an MAE of $2,064, a Median AE of $760, and an R^2 of 0.949. In practical terms, tuition predictions are usually within $3,700 of the actual cost, which is accurate enough for budgeting and comparison purposes. Diagnostic plots, including a predicted vs. actual scatterplot[5] and residuals histogram[6], offer insight into the model's behavior. Feature importance analysis consistently identifies *country* as the most influential predictor[7].

In addition to Random Forest, two linear baselines were evaluated: Linear Regression and Ridge Regression. Both models are simpler and offer greater interpretability through coefficients, but they fell short in predictive accuracy. With above-mentioned configurations for Random Forest method, Linear Regression achieved an RMSE of $4,390, an MAE of $3,019, a Median AE of $1,783, and an R^2 of 0.930, while Ridge Regression yielded an RMSE of $4,445, an MAE of $3,100, a Median AE of $1,953, and an R^2 of 0.928 in cross-validation. By contrast, the Random Forest Regressor reduced the error and improved model accuracy[8], confirming its superior ability to model non-linear patterns in tuition data. This trade-off underscores why Random Forest was selected as the final model: although linear methods are easier to interpret, Random Forest provides substantially greater predictive power for practical tuition estimation.

Finally, the tool provides fairness and segmentation metrics to break down accuracy by groups, like degree level. R^2 was 0.97 for master's programs and 0.91 for bachelor's programs, suggesting the model performs better for certain groups. A toggle for logarithmic scaling was implemented to run predictions with reduced skew from major outliers.

4. Conclusion

This project successfully demonstrated that an interactive dashboard integrating visualization and machine learning can yield actionable insights into the global cost landscape of international education. The exploratory tools and machine learning models effectively clarified cost structures and provided a framework for estimating tuition fees in a transparent, user-driven environment.

The analysis revealed distinct global cost patterns. The clustering tool, for instance, consistently identified three primary cost tiers. The multi-model approach proved particularly insightful, with DBSCAN highlighting the heavily skewed nature of the data by identifying a number of countries as statistical outliers. For prediction, the

[4] Figure 4
[5] Figure 5
[6] Figure 6
[7] Figure 7
[8] Figure 8

Random Forest Regressor performed best, achieving a high degree of accuracy ($R^2 = 0.949$) and confirming that a country's location is the primary driver of tuition variation.

The main limitation of this project is the static, cross-sectional nature of the dataset, which represents a single snapshot in time and does not account for cost inflation or recent policy changes. The dataset also lacks institution-level granularity, and its scope is limited to the available data. Therefore, while the dashboard demonstrates predictive accuracy and provides a powerful exploratory tool, it is best understood as a strong directional guide rather than a definitive financial planner. In summary, the application is a successful proof-of-concept for how data science can increase transparency and help prospective students better anticipate the financial demands of higher education worldwide.

References

1. Harris, C. R., Millman, K. J., van der Walt, S. J., Gommers, R., Virtanen, P., Cournapeau, D., … Oliphant, T. E. (2020). Array programming with NumPy. Nature,
2. The pandas development team. (2024). pandas-dev/pandas: Pandas (Version 2.2.2) [Computer software]. Zenodo. https://doi.org/10.5281/zenodo.3509134
3. Pedregosa, et al., (2011). Scikit-learn: Machine learning in Python. Journal of Machine Learning Research
4. Seabold, S., & Perktold, J. (2010). Statsmodels: Econometric and statistical modeling with Python. In Proceedings of the 9th Python in Science Conference (pp. 92–96).
5. Plotly Technologies Inc. (2015). Plotly: Collaborative data science [Computer software]. https://plotly.com/python/
6. Streamlit Inc. (2023). Streamlit [Computer software]. https://streamlit.io
7. Arvai, K., & contributors. (2021). kneed (Version 0.8.2) [Computer software]. Lindemann, O. (2024). pycountry (Version current at use) [Computer software]. https://pypi.org/project/pycountry
8. Domoritz, T., & contributors. (2021). streamlit-plotly-events [Computer software]. Archibald, R. B., & Feldman, D. H. (2008). Explaining increases in higher education costs. *The Journal of Higher Education, 79*(3), 268–295. https://doi.org/10.1080/00221546.2008.11772099
9. Minor, R. (2023). How tuition fees affected student enrollment at higher education institutions: The aftermath of a German quasi-experiment. *Journal for Labour Market Research, 57*, 28. https://doi.org/10.1186/s12651-023-00354-7 SpringerOpen
10. Shamim A. (2025). *Cost of International Education* [Data set]. Kaggle. Retrieved August, 2025, from https://www.kaggle.com/datasets/adilshamim8/cost-of-international-education

Appendix

Figure 1 - Geographic Distribution of K-Means Clusters (k=3) based on Average Tuition and Rent.

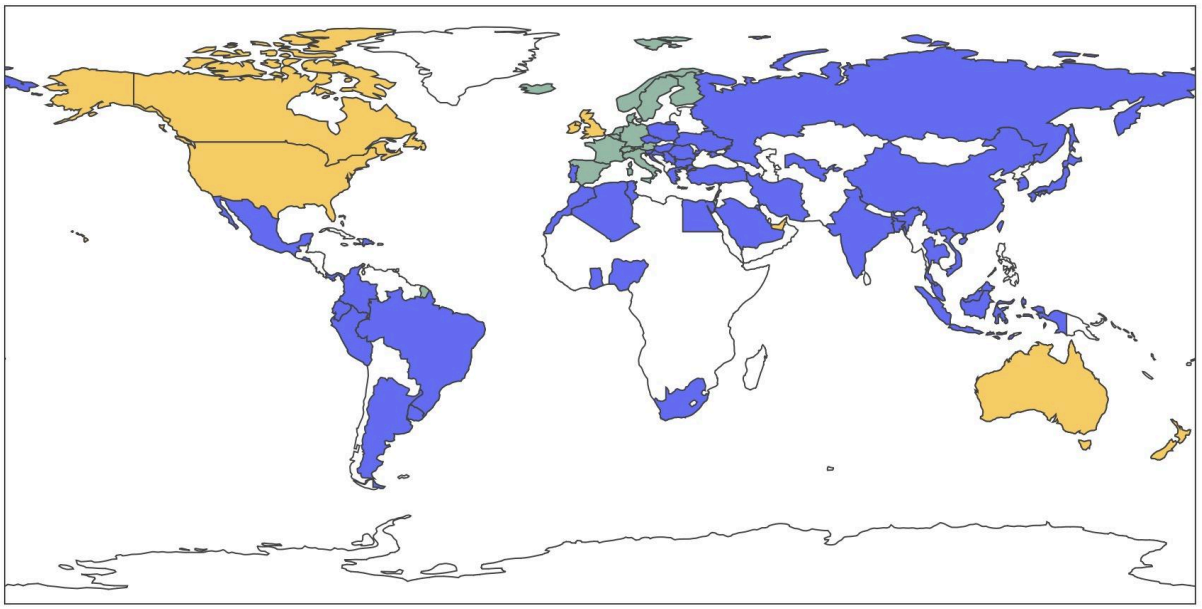


Figure 2 - Dendrogram of Hierarchical Clustering based on Average Tuition and Rent.

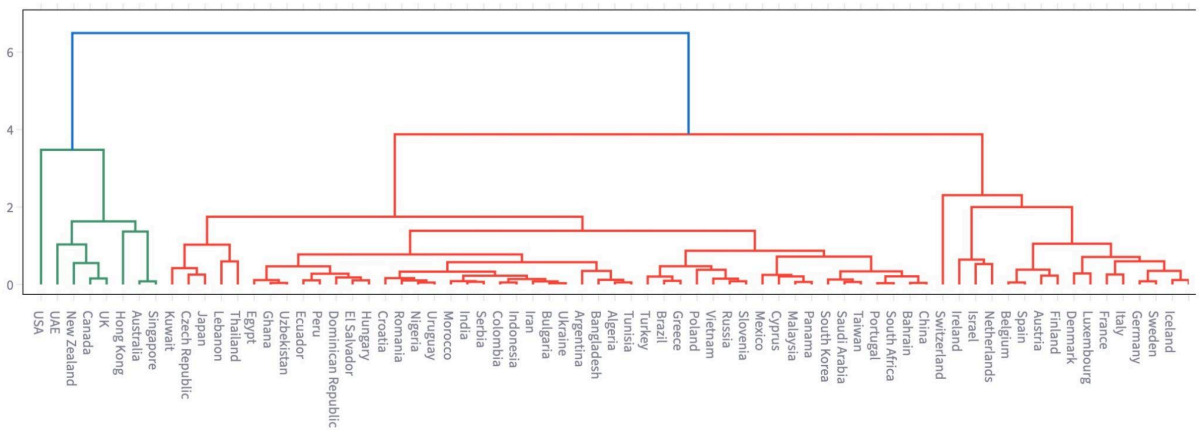


Figure 3 - Figure 6: DBSCAN Clustering Results (eps=0.5) Demonstrating Main Cluster and Highlighting Certain Outliers.

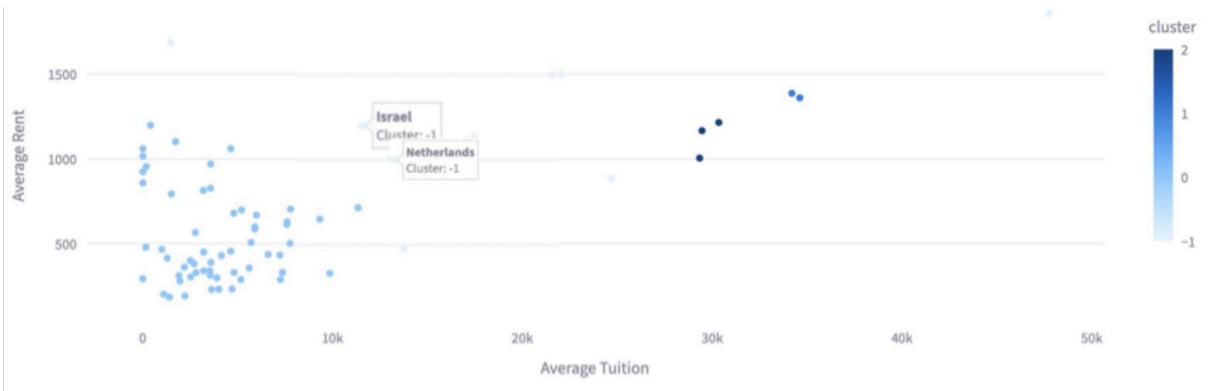


Figure 4 - Interactive prediction form

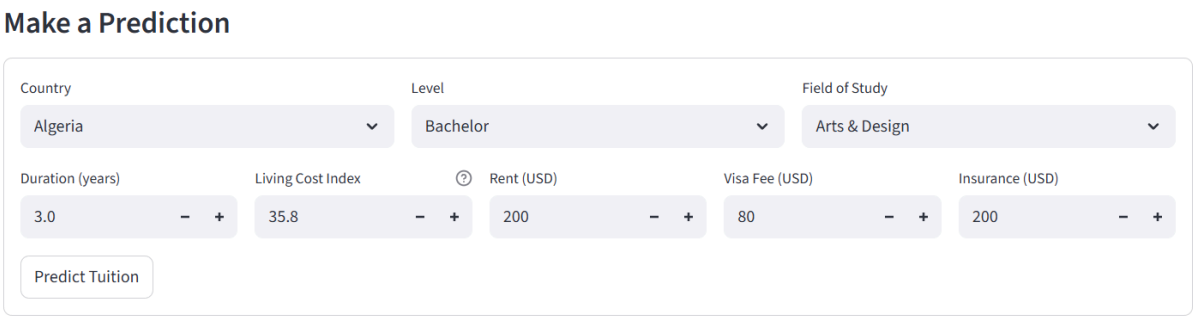


Figure 5 - Predicted vs. Actual Tuition Values

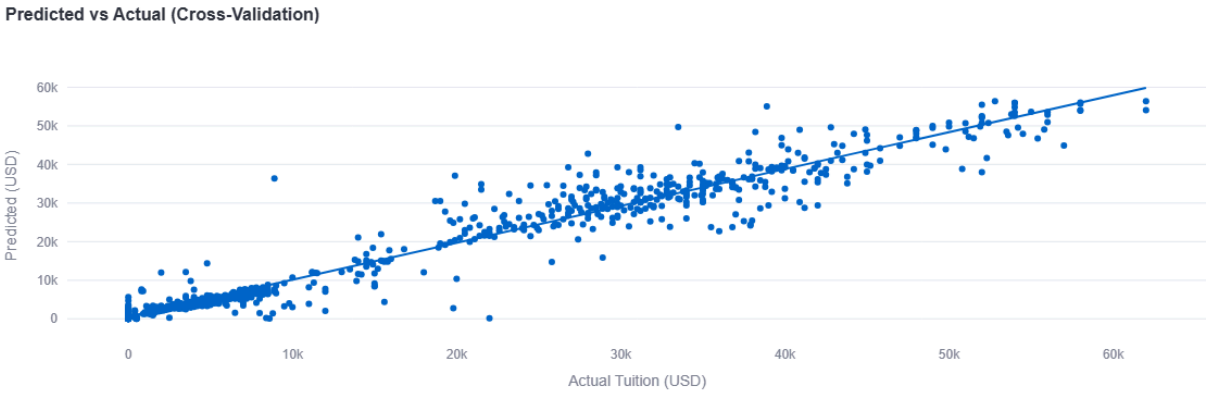


Figure 6 - Distribution of Residuals

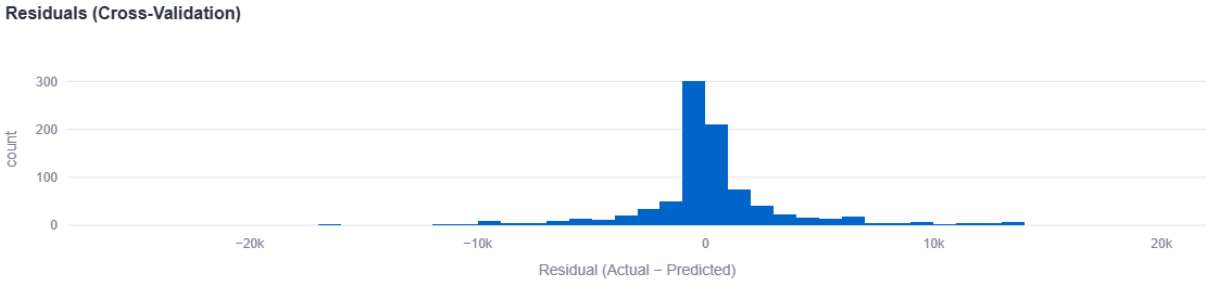
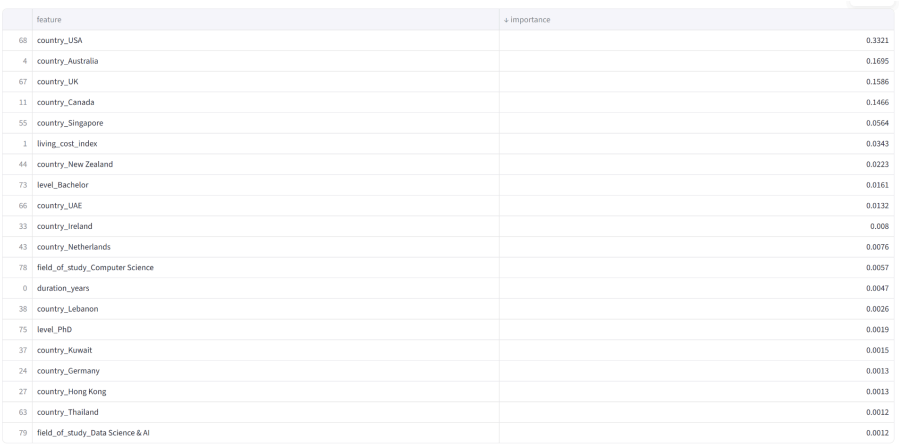


Figure 7 - Feature importances of predictors



	feature	↓ importance
68	country_USA	0.3321
4	country_Australia	0.1695
67	country_UK	0.1586
11	country_Canada	0.1466
55	country_Singapore	0.0564
1	living_cost_index	0.0343
44	country_New Zealand	0.0223
73	level_Bachelor	0.0161
66	country_UAE	0.0132
33	country_Ireland	0.008
43	country_Netherlands	0.0076
78	field_of_study_Computer Science	0.0057
0	duration_years	0.0047
38	country_Lebanon	0.0026
75	level_PhD	0.0019
37	country_Kuwait	0.0015
24	country_Germany	0.0013
27	country_Hong Kong	0.0013
63	country_Thailand	0.0012
79	field_of_study_Data Science & AI	0.0012

Figure 8. Performance Comparison of Regression Models

Model	RMSE (USD)	MAE (USD)	Median AE (USD)	R^2
Random Forest	3,738	2,064	760	0.949
Linear Regression	4,390	3,019	1,783	0.930
Ridge Regression	4,445	3,100	1,953	0.928

Among the three models, the **Random Forest Regressor** achieved the lowest error metrics (RMSE, MAE, and Median AE) and the highest explanatory power ($R^2 = 0.949$). Compared to Linear Regression ($R^2 = 0.930$) and Ridge Regression ($R^2 = 0.928$), Random Forest reduced RMSE by approximately **15%**, demonstrating its superior ability to capture the non-linear relationships driving tuition variation.