BRIDGING EDUCATION YIELDS: INTERNATIONAL TRENDS

An analysis of what aspects of a country affect its passport strength

By: Emre Arslan (earslan), Seyit Metin Barut (sbarut), Can Tulpar (ctulpar), Muhammad Omer Chaudhry (mchaud11)

BACKGROUND

As students from international backgrounds, we were drawn to the idea of working with international travel rights. From our own experience, we know that citizens of some counties enjoy these rights more, while others have to go through layers of bureaucracies and paid applications. As a result, we are aiming to assess what factors contribute to the strength (the number of visa-free travel destinations) of a country's passport. Especially, we wanted to focus on education levels, and economic strength of countries.

HYPOTHESIS

As a group, we test the following hypothesis: The educational and economic standings of a country, evaluated through education index, average IQ, GDP per capita, and import/exports, are directly correlated with the strength of that country's passport.

DATA COLLECTION

We worked with several datasets to test our hypothesis:

- Wikipedia data for passport indices, exports, and imports of countries
- World Population Review data for average IQ and literacy rate
- World Bank data for GDP per capita
- Data Pandas data for education rankings

We have acquired data regarding the mentioned features for world countries by downloading premade CSV files and scraping websites. Afterwards, we cleaned our data by changing all column names that denote the country (like 'country' or 'region') to 'Country,' dropped the years of GDP data we were not interested in (1950-2010), and removed duplicates. Finally, we were able to join the data by country and get the columns we were interested in per country. Though some columns had <null> values, we left dropping them to the analysis part to be performed based on the columns used, and ended up with 597 rows of values (3 years/rows of passport index data per country).

METHODOLOGY

To determine whether there were relationships between features like education index, average IQ, GDP per capita, and import/exports and passport strength, we performed statistical tests and machine learning model predictions. Mainly, we assessed pairwise relationships between features (high/low gdp, literacy, education) and passport strength through two sample t-tests and a chi squared independence test. Then, experimented by predicting passport strength with different combinations features using logistic regression with discretized data, linear regression, and k-means clustering machine learning algorithms, to understand where we can see correlations and groups.

ANALYSIS

- **HYPOTHESIS TESTS**

Claim #1: The education levels and passport indices of countries are not independent. To test this claim, we used a chi-squared independence test to see if there exists a statistical relationship between the two categorical variables.

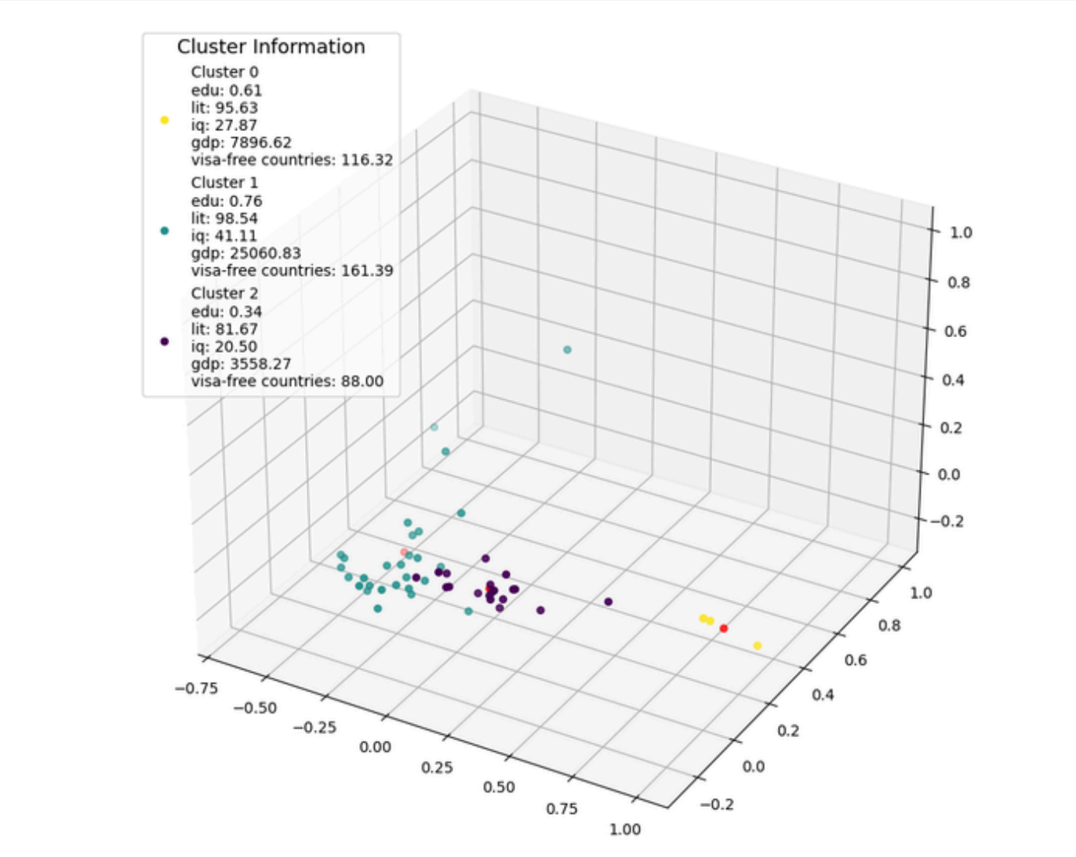
Claim #2: High average literacy rate and low average literacy rate countries have different passport indices. To test the hypothesis, we used a two sample t-test, as we wanted to compare the passport indices among two groups. We categorized the literacy rates higher than the mean as high, and those that are lower than the mean as low.

Claim #3: There is a significant difference in passport index between countries with high GDP per capita and low GDP per capita. We used a two sample t-test, as we wanted to compare the passport indices among two groups. We categorized the GDPs higher than the mean as high, and those that are lower than the mean as low.

EDUCATION V. PASSPORT STRENGTH	HIGH/LOW LITERACY RATE V. PASSPORT STRENGTH	HIGH/LOW GDP PER CAPITA V. PASSPORT STRENGTH
P = 0.89	P < 0.001	P < 0.001
ACCEPT NULL	REJECT NULL	REJECT NULL

For claim #1, we see that since p-value is bigger than 0.05, we fail to reject the null hypothesis and this means that we fail to determine a significant relationship between education and passport strength.
For claim #2 and claim #3, we can see that the p-value is less than 0.001 for both of these tests, so we reject both of the null hypothesis. This indicates that there is a significant difference in the number of visa-free destinations between countries with high/low literacy rate and high/low GDP per capita.

- **MACHINE LEARNING MODELS**
  - **CLUSTERING**

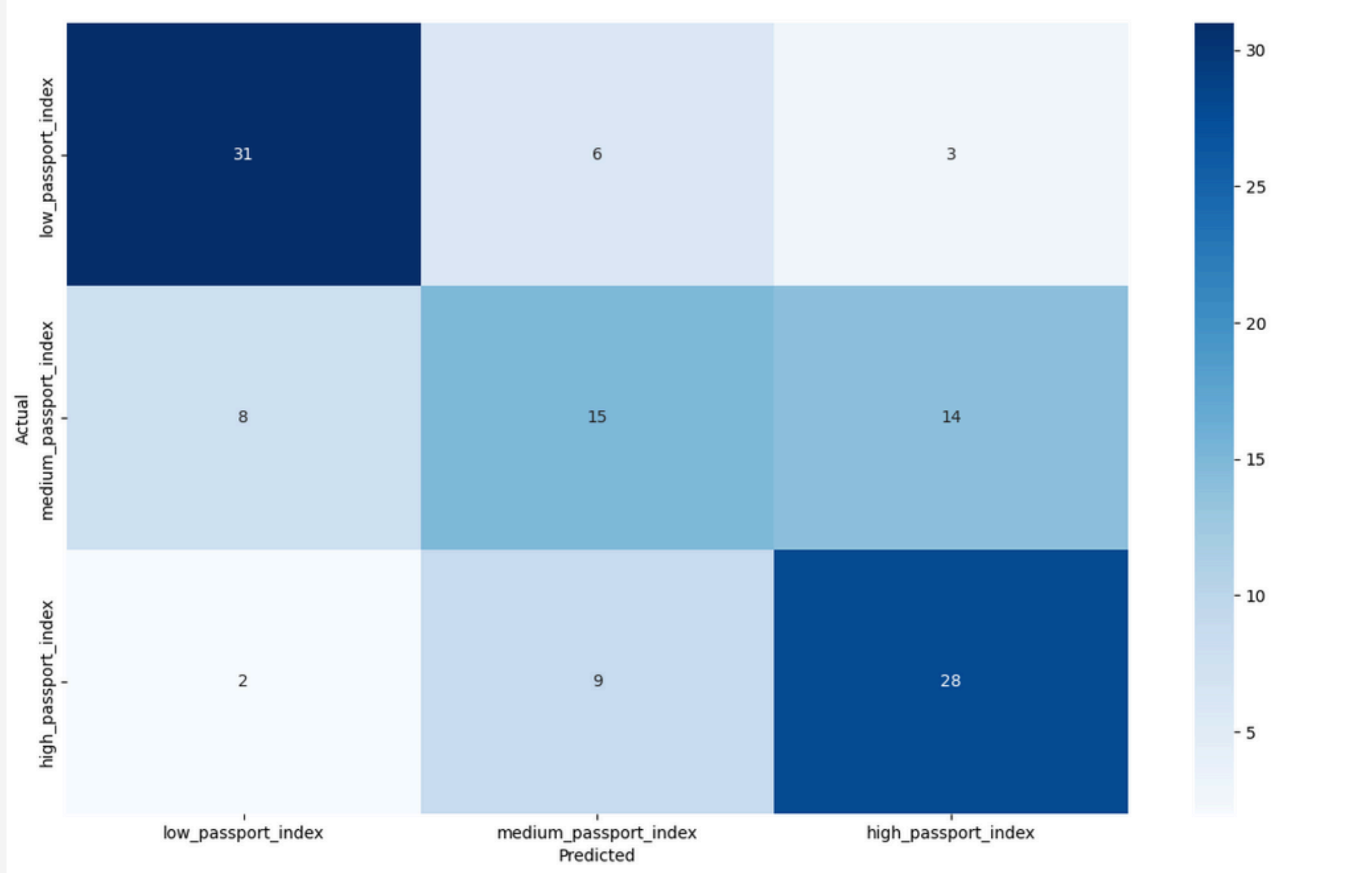


Using KMeans clustering, we analyzed three naturally forming groups. This helped us visualize the data, and reject or reinforce some of our assumptions such as assuming countries fall into three groups: low, mid, and high educational/economic status.

We trained our model using countries' education index, literacy rate, average IQ, and GDP per capita features. We did not include passport strength during clustering. However, after forming the clusters, we calculated "centroids" that included the mean values of all un-normalized features, including the passport strengths. Thus, we had the chance to observe what patterns we could see across clusters, on top of visualizing the clusters themselves.

Our results indicated that while we cannot definitively group countries into three groups, as shown by our low silhouette score used to measure clustering quality, there are quantitative patterns. Each cluster centroid does indeed have low/mid/high feature values and correspondingly, low/mid/high passport strength.

- **LOGISTIC REGRESSION**



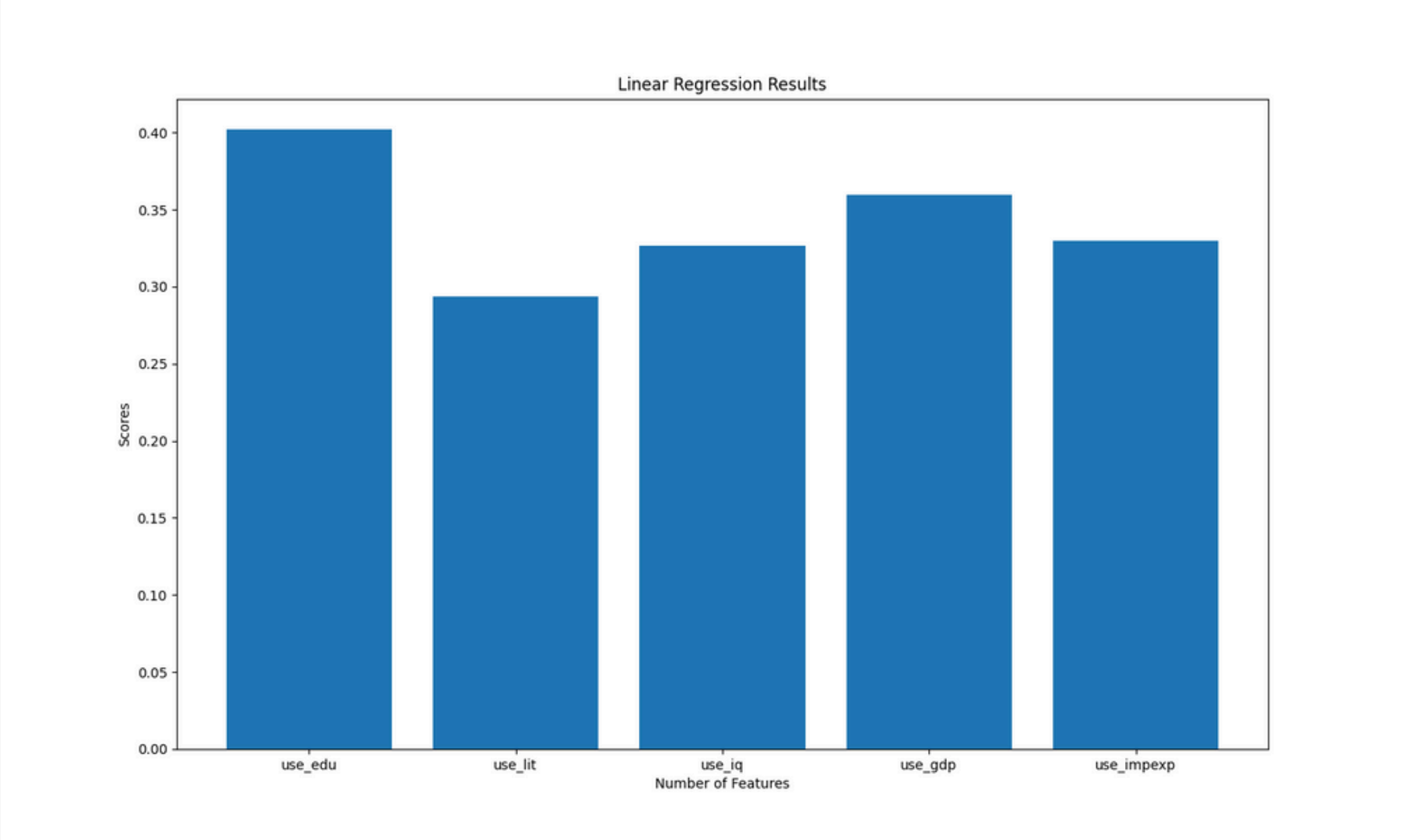
To understand the correlation between the features education index, literacy rate, average IQ, GDP per capita and the label passport strength, we built a logistic regression model as well as a linear regression model. This allowed us to predict the passport strength using the other four features. To effectively train our logistic regression model, we separated passport strength into three categories (class labels), indicating low, medium, and high passport strengths.

After experimentation, we saw that using all four features resulted in low model accuracy. However, using three by either excluding average IQ or literacy rate, we were able to get the confusion matrix seen here. Our model was able to generate true positives reasonably well, indicating a correlation between the three features and passport strength. Furthermore, our other experiments showed that this correlation is stronger than that of passport strength and just one of the features.

More specifically, we can see that our model predicted (31+15+28) correct out of 116 data points, which is a 63.8% accuracy rate. We can also see that our model was best in predicting low and high passport strengths (31 and 28 countries correctly predicted respectively) as opposed to medium strength (15).

- **LINEAR REGRESSION**

This bar graph shows the results of a linear regression analysis, where different features are evaluated for their contribution in predicting the "strength" or power of a country's passport. The y-axis represents a score metric, indicating the average contribution or importance of each feature in the prediction model. We can see that the "education index" was the "strongest" feature in prediction, and "GDP per capita" was the second. The literacy rate of the country seemed to have the least contribution in prediction. The features "average IQ" and "imports/exports" seem to have the same effect.



TAKEAWAYS

From our analysis results, we can see that while education might not directly be related to passport strength, when combined with other factors like GDP it shows a certain trend - countries with more educational opportunities and economic strength tend to have stronger passports.

Hypothesis Testing:
Education and passport strength do not show statistically significant dependence, while the variation of passport strength for countries with high/low literacy rates or GDP per capita is statistically significant.

Machine Learning:
Among the models and combinations of features we used, both linear and logistic regression models performed well when we divided passport strengths into three bins (classes), and when we used GDP and education data. Using four bins and including average IQ also yielded high validation scores. Therefore, we were able to estimate strong relationships between GDP, education, and IQ when the features are considered together.

LIMITATIONS/CHALLENGES

Since we are working with countries, we are limited in the sense that there are only around ~200, and we could have found more features, features with matching years, and possibly more years of passport information to augment and aggregate the data. We also had missing values for some countries in some columns. We could not use data on every single country (we still used most). Also, because we dropped rows if a column was missing an entry when we were joining tables and selecting features we want, we often ended up having with very few data points for our machine learning algorithms. This introduced a certain bias to our models.