**Problem statement and hypothesis**

“Why is it crucial to ensure that nearly 4 billion girls and women around the world have the same chances to receive an education as boys and men? First, education is a human right, enshrined in the Universal Declaration of Human Rights and the United Nations Convention on the Rights of the Child. Second, girls’ education is a strategic development investment – evidence shows that girls’ education brings a wide range of benefits not only for the girls themselves but also for their children and their communities, as well as society at large in terms of economic growth.

The systematic exclusion of girls and women from school and the labor force translates into a less educated workforce, inefficient allocation of labor, lost productivity, and consequently diminished progress in economic development.”[[1]](#endnote-1)

A country's education system often reflects the values of the country's central government. Insofar as a country provides social services for its inhabitants, how much a country provides and how much a people accepts what is provided are critically important features to consider. In other words, if a country does support an educational system for its citizens, knowing how much a country gives to this system is likely indicative of the importance that a system holds for a country. Additionally, Gender Gap in education continues to be an area of research and development, insofar as ensuring that girls are privy to at least the same amount of education that boys have been privy to across countries that have historically not offered equal opportunities.

Therefore, my initial hypothesis is that higher a country’s expenditure on secondary education as a percentage of government expenditure on education percentage, then the higher the school-life expectancy for girls in secondary education.

**Description of your data set and how it was obtained**

The data set I used was obtained from The World Bank[[2]](#endnote-2). The original data set contains approximately 3,000 internationally comparable indicators that, according to the World Bank, can be used to describe education access, progression, completion, literacy, teachers, population, and expenditures. The indicators cover the span of an education cycle from pre-primary through and including tertiary education. The World Bank has a user-friendly databank portal that I leveraged in order to query the data[[3]](#endnote-3).

The data set was made up of continuous variables and my hypothesis has been set up to be a classification problem. Therefore, I needed to create classes to predict. In order to do so, I reviewed the 3,000 features and discovered 10 that I wanted to leverage because of they are aggregate variables that represent underpinnings of the how a country values education as a whole (descriptions included are directly from the Data Dictionary). Data was available across all 10 features for only 27 countries.

**Countries:**

Bulgaria, Burkina Faso, Colombia, Cyprus, Ecuador, Finland, France, Guinea, Guyana, Hungary, Indonesia, Jamaica, Japan, Latvia, Malaysia, Mali, Mexico, Moldova, Netherlands, Peru, Poland, Serbia, Slovenia, Spain, Sweden, Thailand, United States

**Continuous Variables Used for Creating Classes:**

* **All staff compensation as % of total expenditure in secondary public institutions (%)**
  + All staff (teacher and non-teachers) compensation expressed as a percentage of direct expenditure in public educational institutions (instructional and non-instructional) of the specified level of education. Financial aid to students and other transfers are excluded from direct expenditure. Staff compensation includes salaries, contributions by employers for staff retirement programmes, and other allowances and benefits. Divide all staff compensation in public institutions of a given level of education (ex. primary, secondary, or all levels combined) by total expenditure (current and capital) in public institutions of the same level of education, and multiply by 100. For more information, consult the UNESCO Institute of Statistics website: http://www.uis.unesco.org/Education/
* **Duration of compulsory education (years)**
  + Number of years that children are legally obliged to attend school.
* **Expenditure on education as % of total government expenditure (%)**
  + Total general (local, regional and central) government expenditure on education (current, capital, and transfers), expressed as a percentage of total general government expenditure on all sectors (including health, education, social services, etc.). It includes expenditure funded by transfers from international sources to government. Public education expenditure includes spending by local/municipal, regional and national governments (excluding household contributions) on educational institutions (both public and private), education administration, and subsidies for private entities (students/households and other privates entities). In some instances data on total public expenditure on education refers only to the ministry of education and can exclude other ministries that spend a part of their budget on educational activities. The indicator is calculated by dividing total public expenditure on education incurred by all government agencies/departments by the total government expenditure and multiplying by 100. For more information, consult the UNESCO Institute of Statistics website: <http://www.uis.unesco.org/Education/>
* **Expenditure on secondary as % of government expenditure on education (%)**
  + Expenditure on education by level of education, expressed as a percentage of total general government expenditure on education. Divide government expenditure on a given level of education (ex. primary, secondary) by total government expenditure on education (all levels combined), and multiply by 100. A high percentage of government expenditure on education spent on a given level denotes a high priority given to that level compared to others. When interpreting this indicator, one should take into account enrolment at that level, and the relative costs per student between different levels of education. For more information, consult the UNESCO Institute of Statistics website: <http://www.uis.unesco.org/Education/>
* **Expenditure on secondary as % of total government expenditure (%)**
  + Total general (local, regional and central) government expenditure on secondary education (current, capital, and transfers), expressed as a percentage of total general government expenditure on all sectors (including health, education, social services, etc.). It includes expenditure funded by transfers from international sources to government. Divide total government expenditure for a given level of education (ex. primary, secondary, or all levels combined) by total general government expenditure (all sectors), and multiply by 100. A higher percentage of government expenditure on education shows a high government priority for education relative to other public investments. When interpreting this indicator however, one should keep in mind that some governments have more (or less) means and therefore larger (or smaller) overall budgets, and that countries with younger populations may spend more on education in relation to other sector such as health or social security, and vice-versa. For more information, consult the UNESCO Institute of Statistics website: <http://www.uis.unesco.org/Education/>
* **Government expenditure in educational institutions as % of GDP (%)**
  + Total general (local, regional and central) government expenditure in educational institutions (current and capital), expressed as a percentage of GDP. It excludes transfers to private entities such as subsidies to households and students, but includes expenditure funded by transfers from international sources to government. Divide total expenditure in public institutions by the GDP, and multiply by 100. For more information, consult the UNESCO Institute of Statistics website: [http://www.uis.unesco.org/Education/<br><br](http://www.uis.unesco.org/Education/%3cbr%3e%3cbr)>
* **Government expenditure in secondary institutions education as % of GDP (%)**
  + Total general (local, regional and central) government expenditure in educational institutions (current and capital) at a given level of education, expressed as a percentage of GDP. It excludes transfers to private entities such as subsidies to households and students, but includes expenditure funded by transfers from international sources to government. Divide total expenditure in public institutions of a given level of education (ex. primary, secondary, or all levels combined) by the GDP, and multiply by 100. For more information, consult the UNESCO Institute of Statistics website: <http://www.uis.unesco.org/Education/>
* **Government expenditure on education as % of GDP (%)**
  + Total general (local, regional and central) government expenditure on education (current, capital, and transfers), expressed as a percentage of GDP. It includes expenditure funded by transfers from international sources to government. Divide total government expenditure for a given level of education (ex. primary, secondary, or all levels combined) by the GDP, and multiply by 100. A higher percentage of GDP spent on education shows a higher government priority for education, but also a higher capacity of the government to raise revenues for public spending, in relation to the size of the country's economy. When interpreting this indicator however, one should keep in mind in some countries, the private sector and/or households may fund a higher proportion of total funding for education, thus making government expenditure appear lower than in other countries. For more information, consult the UNESCO Institute of Statistics website: http://www.uis.unesco.org/Education/
* **Net enrolment rate, secondary, both sexes (%)**
  + Total number of students in the theoretical age group for secondary education enrolled in that level, expressed as a percentage of the total population in that age group. Divide the number of students enrolled who are of the official age group for secondary education by the population for the same age group and multiply the result by 100. NER at each level of education should be based on enrolment of the relevant age group in all types of schools and education institutions, including public, private and all other institutions that provide organized educational programmes.
* **Pupil-teacher ratio in secondary education (headcount basis)**
  + Average number of pupils per teacher at a given level of education, based on headcounts of both pupils and teachers. Divide the total number of pupils enrolled at the specified level of education by the number of teachers at the same level. In computing and interpreting this indicator, one should take into account the existence of part-time teaching, school-shifts, multi-grade classes and other practices that may affect the precision and meaningfulness of pupil-teacher ratios. When feasible, the number of part-time teachers is converted to a full-time equivalent teachers; a double-shift teacher is counted twice, etc. Teachers are defined as persons whose professional activity involves the transmitting of knowledge, attitudes and skills that are stipulated in a formal curriculum programme to students enrolled in a formal educational institution.

Once I created my classes, I wanted to use gender-specific data about secondary education and see if the gender-specific features could predict the classifications. Thus, I used the following features for the 27 countries identified.

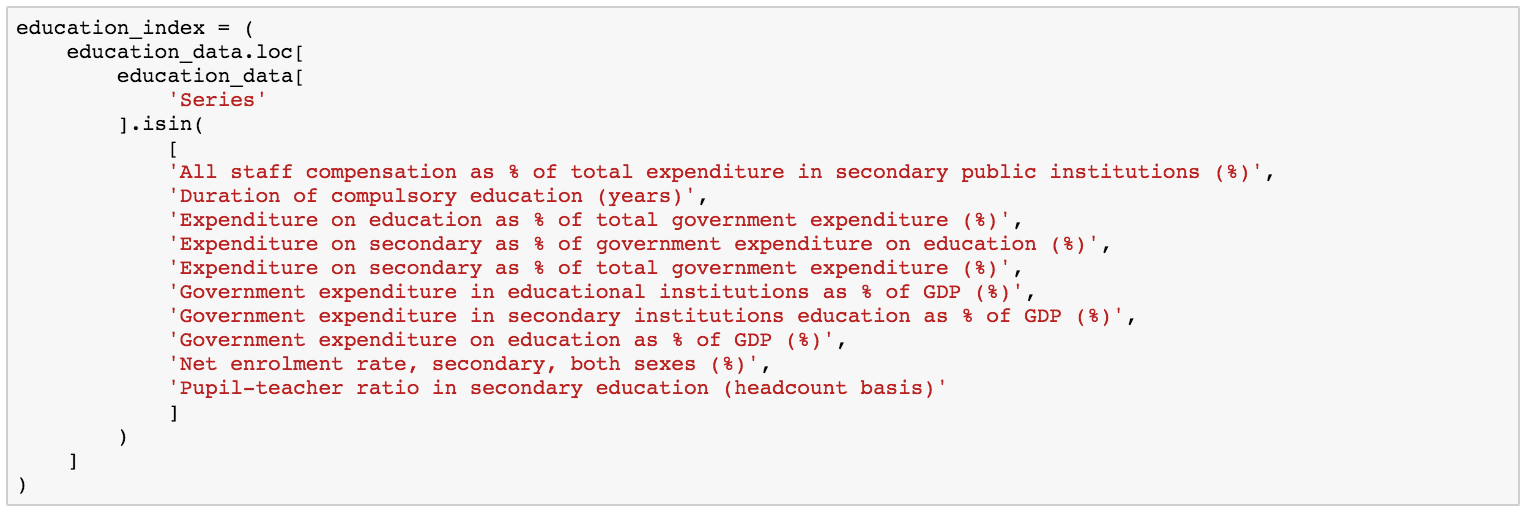
**Continuous Features Used for Class Prediction:**

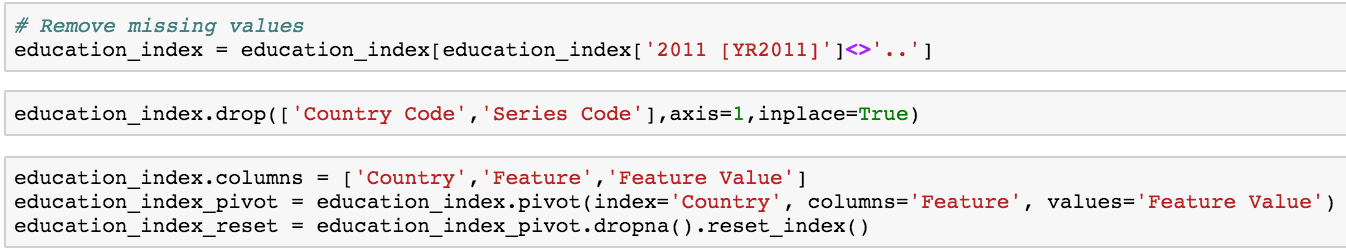
* **Net enrolment rate, secondary, gender parity index (GPI)**
  + Ratio of female net enrolment rate for secondary to the male net enrolment rate for secondary. It is calculated by dividing the female value for the indicator by the male value for the indicator. A GPI equal to 1 indicates parity between females and males. In general, a value less than 1 indicates disparity in favor of males and a value greater than 1 indicates disparity in favor of females.
* **Net enrolment rate, secondary, female (%)**
  + Total number of female students in the theoretical age group for secondary education enrolled in that level, expressed as a percentage of the total female population in that age group. Divide the number of female students enrolled who are of the official age group for secondary education by the female population for the same age group and multiply the result by 100. NER at each level of education should be based on enrolment of the relevant age group in all types of schools and education institutions, including public, private and all other institutions that provide organized educational programmes.
* **Net enrolment rate, secondary, male (%)**
  + Total number of male students in the theoretical age group for secondary education enrolled in that level, expressed as a percentage of the total male population in that age group. Divide the number of male students enrolled who are of the official age group for secondary education by the male population for the same age group and multiply the result by 100. NER at each level of education should be based on enrolment of the relevant age group in all types of schools and education institutions, including public, private and all other institutions that provide organized educational programmes
* **School life expectancy, secondary, gender parity index (GPI)**
  + Ratio of female school life expectancy to the male school life expectancy. It is calculated by dividing the female value for the indicator by the male value for the indicator. A GPI equal to 1 indicates parity between females and males. In general, a value less than 1 indicates disparity in favor of males and a value greater than 1 indicates disparity in favor of females.
* **School life expectancy, secondary, female (years)**
  + Number of years a person of school entrance age can expect to spend within the specified level of education. For a child of a certain age a, the school life expectancy is calculated as the sum of the age specific enrolment rates for the levels of education specified. The part of the enrolment that is not distributed by age is divided by the school-age population for the level of education they are enrolled in, and multiplied by the duration of that level of education. The result is then added to the sum of the age-specific enrolment rates. A relatively high SLE indicates greater probability for children to spend more years in education and higher overall retention within the education system. It must be noted that the expected number of years does not necessarily coincide with the expected number of grades of education completed, because of repetition. Since school life expectancy is an average based on participation in different levels of education, the expected number of years of schooling may be pulled down by the magnitude of children who never go to school. Those children who are in school may benefit from many more years of education than the average.
* **School life expectancy, secondary, male (years)**
  + Number of years a person of school entrance age can expect to spend within the specified level of education. For a child of a certain age a, the school life expectancy is calculated as the sum of the age specific enrolment rates for the levels of education specified. The part of the enrolment that is not distributed by age is divided by the school-age population for the level of education they are enrolled in, and multiplied by the duration of that level of education. The result is then added to the sum of the age-specific enrolment rates. A relatively high SLE indicates greater probability for children to spend more years in education and higher overall retention within the education system. It must be noted that the expected number of years does not necessarily coincide with the expected number of grades of education completed, because of repetition. Since school life expectancy is an average based on participation in different levels of education, the expected number of years of schooling may be pulled down by the magnitude of children who never go to school. Those children who are in school may benefit from many more years of education than the average.

**Description of any pre-processing steps you took**

In order to perform the machine-learning algorithms, there were several steps of pre-processing:

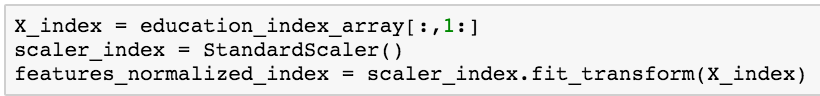
1. Data-Munging & Pre-Processing for Class-Labels
   1. Selecting Data



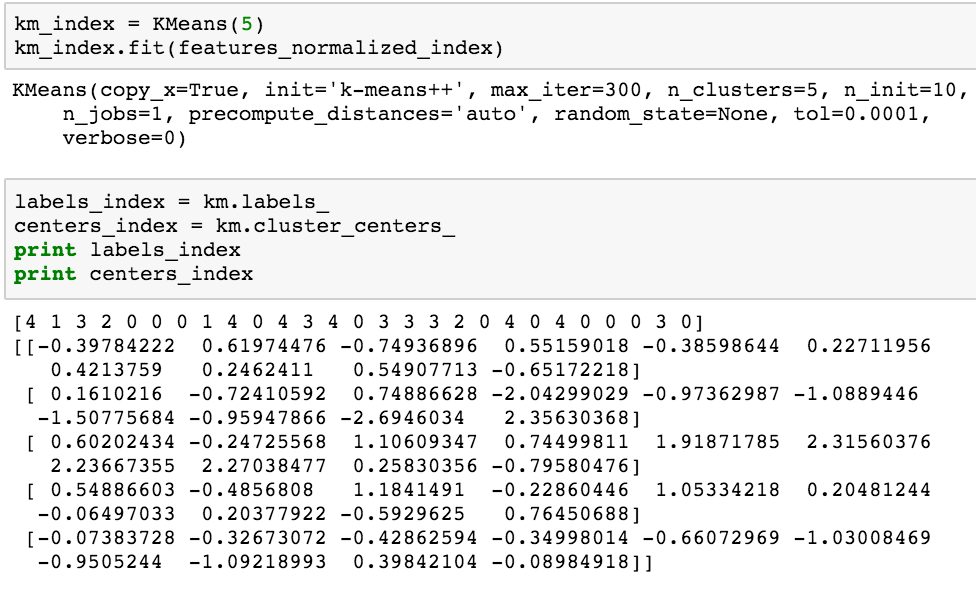
* 1. Removing missing values, dropping unnecessary columns, renaming columns, pivoting the dataframe, and ultimately the index. After these steps I went from 214 countries of data to 27.

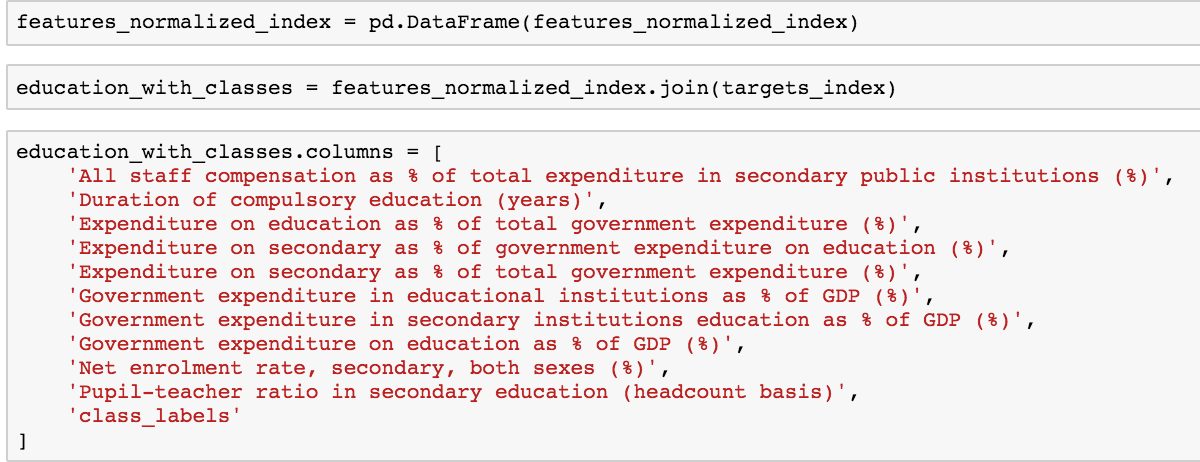
1. Before creating classes, I needed to rescale my data to standardize features by removing the mean and scaling to unit variance. Before scaling, however, I needed to transform my dataframe into an array (which I did using data.**as\_matrix()** in order for the **StandardScaler()** method to process the data) .





1. In order to derive class-labels, I identified that this was an unsupervised learning problem using continuous features. Therefore, I applied a clustering algorithm (KMeans) to create new classes. I used KMeans hoping that it would yield natural groups.



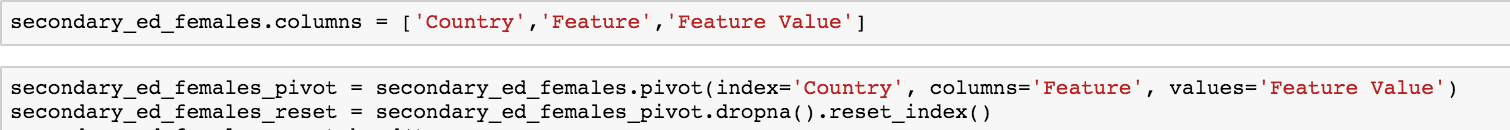
1. With my 5 new class labels, I needed to transform my ndarry back to a dataframe and match each Mean with its respective feature (N.B. I needed to manually rename).

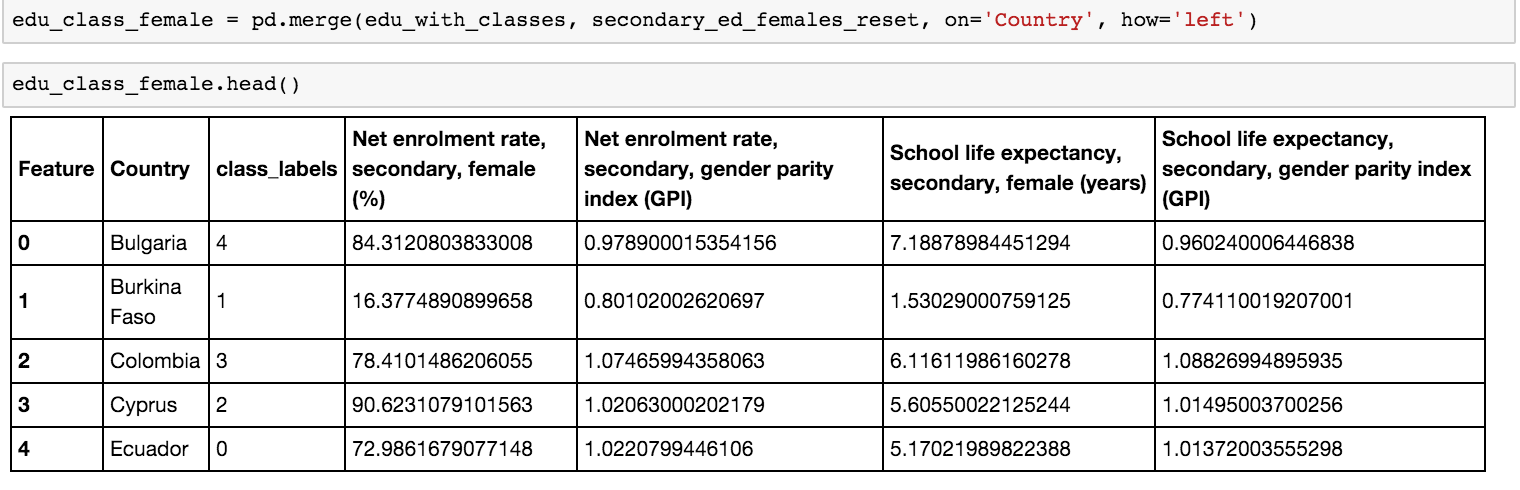
1. To complete the picture, I added the names of the countries back to the dataframe as well:

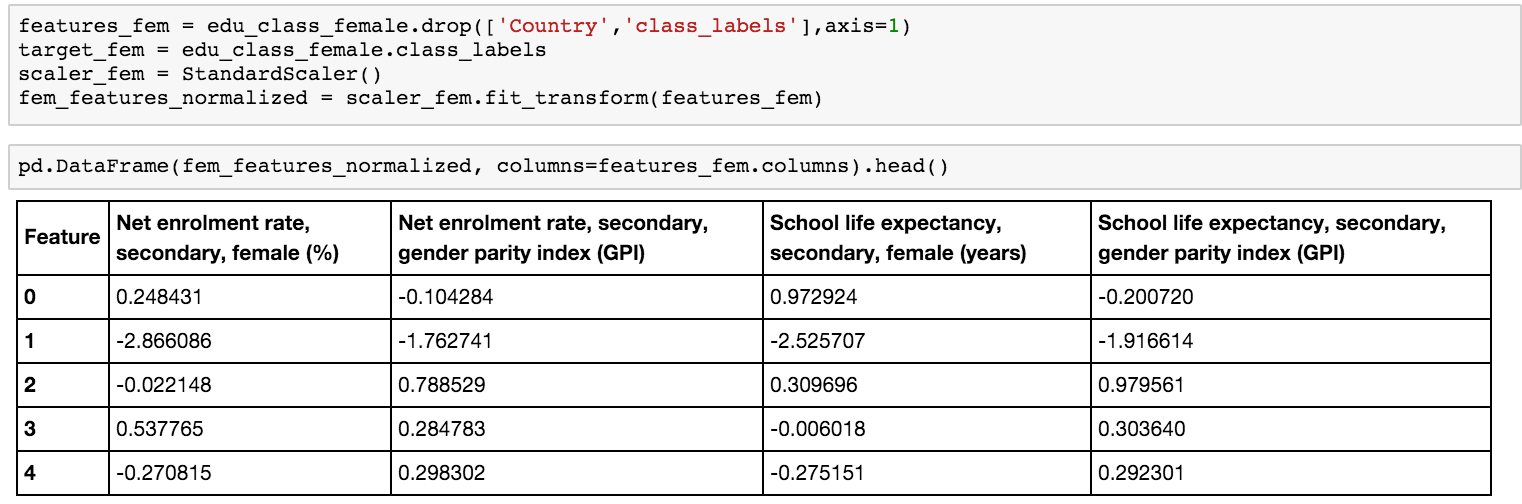


1. Below is the new dataframe, highlighting just the country names and class labels:
2. Data-Munging & Pre-Processing for Features Used in Predicting Classification (steps shown are for gather female-gender data. Same steps were followed for gathering male-gender data. See python code for details)
   1. Selecting Data



* 1. Removing missing values, dropping unnecessary columns, renaming columns, pivoting the dataframe, and ultimately the index.
  2. Merge class-labels with continuous features to be used for predictions:

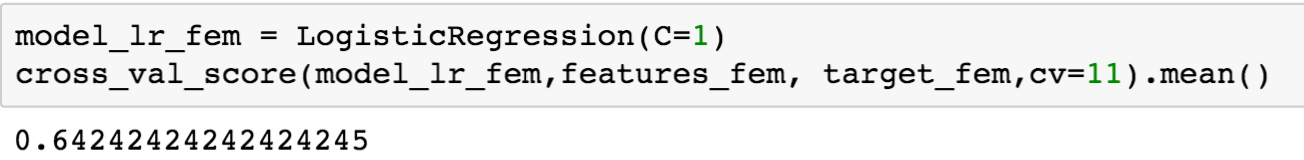


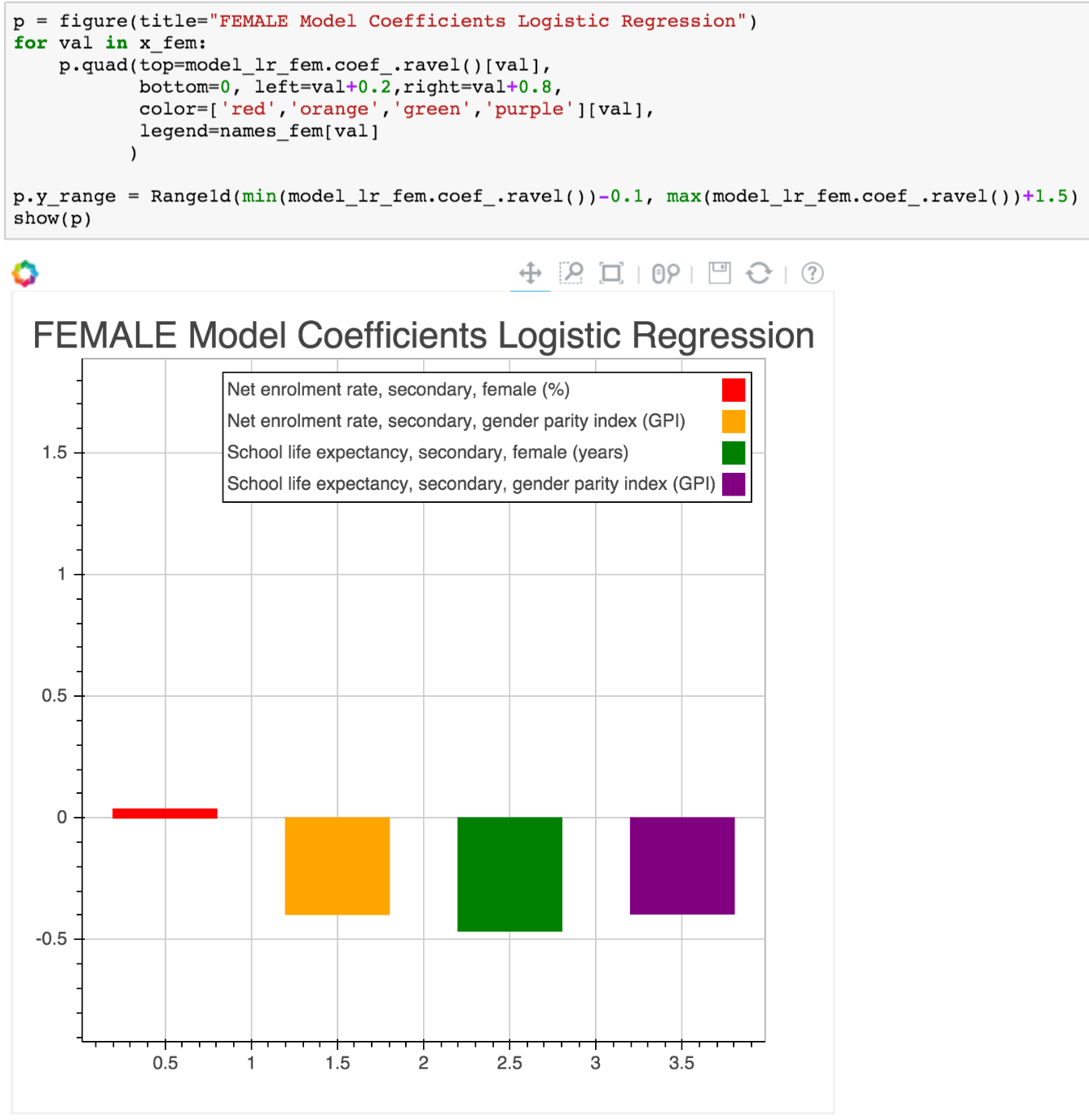
* 1. Before engaging classifiers, I needed to rescale my data to standardize features by removing the mean and scaling to unit variance.

**Details of your modeling process, including how you selected your models and validated them**

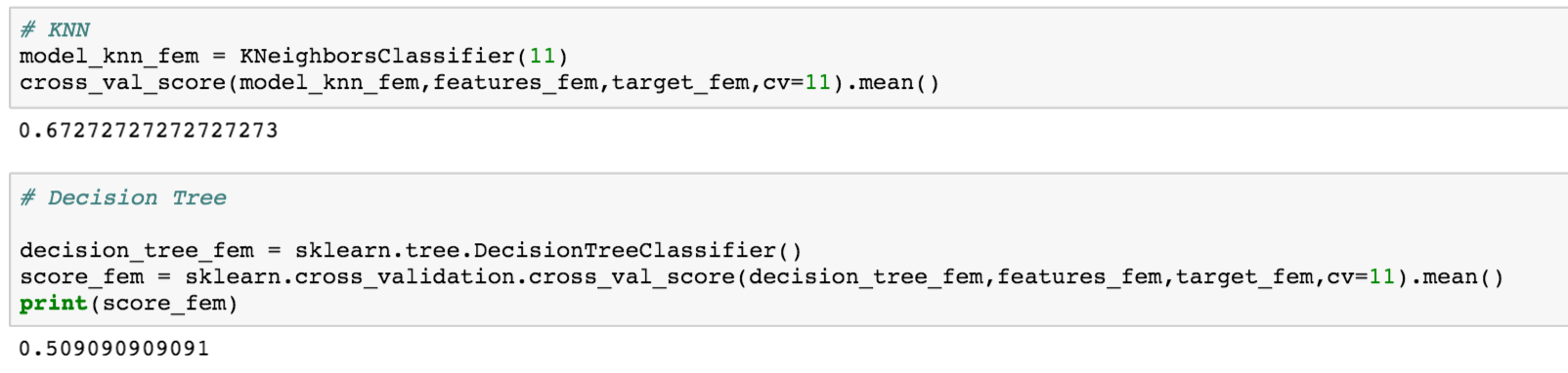
The procedures I used to model were very straightforward applications of 3 classification techniques: Logistic Regression, K-Nearest Neighbors, and Decision Trees. Through trial-and-error I discovered that using 11 Cross-Fold Validations was the optimum number to get the highest accuracy of my model scores, across the 3 techniques.

Below is the code and outcome of each application:

After plotting the features on the bar chart below I also discovered that **Net Enrolment Rate** held the highest predictive power for getting the 64.2% accuracy score:



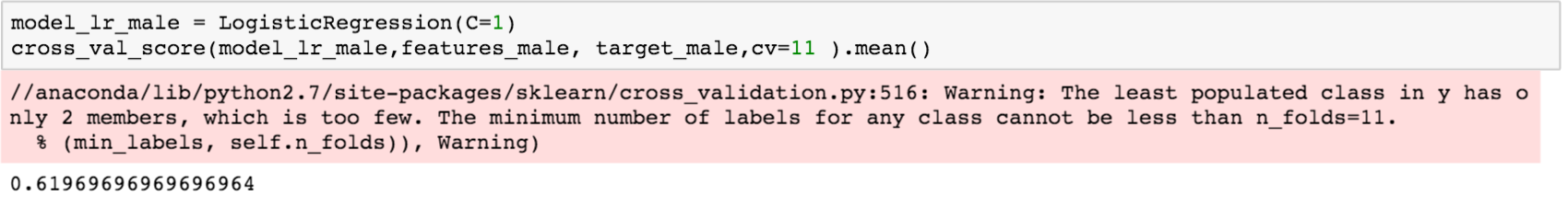
Next, I ran the two remaining classification algorithms, which produced the following results:



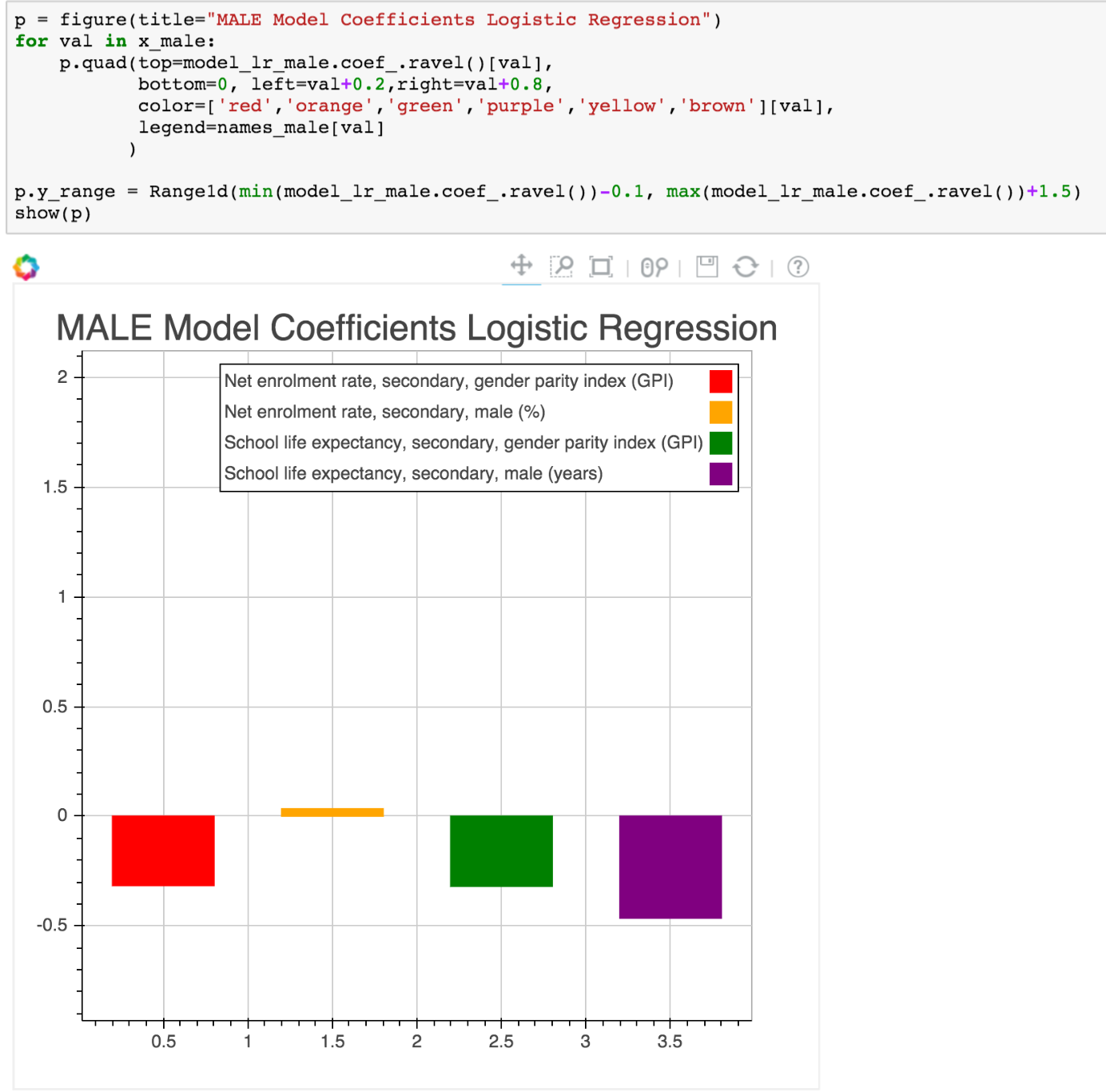
Per the results above, K-Nearest Neighbors yielded the highest accuracy score with 67.2%. This means that, based on the features I used to understand secondary education of females in the 27 countries evaluated the features predicted class membership accurately 67.2% of time.

Likewise, to the procedures followed for females, I used the same steps for the feature spaces I created for secondary education of males, across the 27 countries in question.

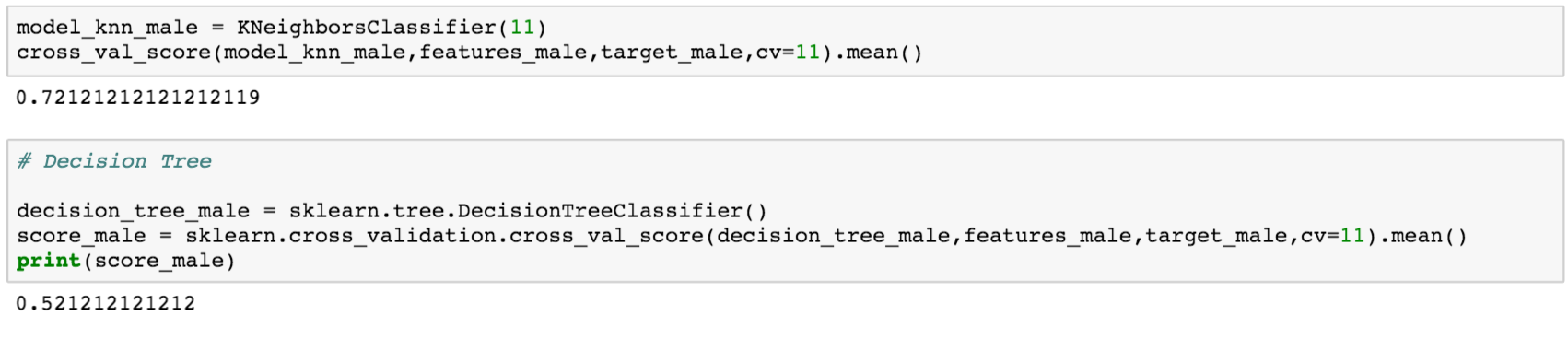
Below is the code and outcome of each application:



After plotting the features on the bar chart below I also discovered that **Net Enrolment Rate** held the highest predictive power for getting the 61.9% accuracy score:



Next, I ran the two remaining classification algorithms, which produced the following results:

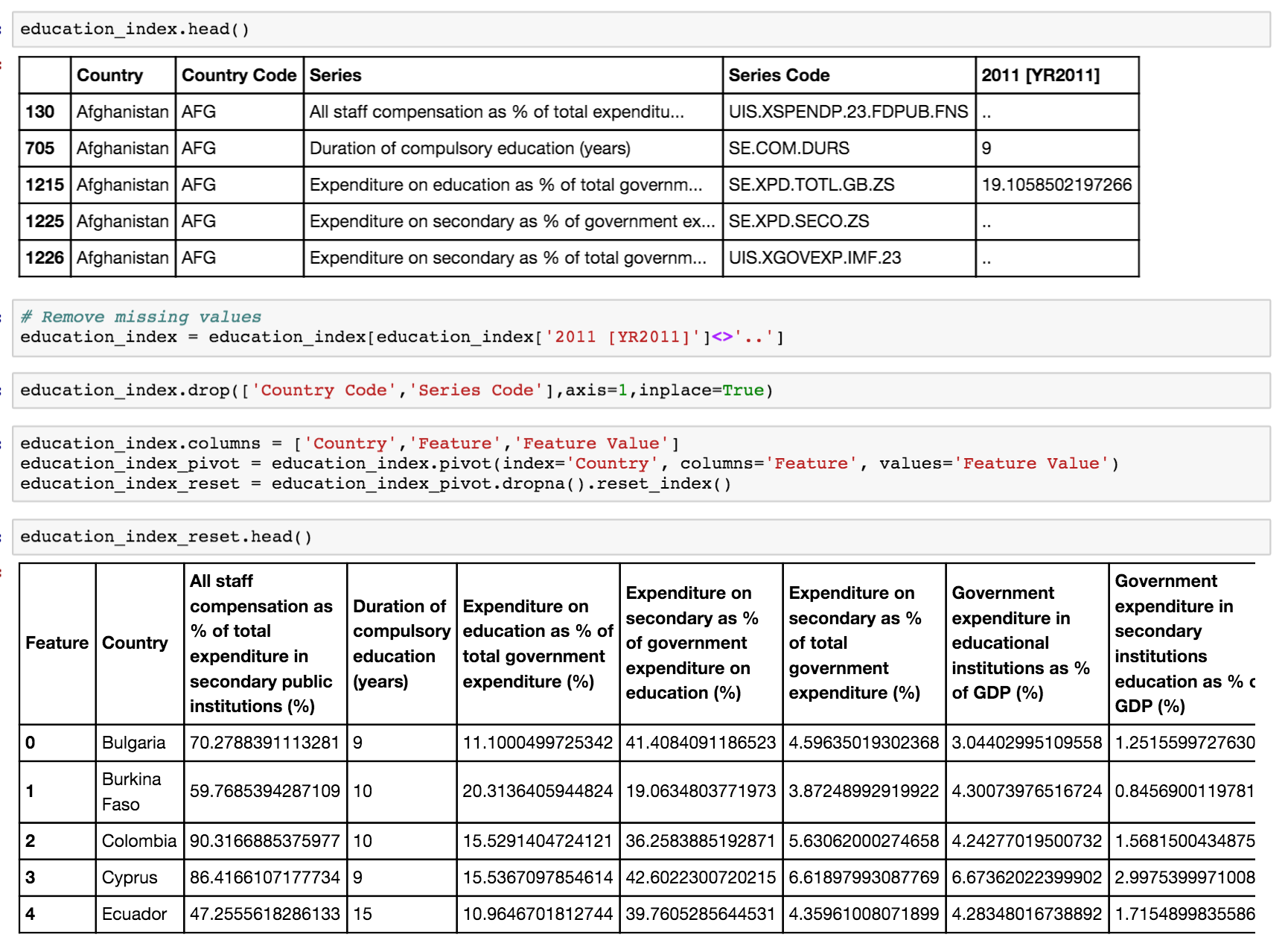


Per the results above, K-Nearest Neighbors yielded the highest accuracy score with 72.1%. This means that, based on the features I used to understand secondary education of males in the 27 countries evaluated the features predicted class membership accurately 72.1% of time.

**Your challenges and successes**

The greatest challenge of the assignment was defining a question. In preparation of settling on this topic, I explored several others, but found the same stopping points along the way: “Does data for this exist? Is the data that does exist useful? Do I have a clear definition of what I’m trying to solve?” Ultimately I’m very happy with this topic.

Munging the data was difficult as well. Eventually I found articles or received answers to questions I posted in stack-overflow[[4]](#endnote-4). In addition to transforms, pivots, and joins datasets another related challenge was identifying and controlling for the structure the data needed to be in in order to move along in answering my question. A perfect example of this is displayed below:



In the picture above it’s clear that the data in its original format was not going to be successfully operated on. Thus, after searching for literature and relevant tutorials, I was able to correctly manipulate and preprocess my data in the necessary form.

When I began the project I set out to “group all countries together with similar education indexes.” In theory this is what I ended up doing to create my class labels, yet in preparation I wasn’t quite sure what this would achieve. I thought the data would “speak” to me. This was a lesson in understanding how clear my question needed to be so that I could tell the data how to move in effort to get answers – or at least proxy answers – to the questions I had. Out of this spurred the notion of creating clusters through the application of the unsupervised learning algorithm of K-Nearest Neighbors. Planning ahead, I wondered what having clusters would “tell me” about a country’s educational profile and thus decided to leverage these clusters as class labels and use new features to predict accuracy of the class labels. The critical motivation of this was that I wanted to know where there are gaps in a country’s education infrastructure. My hope is that with this knowledge I can in some way help in a country or some countries educational reform and evolution.

An exciting success of this project was that after I ran my classifiers on my class labels I discovered that School life expectancy, secondary, (fe)male (years)holds the most predictive power for determining class membership. More specifically, as the project is focused on gender gap in secondary education, I looked into which countries had the lowest coefficients for this feature in order to see which countries essentially had the biggest gap. Guinea and Mali emerged from that exercise. Through additional reading I found that there is much literature documenting how imbalanced these countries are in their respective educating of girls, particularly in secondary education. Although I’m not an industry expert, I was pleased to see that my model made an accurate prediction about the real-world problem I was digging into.

**Possible extensions or business applications of your project**

Using unsupervised learning to understand a problem in the world or in business is exciting as well as daunting. I will likely be using this approach in the work I do at Facebook regarding payments and commerce initiatives that we work on. I will complement this work by reading thoroughly on the topic and/or consulting with researchers who are experts in the payments and commerce fields.

More specifically germane to the topic of my research, I’m planning to extend the investigation and potentially participate in a hackathon geared toward education in the developing world. I’d love to use the framework above or perhaps an even more rigorous one to understand the state of education for girls, set baseline metrics around failures in the system, and even work with policy makers toward improving.

**Conclusions and key learnings**

Education is not egalitarian and it’s imperative that it be reassessed. With resources like the World Bank, there is a deep well of data we can mine in order to understand and improve the world. I’m happy that this course gave me the tools, opportunity, and encouragement to pursue my passion in Data Science, which I plan to use to improve lives however I can.

1. <http://www.worldbank.org/en/topic/education/brief/girls-education> [↑](#endnote-ref-1)
2. <http://data.worldbank.org/data-catalog/ed-stats> [↑](#endnote-ref-2)
3. <http://databank.worldbank.org/data/reports.aspx?source=education-statistics-~-all-indicators> [↑](#endnote-ref-3)
4. <http://stackoverflow.com/users/5808710/levine?tab=questions> [↑](#endnote-ref-4)