Bianca Ida Pedersen 20/1-2024

Titel: Predicting Women STEM Graduates through IT-Linked Social Factors in European Countries

Introduction

The underrepresentation of women in Science, Technology, Engineering and Math (STEM) has long been a concern worldwide. As a country the number of women in STEM not only embodies principles of equity but this number might also be a sign of the societal progress in the country. This project seeks to delve into a dataset on European countries investigating the integration of digital technologies and the prevalence of digital public services as potential predictors for the number of women STEM graduates in a country. Through the data analysis, predictive modelling, and thoughtful interpretation this project aspires to shed light on the interplay between societal aspects and the number of women choosing STEM education.

Research question:

Can the number of women who graduate from STEM field educations in a country be predicted based on social factors related to IT?

Hypothesis:

"The number of women STEM graduates for each county is based on social factors related to IT such as integration of digital technologies and amount of digital public services."

Data collection

For the gathering of the data needed for this project, I turned to Eurostat(ec.europa.eu/Eurostat), a database managed by the European Union which provides statistical information on Europe. The Eurostat database is a user-friendly open data source where I was able to collect the data freely. The

Eurostat database website[1] was a reliable go-to source due to its comprehensive collection of statistics covering various fields of subjects of the European countries. By using Eurostat, I received access to a diverse range of statistical information. This was important for the project if it at some point needed more data for the project analysis. Since I ended up adding more data to my analysis, I was able to access it more easily to the project through Eurostat than it would have been through another database.

To do this analysis I started out retrieving data on women graduating in STEM field educations.

Through the Eurostat database, I was able to access data on graduates based on sex and field of education[2]. I filtered the data by selecting 'women' as sex and I selected educations within the fields of STEM. This gave me a TSV file with the countries and their number of women graduates in the year 2021. This number was not fair and would not be good enough for my analysis since every European country has a different population. I, therefore, needed to process this data to show a normalized number which was scaled to account for the size of the population of each European country. I gathered data on each country's population in the same year 2021[3]. The population was then used to calculate the number of women STEM graduates per 1000 people for each country.

For the second part of my data collecting, I needed to find data which could be used to predict the amount of women STEM graduates. Through Eurostat, it was possible to find a lot of data connected to IT due to their Digital Economy and Society Index (DESI) where they focus Europe's digital performance which they base on four dimensions: Human capital, Connectivity, Integration of digital technology and Digital public services[4].

---

[1] https://ec.europa.eu/eurostat (20/1-2024)
[2] https://ec.europa.eu/eurostat/databrowser/view/educ_uoe_grad02/default/table?lang=en (20/1-2024)
[3] https://ec.europa.eu/eurostat/databrowser/view/tps00001/default/table?lang=en. (20/1-2024)
[4] https://digital-decade-desi.digital-strategy.ec.europa.eu/datasets/desi-2022/charts (17/1-2024)

Together with the data collected on women STEM graduates and the level of AI use in enterprises the initial features selected for the analysis were:

- Internet use (percentage of individuals who used the internet "*within the last 3 months*")[5]
- Internet access (percentage of individuals who accessed the internet away from work and home/mobile internet access)[6]

Exploratory data analysis:

In some of the initial stages of this research, a crucial step involved conducting an Exploratory Data Analysis. This would help to illuminate key insight into the variables chosen. I used Python to create histograms showing each country and the number of women STEM graduates per 1000 people in these European countries as shown in Figure 1.

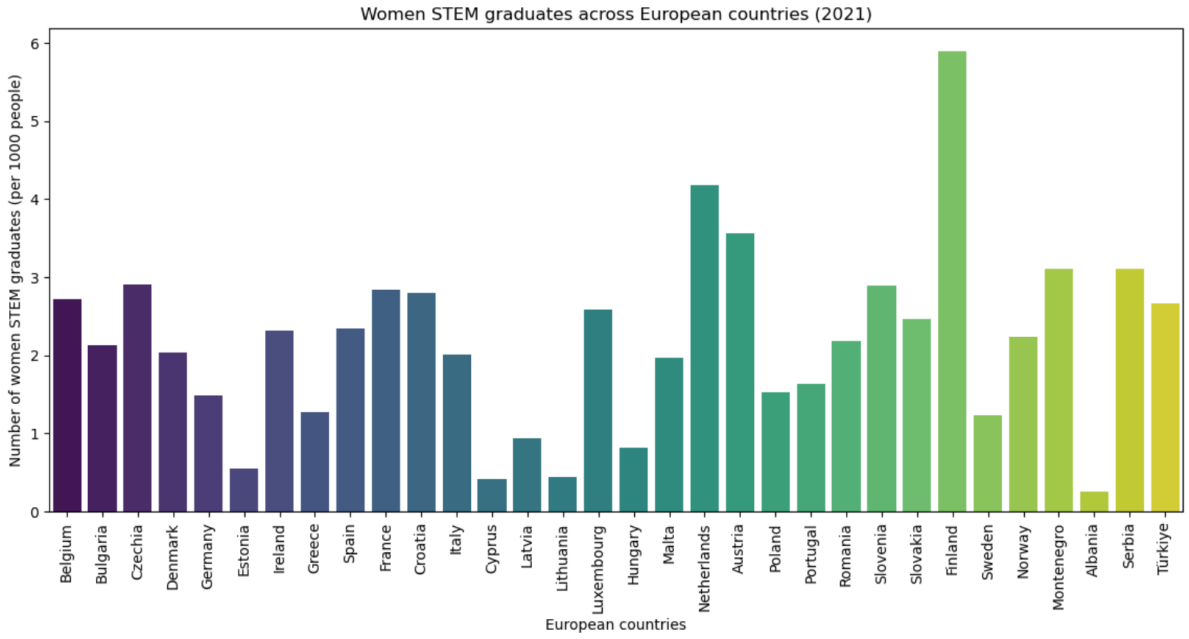

Figure 1 - Women STEM graduates across European countries (2021) - made with Python.

[5] https://ec.europa.eu/eurostat/databrowser/view/tin00028/default/table?lang=en (20/1-2024)
[6] https://ec.europa.eu/eurostat/databrowser/view/isoc_ci_im_i/default/table?lang=en (19/1-2024)

I did the same plot on the countries and the level of artificial intelligence use in enterprises as shown in Figure 2. These histograms visually showcased the distribution of these variables across different countries.

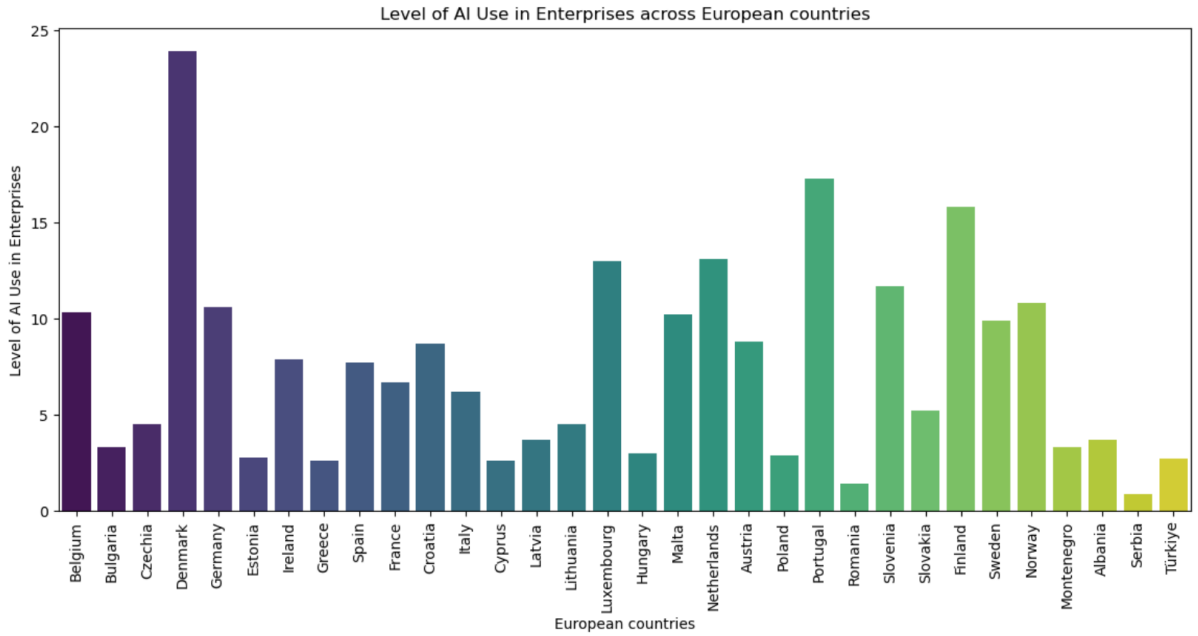

Figure 2 - Level of AI Use in Enterprises across European countries (2021) - made with Python

The purpose of creating these histograms is to visualize the spread and concentration of women STEM graduates and AI adoption levels so that I could identify any potential outliers, observe any trends and highlight variations in the data. I did the same later in the project when I realized that it would benefit my analysis to add more features. This will be explained later in this report. This helped me lay the groundwork for subsequently formulating my hypothesis and understanding my data.

Models and results

Linear regression

This project focuses on investigating the predictive relationship between women STEM graduates and some key social indicators, including the integration of digital technologies and digital public

services across European countries. I therefore chose to employ Linear regression to model this potential predictive link. Linear regression is suited for this analysis since it will provide a straightforward means to quantify links between the features chosen.

Initial Linear Regression Model

Early in the employment of this model, there were only 3 features to predict from the Level of AI use in enterprises, Internet access, and Internet use. The results from the Linear regression were:

Root Mean Squared Error (RSME): 1.013

This result of RMSE shows the average magnitude of errors between the predicted and the actual values. Having RSME suggest that the predictions deviate by approximately 1.01 units. Based on the range of the target variable (number of women STEM graduates) this means that the models' predictions deviate by approximately 18%. Considering this the RSME is reasonable, and it could suggest that the linear model predictions are relatively accurate.

Mean Absolute Error (MAE): 0.776

Just like RMSE MAE provides an indication of the accuracy of the linear regression model. The lower this number is the better predictive accuracy the model has. Having an MAE of 0.77 the average absolute difference between predicted and actual values is approximately 0.77 units. Given the same context of the range of the target variable as in RSME, this suggests that on average the linear regression model predictions deviate with approximately 14% of the total range. Again, this shows that the obtained errors seem reasonable and not excessively large.

Scatterplot:

As a part of the exploration of this linear regression model, I made a scatterplot (figure 3). This scatterplot is meant to provide a visual representation of how well this model aligns with the

observed data. As seen in Figure 3 some of the points are placed relatively close to the linear regression line. There are some exceptions where the points are placed quite far from the line. These outliers could suggest that the model struggles to make accurate predictions for some instances. The points are spread unevenly along the regression line which could indicate that the model's predictive power varies across the high and low numbers of women STEM graduates. This made me consider refining the model by including additional features.

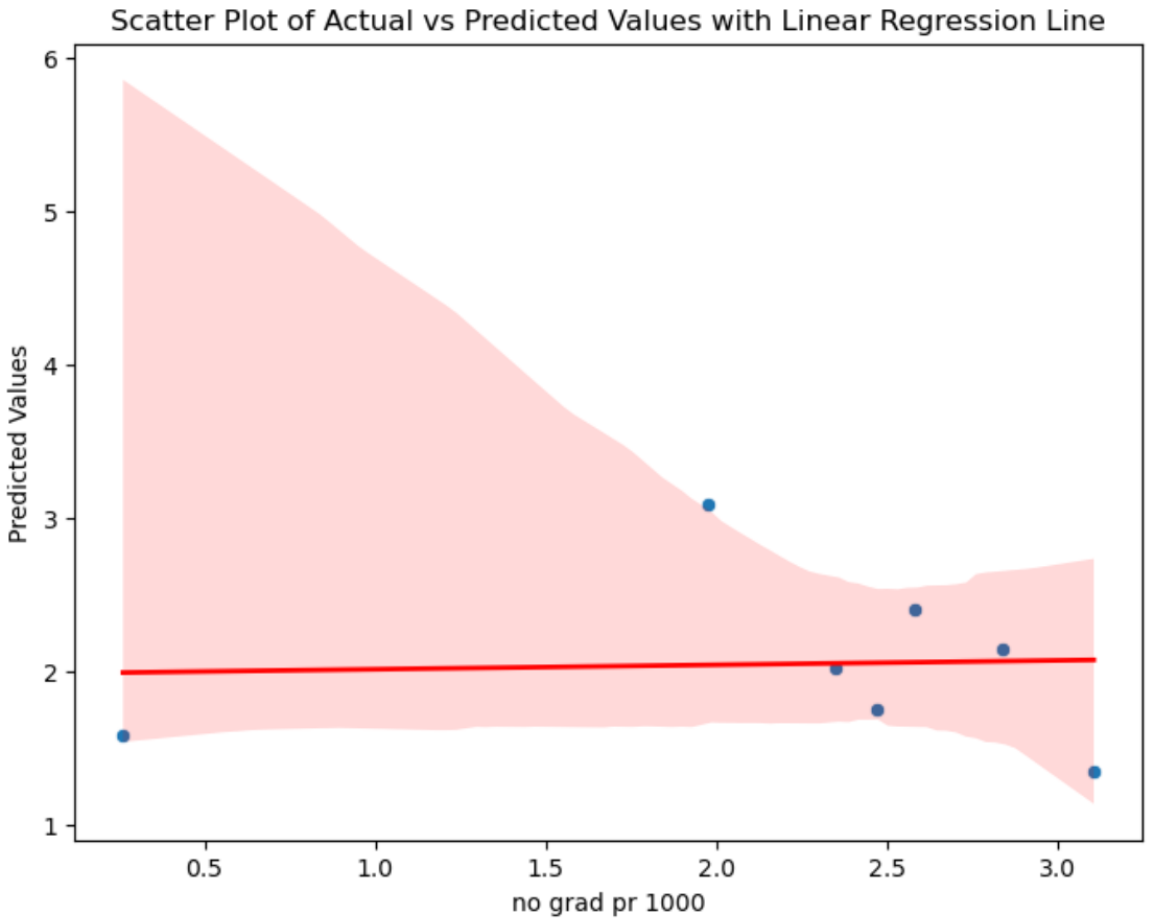

Figure 3 - Scatterplot with Linear regression line - Made with Python

In summary, the results of this linear regression model suggest that the model is providing predictions that are reasonably accurate. You could also argue that the model could use some improvements to give better predictions. This leads the analysis to the next step – adding more features.

Refined model

To optimize and refine the linear regression model I wanted to add more features. I added features within the same area of interest by exploring data on the Eurostat database, focusing on social factors related to IT such as the integration of digital technologies and the amount of digital public service. This resulted in 3 additional features to be added to the model:

- E-Government use (percentage of individuals who obtained information from public authorities' websites)[7]
- Number of people who have never used the internet (percentage of individuals)[8]
- People with basic or above basic digital skills (percentage of individuals) [9]

Adding these 3 features gave the following results:

Root Mean Squared Error: 0.56

This suggests that this model's predictions on average deviate by approximately 0.56 units. Considering the total range of the target feature, this means that the model predictions deviate by approximately 10%

Compared to the results from the previous model these results suggest that the introduction of new features has improved the models' predictive accuracy.

Mean Absolute Error: 0.51

This result indicates that the linear model's predictions on average deviate by approximately 0.5 units. Again, considering the total range of the target feature, this suggests that the model's predictions deviate by approximately 9%

[7] https://ec.europa.eu/eurostat/databrowser/view/isoc_ciegi_ac/default/table?lang=en (17/1-2024)
[8] https://ec.europa.eu/eurostat/databrowser/view/tin00093/default/table?lang=en (20/1-2024)
[9] https://ec.europa.eu/eurostat/databrowser/view/tepsr_sp410/default/table?lang=en (20/1-2024)

Compared to the previous model the MAE is improved. This suggests that adding the new features improved the models' performance.

Scatter plot:

After creating the scatterplot for the refined model, the points are still not on the regression line, as seen in Figure 4. Even though the points are not on the line they are now, compared to the first scatterplot (Figure 3), closer to the regression line and there is only one real outlier.

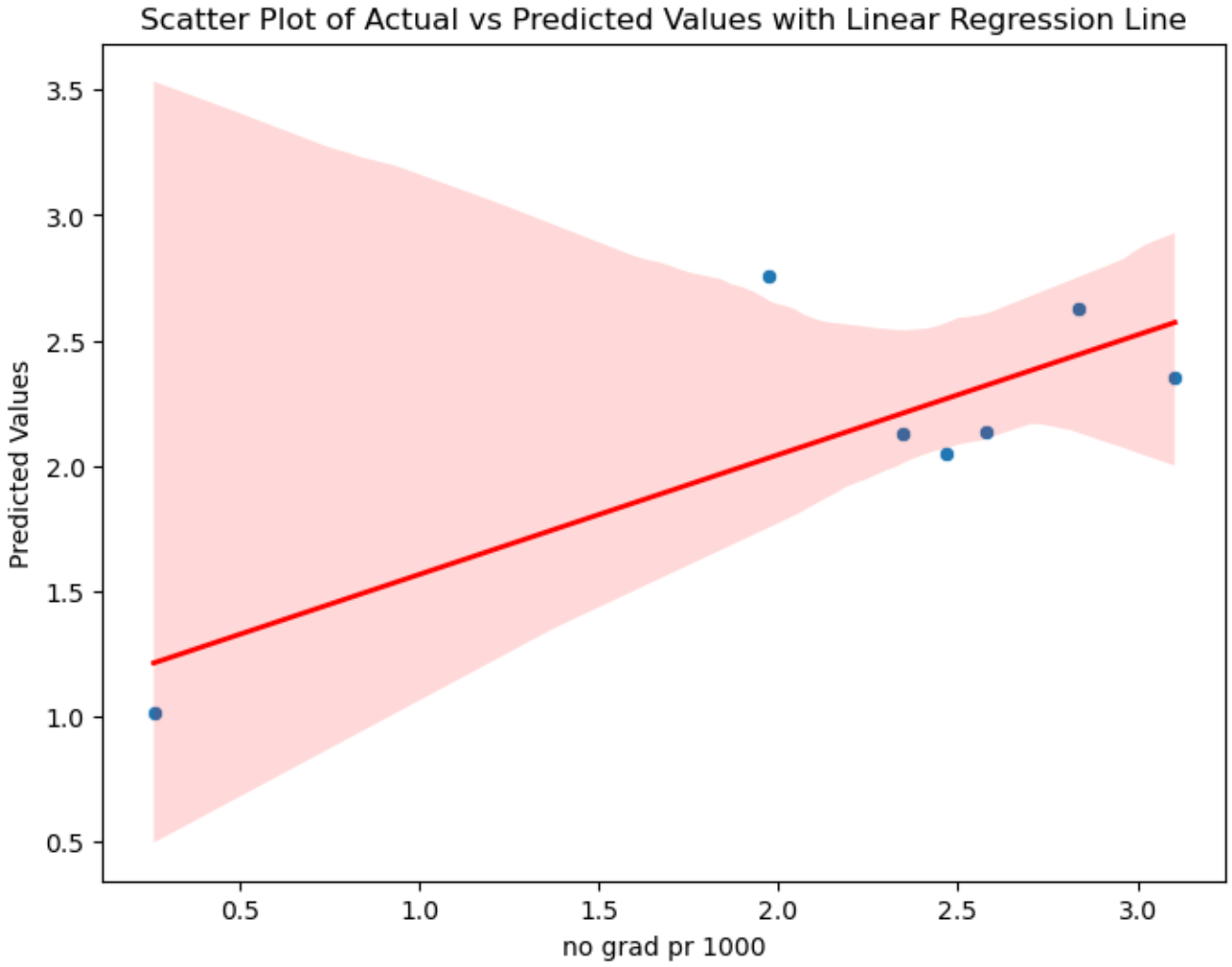

Figure 4 - Scatterplot with linear regression line - Made with Python

Based on this we can conclude that the refined linear regression model has an acceptable performance. There could still be room for improvements to the model.

Random forest

I also experimented with using the Random Forest regression model for this analysis, but the results were too bad. Even with different refinements, the model did not perform well enough to even consider using it for this analysis.

This could be caused by different things, for example, by irrelevant or noisy features. Since the selected features are not a perfect fit for linear regression this could be a sign that they are not necessarily good enough for random forest regression. This could affect the model's performance.

A more plausible reason could be that random forest regression generally benefits from large amounts of data. Having a small dataset, the random forest regression may struggle to learn the patterns and relationships. This will also lead to poor performance of the model.

Discussion

For the linear regression model, I used 6 feature variables to predict from. This was done to try to better the performance of the model. It did help the model's performance by adding features, but this also made the model more complex. Adding more features will increase the risk of overfitting. This is because the model will start to capture noise in the training data rather than capturing patterns. Increasing the features could also impact the simplicity and interpretability by challenging the understanding of the impact of each feature on the target variable. Balancing the predictive accuracy and model complexity is therefore important.

Limitations

One of the limitations of this project was that I could not access a large amount of data. I was only able to access data from 32 European countries. This affected the performance of the models I chose

to work with and might even be the reason I was not able to use models such as Random Forest Regression.

Future work

For future work, it could maybe have a positive impact on the model to remove potential outliers from the datasets and the model. If possible, more relevant data could be added.

Conclusion

Through the findings derived from the linear regression model, Root Mean Squared Error and Mean Absolute Error, a link can be made between social factors linked to Information Technology and the choices of education in STEM fields made by women. This is based on factors like the level of AI use in enterprises, Internet access, Internet use, E-Government use, Number of people never having used the internet and having good digital skills.

In conclusion, this project not only shows the efficiency of predicting the number of women STEM graduates for a country based on IT-related indicators but also promotes for further inclusive and informed research on the topic of gender diversity in STEM educations.

Similar analysis in another course

I have been taking a course in GIS next to this course. In the final project of this course, I have been using data from the same database (Eurostat.com) as the data used in this project. The analysis and research questions are not the same in that course as it is in this course. In the GIS course, I analyze the data through an application called QGIS which includes the use of Vector data which I collected from another website and using this to show the spatial distribution of women graduating in stem field educations on a map. It is a comparative analysis of different sets of data. I did not use the same analyses and I handled the data differently in the two projects. The similarity lies in the data collection through the Eurostat database. I just wanted to make this clear to avoid any misunderstandings in this matter.