Data Engineering 1: SQL and Different Shapes of Data
Term Project 2

Artem Timonov, Işıl Oral, Saad Joiya, Vasilisa Vashchenko
December 4, 2024

1 Introduction

In this project, we aim to build a machine learning algorithm to predict happiness scores of European countries. Happiness score by country reported by the Helliwell et al. (2023) enriched with several social, environmental, and economic parameters. We apply various data engineering techniques to build a project in Knime that we use as our central platform for executing the ETL processes. Afterwards, we implement a random forest to predict the happiness value based on the chosen predictors.

2 Data

The project used diverse data sources to create a comprehensive data set for analysis. The primary source was the Helliwell et al. (2023), which provided yearly data from 2015 to 2023 through World Happiness Report 2023, including variables such as happiness scores, GDP per capita, social support, and perceptions of corruption. This data was accessed via Kaggle (Islam, 2023) and stored in an AWS S3 bucket. Additional data from the World Bank (Group, n.d.) included variables like unemployment rates, internet access percentages, and urban population percentages for European countries during the same period. Further enrichment was achieved using four Eurostat APIs, which added variables such as pollution levels (Eurostat, n.d.-b), house price-to-income ratios (Eurostat, n.d.-d), mobile internet usage (Eurostat, n.d.-a), and population (Eurostat, n.d.-c) statistics.

The dataset was stored in a MySQL database hosted on AWS RDS for easy access and collaboration. An IAM user policy enabled SageMaker access to the S3 bucket, while a security group configured with an EC2 instance managed inbound and outbound rules for the RDS. The final dataset included happiness-related variables from the World Happiness Report, with additional variables from the World Bank stored in a standalone CSV file and later integrated into the workflow. Data preparation began by ingesting nine CSV files—one for each year—from S3 bucket into Jupyter Notebook hosted on Sagemaker. These were then consolidated into a single dataset. The year of each observation (csv name prefix) was appended as a separate column. The data was then filtered to retain only European countries with complete records for all nine years. The final dataset was pushed to MySQL database on RDS.

3 ETL Processes and Model

After succesfully carrying the final data set to Knime, first, in order to standardize the data, a string replacer node in Knime was used to align country names (e.g., "Slovak Republic" to "Slovakia"). Furthermore, while including APIs to enrich our dataset, an HTML query in combination with a python script using BeautifulSoup package were used to extract ISO 3166-1 alpha-2 country codes from Wikipedia (2024), merging it with the aggregated dataset. Then, for all the four variables, the APIs were handled in three steps. First, a column was added via String Manipulation that stored an API request for specific country and year. After that, the GET Request node was used to access the api request for every value in the column. Finally, JSON Path node unpacked the obtained value and converted it to the double format. Afterwards, with the help of the Expression node, a variable named 'mob_internet_per_capita_gb' is built from mobile data usage and population.

The ETL pipeline ensured that all data sources were harmonized into a single table, ready for analysis. The explained ETL part of our Knime

Workflow can be seen in Figure 1. The final dataset including APIs is depicted in ERR diagram at Appendix.

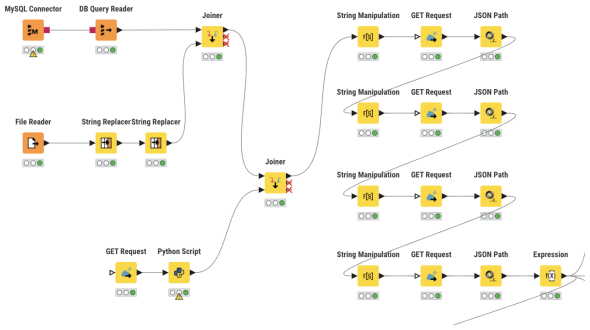

Figure 1: Knime Workflow

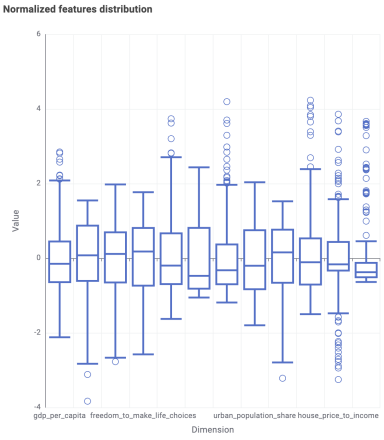

Figure 2: Normalized features

The analysis focused on deriving meaningful insights from the correlations between variables and predicting happiness scores. We used Knime tools for data imputation, normalization, partitioning, correlation analysis, and machine learning.

To address missing values, linear interpolation was used. This approach interpolated missing values based on adjacent observations. For instance, pollution data was unavailable for the years 2021 and 2022, and smaller countries like Bosnia and Herzegovina exhibited significant gaps in multiple variables. Then, normalization was applied to bring numeric values to a common range, ensuring comparability and preparing the data for machine learning.

We detect strong correlation between several predictors. In order to mitigate the challenges posed by multicollinearity, we chose a random forest regressor for prediction due to its robustness against multicollinearity and ability to reduce overfitting through an ensemble of decision trees.

The model was trained on historical data and tested on an out-of-time sample consisting of the last year in the data. This partitioning structure simulated a real-world prediction scenario by ensuring temporal separation between training and testing data. The predictions of our model for the happiness scores in 2023 were closely aligned with the actual values. The final model's predictions achieve an R-sq. above 0.7 indicating good predictive quality of the features used in the model for the target. However, we observed a trend in error distribution: the errors are not randomly distributed but highlight the tendency of the model to predicted close to the average. As such, higher target values are under-predicted while lower values are overshot.

4 Results and Conclusion

Integration of diverse data sources, careful data processing, and an efficient ETL process allowed us to build a comprehensive dataset for further analysis. The project includes three distinct types of data input, as well as a broad variety of techniques of data pre-processing. In our case, a random forest model provided close-to-real predictions for 2023 happiness scores, yet indicating a systematic tendency in the errors of the model. The pipeline created through Knime ensures reproducibility, offering a strong foundation for further analysis of happiness and socioeconomic factors.

References

Eurostat. (n.d.-a). Mobile broadband internet traffic (within the country) [isoc_tmi$defaultview] [Accessed: 2024-11-28]. https://ec.europa.eu/eurostat/databrowser/bookmark/c815e093-d12b-4bd6-b2dc-dbcd3e970bba?lang=en

Eurostat. (n.d.-b). Pollution, grime or other environmental problems [ilc_mddw02]. https://doi.org/10.2908/ILC_MDDW02

Eurostat. (n.d.-c). Population on 1 january [tps00001]. https://doi.org/10.2908/TPS00001

Eurostat. (n.d.-d). Standardised house price-to-income ratio – annual data [tipsho60] [Accessed: 2024-11-28]. https://ec.europa.eu/eurostat/databrowser/bookmark/3e391ba8-1b5a-4adf-a32f-6a9c02e21b73?lang=en

Group, W. B. (n.d.). World bank group databank [Accessed: 2024-11-17]. %5Curl%7Bhttps://databank.worldbank.org/%7D

Helliwell, J. F., Layard, R., Sachs, J. D., Aknin, L. B., De Neve, J.-E., & Wang, S. (2023). Statistical appendix for "world happiness, trust and social connections in times of crisis," chapter 2 of world happiness report 2023. In J. F. Helliwell, R. Layard, & J. D. Sachs (Eds.), *World happiness report 2023 (11th ed.)* Sustainable Development Solutions Network.

Islam, S. (2023, September). World happiness report (till 2023) [Accessed: 2024-11-10]. https://www.kaggle.com/datasets/sazidthe1/global-happiness-scores-and-factors

Wikipedia. (2024). List of iso 3166 country codes [Accessed: 2024-11-28]. https://en.wikipedia.org/wiki/List_of_ISO_3166_country_codes

Appendix

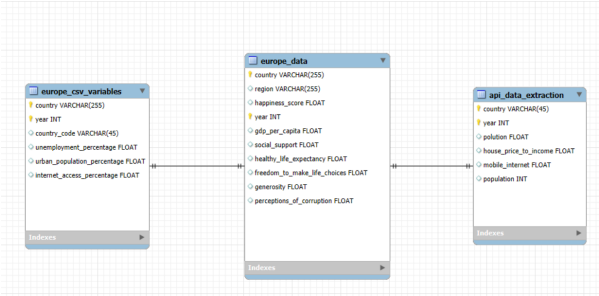

Figure 3: ERR Diagram