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

Artem Timonov, Işıl Oral, Saad Joiya, Vasilisa Vashchenko

Central European University

December 4, 2024

Introduction

- Machine learning algorithm to predict happiness score of the countries
- Data sources enriched by APIs
- Knime, low-code data analytics tool, as the central platform

Data Sources

- Yearly data of 'World Happiness Report' from 2015 to 2023 retrieved from Kaggle
  - Variables include happiness rating, gdp per capita, social support, health life expectancy, freedom to make life choices and perception of corruption
  - Hosted in AWS RDS
- World Bank data
  - Variables include unemployment percentage, urban population percentage and internet access percentage
  - Present as stand alone csv file
- Eurostat APIs
  - Standardised house price-to-income ratio, Pollution, Population, Mobile internet traffic
  - Accessed and merged in KNIME

ETL Processes

- wrangling csvs in s3 using Sagemaker and hosting output in RDS
- using string replacer in Knime to ensure data compatibility across data sets
- Python script and HTML query for integration of API
- normalization variables
- handling missing data

Happiness Distribution by Year

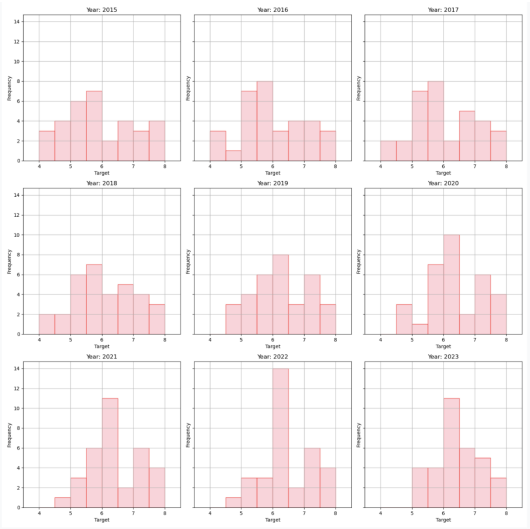
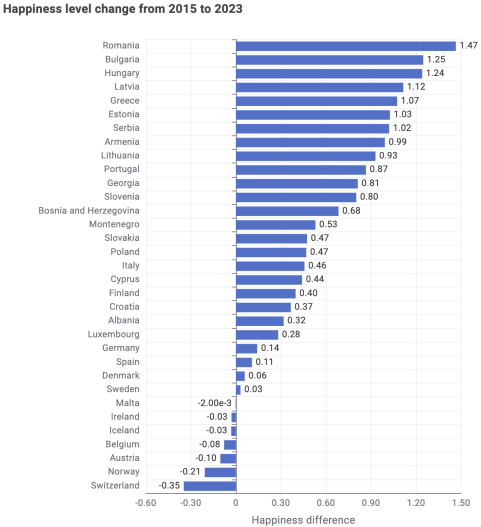
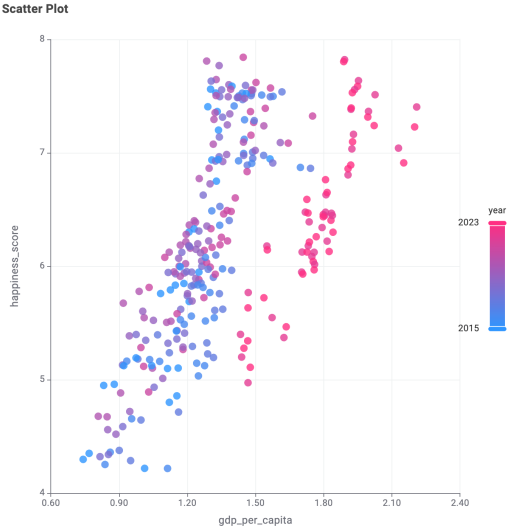


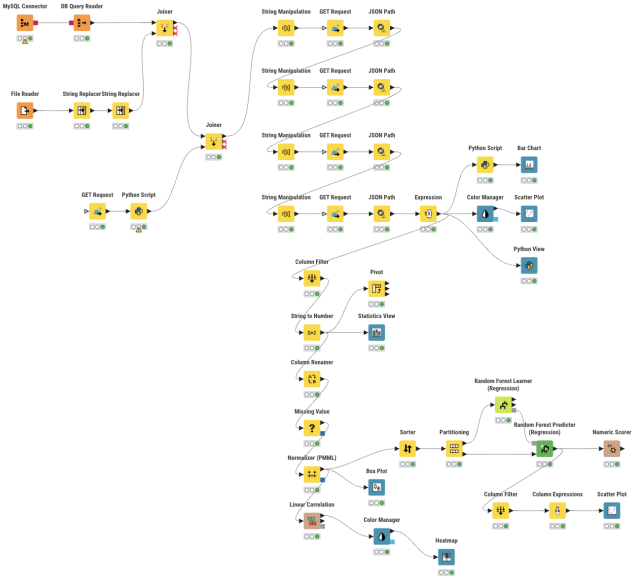
Figure: Histograms for happiness per year

More Visual Analysis





KNIME

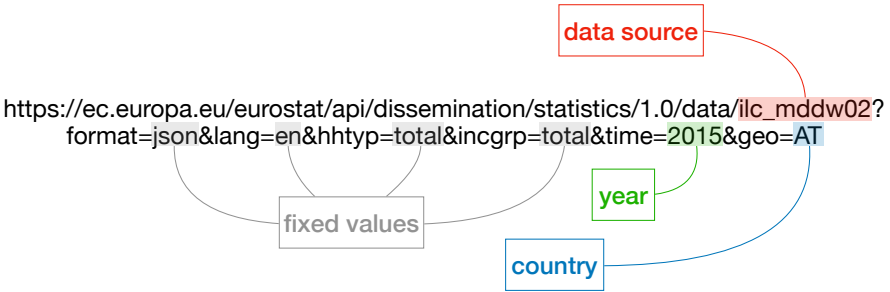


to the Knime extensions

to the normalized variables

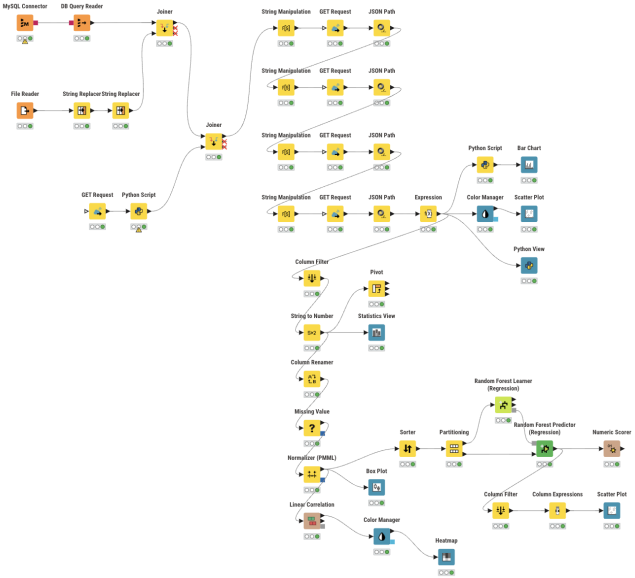
to the heatmap

API requests



- API request structure based on a specific example

KNIME



to the Knime extensions

to the normalized variables

to the heatmap

Normalization of variables

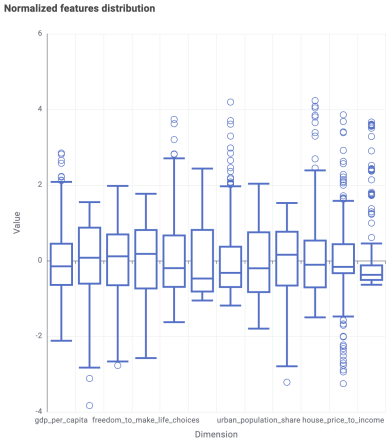


Figure: Normalized features

Back

Random Forest

RowID	Prediction (happiness_score) *Number (double)*
R^2	0.729
mean absolute error	0.296
mean squared error	0.131
root mean squared error	0.362
mean signed difference	0.015
mean absolute percentage error	0.047
adjusted R^2	0.729

Figure: Statistics of predicted outcome

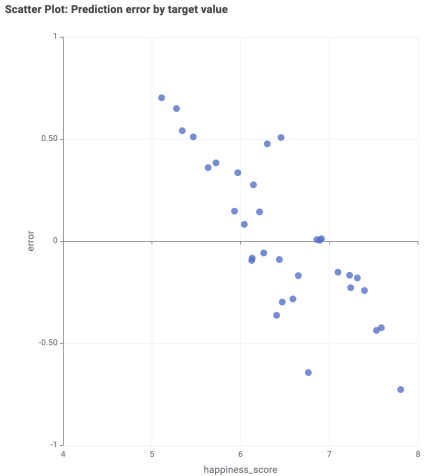


Figure: Prediction errors

Conclusion

- Diverse data sources: World Happiness Report 2023, World Bank, Eurostat APIs
- Data processing with AWS and careful ETL process in Knime
- "Random Forest Model" as machine learning algorithm to predict the last year of the available data
- Test on the actual data

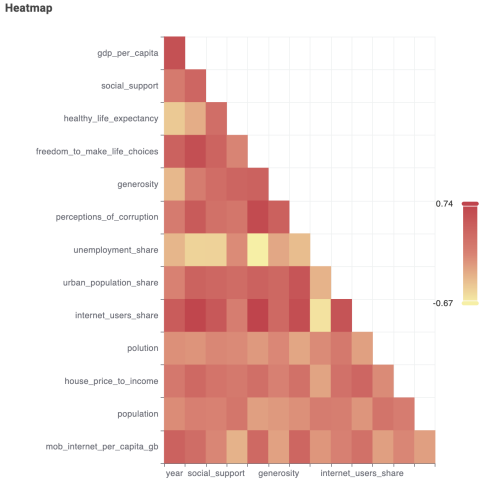
References

- Databank. World Bank Group DataBank. (n.d.). https://databank.worldbank.org/
- Eurostat. Standardised house price-to-income ratio – annual data [tipsho60]. Available at: https://ec.europa.eu/eurostat/databrowser/bookmark/3e391ba8-1b5a-4adf-a32f-6a9c02e21b73?lang=en
- Eurostat. Pollution, grime or other environmental problems [ilc_mddw02]. https://doi.org/10.2908/ILC_MDDW02.
- Eurostat. Population on 1 January [tps00001]. https://doi.org/10.2908/TPS00001
- Eurostat. Mobile broadband internet traffic (within the country) [isoc_tmi$defaultview]. Available at:https://ec.europa.eu/eurostat/databrowser/bookmark/c815e093-d12b-4bd6-b2dc-dbcd3e970bba?lang=en
- 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 World Happiness Report 2023 (11th ed.). (Eds.). Sustainable Development Solutions Network.
- Islam, S. (2023, September 9). World happiness report (till 2023). Kaggle. https://www.kaggle.com/datasets/sazidthe1/global-happiness-scores-and-factors
- Wikimedia Foundation. (2024, October 26). List of ISO 3166 country codes. Wikipedia. https://en.wikipedia.org/wiki/List_of_ISO_3166_country_codes

Knime Extensions & GitHub link

- KNIME Database Extension
- KNIME Python Script Extension
- GitHub Repository to our project

Appendix



Back