README

Data Science Project on Data Science Salaries

Nicholas Nagy

2023-04-04

- **ClassificationModel/**: The folder contains a Python-based machine learning application that uses a Decision Tree Classifier from the sklearn library. The purpose of this application is to predict employment type based on the salary and the remote work ratio. The data used for training and testing the model is read from a CSV file named - **"ds_salaries.csv"**.
- **data/**: Folder containing the data that was used in this project
  - **"ds_salaries.csv"**: The data set that was used for this project. The data was cited in the main project summary.
- **doc/**: Folder containing the main RMarkdown file that was created for this project and any other possible documents.
  - **"Final Project Summary.Rmd"**: The project summary file in a RMarkdown format.
  - **"Final-Project-Summary.html"**: The project summary file in an html file format.
- **figs/**: Folder containing all the plots and figures produced during the project.
  - **"plot1_1.png"**: The top half of the first plot presented in the project summary.
  - **"plot1_2.png"**: The bottom half of the first plot presented in the project summary.
  - **"plot2.png"**: The the second plot presented in the project summary.
  - **"plot2_colourblind_test.png"**: The second plot run through a cvd grid to show how effective the colour palette is.
  - **"plot3_1.png"**: The top left of the third plot presented in the project summary.
  - **"plot3_2.png"**: The top right of the third plot presented in the project summary.
  - **"plot3_3.png"**: The bottom section of the third plot presented in the project summary.
  - **"plot3_3v2.png"**: The second version of the bottom section of the third plot presented in the project summary.
- **lit/**: Folder containing a BibTeX Database which holds the sources used in this project.
  - **"apa.csl"**: A file containing formatting used by RMarkdown to cite sources from the BibTeX Database.
  - **"Data Science Salary Analysis Bibliography.bib"**: A file containing all the sources used in the project.
  - **"Data Science Salary Analysis Bibliography.bib.sav"**: A backup file for the sources used in the project.
- **output/**: Folder containing all the tables produced during the project.
  - **"com_Location_counts.csv"**: Contains the countries that had more than 15 entries in the dataset.

  - **“ds_median_data.csv”**: Contains the salary median data grouped by company size.
  - **“emp_type.csv”**: Contains data with only an experience level of EN (Entry-level / Junior).
  - **“med_us_data.csv”**: Contains the salary median data grouped by percentage of remote workers. There is also some uncertainty statistics included with a randomized y value at which this information is included in the plot.
  - **“median_country_data.csv”**: Contains the median salary for each country above the 15 dataset entry limiter. There are also some uncertainty statistics included with these values.
  - **“median_emp_type.csv”**: Contains the median salary grouped by different employment types. This also only looks at EN experience (Entry-level / Junior). There are also some uncertainty statistics included with these values.
  - **“used_data_plot3.csv”**: Contains the data used in the third plot in the project summary. Contains data entries of countries above the 15 occurrences in the original dataset with more legible columns.

- **Data Science Salary Analysis.Rproj**: The R project