Medical Cost Analysis

Dataset: https://www.kaggle.com/datasets/mirichoi0218/insurance

In this project, you will be trying to develop an end-to-end data science application using the dataset given above. The aim of the project is to estimate the approximate cost of a person's health insurance based on the given variables. While creating the project, try to follow the instructions below and make sure that the project is unique.

1. Creating a Google Colaboratory File

- Make sure your project has .ipynb extension.
- Make sure that there are comment lines explaining the details in your project.
- When submitting the project, submit the cells of this .ipynb file so that the cells are run and the results are visible.

2. Importing Required Libraries

- Import the required libraries for the project to the Colab environment.
- Import Pandas, NumPy, Seaborn, Matplotlib and Sklearn libraries for data analysis

3. Perform An Exploratory Data Analysis

- Analyze the data and draw meaningful conclusions from the data.
    - Examine the distribution of Bmi (Body Mass Index)
    - Examine the relationship between “smoker” and “charges”
    - Examine the relationship between “smoker” and “region”.
    - Examine the relationship between “bmi” and “sex”.
    - Find the "region" with the most "children".
    - Examine the relationship between “age” and “bmi”.
    - Examine the relationship between “bmi” and “children”.
    - Is there an outlier in the "bmi" variable? Please review.
    - Examine the relationship between “bmi” and “charges”.
    - Examine the relationship between “region”, “smoker” and “bmi” using bar plot.

- Try to use data visualization techniques as much as possible while examining the data.

- Please add the meanings you deduced from the analyzes as a comment line.

4. Data Preprocessing

- In this section, prepare the data you have, for training the model.
- Use Label Encoding and One-Hot Encoding techniques to deal with categorical variables.
- Split your dataset into X_train,X_test, y_train, y_test.
- Scale the dataset by normalizing it(Min-Max Scaling or Standard Scaling).

5. Model Selection

- Select several regression models and train them with the preprocessed data.
- Examine the performances of the selected models using cross validation.
- Choose the best performing model

6. Hyper-parameter Optimization

- Optimize the hyper-parameters of the model selected in the previous step.
- Optimize parameters with Grid Search. (Grid Search or Randomized Search)

7. Model Evaluation

- Evaluate the optimized model using regression model evaluation metrics. (Ex. Mean Squared Error, Mean Absolute Error etc.)

8. Project Delivery

- For the project, you need to prepare a code file with the extension of .ipynb and run all the cells.
- You need to add these files that you have prepared to a GitHub repo and add the link of this repo to the form that is given down below.
- The project will be done as a team or individual. The teams created should be a maximum of 3 people.
- Form Link: https://forms.gle/2apBpbawauTLcA817
- **Deadline: 26.08.2023 - 23:59**