Insurance charge prediction

Suppose you are the CEO of a medical insurance company and want to build a system to predict member charges. You would like to use this data to help you decide acceptance of applications for coverage under medical insurance by looking at the predicted charges.

The file insurance.csv contains the dataset for our linear regression problem. The first column is the age of the person, the second column is the person sex, the third column is bmi, the fourth column is number of children, the fifth column is if the person smokes or not, and the six column is the region where the person lives.

Download the file insurance.csv, save it in a separate folder. Then create a program in the same folder where insurance.csv resides that will use linear regression and random forest to predict the charges of an applicant.

You will need to clean the data first:

- Handle missing data, try removing the missing data one time, and then try filling the data with the average or mode (if categorical feature).
- Handle outlier, remove any outlier; use either histogram or box plot method.
- Convert categorical features using either get_dummies or map encoding.
- Scale the data using min-max scaler: (X-X.min())/(X.max()-X.min())

Finally, use linear regression, polynomial regression and random forest and display for each method; the training and testing errors.

For polynomial regression, do not exceed polynomial degree of 5. More information is in this web page link: https://scikit-learn.org/stable/modules/generated/sklearn.preprocessing.PolynomialFeatures.html

For random forest, make sure to tune the algorithm by adjusting the max_depth, max_features, and criterion ‘geni’ or ‘entropy’.

What to submit: create a git-hub repository called medical insurance, make sure it is a public repository and upload the data set and the program .ipynb. Then copy and paste the repository link into Canvas.