**Problem-2**

**Create a linear regression model to predict the charges billed based on the health insurance data.**

Please find Jupiter-notebook with name ‘**problem-2/insurance\_data-linear\_regression.ipynb**’

View with result in ‘**problem-2/insurance\_data-linear\_regression.html**’

**Perform following activities on the given dataset:**

* **Perform exploratory data analysis and Data preprocessing (like problem 1)** 
  + **First 10 rows of the data:** Done
  + **5-point summary:** Done show ('min', '25%', '50%', '75%', 'max')
  + **Information about the column (data types):** Done show data after encoded.
  + **Number of outliers (extra points):** Done with Outliers-IQR and Seaborn-Boxplot
  + **Any missing value:** Find with isnull() not find any
  + **Correlation between variables:** Done with Pandas-corr() and heatmap graph
  + **Distribution of the data:** Done with Seaborn-Histplot
* **Feature engineering**

In insurance data sex(gender), region and smoke value store in form of string. Encode this data as follows:

* sex (male: 1 and female: 0)
* smoke (yes: 1 and no: 0)
* region (northeast: 0, northwest:1, southeast: 2, southwest:3)

Note: There is no missing value so no need to fill or remove data.

* **Train Test split (80-20)**

Perform with *sklearn.model\_selection import train\_test\_split* function.

* **Use appropriate evaluation metrics (RMSE/R-squared) to validate if the model is performing well.**

Presenting with both technic RMSE, R-squared and additional with statsmodels lib.

* **Equation of the model.**

Show weight and Intercept of model.

* **Provide names of significant and insignificant variables (If any insignificant variables and there, then check the impact on model after removing them**).

Statsmodels summary of model-3, p-Value (P>|t|) of gender(sex) is 0.906 which is greater than 0.05 so it is in insignificant relationship with the response variable in the model. After removing the gender model accuracy increases by 1%.

**Deliverables**

• Submit separate Jupiter-notebooks: “**insurance\_data-linear\_regression.ipynb”**

• Final model pickle file: “**insurance\_model.pkl”**