Name: Vikrant Singh Jamwal

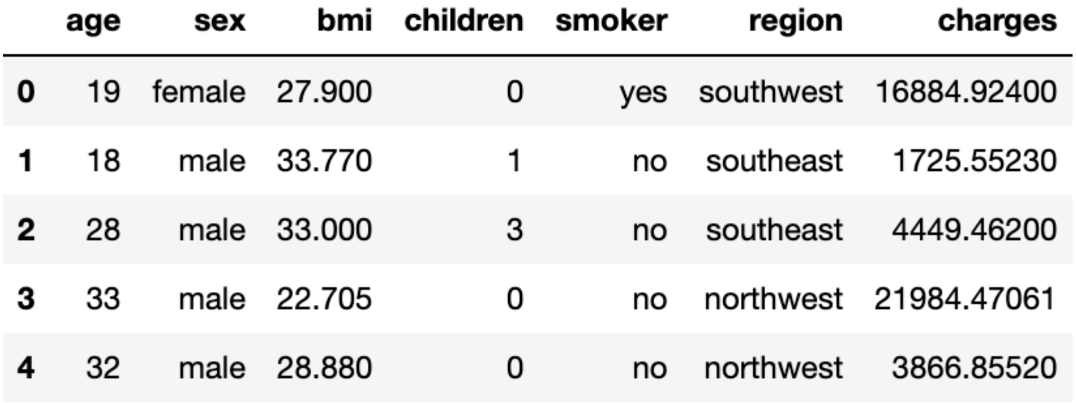
Student ID: 23104534 Class: 1MAI1

ML Assignment 3

1. Problem Statement

Train dataset contains features age, sex, bmi, children, Smoker, region and charges of each employee. The objective is to predict weather the price given by the insurance company after addition of 100 new employees is reasonable or not.

Table1: First 5 rows of Training dataset



	age	sex	bmi	children	smoker	region	charges
0	19	female	27.900	0	yes	southwest	16884.92400
1	18	male	33.770	1	no	southeast	1725.55230
2	28	male	33.000	3	no	southeast	4449.46200
3	33	male	22.705	0	no	northwest	21984.47061
4	32	male	28.880	0	no	northwest	3866.85520

2. Regression Algorithm

For this problem, regression model is suitable as the objective is to predict the “charges” column (dependent variable) which is continuous in nature. Hence, statistical modeling technique like Regression is used as it helps understanding the relation and strength between the variables.

Two algorithms used are:

1. SGD Regressor
   a. Linear (Degree- 1)
   b. Non-linear (Degrees – 2, 3 and 4)

3. Data Exploration and Wrangling

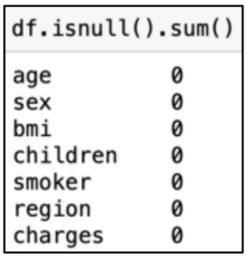
- Shape - 1238 rows and 7 columns

- Column types
  - “bmi”, “charges” – Float
  - “sex”, “smoker”, “region” – Object
  - “age”, “children” – INT

- Independent Variables – {“age”, “sex”, “bmi”, “children”, “region”}

- Dependent Variable (Target) – {“charges”}

- **Preprocessing:**



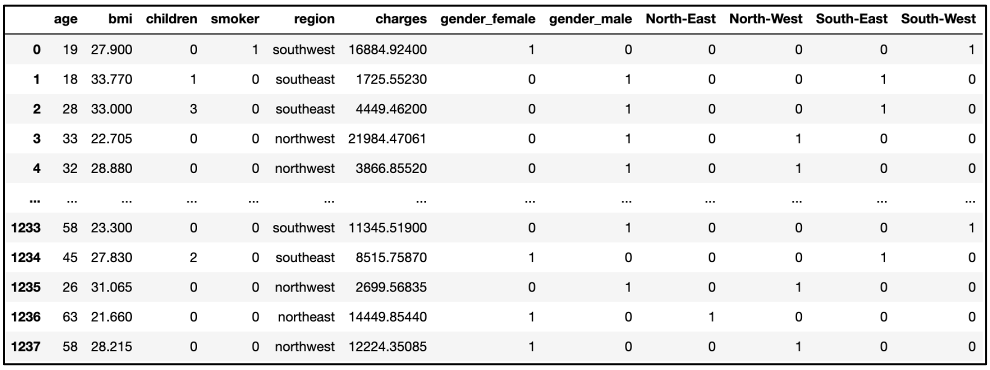
```
df.isnull().sum()

age         0
sex         0
bmi         0
children    0
smoker      0
region      0
charges     0
```

Missing values: There were no missing values present in the dataset.

Feature Scaling: As in this assignment, we have used Stochastic Gradient Decent model, scaling the features helps reach the convergence faster as there will be numerical stability due to similar scales. We have used Min Max Scaler feature of sklearn's pre-processing library.

Feature Encoding: As features with nature as "objects" cannot be statistically analysed, hence we used one-hot encoding method to convert features {"sex", "smoker" and "region"} to numeric type.



	age	bmi	children	smoker	region	charges	gender_female	gender_male	North-East	North-West	South-East	South-West
0	19	27.900	0	1	southwest	16884.92400	1	0	0	0	0	1
1	18	33.770	1	0	southeast	1725.55230	0	1	0	0	1	0
2	28	33.000	3	0	southeast	4449.46200	0	1	0	0	1	0
3	33	22.705	0	0	northwest	21984.47061	0	1	0	1	0	0
4	32	28.880	0	0	northwest	3866.85520	0	1	0	1	0	0
...	...	...	...	...	...	...	...	...	...	...	...	...
1233	58	23.300	0	0	southwest	11345.51900	0	1	0	0	0	1
1234	45	27.830	2	0	southeast	8515.75870	1	0	0	0	1	0
1235	26	31.065	0	0	northwest	2699.56835	0	1	0	1	0	0
1236	63	21.660	0	0	northeast	14449.85440	1	0	1	0	0	0
1237	58	28.215	0	0	northwest	12224.35085	1	0	0	1	0	0

Feature Selection: Plotting correlation bar graph of each feature. We found "smoker" has high positive corelation with "charges", this implies that "smoker" feature is an important aspect to predict charges. {"children", "region" and "sex"} shows negligible corelation with charges and are dropped as they will not affect the prediction.

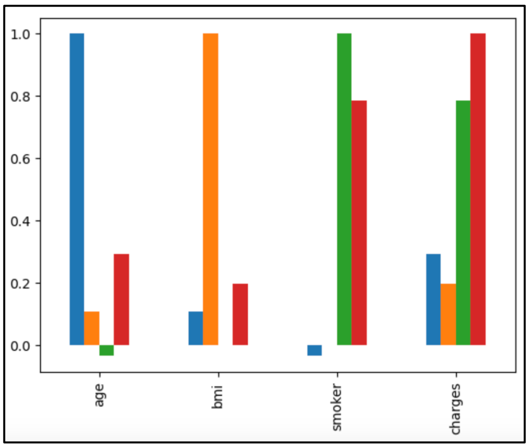


Fig1. High Correlation Features

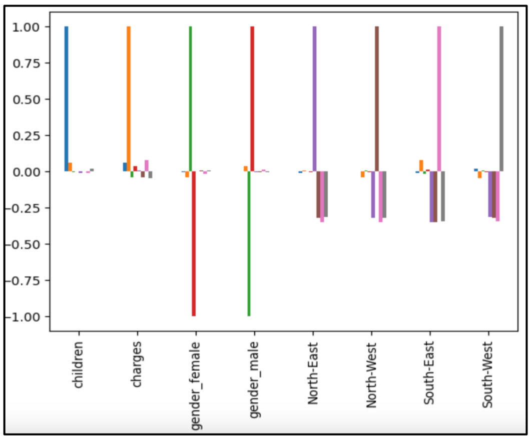


Fig2. Low Correlation Features

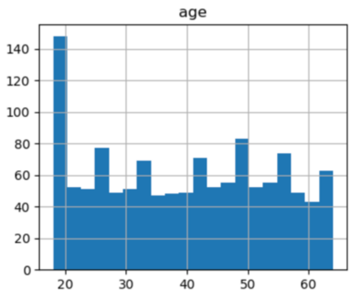
Features after Feature Selection:

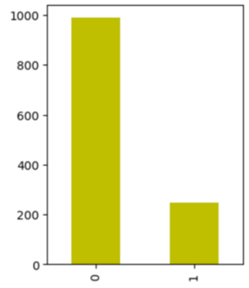
Independent Variable – {"age", "bmi", "smoker"}
Dependent Variable – {"charges"}

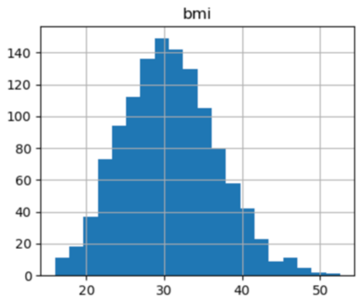
- **Insights:**

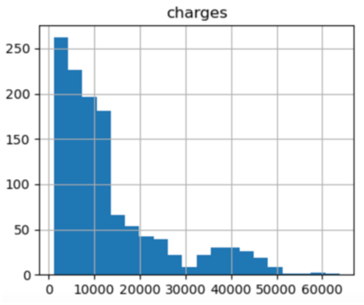
Distribution of data of each feature:

- Age is distributed uniformly among different age groups, but there are more than 140 employees working with age less than 20.
- Most of the employees have bmi around 30 and count decreases either direction.
- Majority of the charges are under 20,000.
- 80% of the employees are Non-Smokers while remaining 20% employees smoke.

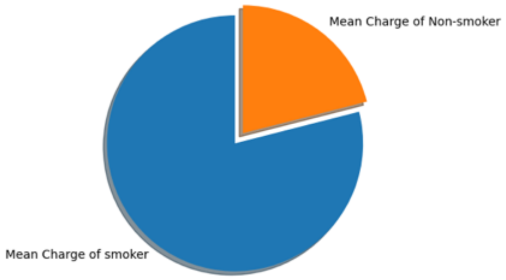








- Mean charges given by a smoker is more than 3 times than the mean charges given by a non-smoker. This implies that company charges more for employees who smokes.



4. SGD Regressor

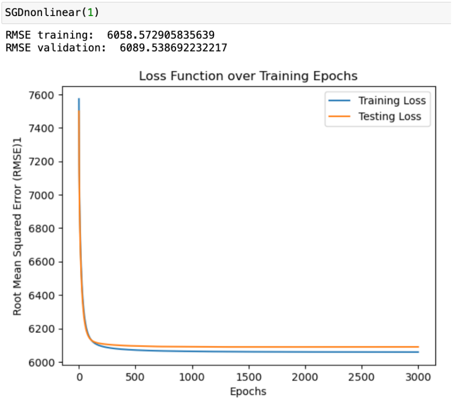
SGD is an iterative optimization algorithm that updates the model parameters using a small subset of the training data, also known as batches, at each iteration. This is in contrast to traditional gradient descent, which uses the entire dataset in each iteration. The stochastic nature of SGD can lead to faster convergence, especially for large datasets. (source: https://scikit-learn.org/stable/modules/sgd.html)

Modified the sklearn’s library function SGDRegressor() as:

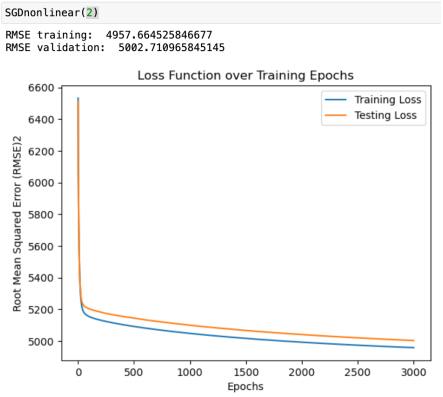
- Used K-fold with K=5
- Used 3000 epochs/iterations manually by partially fitting the dataset into the SGDRegressor and finding the RMSE value for each iteration as well as each fold.
- Created a numpy array by adding each epoch's RMSE of every fold.
- Also used Polynomial Features () to transform the data into polynomial of degree (2,3 and 4) and then applied SGD to find minimum Error at convergence.

From our defined function SGDnonlinear(degree) we plotted convergence graph for:

- Degree = 1 (RMSE training: 6058, RMSE validation: 6089)
- Degree = 2 (RMSE training: 4957, RMSE validation: 5002)

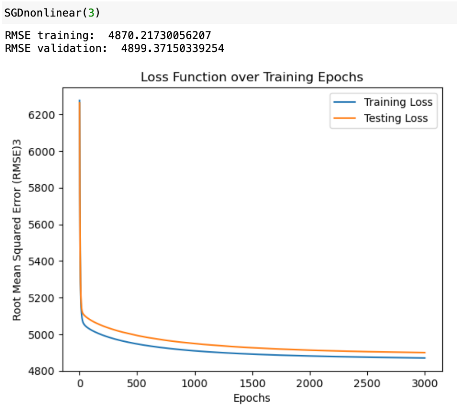


```
SGDnonlinear(1)
RMSE training:  6058.572905835639
RMSE validation:  6089.538692232217
```

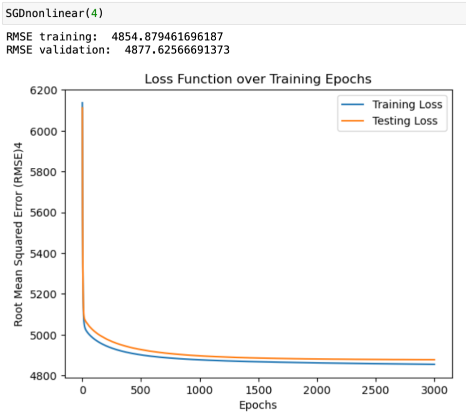


```
SGDnonlinear(2)
RMSE training:  4957.664525846677
RMSE validation:  5002.710965845145
```

- Degree = 3 (RMSE training: 4870, RMSE validation: 4899)
- Degree = 4 (RMSE training: 4854, RMSE validation: 4877)



```
SGDnonlinear(3)
RMSE training:  4870.21730056207
RMSE validation:  4899.37150339254
```



```
SGDnonlinear(4)
RMSE training:  4854.879461696187
RMSE validation:  4877.62566691373
```

Hence, **best model is 3-degree polynomial SGD Regressor** as it gave the least RMSE value of 4877.62 on validation data.

5. Evaluation on Test Set

- Test data is pre-processed, feature selected and normalised as done in Train dataset. Three-degree polynomial SGD Model is fitted and "charges" columns are predicted.

Price Given by Insurance Company: 17,755,825 Euros

Price Predicted by the Best model: 17,708,561 Euros

Price Difference: 47,264 Euros

Insurance Company is asking for 47,264 Euros more, which is not reasonable as every employee has to pay on an average of 35.5 Euros extra.