Medical Insurance Price Prediction

Karthik Kanaparthi
FO62975

Introduction:

- Medical insurance is essential because it shields people from the exorbitant expenses of healthcare. It guarantees that people can receive essential medical care without having to worry about paying excessive costs, which improves health and financial stability.

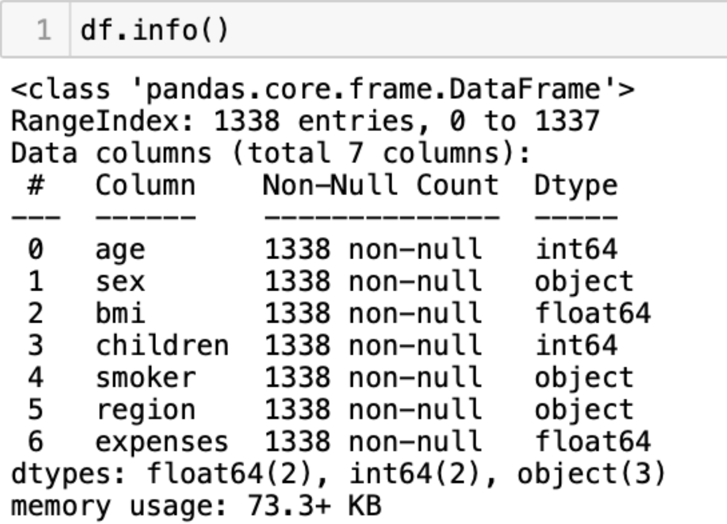
How medical insurance helps mitigate financial risks associated with healthcare costs?

- Financial Coverage for Expensive Treatments: A variety of medical costs, such as pricey procedures, treatments, and hospital stays, are covered by medical insurance. Insurance protects people from the potentially disastrous financial effects of unforeseen medical issues by paying for these costs.
- Predictable Expense Management: Medical insurance helps people better plan and manage their finances by converting unpredictable and potentially enormous medical costs into regular, manageable insurance premium payments. This lowers the risk of financial instability due to health-related expenses.

Data Source:

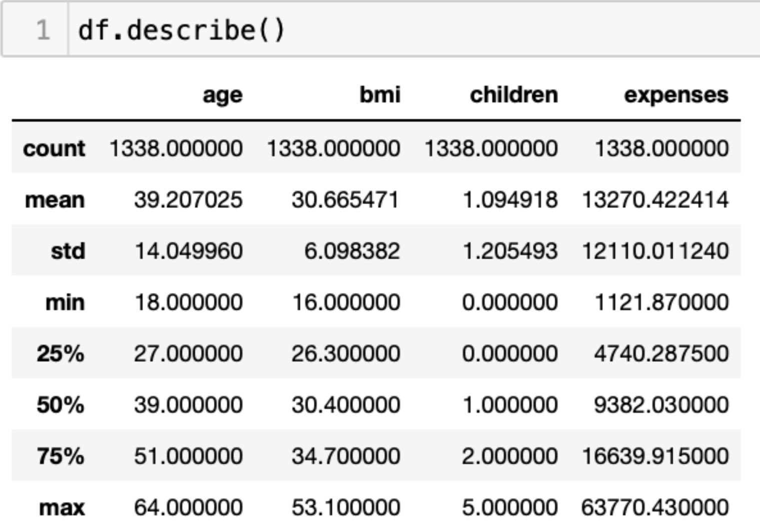
- Data set: Kaggle
- Size and shape of Data set: 51 KB, 1388(rows) x 7(columns/features)
- Time period of data set: 2012 – 2016
- **Data Contains the following columns:**
  - Age
  - Sex
  - BMI (Body Mass Index)
  - Children
  - Region
  - Expenses

Data Source:



```
df.info()
```

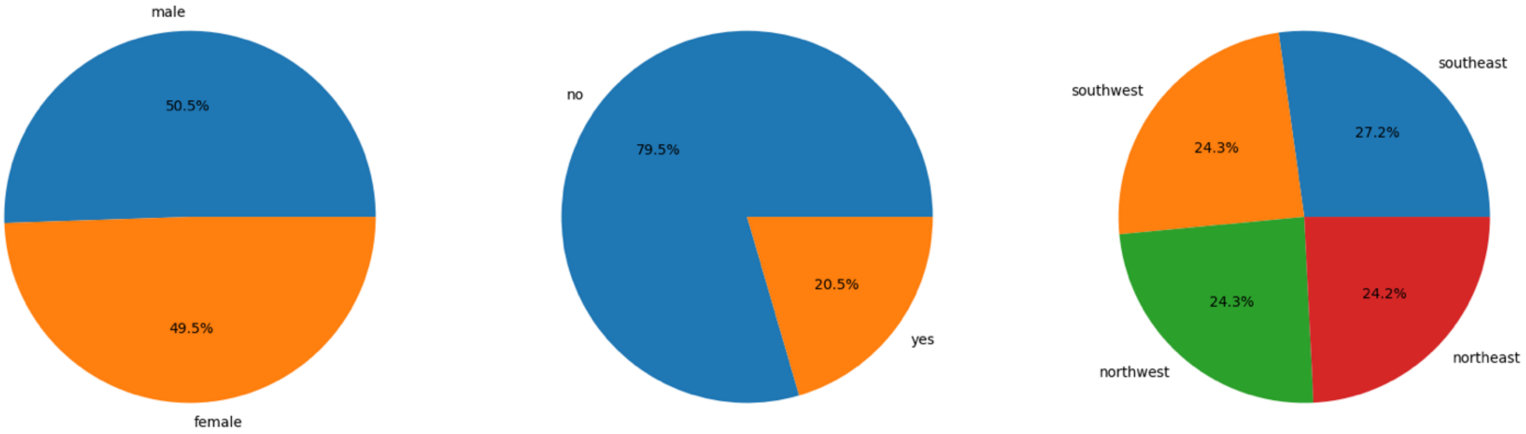
```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 1338 entries, 0 to 1337
Data columns (total 7 columns):
 #   Column    Non-Null Count  Dtype  
---  ------    --------------  -----  
 0   age       1338 non-null   int64  
 1   sex       1338 non-null   object 
 2   bmi       1338 non-null   float64
 3   children  1338 non-null   int64  
 4   smoker    1338 non-null   object 
 5   region    1338 non-null   object 
 6   expenses  1338 non-null   float64
dtypes: float64(2), int64(2), object(3)
memory usage: 73.3+ KB
```



```
df.describe()
```

	age	bmi	children	expenses
count	1338.000000	1338.000000	1338.000000	1338.000000
mean	39.207025	30.665471	1.094918	13270.422414
std	14.049960	6.098382	1.205493	12110.011240
min	18.000000	16.000000	0.000000	1121.870000
25%	27.000000	26.300000	0.000000	4740.287500
50%	39.000000	30.400000	1.000000	9382.030000
75%	51.000000	34.700000	2.000000	16639.915000
max	64.000000	53.100000	5.000000	63770.430000

EDA:

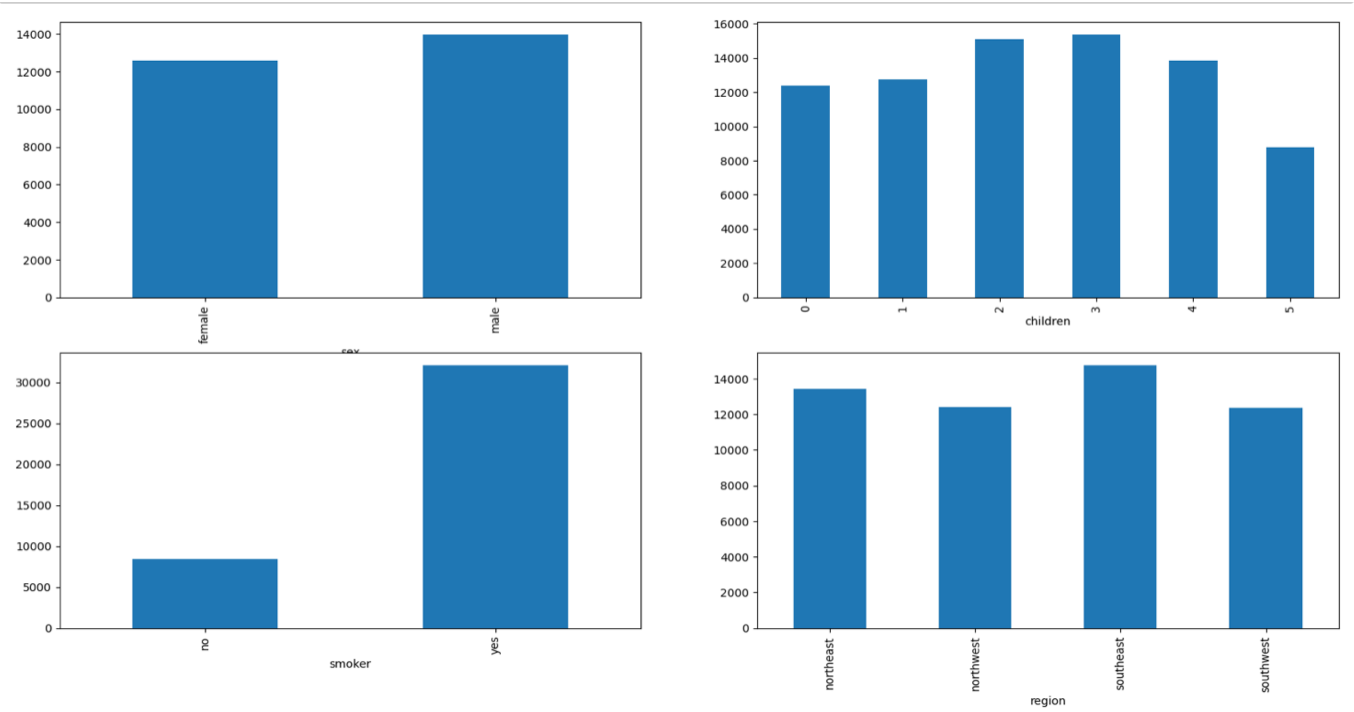


Pie chart for the sex, smoker, and region column

While the sex and region columns of the data are equally distributed, we can see an 80:20 ratio in the smoker column.

EDA:

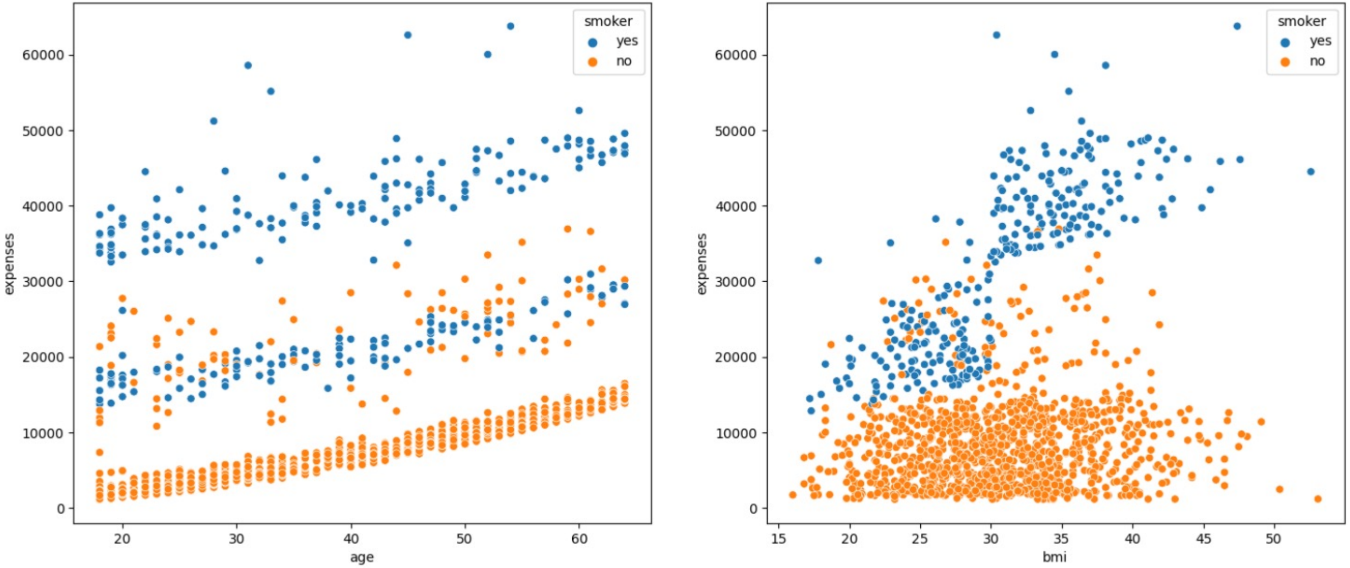
Comparison between expenses paid between different groups



Males pay slightly more in fees than females do, but the difference is not very large. The premium that smokers pay is roughly three times what non-smokers pay. The four specified regions have roughly the same prices.

EDA:

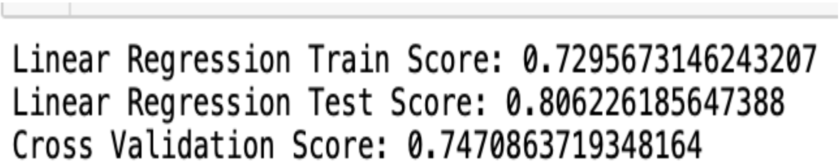
Scatter plot of the expenses paid v/s age and BMI respectively



Here, there is a definite distinction between the fees that smokers must pay. Additionally, we can see that premium costs increase along with an individual's advancing years.

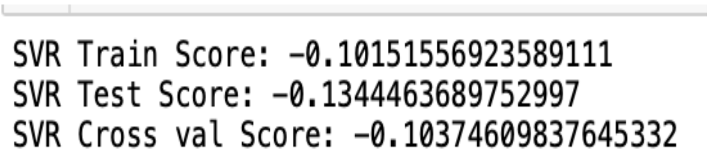
Model Development:

Linear regression:



```
Linear Regression Train Score: 0.7295673146243207
Linear Regression Test Score: 0.806226185647388
Cross Validation Score: 0.7470863719348164
```

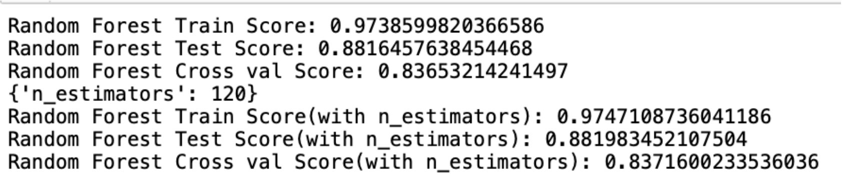
SVM:



```
SVR Train Score: -0.10151556923589111
SVR Test Score: -0.1344463689752997
SVR Cross val Score: -0.10374609837645332
```

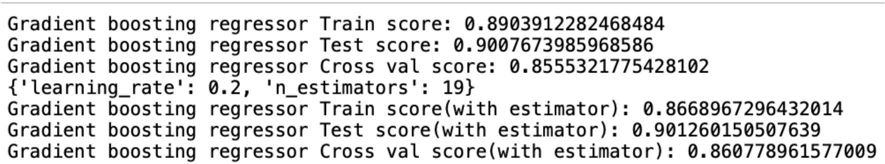
cont.

Random Forest Regressor:



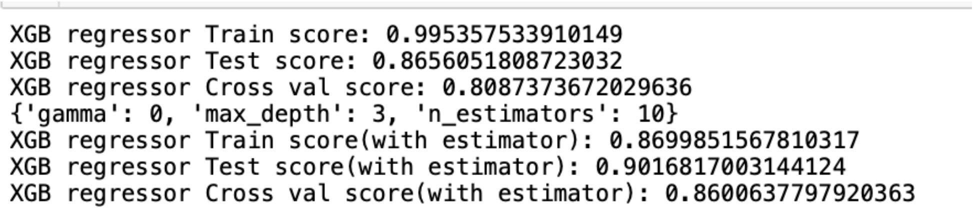
```
Random Forest Train Score: 0.9738599820366586
Random Forest Test Score: 0.8816457638454468
Random Forest Cross val Score: 0.83653214241497
{'n_estimators': 120}
Random Forest Train Score(with n_estimators): 0.9747108736041186
Random Forest Test Score(with n_estimators): 0.881983452107504
Random Forest Cross val Score(with n_estimators): 0.8371600233536036
```

Gradient Boost regressor:



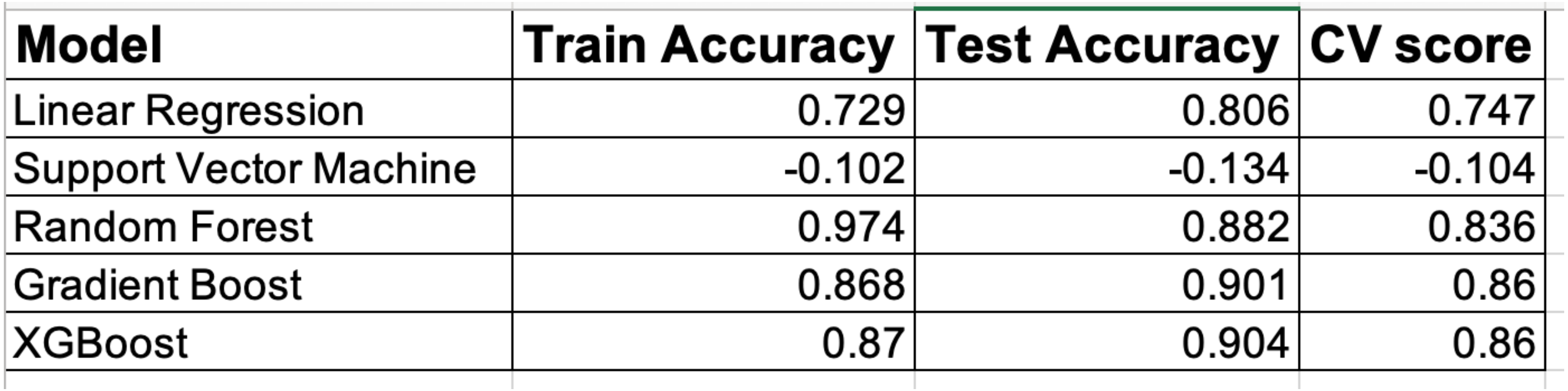
```
Gradient boosting regressor Train score: 0.8903912282468484
Gradient boosting regressor Test score: 0.9007673985968586
Gradient boosting regressor Cross val score: 0.8555321775428102
{'learning_rate': 0.2, 'n_estimators': 19}
Gradient boosting regressor Train score(with estimator): 0.8668967296432014
Gradient boosting regressor Test score(with estimator): 0.901260150507639
Gradient boosting regressor Cross val score(with estimator): 0.860778961577009
```

XGB Regressor:



```
XGB regressor Train score: 0.995357533910149
XGB regressor Test score: 0.8656051808723032
XGB regressor Cross val score: 0.8087373672029636
{'gamma': 0, 'max_depth': 3, 'n_estimators': 10}
XGB regressor Train score(with estimator): 0.8699851567810317
XGB regressor Test score(with estimator): 0.9016817003144124
XGB regressor Cross val score(with estimator): 0.8600637797920363
```

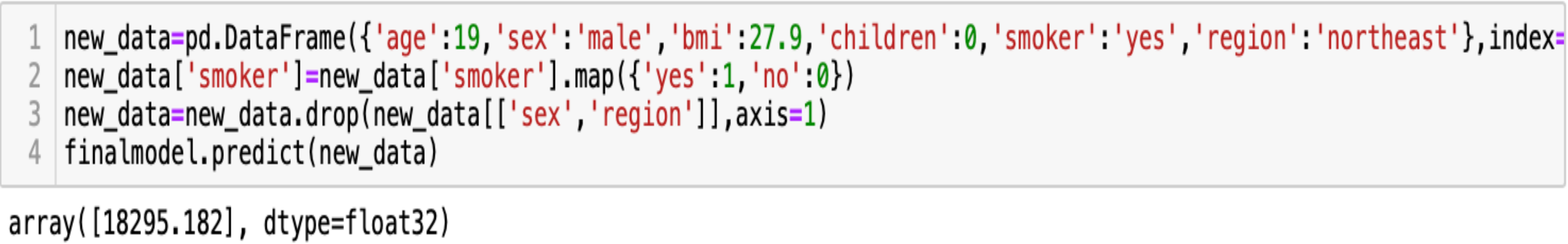
Comparison of all models:



Model	Train Accuracy	Test Accuracy	CV score
Linear Regression	0.729	0.806	0.747
Support Vector Machine	-0.102	-0.134	-0.104
Random Forest	0.974	0.882	0.836
Gradient Boost	0.868	0.901	0.86
XGBoost	0.87	0.904	0.86

Out of all models compared, XGBoost regressor has the highest accuracy for all the parameters which gives you consistency in predicting the expenses.

Prediction:



```
new_data=pd.DataFrame({'age':19,'sex':'male','bmi':27.9,'children':0,'smoker':'yes','region':'northeast'},index=
new_data['smoker']=new_data['smoker'].map({'yes':1,'no':0})
new_data=new_data.drop(new_data[['sex','region']],axis=1)
finalmodel.predict(new_data)
```

```
array([18295.182], dtype=float32)
```

- 18295.182 shows the price prediction of the above defined parameters.

Conclusion:

- The XGBoost model has the highest accuracy of all the models, which indicates that its predictions are more accurate than those of the other models.

- In summary, the combination of Python and machine learning models for predicting medical insurance prices has shown to be a game-changer, radically improving the precision and effectiveness of premium estimates. These models, which are powered by complex algorithms, have proven to be more adept at analyzing a wide range of datasets and taking into account several factors, leading to more accurate and customized insurance pricing.

Thank you