project cs401-machine learning

Insurance Premium Prediction

Submitted by:

Name: Abhishek Kumar Pandey (20250154) and Manasi Tondulkar (20250123)

# Context

The insurance.csv dataset contains 1338 observations (rows) and 7 features (columns). The dataset contains 4 numerical features (age, BMI, children, and expenses) and 3 nominal features (sex, smoker, and region) that were converted into factors with numerical values designated for each level.

# Inspiration

The purposes of this exercise to look into different features to observe their relationship, and build a model based on several features of individual such as age, physical/family condition, and location against their existing medical expense to be used for predicting future medical expenses of individuals that help medical insurance to decide on charging the premium.

# Acknowledgments

Insurance.csv file is obtained from the Machine Learning course website (Spring 2017) from Professor Eric Suess at:

<http://www.sci.csueastbay.edu/~esuess/stat6620/#week-6>.

# Contribution

**Mansi:**

Data preparation and pre-processing.

Model Building: Multiple Linear Regression/ Polynomial Regression/ Decision Tree

Model performance using R2 score, RMSE, and k-fold validation of the above models.

Documentation for Report.

**Abhishek:**

Exploratory Data Analysis and finding correlation.

Model Building: Random Forest Regressor/ Support Vector Regressor.

Model performance using R2 score, RMSE, and k-fold validation of the above models.

Documentation for Report.

# Goal of the Regression Analysis

1. Divide the features into two sections: numerical and categorical to clean and transform each appropriately.
2. For predictor variables:

a) Create a data frame and check for missing data/ unique values.

b) Represent data visually in an appropriate format.

c) Convert categorical variables to represent in a binary format.

d) Dividing the data frame into train test split to feed it into the model.

e) Then, perform the Standardization of the predictor variables so that the values are normalized.

1. Build different Machine learning models, fit and predict the relevant features.
2. Comparing different models to measure the accuracy of the fit.

# Data Preparation

The CSV file **‘insurance.csv'** is imported into the Jupyter notebook and data preprocessing tasks like checking for null values and NA values in the dataset are done. It is seen that the dataset doesn’t contain any null/NA values. The shape of the **‘insurance.csv’** dataset is (1338,7).

Also, **Searborn** is used to check the correlation between the variables.

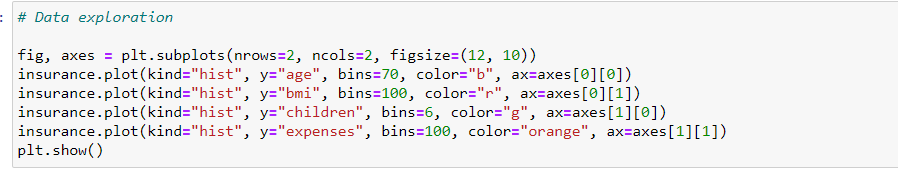
Using **StandardScaler** from **sklearn** library is used to standardize the values.

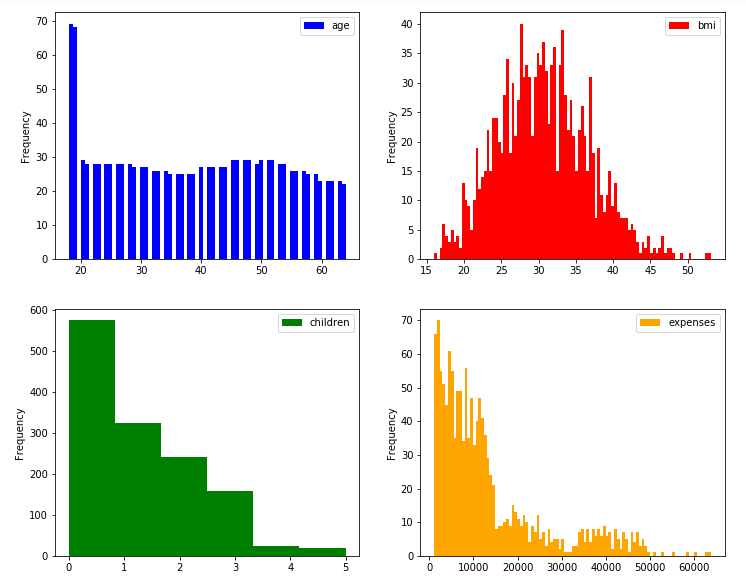
Also, **train\_test\_split** is being used again from the **sklearn** library to divide the dataset into X\_train, X\_test, y\_train, y\_test respectively with training set split into 80% of the dataset and test set into 20% of the remaining dataset.

# Exploratory Data Analysis(EDA)

**Fig.1**

Checking the skewness of the dataset using matplotlib library.



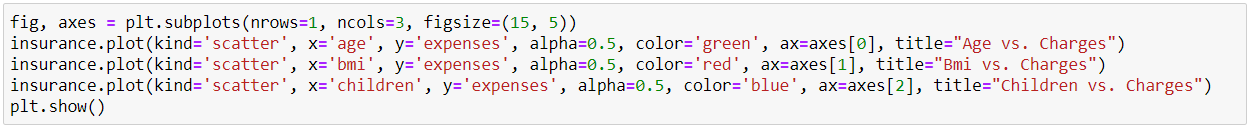
  
 Fig.1

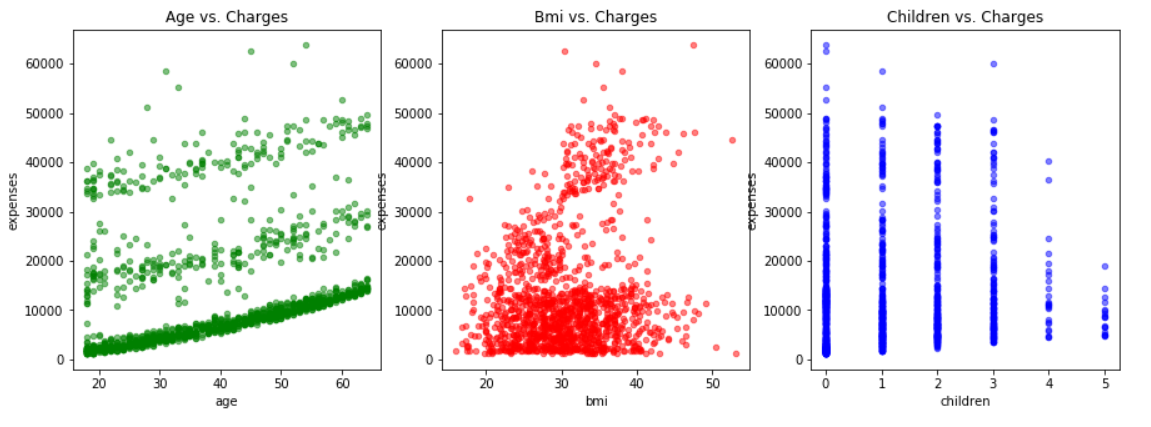
We can observe that apart from bmi other predictors are right skewed which violets the normality assumption in regression.

The plot for bmi shows the variable is perfect Gaussian, which satisfies the normality assumption.

**Fig.2**

Checking the association between the response variable expenses and the predictors.



  
Fig.2

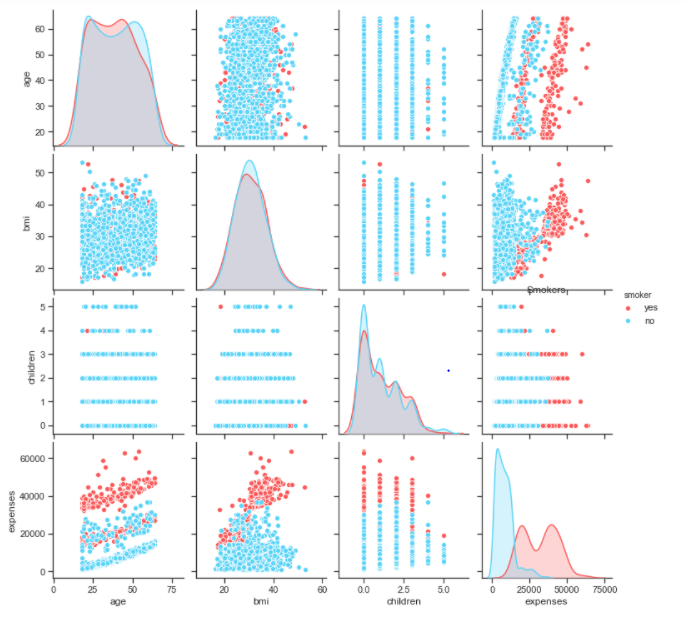
As we can see in the above graphs, there is strong association between expenses and age. The graph for bmi shows that there are more people with bmi between 20 and 35 with minimum expense compared to others.

People with more than 3 children has low expenses compared to others with expense being highest for people with no child.

**Fig3:**

Pair plot to analyze association between all predictors and response, dividing the data into smoker and non-smoker category.

The diagonal plots are the gaussian distribution showing the normality of the variable.

  
 Fig.3

Looking at each plot above, the expense for smokers are high than for non-smokers. The distribution for smokers has high variance and the data is spread throughout the expenses margin. Whereas, the distribution for non-smokers is highly skewed. Also, age and bmi seems to have a good relation with the target variable.

# Model Selection/Evalution

As the expenses variable, is a continuous variable, we would use the regression method.

Selected Models:

* Multiple Linear Regression
* Polynomial Regression
* Decision Tree Regressor
* Random Forest Regressor
* Support Vector Regressor

**Multiple Linear Regression:**

Linear regression is a linear approach to modeling the relationship between a scalar response and explanatory variables. As we have more than one explanatory variables, we are using Multiple Linear regression.

Multiple linear regression is based on few assumptions such as, there is a linear relationship between the dependent and independent variables, independent variables are not highly correlated to each other and residuals should be normally distributed. MLR examines how multiple independent variables are related to one dependent variable. Once each of the independent factors has been determined to predict the dependent variable, the information on the multiple variables can be used to create an accurate prediction on the level of effect they have on the outcome variable. The model creates a relationship in the form of a straight line (linear) that best approximates all the individual data points. The coefficient of determination (R-squared) is a statistical metric that is used to measure how much of the variation in outcome can be explained by the variation in the independent variables. As R-square always increase on adding extra variables, for MLR, we use adjusted R-square which also considers the number of explanatory variables in the model. The end goal of the linear regression is to find the best fit line for dependent and independent variables by minimizing the mean square error and increasing adjusted R-square.

**Polynomial Regression:**

As we have very low accuracy in multiple linear regression, we can say that there might not be a perfect linear relationship between response and predictors. Which is why we try to fit polynomial regression where we find if there are higher-order relationships between X and Y, beyond the linear relationships. Polynomial Regression models are usually fit with the method of least squares. The least square method minimizes the variance of the coefficients, under the Gauss Markov Theorem. Polynomial regression is based on few assumptions such as, there is a relationship between the dependent variable and any independent variable is linear or curvilinear, independent variables are independent of each other and the errors are independent, normally distributed. In this dataset, we have taken degree 3, which means the equation for polynomial regression will be cubic equation. Same as multiple linear regression, the end goal of polynomial regression is to find the best fit line for dependent and independent variables by minimizing the mean square error.

**Decision Tree Regressor:**

One of the statistical modeling methods used in analytics, data processing and machine learning is decision tree learning. A decision tree is used (as a statistical model) to go from an item's observations (represented in the branches) to conclusions about the goal value of the item (represented in the leaves). Tree structures where a discrete range of values can be taken by the target variable are called classification trees. Leaves represent class labels in these tree systems, and branches represent conjunctions of characteristics that correspond to certain class labels. Decision trees where constant values (typically real numbers) can be taken by the target variable are called regression trees. In this dataset, the decision tree regressor selected is having max\_depth=5 which means that the depth or the length of the tree is till 5 starting from the root node. This has been selected for model prediction as the decision trees are considered as on the good predictors for regression as well classification.

**Random Forest Regressor:**

Random forests or random decision forests are an ensemble learning tool for classification, regression and other tasks that function by creating a number of decision trees at training time and generating the class that is the class mode (classification) or the individual trees' mean/average predictor (regression). Random forest is on the upper edge with its parent algorithm Decision Trees. Hence, it has even better performance compared to decision trees. In this dataset, the Random Forest regressor selected is having max\_depth=5 which means that the depth or the length of the tree is till 5 starting from the root node. This has been selected for model prediction as the random forest are considered as on the good predictors for regression as well classification.

**Support Vector Regressor:**

Support-vector machines (SVMs, also support vector networks) in machine learning are supervised learning models with associated learning algorithms that analyze classification and regression analysis results.

It can be used for both classification and regression problems

In our dataset, we have used Regularization Parameter C, kernel and gamma with C value equals 1000 which is high and signifies that it will choose smaller margin hyperplane and gamma value=auto parameter signifies how far the influence of single training example is. In our case, gamma is set to auto so it will automatically adjust the value depending upon the points.

# 3. Predicting Insurance premium

From the prediction performance table below, we can see that the **Multiple Linear Regression** doesn’t perform well with the data. The training/testing accuracy is in negative. Also, the RMSE for train/test is too high which signifies high bias and low variance (Underfit).

**Polynomial Regression** shows a got training/test accuracy at 84.71% and 86.66% respectively. Also, 10-fold score is around 83% and RMSE value is low.

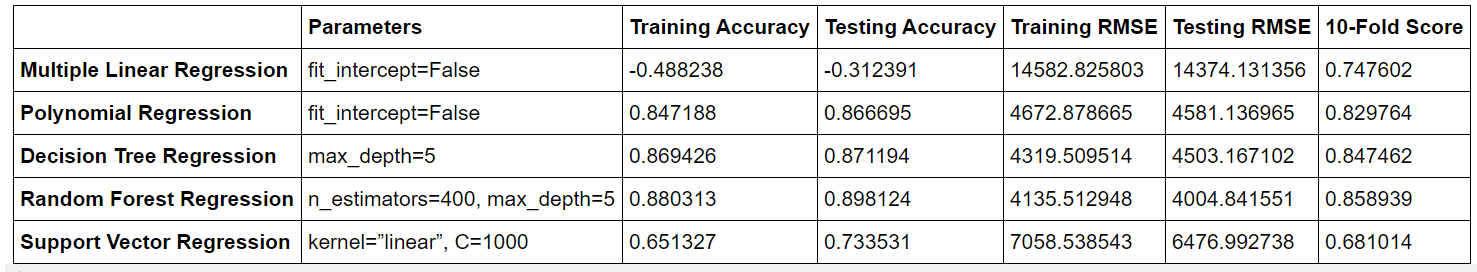
**Decision Tree** performs nicely with training/test accuracy of 87% approx. and low RMSE value.

Similarly, **Random Forest Regressor** works very well with both training/testing/10-fold cross validation score of approximately 88%, 89% and 86% respectively.

On the other hand, **Support Vector Regressor** didn’t perform quite well on the dataset as seen from the Training/Testing/10-fold score of approximately 65%, 73%, and 68% respectively. Also, the RMSE value for testing/training data is too high compared to Random Forest and Decision Tree.

Comparing all the five models, we can say that the Random Forest Regressor performs best compared to other predictors as the accuracy score is highest for the same and low RMSE value.

Below table show consolidated output from all the machine learning algorithm applied:



# Conclusion

So, from the above machine learning project based on the insurance premium, we can conclude that the Random Forest Regressor works best with the dataset provided and could be best for predicting the future values of insurance based on the given parameters.