**Doctors Fee Prediction Analysis**

Choosing a doctor and building a trusting relationship is not simple. Patients expect their doctor to prescribe the best and least costly drug for their conditions, not drugs that the doctor is paid to promote. But now patients will have three steps to take in order to get that assurance: first, they can do some research on the drug and try to make a decision to take a drug based on the evidence.Second, they can research some database to learn if their doctor is one of the high-rollers on the Pharma gravy train. And lastly if so, will they decide to stay with that doctor or begin the process all over again to find a new one?' Do consumers want to do all that, and it is fair to ask that they should?

Databases are that disclosures might make a difference to employers who decide what insurance most citizens can have and to insurers that decide which doctors and hospitals consumers can use. After all, insurance companies are the gatekeepers to their networks, and employers often insist that certain doctors and hospitals are included. Databases found that many of the doctors who received the most money were not necessarily the top practitioners in their fields. Some were not even board certified in their specialties. Others had been sanctioned by state medical boards or received warning letters from the FDA. They had done medical research, but had not bothered to obtain required consent forms for patients taking part in clinical trials.

Transparency is a wonderful thing, prized particularly by doctors both because it makes their job easier and plays to their belief that public behavior would change if people truly understood what was happening in their environment.But behavior is a sticky thing. I wouldnt change physicians if I learned my doctor was receiving speech fees from a drug firm whose product it endorsed just like voters dont oust their elected officials because they are taking oodles of money from the special interests.

My assumption is that physicians who bang the drums for stents or statins at medical meetings are doing so because they're happily and successfully using these tools in their practice and want to share their success record with others.I dont think my liberal politicians would change their views if their contributions dried up. And I think in the absence of an iota of evidence to the contrary that doctors who collect speaking fees are promoting practices they believe improve public health.

When setting a fee schedule, consistency is important so you can get a true idea of what your accounts receivables are at any time. If you have a bevy of legacy fees that are different multiples of Medicare allowables, your AR picture becomes hazy. In addition to consistency, another goal should be minimizing money left on the table by charging above the allowable always.Know, at minimum, what Medicare allowables are. If you’re charging less than what Medicare allows, you may develop a false sense of prosperity since you’re collecting 100% of what your billing commercial payers, many of whose allowables are higher than Medicare’s.

And, of course, if you’re charging a payer less than the allowable, you have no sure way of knowing how much you should have billed out.

You should make revisiting your fee schedule a regular practice to make sure your billed charge is higher than the allowed amount. If your billed charge is equal to the allowed amount, you’ve billed too little and left money on the table.If you can review and revise your fee schedule every six months, that’s probably ideal. You shouldn’t go longer than a year, however. Reviewing an allowed vs. paid report should be reviewed on a monthly basis.

Healthcare market dynamics change, your liability insurance rates change, and patient volumes change frequently, and keeping up with where your fee schedule should be is essential to collecting sufficient revenue to stay in business and invest in the agency’s future. If you’re seeing that your payers are regularly allowing 100% of your charges, it’s time to modify your fee schedule. Check what Medicare and private payers are paying compared to your charges.Perhaps the simplest way to set fee schedules is to use a percentage of what Medicare allows. For example, family practices may charge 150% to 200% of what Medicare allows, and specialists may charge 300% of what Medicare allows. But the percentage you arrive at should be based on your own payer contracts and what other practices in your area are charging.

Another good starting place for most practices is conducting a cost study, either in-house or with the help of a consultant. Ultimately you want a list of services, each of which is assigned a proportion of overhead and margin as well as the base cost.

Data Science field is evolving in all departments.The Data Analysis team has made analyses and prediction the various fields.Here we are going to predict the doctors fee consultation,from Various departments.

Doctor Fee Prediction Analysis comes under Supervised Learning and it's a Regression Problem.The prediction and model we select to solve the problem depends on the target variable.Since we are going to calculate the doctors fee and its a continuous value, we are going to solve using the regression model. By learning new facts and new skills we are going to solve the problems more efficiently and accurately.

In this article we are going to see the following steps :

Step 1: Problem Definition

Step 2: Data Analysis

Step 3: EDA Concluding Remarks

Step 4: Pre-processing Pipeline

Step 5: Building Machine Learning Models

Step 6:Concluding Remarks

Step 1: Problem Definition:

Our task is to analyse key parameters that the doctors of various departments are collecting their fees . We are going to check the features and provide the prediction according to the dataset given.Comparing the models and selecting the best model for this task accuracy is measured .Based on the characteristics of dataset we are going to calculate the "Doctors Fee Prediction" by considering the metrics used in the model.

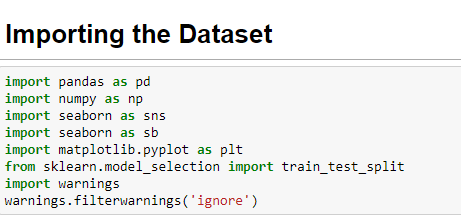
Step 2: Data Analysis

The dataset for this Regression Problem is taken from the records and doctors important details randomly.

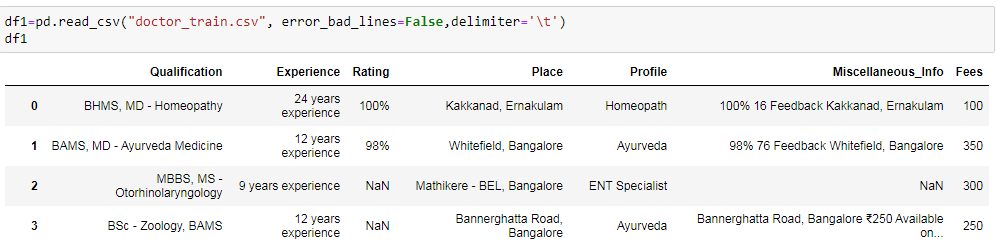
Note: You can find the dataset in the link below.

<https://github.com/dsrscientist/Data-Science-ML-Capstone-Projects/blob/master/Doctor_fee_consultation.zip>

The Data Analysis is started up by using the basic libraries.We are going to use Numpy,Pandas,Seaborn Visualization and Matplotlib,Encoder,Standard Scaler Method etc.



For this exercise, the data set (.csv format) is downloaded and read into the Jupyter notebook and stored in a Pandas DataFrame.



Step 3: Exploratory Data Analysis Concluding Remarks:

Once data collection is done we perform several steps to explore the data.In this step we are going to an understanding of the data structure ,doing initial preprocessing ,clean the data,identify the patterns and features of the data to finally build and validate hypothesis.

Exploring given data:

In this part of EDA the data frame is evaluated for number of columns,data types and null values in the dataset.The aim of this step are to get a general understanding of the data set,check domain knowledge and first get the dataset idea on topics for further investigatation.Here we use some basic pandas library.

Defining Each Features in the Dataset.

By Analyzing the columns and their unique values we are grouping them into category accordingly:

Doctors Information on the Dataset:

Size of training set: 5961 records

Size of test set: 1987 records

Training Dataset Features:

Qualification: Qualification and degrees held by the doctor.Herewe have the departments like Homeopathy,Ayurveda,Otorhinolaryngology,BAMS,ENT,MBBS,Orthodontics.

Experience: Experience of the doctor in a number of years,doctors are experienced from 12 years to 33 years.

Rating:Doctors meet the patients daily ,so the ratings are given by the patients.

Profile: Type of the doctor,depends on which department they have completed their degree.

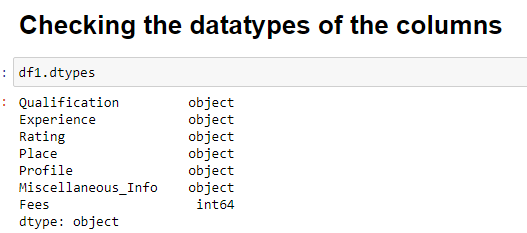
Miscellaneous\_Info: Extra information about the doctor,like address of the doctor and department which they provide treatment.

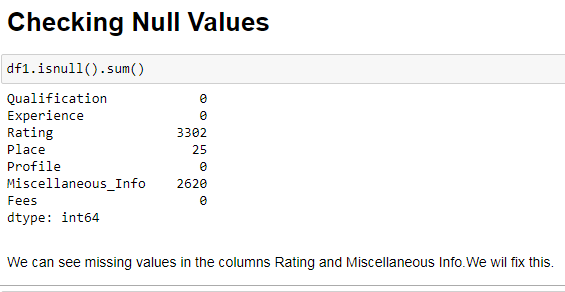
Fees: Fees charged by the doctor (Dependent Variable)

Place: Area and the city where the doctor is located.

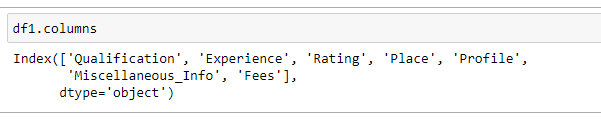
Step 3: EDA Concluding Remarks:

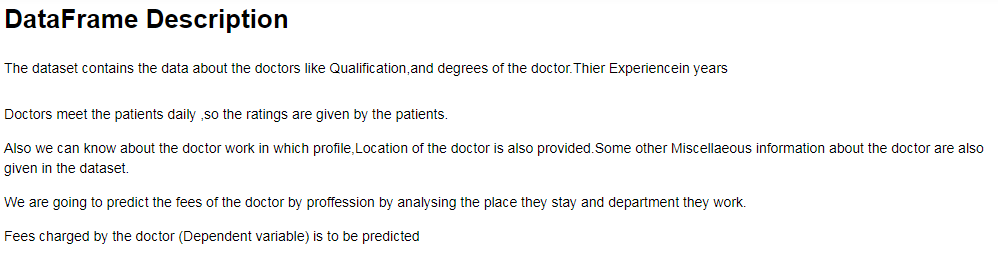
We are going to get printed all the column names,data types and shape of the dataset.



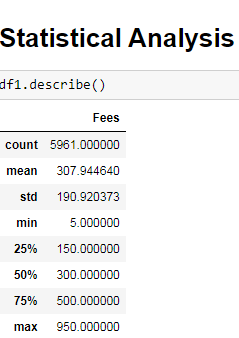


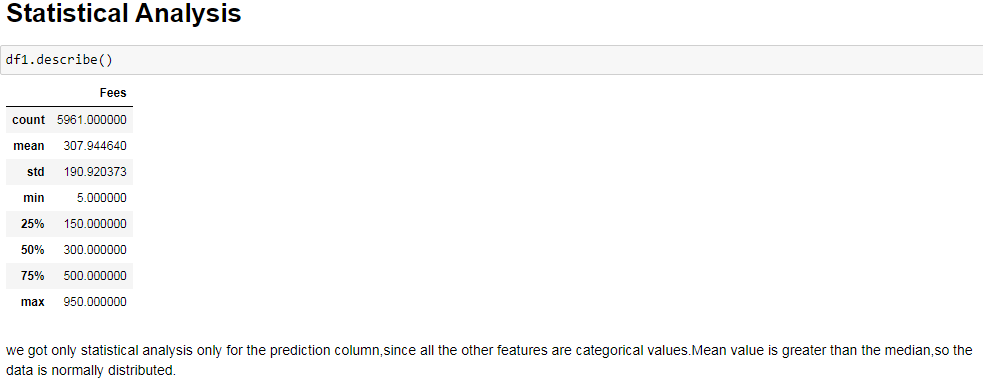
We have null values in the Rating and Miscellaneous column and the datatype is in object type,we shall handle this.



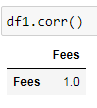


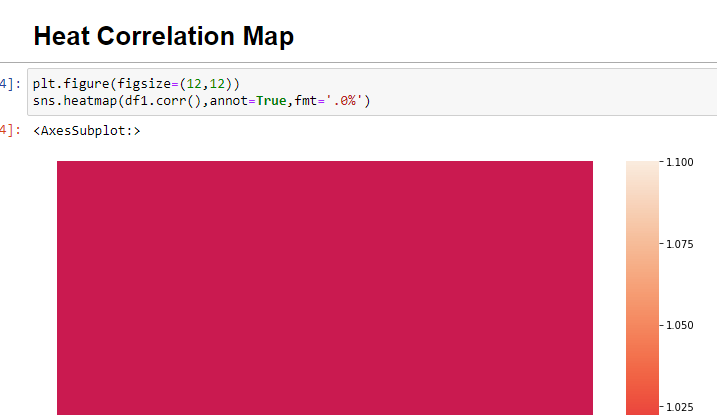
Statistical Analysis of the Dataset:





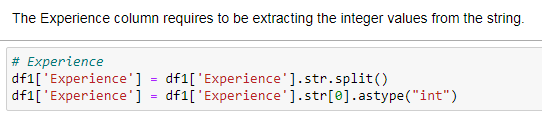
We got only statistical analysis only for the prediction column,since all the other features are categorical values.Mean value is greater than the median,so the data is normally distributed.





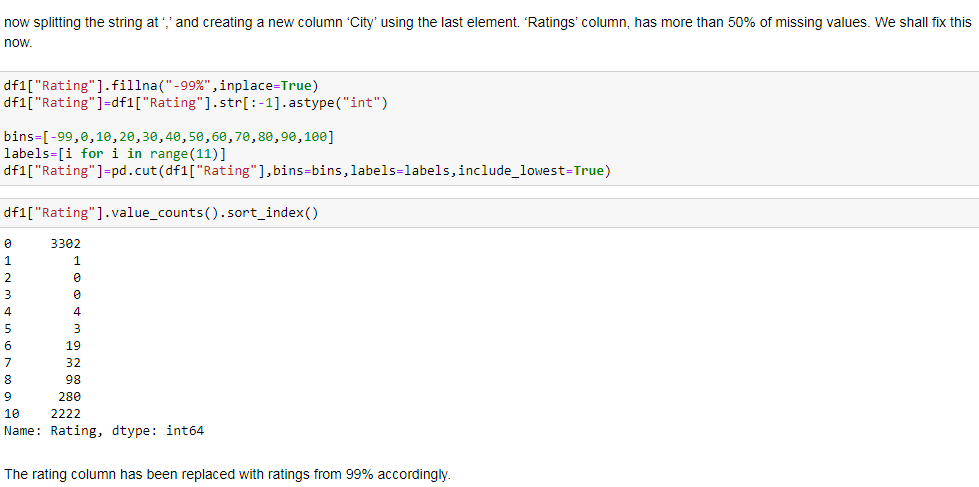
Step 4: Pre-processing Pipeline:

The Experience column is in object type so we are extracting the integer values from the string.



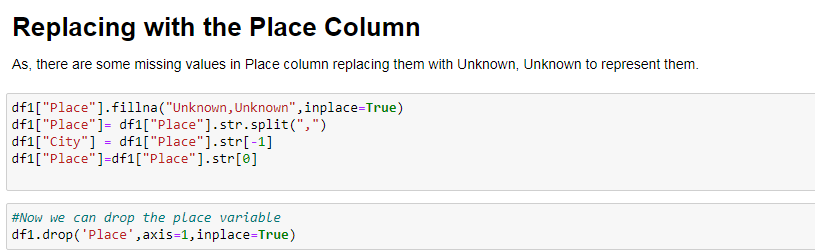
Now we are replacing the Rating column ,It has more than 50% missing values.

Missing values were replaced with -99% to differentiate them. Then, assuming a rating of 91% has no significant difference as a rating of 99%, so we group them into bins of size 10. Missing values will fall under class 0 while 0-9% will be class 1, 10–19% will be class 2,and so on .



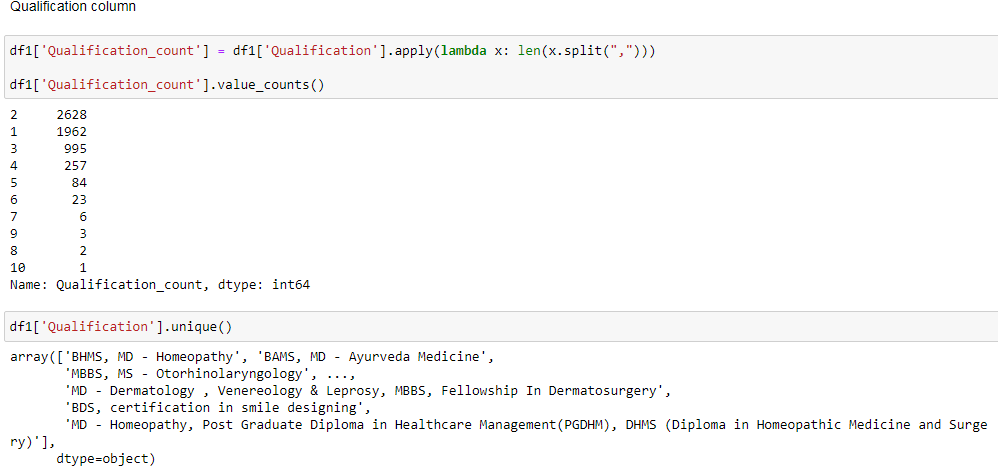
In the Place column we can easily do by separating the City from the area.

Before extraction, we should replace missing values in this column with the string 'Unknown' to represent them. we shall give missing values a separate class instead of relying on missing values imputation technique like mean/median/mode. For example in this dataset, some regions might not have listed down their location during data collection but they could have come from the same region. Next, splitting the string at ‘,’ and creating a new column ‘City’ using the last element of the list.



For the ‘Qualification’ columns, it consists of various qualifications of the doctor without any particular reporting method. so we shall do a normal split and try to get an idea of the frequency of the different terms appearing in this column.

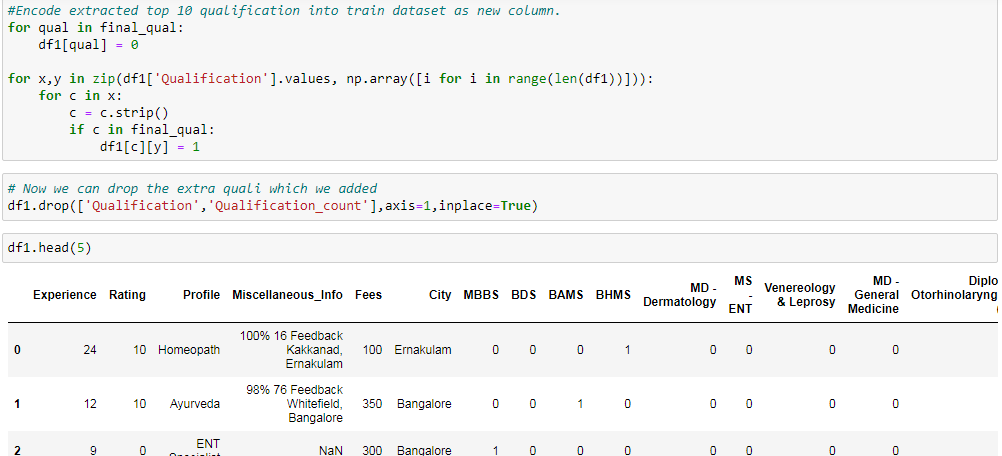
so we shall do the simplest approach and simply identify the top 10 qualifications that occurs the most.



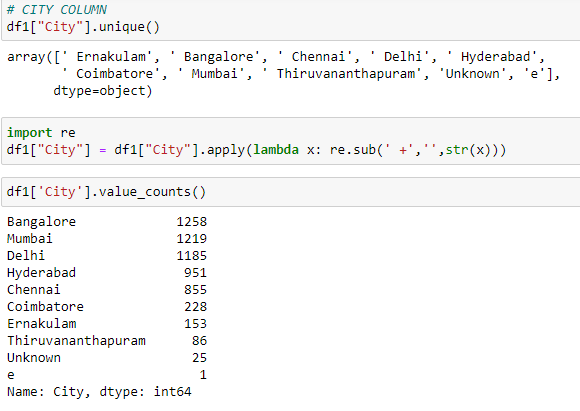
Here there are lot of Qualifications mentioned but most of doctors only has 1 or 2 qualification.



we got as some dummies variables for the 10 highest frequency qualification in the dataset.

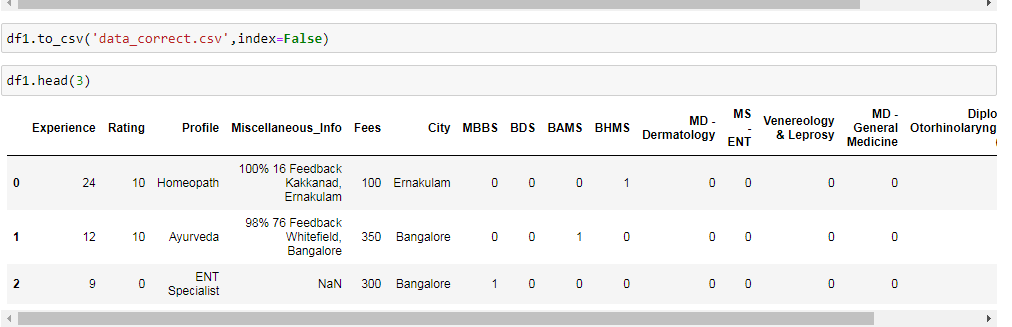


we are going to get the unique value counts of city column and then replace them from object to string type values and getting to print the value counts of city.Then replace the nan values in city column with sector5.



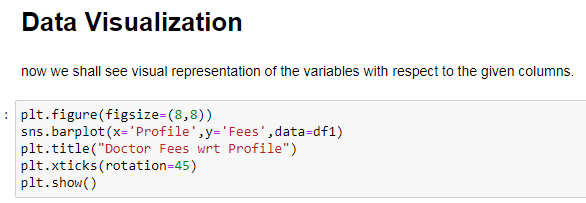


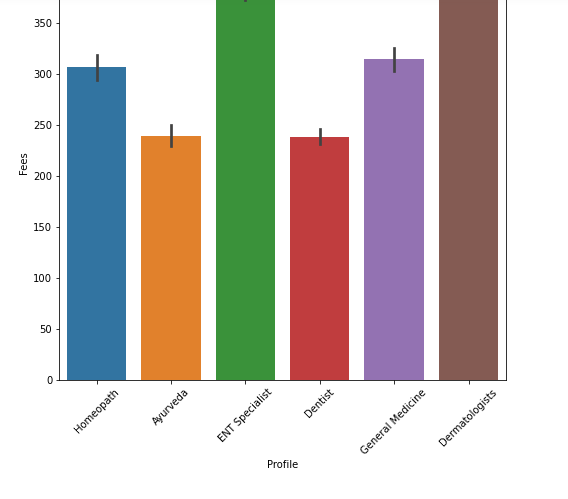
Finally the city column is replaced with the missing values.Now we are done with the data preprocessing of the training dataset df1.

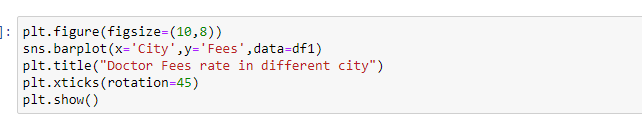


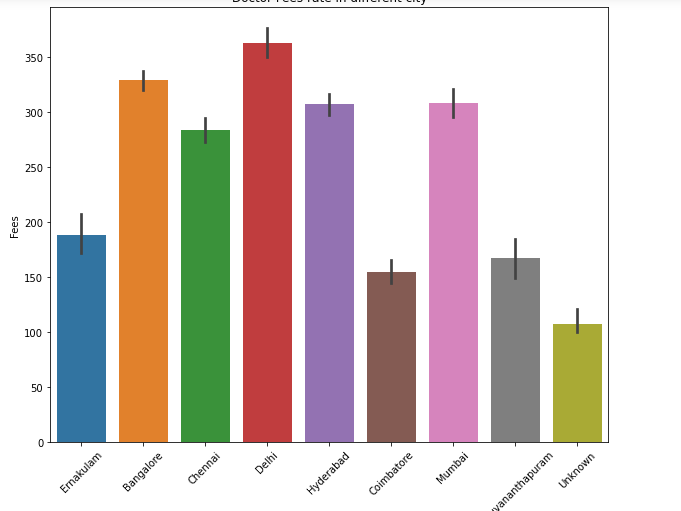
We shall see the data visualization of the training dataset.

Data Visualization: Plotting the profile features of the doctor specialization with fees values.we can see the specialist doctors charge more fees than the general medicine doctors.

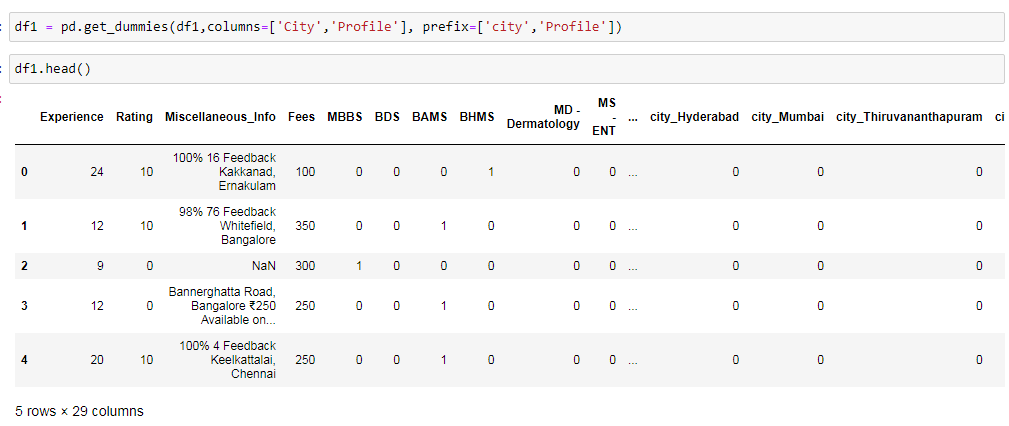




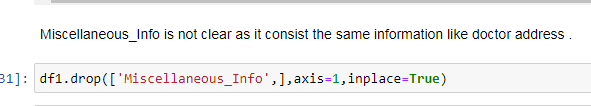
Doctors fees is plotted with different city values.since the cities of Delhi, Banglore, Hyderabad, Mumbai and Chennai come under Tier 1 city,so the fees in these cities are very high as compared to Tier 2 and Tier 3 city. Fees of the specialist like ENT and Dermatologist is high than Homeopathy and General medicine.



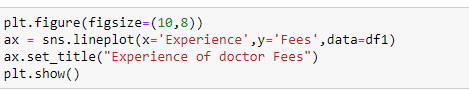
Getting dummy values in the City and Profile column as below:

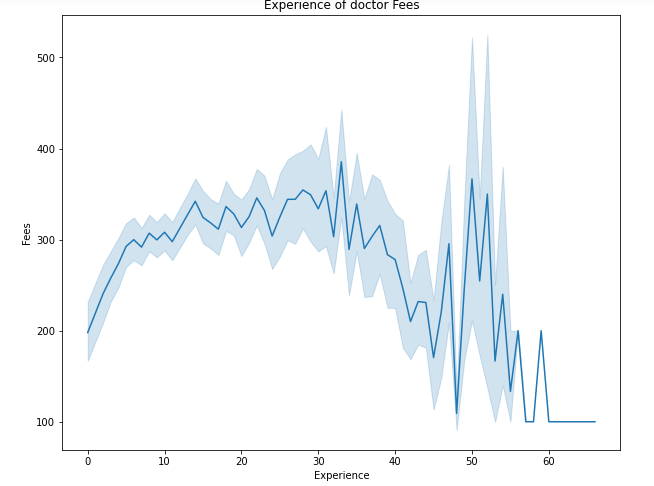


We are dropping the Miscellaneous-info as it consists of more number of missing values and some information like doctor address and general treatments given by the doctor.

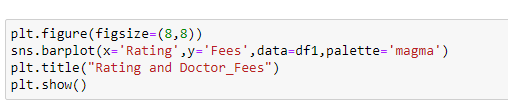


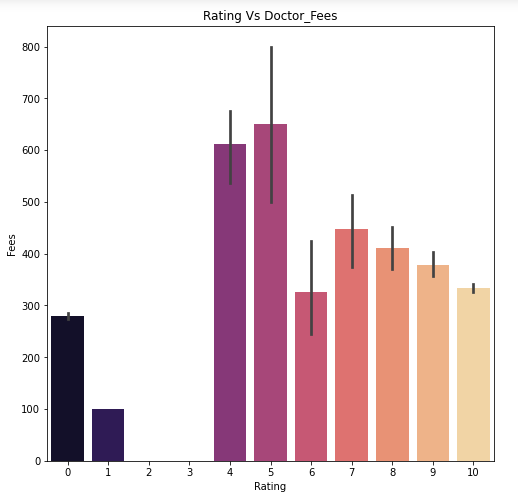
Plot against the experience of the doctor and fees.





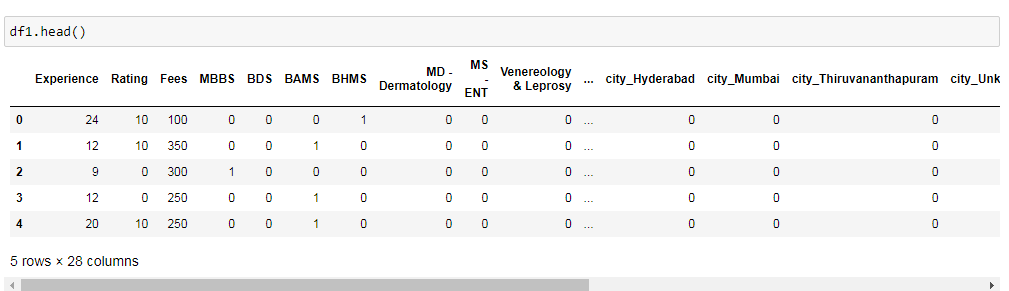
Plot against the feature of rating and doctor fees.



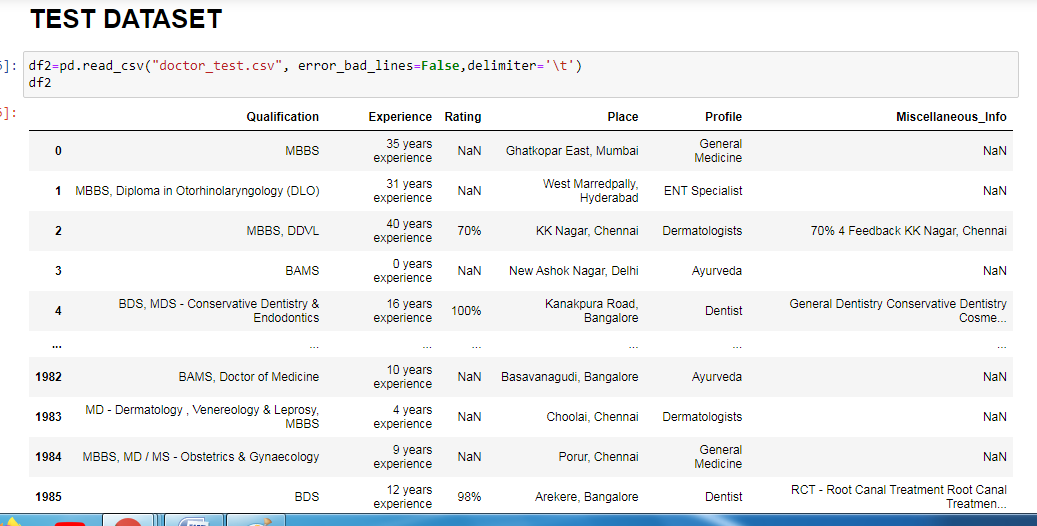


we can see rating of the doctor according to the fees they are getting.we can see the fees of 350 has more number of ratings.

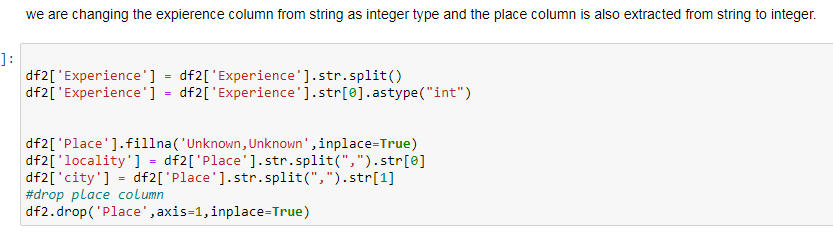
The rating column is grouped in the bins of 10,the ratings are also given in accordance with the number of years of expierence they have.



We have done preprocessing of the training dataset,now we shall repeat the process with Test data set.

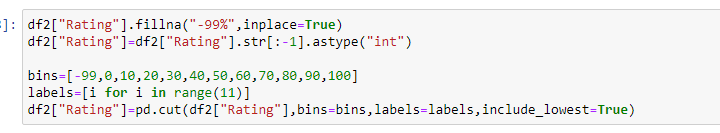


Experience column,Place column is being extracted from string to integer type.



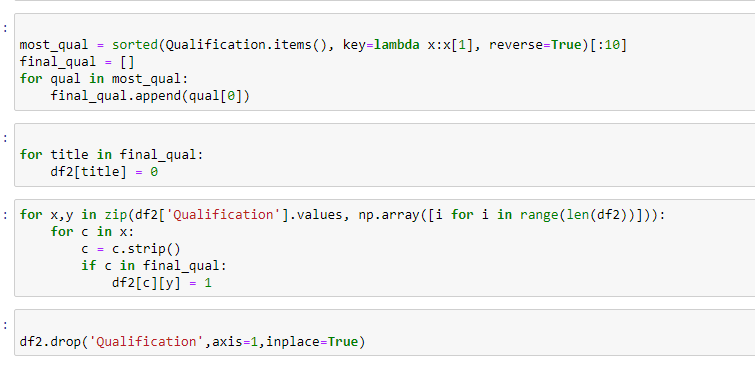
Now we are replacing the Rating column ,It has more than 50% missing values.

Missing values were replaced with -99% to differentiate them. Then, assuming a rating of 91% has no significant difference as a rating of 99%, so we group them into bins of size 10. Missing values will fall under class 0 while 0-9% will be class 1, 10–19% will be class 2,and so on .



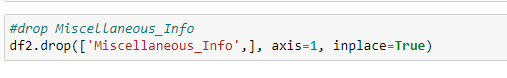
For the ‘Qualification’ columns, it consists of various qualifications of the doctor without any particular reporting method. so we shall do a normal split and try to get an idea of the frequency of the different terms appearing in this column.

so we shall do the simplest approach and simply identify the top 10 qualifications that occurs the most.



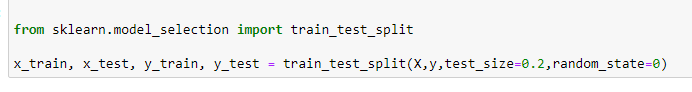


We are dropping the Miscellaneous-info as it consists of more number of missing values and some information like doctor address and general treatments given by the doctor.



Train Test Split:

In train-test-split model training the data is split into 80% training data and 20% testing data.Here our target variable column "Churn" is defined as the class "y" and the remaining columns as the feature "x"**.**



Step 5: Building Machine Learning Models:

Here now we are going to do the Model Building after splitting the train and test data.We can see the evaluation metrics and its parameters which we are going to use in our model building.

In Machine Learning, performance measurement is an important task.

This data comes under Supervised Machine Learning and typical Regression Problem.So the parameters for measuring accuracy involved here are:

1.Standard Scaler:

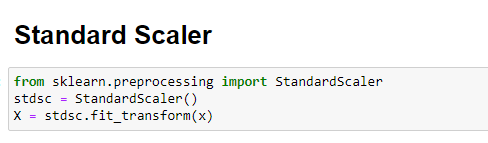
Many machine learning algorithms perform better when numerical input variables are scaled to a standard range.This includes algorithms that use a weighted sum of the input, like linear regression, and algorithms that use distance measures, like k-nearest neighbors.

The two most popular techniques for scaling numerical data prior to modeling are normalization and standardization. Normalization scales each input variable separately to the range 0-1, which is the range for floating-point values where we have the most precision. Standardization scales each input variable separately by subtracting the mean (called centering) and dividing by the standard deviation to shift the distribution to have a mean of zero and a standard deviation of one.

Data scaling is a recommended before modelling step when working with many machine learning algorithms.Data scaling can be achieved by normalizing or standardizing real-valued input and output variables.

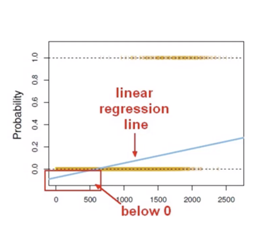
StandardScaler follows Standard Normal Distribution (SND). Therefore, it makes mean = 0 and scales the data to unit variance.

MinMaxScaler scales all the data features in the range [0, 1] or else in the range [-1, 1] if there are negative values in the dataset. This scaling compresses all the inliers in the narrow range [0, 0.005].



2.Linear Regression:

Linear regression is the simplest and most extensively used statistical technique for predictive modelling analysis. It is a way to explain the relationship between a dependent variable (target) and one or more explanatory variables(predictors) using a straight line. There are two types of linear regression - Simple and Multiple.



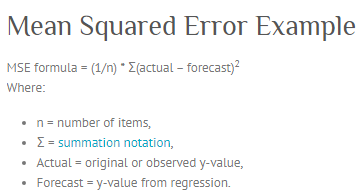
The linear regression line is below 0.

Linear regression is only dealing with continuous variables instead of Bernoulli variables. The problem of Linear Regression is that these predictions are not sensible for classification since the true probability must fall between 0 and 1, but it can be larger than 1 or smaller than 0. Noted that classification is not normally distributed which is violated assumption 4: Normality. Moreover, both mean and variance depend on the underlying probability. Any factor that affects the probability will change not just the mean but also the variance of the observations, which means the variance is no longer constantly violating the assumption .

2.Mean\_squared\_error:

The mean squared error (MSE) tells you how close a regression line is to a set of points. It does this by taking the distances from the points to the regression line (these distances are the “errors”) and squaring them. The squaring is necessary to remove any negative signs. It also gives more weight to larger differences. It’s called the mean squared error as you’re finding the average of a set of errors. The lower the MSE, the better the forecast.

The RMSE is the square root of the variance of the residuals. ... Lower values of RMSE indicate better fit. RMSE is a good measure of how accurately the model predicts the response, and it is the most important criterion for fit if the main purpose of the model is prediction.



General steps to calculate the MSE from a set of X and Y values:

Find the regression line.

Insert your X values into the linear regression equation to find the new Y values (Y’).

Subtract the new Y value from the original to get the error.

Square the errors.

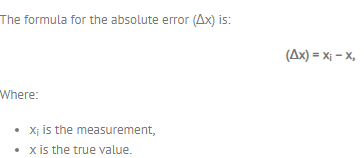
Add up the errors (the Σ in the formula is summation notation).

Find the mean.

3.Mean Absolute Error:

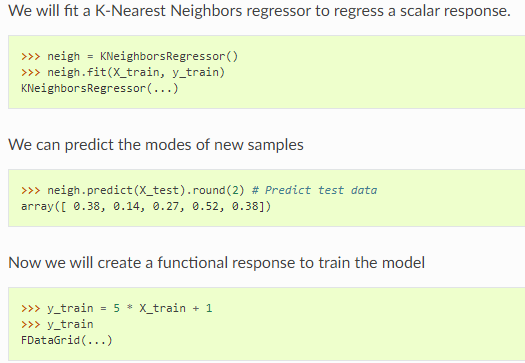
Absolute Error is the amount of error in your measurements. It is the difference between the measured value and “true” value. For example, if a scale states 90 pounds but you know your true weight is 89 pounds, then the scale has an absolute error of 90 lbs – 89 lbs = 1 lbs.

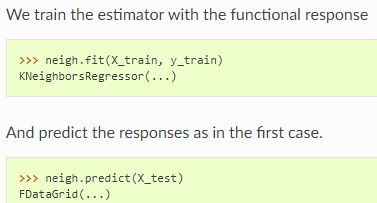
This can be caused by your scale not measuring the exact amount you are trying to measure. For example, your scale may be accurate to the nearest pound. If you weigh 89.6 lbs, the scale may “round up” and give you 90 lbs. In this case the absolute error is 90 lbs – 89.6 lbs = .4 lbs.



4.KNeighbors Regressor:

Regression based on k-nearest neighbors. The target is predicted by local interpolation of the targets associated of the nearest neighbors in the training set. Number of neighbors to use by default for kneighbors queries. weights{'uniform', 'distance'} or callable, default='uniform' weight function used in prediction.



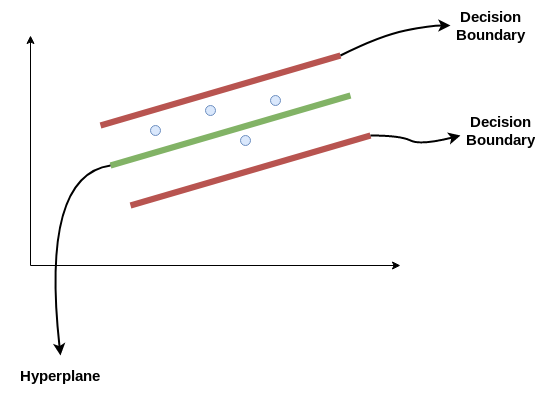


5.Support Vector Machine:

Support Vector Machine can also be used as a regression method, maintaining all the main features that characterize the algorithm (maximal margin). ... In the case of regression, a margin of tolerance (epsilon) is set in approximation to the SVM which would have already been requested from the problem.

Kernel: A kernel helps us find a hyperplane in the higher dimensional space without increasing the computational cost. Usually, the computational cost will increase if the dimension of the data increases. This increase in dimension is required when we are unable to find a separating hyperplane in a given dimension and are required to move in a higher dimension.

Hyperplane: This is basically a separating line between two data classes in SVM. But in Support Vector Regression, this is the line that will be used to predict the continuous output.



Decision Boundary: A decision boundary can be thought of as a demarcation line (for simplification) on one side of which lie positive examples and on the other side lie the negative examples. On this very line, the examples may be classified as either positive or negative. This same concept of SVM will be applied in Support Vector Regression .

6.KNeighborsRegressor:

Regression based on k-nearest neighbors. The target is predicted by local interpolation of the targets associated of the nearest neighbors in the training set. Number of neighbors to use by default for kneighbors queries. weights{'uniform', 'distance'} or callable, default='uniform' weight function used in prediction.

A simple implementation of KNN regression is to calculate the average of the numerical target of the K nearest neighbors. Another approach uses an inverse distance weighted average of the K nearest neighbors. KNN regression uses the same distance functions as KNN classification.

Calculating is like determining parameter K = number of nearest neighbors.

Calculate the distance between the query-instance and all the training samples.

Sort the distance and determine nearest neighbors based on the K-th minimum distance.

7.Random Forest Regressor:

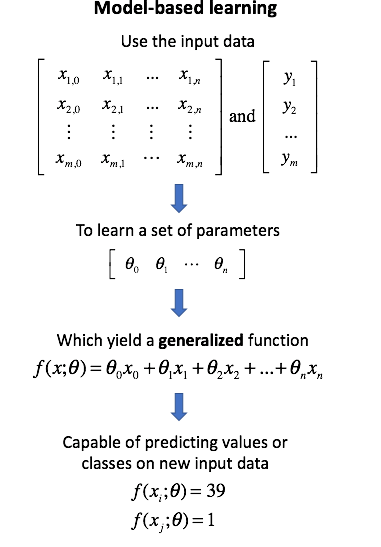
A random forest regressor. A random forest is a meta estimator that fits a number of classifying decision trees on various sub-samples of the dataset and uses averaging to improve the predictive accuracy and control over-fitting. ... The number of trees in the forest.

The algorithm operates by constructing a multitude of decision trees at training time and outputting the mean/mode of prediction of the individual trees.

8.HyperParameter Tuning:

Hyperparameter optimization or tuning is the problem of choosing a set of optimal hyperparameters for a learning algorithm. A hyperparameter is a parameter whose value is used to control the learning process. By contrast, the values of other parameters (typically node weights) are learned.

Hyperparameters are not model parameters and they cannot be directly trained from the data. Model parameters are learned during training when we optimize a loss function using something like gradient descent.The process for learning parameter values is shown generally below.

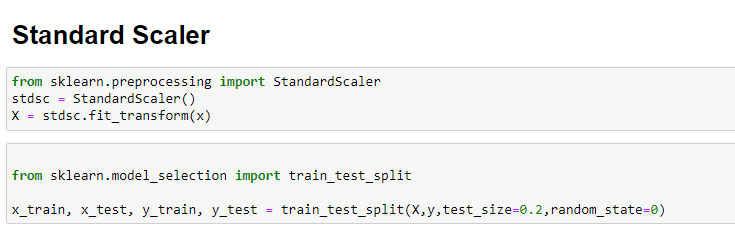


Whereas the model parameters specify how to transform the input data into the desired output, the hyperparameters define how our model is actually structured.

We have seen all the definition and Model Building Algorithms in short:Now we shall proceed with the model building for our dataset.

First we shall apply Standard Scaler to our dataset and then the model building approach.

**Step 5: Building Machine Learning Models:**

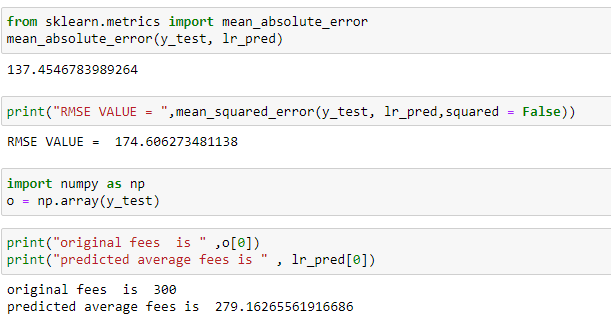
****

**Linear Regression:**

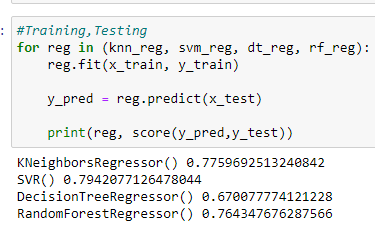
****

Mean absolute error calculated is 137.4

RMSE value is 174.6

****

****

****

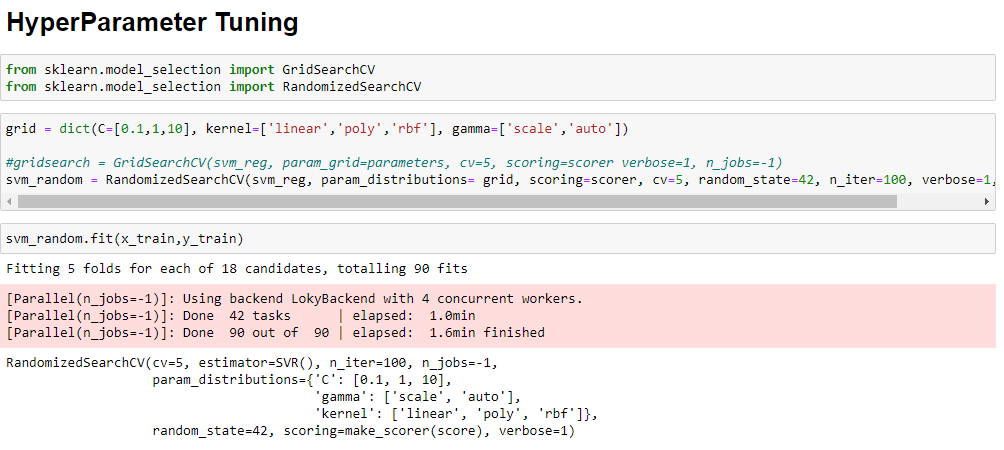
We can see the scores of the model algorithms

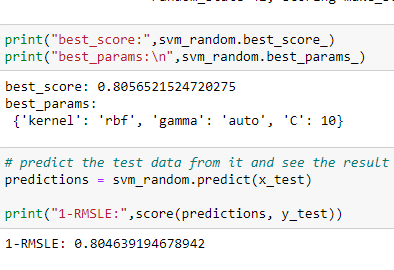
KNeighborsRegressor:0.77

SVR:0.79

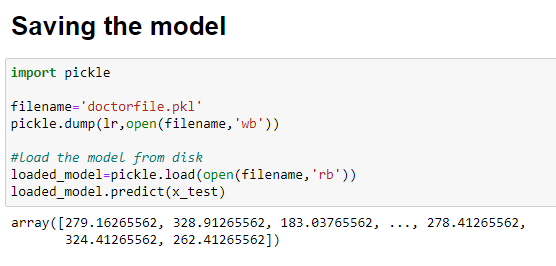
DecisionTreeRegressor:0.67

RandomForestRegressor:0.76

****

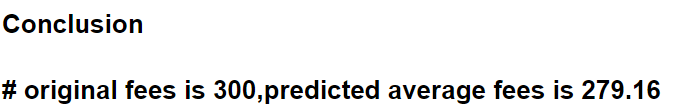
****

**Hyper parameter best score is 0.80**

****

**7.Concluding Remarks:**

Finally we shall save our predictions in pickle file.We have learnt to build a complete machine learning project. In the process, we built custom transformers that can be used with sklearn’s Pipeline class. We also learned to fine-tune our model and save it for further use.

****

**Final Conclusion :**Our Best model is SVM which gives 0.79 score.

We have printed our prediction and saved the prediction in pickle file named as doctorfile.pkl

**This Article about Doctors Fees Prediction Analysis is written by myself,**

**M.Kavitha,**

**Data Trained,**

**Batch number:1829.**

Reference was taken on the Data Analysis model done by myself.

Please click the link below to see the model solved.

<https://github.com/kavi4m/Data-Trained-Practice-Project/blob/main/Doctor%20Fee%20(3).ipynb>