**PROJECT REPORT**

***Topic: Build a Model that Classifies the Side Effects of a Drug***

**Submitted by**

**Arjun Kumar N S**

|  |  |
| --- | --- |
| Internship Project Title | Classification Model - Build a Model that Classifies the Side Effects of a Drug |
| Name of the Company | TCS ION |
| Name of the Industry Mentor | Himalaya Aashish |
| Name of the Institute | ICT Academy of Kerala |

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| **Start Date** | **End Date** | **Total Effort (hrs.)** | **Project Environment** | **Tools used** |
| 22-02-2021 | 23-03-2021 | 125 | Windows 10, Jupyter Notebook | MS Excel and Python 3.8 |

**Acknowledgement**

*“I am very much indebted to all the mentors of ICT Academy of Kerala and TCS industry mentor Himalaya Aashish for their valuable advices, suggestions and criticism during the course of our internship ".*

Place: Thiruvananthapuram Arjun Kumar N S

Date: 22/03/2021

**Abstract**

A drug, or pharmaceutical, is a substance used to prevent or cure a disease or ailment or to alleviate its symptoms. Drugs can be taken orally, via a skin patch, by injection, or via an inhaler, to name the most common methods. The development of new and improved drugs, or pharmaceuticals, is a complex and costly business.

Here we are using machine learning model to predict or classify the side effects of particular Drug called Capoten [Captopril]. Capoten (captopril) is an angiotensin converting enzyme (ACE) inhibitor prescribed for treating high blood pressure, heart failure, and for preventing kidney failure due to high blood pressure and diabetes. Capoten is available as a generic drug.

The dataset have been created in Microsoft Excel to meet the requirements of problem statement.

Following are the columns of the dataset-

* Name
* Age
* Gender
* Race
* Side effects

The features Age, Gender and Race are contributing factors or dependent values from which the Machine predict the type of Side Effect and I set the target variable as Side Effects.

**Table of Contents**

[1. Objective 6](#_Toc67407620)

[2. Introduction 6](#_Toc67407621)

[3. Problem Statement 7](#_Toc67407622)

[4. Methodology 7](#_Toc67407623)

[5. Pre-Processing 8](#_Toc67407624)

[6. Label encoding 11](#_Toc67407625)

[7. Data Visualization 11](#_Toc67407626)

[8.Feature Engineering 14](#_Toc67407627)

[9. Splitting the dataset into Target and Features 15](#_Toc67407628)

[10. Splitting x and y into Training set and testing set 16](#_Toc67407630)

[11. Data Modelling 16](#_Toc67407631)

[12. Hyper Parameter Tuning 17](#_Toc67407632)

[13. Data Evaluation 18](#_Toc67407633)

[14. Model Comparisons 22](#_Toc67407634)

[15. Result Analysis 25](#_Toc67407635)

[16. Conclusion 26](#_Toc67407636)

[17. Challenges Faced 27](#_Toc67407637)

[18. Enhancement Scope](#_Toc67407638) 27

[19. References](#_Toc67407639) 27

[20. Link to the code and executable file 28](#_Toc67407640)

**1. Objective**

The objective of this Internship project is to build a classification model that classifies the side effects of Capoten drug based on Age, Gender and Race.

**2. Introduction**

A drug is any substance that causes a change in an organism's physiology or psychology when consumed.In pharmacology, a drug is a chemical substance, typically of known structure, which, when administered to a living organism, produces a biological effect. A pharmaceutical drug, also called a medication or medicine, is a chemical substance used to treat, cure, prevent, or diagnose a disease or to promote well-being.

Side effects happen when a treatment causes a problem because it does more than treat the target issue. The impact can range from minor to severe and life-threatening.

This Data Science and Analytics internship will help to create a thorough knowledge in creating a dataset with 4 lakh instances based on the patient and a particular drug details containing the features Name, Age, Gender, Race and Side effects.

The drug I chose is **Capoten (Captopril)** which is used to treat high blood pressure (Hypertension), Lowering high blood pressure helps prevent strokes, heart attacks, and kidney problems. It is also used to treat heart failure, protect the kidneys from harm due to diabetes, and to improve survival after a heart attack.

**3. Problem Statement**

Here we are predicting the side effects of **Capoten (Captopril)**drug.This is a Classification problem. It is a supervised Learning Method since we are predicting a labelled output.

**Dataset Description**

The dataset contains 4, 00,000 rows and 5 columns.

Explanation of the features of dataset is as follows:

**Name**: It provides the name of each patient who have tried out the drug. It has duplicate string values and is of Object data type.

**Age**: Age of respective person who have tried out the drug is given here. It is of Float data type

**Gender**: Gender of the respective people who consumed drug are provided here and it is of binary categorical and of Object data type.

**Race**: Race of every person who used the drug is provided here. It is in categorical and object data type.

**Side Effects**: The type of side effect that happens to each person is provided. It is in categorical and object data type. It is used as the target value for prediction.

**4. Methodology**

*Data Collection 🡪 Data Preparation🡪Data Cleaning/Sanitize*

*🡪 Data Pre-Processing🡪 Data Visualization 🡪 Data*

*Modelling🡪 Data Evaluation*

**5. Pre-Processing**

On Jupyter Notebook, imported Pandas and Numpy libraries in order to the project. Read the dataset into the environment using Pandas library and did the pre-processing steps as follows:

* Checked the descriptive statistics.
* Checked and Handled Missing values.
* Checked and Handled Outliers.

**Handling Missing values**: Found missing values in the attributes ‘Gender, Race & Age’ using the code:

***data.isna().sum()***

Handled identified missing values using:

* Filled the attributes ‘Gender and Race’ with their respective Modes
* Filled the attribute ‘Age’ with its Median

**Handling Outliers:** Identified the presence of outliers from plotting a boxplot of attribute ‘Age’.

**Box Plot:**

Box plots (also called box-and-whisker plots or box-whisker plots) give a good graphical image of the concentration of the data. They also show how far the extreme values are from most of the data.

A box plot is constructed from five values: the minimum value, the first quartile, the median, the third quartile, and the maximum value. We use these values to compare how close other data values are to them.

To construct a box plot, use a horizontal or vertical number line and a rectangular box. The smallest and largest data values label the endpoints of the axis. The first quartile marks one end of the box and the third quartile marks the other end of the box. Approximately the middle 50 percent of the data fall inside the box. The “whiskers” extend from the ends of the box to the smallest and largest data values.

The median or second quartile can be between the first and third quartiles, or it can be one, or the other, or both.

Since the outliers obtained by using **Inter Quartile Range\*** is less than 1% of the data set, dropping them will not affect the accuracy of our analysis, so droppedthose outlier after finding their respective index rows.

**\*Inter Quartile Range (IQR):**

IQR is used to measure variability by dividing a data set into quartiles. The data is sorted in ascending order and split into 4 equal parts. Q1; Q2; Q3 called first, second and third quartiles are the values which separate the 4 equal parts.

IQR tells where the majority of data lies and is thus preferred over range. It can be used to identify outliers in a data set and gives the central tendency of the data.

* Q1 represents the 25th percentile of the data. It is defined as the middle number between the smallest number and the median of the data set.
* Q2 represents the 50th percentile of the data. It is the median of the data.
* Q3 represents the 75th percentile of the data. It is the middle value between the median and the highest value of the data set.

IQR is the range between the first and the third quartiles namely Q1 and Q3:

IQR = Q3-Q1

The data points which fall below Q1-1:5\*IQR or above Q3 + 1:5\*IQR are outliers.

The interquartile range IQR tells us the range where the bulk of the values lie.

There were **26 outlier values**.

* Handled the outlier using concepts of **Inter Quartile Range** by identifying the Upper and lower limits and obtained the index values of the outliers.
* After finding the index rows of outliers, dropped those rows because the outliers obtained was considerably negligible compared with the size of whole dataset.

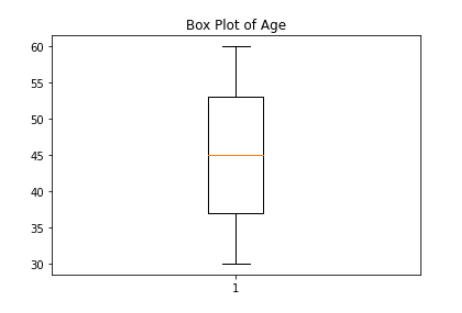


Fig-1 Box plot before removal of outliers in data

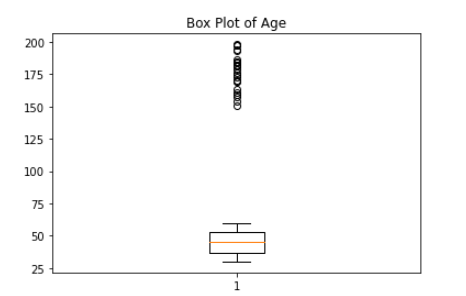


Fig-2 Box plot after removing the outliers

**6. Label encoding**

In the dataset except the attribute ‘Age’ all other attributes are in categorical nature and object data type, so need to convert them into machine readable format in order to apply in the classification models.

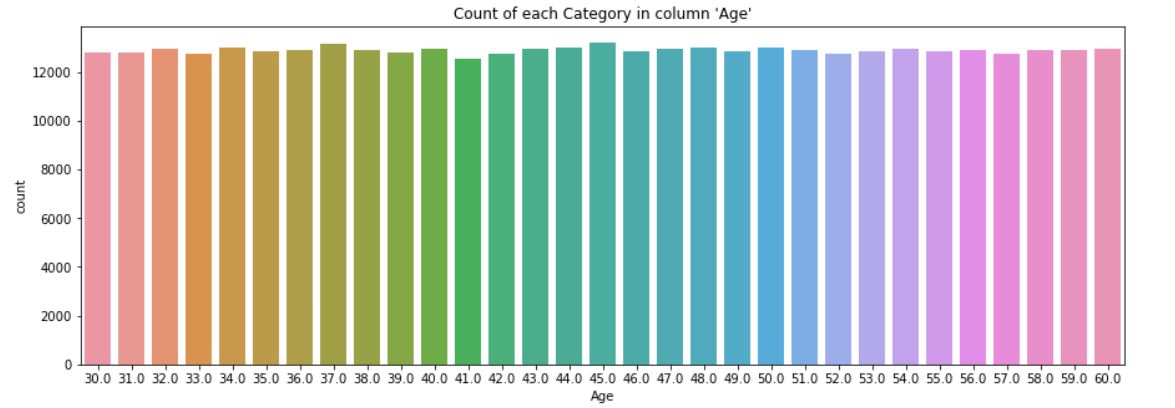
Applied Label encoding to the attributes ‘Gender’ and ‘Race’ from **Label encoder** in **scikit learn library** under preprocessing class. Did not label encode the attribute ‘Name’ because it will not contribute to the target (Side Effects).

**7. Data Visualization**

Visualized count plots of each features to obtain the count of each categories present in those attributes.

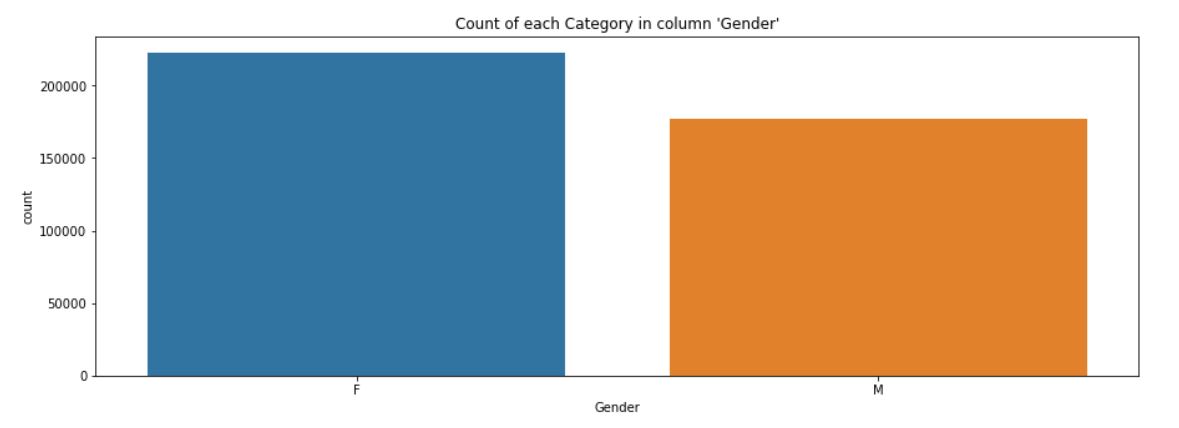
Visualized count plots of each features to obtain the count of each categories present in those attributes.

**Count plot of the feature Age**

****

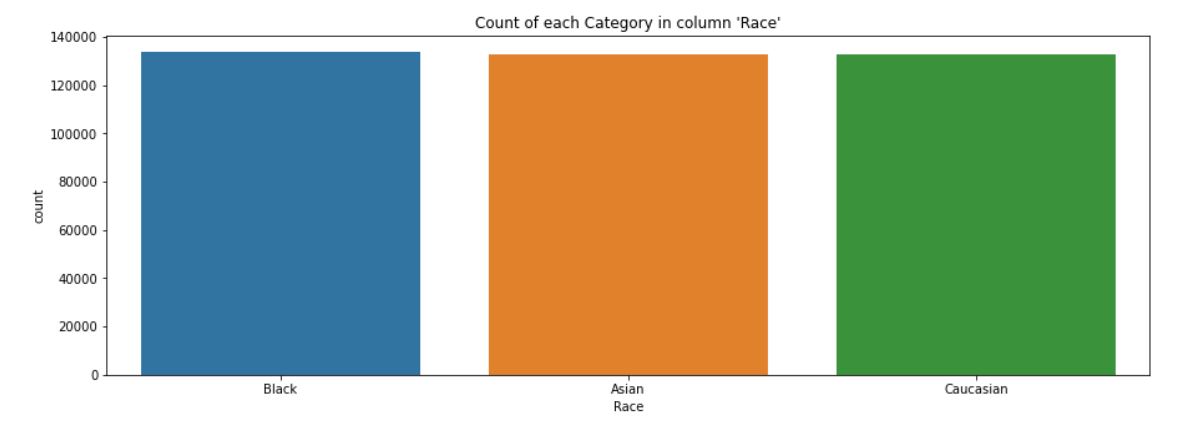
Count of each unique age is in symmetrical nature, all the ages counted above 12,000 nos. The range of age is from 30 to 60. From the above plot we can find that the patients with age of 37 and 45 experienced more Side effects than other aged patients. Therefore we can analyze that high blood pressure occurs more people whose age is at 37 and 45.

**Count plot of the feature Gender**

****

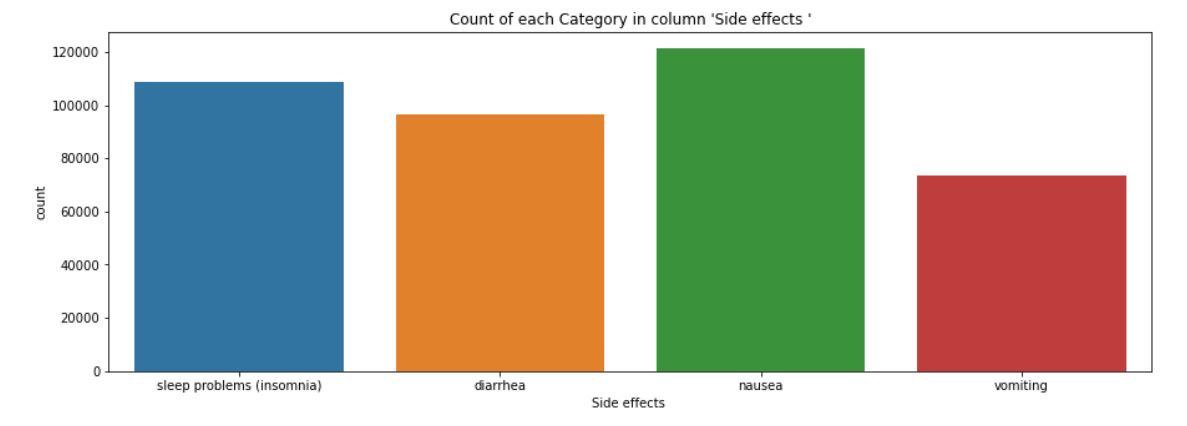
From the above plot we can identify that Side effects of the drug Capoten affects most in Female patients. Counted above 2 lakh instances while the Male patients was between 1.5 lakh and 2 lakh. By this graph it’s clear that the drug is mostly used by females than male.

**Count plot of the feature Race**

****

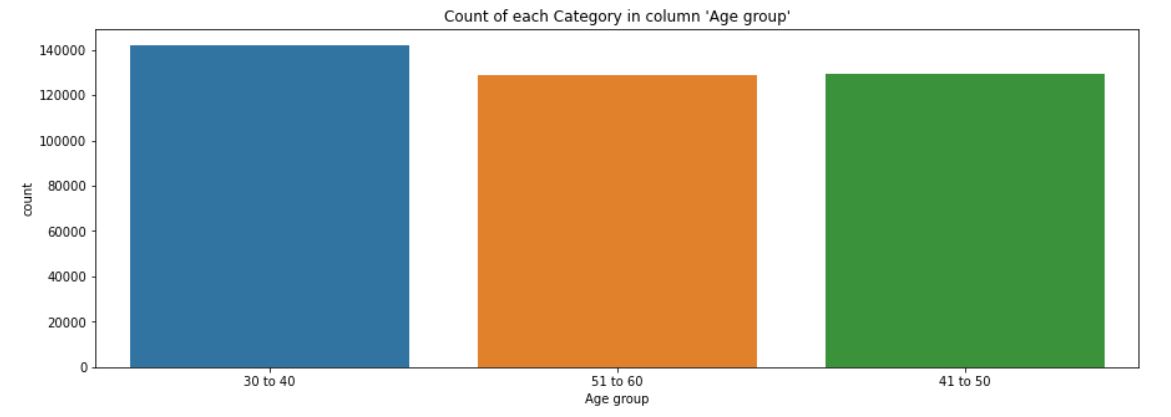
The above plot depicts the count plot of the feature Race, were the races taken into consideration was Black, Asian and Caucasian and these three was equally distributed among the 4 lakh instances (~1.3 lakh each)

**Count plot of the feature Side Effects**

****

From the above count plot, we can understand the Side Effect category named ‘Nausea’ was one of the most occurred side effects to all the patients (occurred almost to 1.2 lakh patients). The category ‘Sleep problems (insomnia)’ also affected above 1 lakh patients which is also another mostly occurred side effects to the patients.

**Count plot of the feature Age group**

****

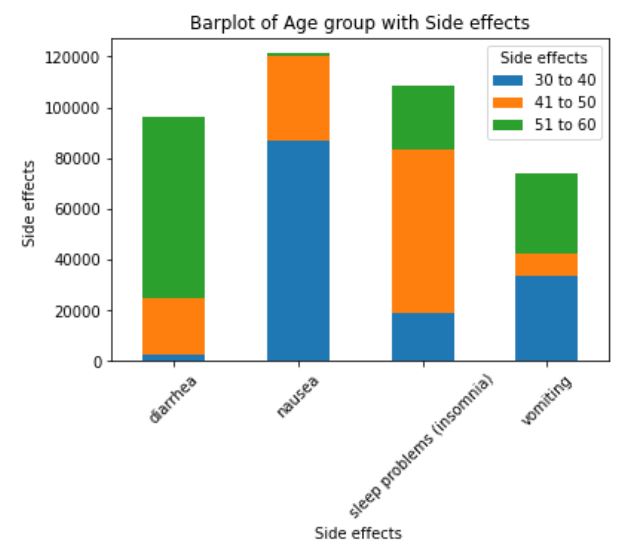
From the above graph we can identify that the age group of 30 to 40 affected most side effects than other age group. That is people whose age is between 30 and 40 tend to have more hyper tension and they are the most side effected age group who happen to consume Capoten drug.

**8. Feature Engineering**

After the data visualization, I could identify that the distribution of the attribute ‘Age’ is from 30 to 60. So I grouped the age into 3 groups as: Age from the range of ’30 to 40’, ’41 to 50’, and ’51 to 60’ using define (def.) function.

After grouping the age into three categories, reinserted the obtained values into a new feature named ‘Age group’ with the dataset.

**Stacked Bar Plot of Side effects and Age group**



From the stacked bar plot we can find that the age group ‘30 to 40’ is highly affected to all the side effects and the nausea is the most occurred side effects to the patients. Least side effect affected age group is ’51 to 60’ and the least occurred side effect is vomiting.

**9. Splitting the dataset into Target and Features.**

Here the target is ‘Side Effects’ & stored in the variable ‘y’ and the features include ‘Age, Race, Gender ’ which is stored in variable ‘x’.



Here we are dropping the ‘Age group” attribute which was formed by using feature engineering, as it was created for the sole purpose of visualization only and also it is considered for model building.

**10. Splitting x and y into Training set and testing set.**

Assigned 80% of data to the training set and 20% of the data into testing test using **train\_test\_split** module from **Sci-kit learn library** in order to train the models.

**11. Data Modeling**

It is a `Supervised Learning' method since we are predicting labelled output with the labelled input. The response we are getting here is in a Categorical nature.

We tried different classification models for predicting the landslides: Logistic regression, Decision Tree, Random Forest, KNN, Extreme Gradient Boost, and Gradient Boost Classifier. From these classification model I chose Decision Tree Classifier as our Final Model. The details of the Decision Tree Classifier Model is explained below.

**Decision Tree Classifier**

Decision Tree is a supervised learning technique that can be used for both classification and Regression problems, but mostly it is preferred for solving Classification problems. It is a tree-structured classifier, where internal nodes represent the features of a dataset, branches represent the decision rules and each leaf node represents the outcome.In a Decision tree, there are two nodes, which are the Decision Node and Leaf Node. Decision nodes are used to make any decision and have multiple branches, whereas Leaf nodes are the output of those decisions and do not contain any further branches. The decisions or the test are performed on the basis of features of the given dataset.

It is called a decision tree because, similar to a tree, it starts with the root node, which expands on further branches and constructs a tree-like structure.

**How does the Decision Tree algorithm Work?**

In a decision tree, for predicting the class of the given dataset, the algorithm starts from the root node of the tree. This algorithm compares the values of root attribute with the record (real dataset) attribute and, based on the comparison, follows the branch and jumps to the next node.

For the next node, the algorithm again compares the attribute value with the other sub-nodes and move further. It continues the process until it reaches the leaf node of the tree. The complete process can be better understood using the below algorithm:

* **Step-1:** Begin the tree with the root node, says S, which contains the complete dataset.
* **Step-2:** Find the best attribute in the dataset using **Attribute Selection Measure (ASM).**
* **Step-3:** Divide the S into subsets that contains possible values for the best attributes.
* **Step-4:** Generate the decision tree node, which contains the best attribute.
* **Step-5:** Recursively make new decision trees using the subsets of the dataset created in step-3. Continue this process until a stage is reached where you cannot further classify the nodes and called the final node as a leaf node.

**12. Hyper Parameter Tuning**

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

The same kind of machine learning model can require different constraints, weights or learning rates to generalize different data patterns. These measures are called hyper-parameters, and have to be tuned so that the model can optimally solve the machine learning problem. Hyper-parameter optimization finds a tuple of hyper-parameters that yields an optimal model which minimizes a predefined loss function on given independent data. The objective function takes a tuple of hyper parameters and returns the associated loss.

**Decision Tree Classifier Hyper parameter tuning:**

*Criterion:*The function to measure the quality of a split. Supported criteria are “gini” for the Gini impurity and “entropy” for the information gain.

*Max depth:* The maximum depth of the tree. If none, then nodes are expanded until all leaves are pure or until all leaves contain less than min samples split samples.

In the decision tree classifier model, Tried the model by changing the parameter named ‘Criterion’ with ‘gini and entropy’ and changing the parameter ‘max-depth’ with ‘2, 4, 6, 8, 10, 12’ and checked for each iterations in each of the parameters and found that the scores are high at either of the criterion as gini or entropy as 12 in max depth.

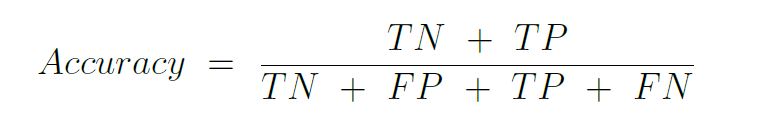
**13. Data Evaluation**

After prediction, we got the highest scores in **Accuracy Scores, Recall Scores, Precision Score, and F1 Score** for the classification models, Decision Tree classifier, Random Forest, Gradient Boost Classifier, and Extreme Gradient Boost Model.

Among these models, I chose **Decision Tree Classifier** as our model for predicting the side effects since it took less run time for the results compared with other models.

**Accuracy Score**

Accuracy represents the number of correctly classified data instances over the total number of data instances.



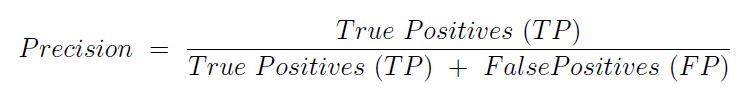
Accuracy may not be a good measure if the dataset is not balanced (both negative and positive classes have different number of data instances).

**Precision- Recall**

Precision and Recall helps us further understand how strong the accuracy shown holds true for a particular problem.

**Precision** (also called positive predictive value) is the fraction of relevant instances among the retrieved instances, while Recall (also known as sensitivity) is the fraction of the total amount of relevant instances that were actually retrieved.

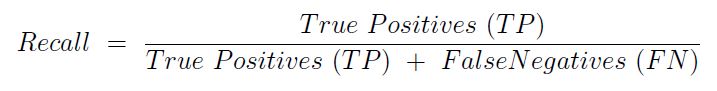
Both precision and recall are therefore based on an understanding and measure of relevance.



Precision should ideally be 1 (high) for a good classifier. Precision becomes 1 only when the numerator and denominator are equal i.e., TP = TP + FP, this also means FP is zero.

As FP increases the value of denominator becomes greater than the numerator and precision value decreases (which we don't want).

**Recall** in simple terms means, what percentage of actual positive predictions were correctly classified by the classifier.



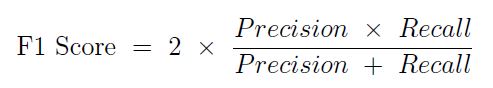
Recall should ideally be 1 (high) for a good classifier. Recall becomes 1 only when the numerator and denominator are equal i.e., TP = TP + FN, this also means FN is zero. As FN increases the value of denominator becomes greater than the numerator and recall value decreases (which we don't want).

As default Sci-kit learn uses binary average for the calculation of Precision and Recall in binary classifications.

Since our target is multiclass/multilabel we can't use binary as average for the calculation, so here I used weighted average because it calculate metrics for each label, and find their average weighted by support. (The number of true instances for each label.)

**F1 Score**

In statistical analysis of binary classification, the F1 score (also F-score or F-measure) is a measure of a test's accuracy. It is calculated from the precision and recall of the test, where the precision is the number of correctly identified positive results divided by the number of all positive results, including those not identified correctly, and the recall is the number of correctly identified positive results divided by the number of all samples that should have been identified as positive.



F1 Score becomes 1 only when precision and recall are both 1. F1 score becomes high only when both precision and recall are high. F1 score is the harmonic mean of precision and recall and is a better measure than accuracy.

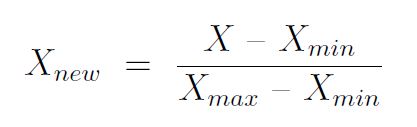
As default Sci-kit learn uses binary average for the calculation of F1 Score in binary classifications.

Since our target is multiclass/multilabel we can't use binary as average for the calculation, so here I used weighted average because it calculate metrics for each label, and find their average weighted by support. (The number of true instances for each label.)

Then **scaled the data using the Normalizing method** from the Sci-kit learn library so that the data set values will fall in a smaller range (0-1)

**Normalization**

In statistics and applications of statistics, normalization can have a range of meanings. In the simplest cases, normalization of ratings means adjusting values measured on different scales to a notionally common scale, often prior to averaging. In more complicated cases, normalization may refer to more sophisticated adjustments where the intention is to bring the entire probability distributions of adjusted values into alignment. A different approach to normalization of probability distributions is quantile normalization, where the quantiles of the different measures are brought into alignment.

****

Where;

* X: It is a set of the observed values present in X.
* Xmin: It is the minimum values in X
* Xmax: It is the maximum values in X

Normalization refers to rescaling real-valued numeric attributes into a 0 to 1 range.

Data normalization is used in machine learning to make model training less sensitive to the scale of features. This allows our model to converge to better weights and, in turn, leads to a more accurate model.

Normalization makes the features more consistent with each other, which allows the model to predict outputs more accurately.

Python provides the preprocessing library, which contains the normalize function to normalize the data. It takes an array in as an input and normalizes its values between 0 & 1. It then returns an output array with the same dimensions as the input.

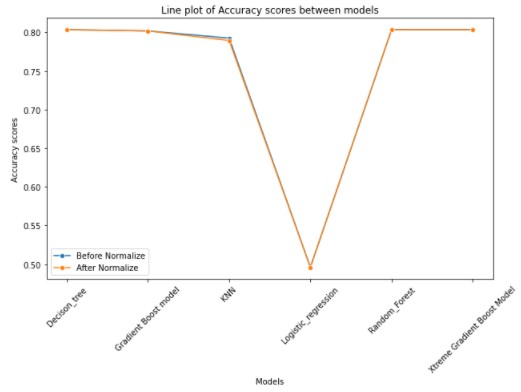
After Normalization, rechecked the Scores for all the models and there was no improvements in the scores after scaling.

**The evaluation criteria I followed was the score of F1**, because in most real-life classification problems, imbalanced class distribution exists and thus F1-score is a better metric to evaluate our model.

Also F1 score is the harmonic mean of precision and recall and is a better measure than accuracy.

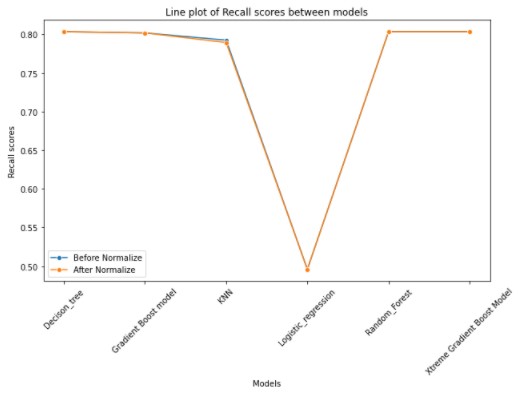
**14. Model Comparisons**

**Line plot of Accuracy Scores between Models**

****

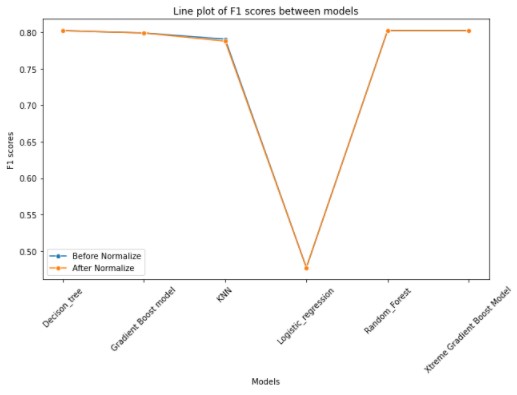
From the above graph , we are able to figure out that most of the algorithms have similar trend of accuracy except Logistic Regression where accuracy was low compared to other models.

**Line plot of Recall Scores between Models**

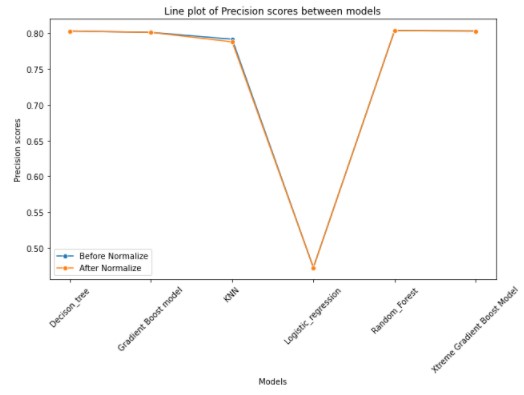
****

Here we can see that the Recall scores are having same trend as Accuracy , Decision Tree models has highest Recall score and lowest Recall score is obtained in Logistic regression model.

**Line plot of F1 Scores between Model**

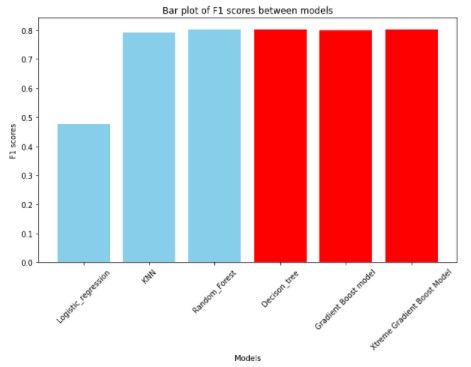
****

Here we can see that the F1 scores are having same trend as Accuracy , Decision Tree models has highest F1 score and lowest F1 score is obtained in Logistic regression model.

**Line plot of Precision Scores between Models**

Here we are able to witness from the graph that the precision score of Decision tree model is higher compared to other models.

**Bar plot of F1 scores between Models**

****

From this Barplot , the representation of different F1 scores are provided for analysis. Here Logistic regression has the lowest F1 score compared to other models.

**15. Result Analysis**

From the model evaluations and model comparison steps, we can find that the classification models, Random Forest, Decision Tree, Gradient Boost model, and Extreme Gradient Boost Model gives the highest and equal F1 scores. Therefore based on the execution time of the program code, I chose the Classification Model Decision Tree classifier as the final model for predicting the classification of Side effects of the drug Capoten.

# 16. Conclusion

Captopril, sold under the brand name Capoten among others, is an angiotensin-converting enzyme (ACE) inhibitor used for the treatment of hypertension and some types of congestive heart failure.Captopril's main uses are based on its vasodilation and inhibition of some renal function activities. These benefits are most clearly seen in: 1) Hypertension 2) Cardiac conditions such as congestive heart failure and after myocardial infarction 3) Preservation of kidney function in diabetic nephropathy.

Additionally, it has shown mood-elevating properties in some patients. This is consistent with the observation that animal screening models indicate putative antidepressant activity for this compound, although one study has been negative. Formal clinical trials in depressed patients have not been reported. It has also been investigated for use in the treatment of cancer.

By doing this internship, apart from Side effect classification prediction I was able to narrow out certain points from data exploration and visualization.

The patients with age of 37 and 45 experienced more Side effects than other patients. Therefore we can analyze that high blood pressure occurs more people whose age is at 37 and 45. Side effects of the drug Capoten affects most in Female patients. Counted above 2 lakh instances while the Male patients was between 1.5 lakh and 2 lakh. Therefore the drug is mostly used by females than male. The age group ‘30 to 40’ is highly affected to all the side effects and the nausea is the most occurred side effects to the patients. Least side effect affected age group is ’51 to 60’ and the least occurred side effect is vomiting. People whose age is between 30 and 40 tend to have more hyper tension and they are the most side effected age group who happen to consume Capoten drug.

**17. Challenges Faced**

The main challenges I faced while doing this internship was with creating the dataset and cleaning the dataset. The data set created contains 4 lakh instances and 5 features. I made sure the created dataset contains missing values and outliers using MS Excel in order to apply pre-processing concepts in python.

# 18. Enhancement Scope

As further research, by collecting the details of the medical conditions, health parameters, diet, job (to find whether the job will affect the health of a patient) and location of the patients will be taken as an contributing features to the target variable and we can obtain different insights from relation between the features with the target.

Building an application to predict the classification of side effects will be another area which can be further done after this internship.

# 19. References

* Classification Algorithms: <https://youtu.be/ppXFoltcX7A>
* Data preprocessing: <https://youtu.be/NBm4etNMT5k>
* Training and Testing Data: <https://youtu.be/fwY9Qv96DJY>
* Classification and Regression: <https://youtu.be/TJveOYsK6MY>
* 7 Types of Classification Algorithms - Analytics India Magazine:<https://analyticsindiamag.com/7-types-classification-algorithms/>
* Machine learning classifiers: <https://towardsdatascience.com/machine-learning-classifiers>
* Train test split for evaluating machine learning algorithms: <https://machinelearningmastery.com/train-test-split-for-evaluating-machine-learning-algorithms/>
* Data pre-processing, Python: <https://www.kdnuggets.com/2020/07/easy-guide-data-preprocessing-python.html>

# 20. Link to the code and Executable file