**NIT’S and IIT’S 2016 cut-off’s**

*A report submitted in partial fulfillment of the requirements for the Award of Degree of*

**BACHELOR OF TECHNOLOGY**

**in**

**COMPUTER SCIENCE AND BUSINESS SYSTEMS ENGINEERING**

**By**

**Name of student: P.B.V.Ashritha**

**Regd. No.: 20B91A5753**

**Under Supervision of Mr. Gundala Nagaraju**

**Henotic Technology Pvt Ltd, Hyderabad**

**(Duration: 7th July, 2022 to 6th September, 2022)**



**DEPARTMENT OF INFORMATION TECHNOLOGY**

**SAGI RAMA KRISHNAM RAJU ENGINEERING COLLEGE**

(An Autonomous Institution)

Approved by AICTE, NEW DELHI and Affiliated to JNTUK, Kakinada

CHINNA AMIRAM, BHIMAVARAM,

ANDHRA PRADESH

Table of Contents

[**1.0** **Introduction** 4](#_Toc110502057)

[**1.1.** **What are the different types of Machine Learning?** 4](#_Toc110502058)

[**1.2.** **Benefits of Using Machine Learning in NIT’S and IIT’S** 4](#_Toc110502059)

[**1.3.** **About Industry** 5](#_Toc110502060)

[1.3.1 AI / ML Role in NIT’S and IIT’S 5](#_Toc110502061)

[**2.0** **NIT’S and IIT’S Claims (your internship project)** 6](#_Toc110502062)

[**2.1.** **Internship Project - Data Link** 7](#_Toc110502064)

[**3.0** **AI / ML Modelling and Results** 8](#_Toc110502065)

[**3.1.** **Your Problem of Statement** 8](#_Toc110502066)

[**3.2.** **Data Science Project Life Cycle** 8](#_Toc110502067)

[3.2.1 Data Exploratory Analysis 9](#_Toc110502068)

[3.2.2 Data Pre-processing 9](#_Toc110502069)

[3.2.2.1. Check the Duplicate and low variation data 9](#_Toc110502070)

[3.2.2.2. Identify and address the missing variables 10](#_Toc110502071)

[3.2.2.3. Handling of Outliers 10](#_Toc110502072)

[3.2.2.4. Categorical data and Encoding Techniques 10](#_Toc110502073)

[3.2.2.5. Feature Scaling 10](#_Toc110502074)

[3.2.3 Selection of Dependent and Independent variables 10](#_Toc110502075)

[3.2.4 Data Sampling Methods 10](#_Toc110502076)

[3.2.4.1. Stratified sampling 10](#_Toc110502077)

[3.2.4.2. Simple random sampling 11](#_Toc110502078)

[3.2.5 Models Used for Development 11](#_Toc110502079)

[3.2.5.1. Model 01 11](#_Toc110502080)

[3.2.5.2. Model 02 11](#_Toc110502081)

[3.2.5.3. Model 03 11](#_Toc110502082)

[3.2.5.4. Model 04 11](#_Toc110502083)

[3.2.5.4. Model 05 11](#_Toc110502083)

[3.2.5.4. Model 06 11](#_Toc110502083)

[3.2.5.4. Model 07 11](#_Toc110502083)

[3.2.5.4. Model 08 11](#_Toc110502083)

[3.2.5.4. Model 09 11](#_Toc110502083)

[3.2.5.5. Model 10 11](#_Toc110502084)

[**3.3.** **AI / ML Models Analysis and Final Results** 12](#_Toc110502085)

[3.3.1 Different Model codes 12](#_Toc110502086)

[3.3.2 Random Forest Python Code 12](#_Toc110502087)

[3.3.3 Extra Trees Python code 12](#_Toc110502088)

[**4.0** **Conclusions and Future work** 13](#_Toc110502089)

[**5.0** **References** 14](#_Toc110502090)

**Abstract**

In India, every year lacs of students sit for competitive examinations like JEE Advanced, JEE Mains, NEET, etc. These exams are said to be the gateway to get admission into India's premier Institutes such as IITs, NITs, AIIMS, etc. Keeping in mind that the competition is tough as lacs of students appear for these examinations, there has been an enormous development in Ed Tech Industry in India, fortuning the dreams of lacs of aspirants via providing online as well as offline coaching, mentoring, etc.

[Data Science,](https://www.analyticsinsight.net/quantum-computing-in-data-science-the-weird-yet-profitable-duo/) an evolving technology, has been the most talked-about term in the technology and information industry in recent years. One of the most common career options for IT professionals in data science. The eminent educational institute, the [Indian Institute of Technology or IIT, is offering several PG and UG courses in Data Science](https://www.analyticsinsight.net/10-brand-new-data-science-courses-by-iits-and-iims-for-2022/) with a practical curriculum. This will help aspiring data scientists to get recruited in reputed organizations around the world with lucrative salary packages. IIT Madras, IIT Guwahati, IIT Hyderabad, IIT Mandi, IIT Bombay, IIT Kharagpur, and IIT Kanpur are offering [Data Science courses](https://www.analyticsinsight.net/top-10-python-data-science-courses-you-should-take-up-in-2022/) while the other [IITs](https://www.analyticsinsight.net/innovations-top-iits-unravelling-artificial-intelligence-initiatives/)are offering certificate programmes in Data Science. Let’s take a look at the top IITs and NITs offering data science courses in 2022. The National Institutes of Technology (NITs) are governed by the National Institutes of Technology Act, 2007 whereas Indian Institutes of Technology (IITs) are governed by the Institutes of Technology Act, 1961.

# **1.0 Introduction**

**NIT Cut off** is the minimum marks or ranks that the candidates must secure for admission to BTech courses. Joint Seat Allocation Authority (JoSAA) will release NIT cutoff scores after each round of counselling. The cutoff of NIT is released in the form of opening and closing ranks as per the candidate category. The cutoff scores of NITs will be different for all categories depending on various factors determined by the authority. All the engineering aspirants seeking admission into 31 [NITs in India](https://www.shiksha.com/b-tech/articles/nits-in-india-blogId-20003) will have to score more than or equal to NIT cut off scores.

Only the eligible candidates who have secured the minimum NIT cutoff marks would be able to apply for admission in the respective institute as per their JEE Main 2022 ranks. National Testing Agency (NTA) conducted [JEE Main 2022](https://www.shiksha.com/engineering/jee-main-exam) twice- First session exam was held in June and the second session was in July.

With the declaration of JEE Main result 2021, the institute-wise allotment of seats for admission to BTech courses will be released by JoSAA. Firstly, candidates who fall within the JoSAA opening & closing ranks for the preferred college and course are shortlisted for admission. Secondly, shortlisted candidates must participate in JoSAA counselling process for seat allotment at the reporting institute for document verification and payment of fees. Candidates are called as per the JEE Main 2021 All India Rank (AIR) and availability of seats.

**IIIT Cut-off 2022:**The JEE Main IIIT cut-off will be released online by the Joint Seat Allocation Authority (JoSAA). After completing the JEE Main exam, the NTA will announce the JEE Main 2022 cut-offs. These cut-offs will be available to candidates by opening and closing ranks. The IIIT cut-offs will help aspirants understand whether or not they are eligible for admission to different courses they have applied for. Also, the IIIT admission cut-offs are highly influenced by some determining factors that we have furnished in this article.

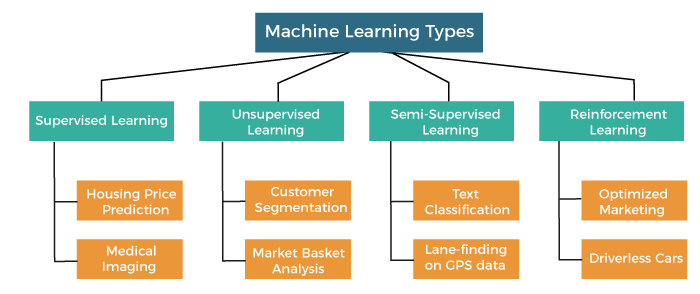
The IIIT cut-off is the minimum score required for admission to various IIIT programmes. The IIIT cut-off for top engineering colleges is established by a number of factors, including seat availability, overall student enrollment, exam difficulty level, and other variables. Continue reading to learn more about the previous years’ cut-offs, institute-wise cut-offs, branch-wise cut-offs, and other crucial updates and trends on the IIIT 2022 cut-off.

## **What are the different types of Machine Learning?**

**Machine learning is a subset of AI, which enables the machine to automatically learn from data, improve performance from past experiences, and make predictions**. Machine learning contains a set of algorithms that work on a huge amount of data. Data is fed to these algorithms to train them, and on the basis of training, they build the model & perform a specific task.

These ML algorithms help to solve different business problems like Regression, Classification, Forecasting, Clustering, and Associations, etc.

Based on the methods and way of learning, machine learning is divided into mainly four types, which are:

1. Supervised Machine Learning
2. Unsupervised Machine Learning
3. Semi-Supervised Machine Learning
4. ReinforcementLearning

## **1. Supervised Machine Learning**

As its name suggests, [Supervised machine learning](https://www.javatpoint.com/supervised-machine-learning) is based on supervision. It means in the supervised learning technique, we train the machines using the "labelled" dataset, and based on the training, the machine predicts the output. Here, the labelled data specifies that some of the inputs are already mapped to the output. More preciously, we can say; first, we train the machine with the input and corresponding output, and then we ask the machine to predict the output using the test dataset.

### **Categories of Supervised Machine Learning**

Supervised machine learning can be classified into two types of problems, which are given below:

* **Classification**
* **Regression**

### **a. Classification**

Classification algorithms are used to solve the classification problems in which the output variable is categorical, such as "**Yes" or No, Male or Female, Red or Blue, etc**. The classification algorithms predict the categories present in the dataset. Some real-world examples of classification algorithms are **Spam Detection, Email filtering, etc.**

### **b. Regression**

Regression algorithms are used to solve regression problems in which there is a linear relationship between input and output variables. These are used to predict continuous output variables, such as market trends, weather prediction, etc.

## **2. Unsupervised Machine Learning**

[Unsupervised learnin](https://www.javatpoint.com/unsupervised-machine-learning)g is different from the Supervised learning technique; as its name suggests, there is no need for supervision. It means, in unsupervised machine learning, the machine is trained using the unlabeled dataset, and the machine predicts the output without any supervision.

### **Categories of Unsupervised Machine Learning**

Unsupervised Learning can be further classified into two types, which are given below:

* **Clustering**
* **Association**

### **1 . Clustering**

The clustering technique is used when we want to find the inherent groups from the data. It is a way to group the objects into a cluster such that the objects with the most similarities remain in one group and have fewer or no similarities with the objects of other groups. An example of the clustering algorithm is grouping the customers by their purchasing behaviour.

### **2. Association**

Association rule learning is an unsupervised learning technique, which finds interesting relations among variables within a large dataset. The main aim of this learning algorithm is to find the dependency of one data item on another data item and map those variables accordingly so that it can generate maximum profit. This algorithm is mainly applied in **Market Basket analysis, Web usage mining, continuous production**, etc.

## **3. Semi-Supervised Learning**

**Semi-Supervised learning is a type of Machine Learning algorithm that lies between Supervised and Unsupervised machine learning**. It represents the intermediate ground between Supervised (With Labelled training data) and Unsupervised learning (with no labelled training data) algorithms and uses the combination of labelled and unlabeled datasets during the training period.

## **4. Reinforcement Learning**

## **Reinforcement learning works on a feedback-based process, in which an AI agent (A software component) automatically explore its surrounding by hitting & trail, taking action, learning from experiences, and improving its performance.**

# 

# **1.2 Benefits of Using Machine Learning in NIT’S and IIT’S**

Banks can benefit from ML since it helps adopt great management in an organization, enhance customer satisfaction, get simplified deployment and operations — let's make a fair balance by describing the benefits.

1. **Improved decision making**

One of the benefits of machine learning in banking is improved decision making. As compared to traditional methods, artificial intelligence helps banks to calculate credit scores accurately. The main reason ML can do this is that it can provide an objective evaluation without any bias. The huge amount of data collected from the potential borrower assists banks in making better decisions.

**2.** **Better risk management**

AI and ML reduce risks for both customers and banks through accurate reporting. Artificial intelligence can also make predictions based on transaction history after giving credit to customers. Employees have more insights into credit risk testing. Early detection of errors and the availability of potential future risks helps the banking industry to prepare in advance.

**3.** **Prevention of fraud**

Credit card fraud is a huge problem in the banking industry. Machine learning for banking can significantly lower the number of fraudulent activities. The majority of fraud occurs when customers pay for products, whether online or offline. Machine learning in banking prevents this from happening in several ways. For example, facial recognition can be used to confirm the person using a credit card is the owner.

**4.** **Improved customer experience**

With technology changing almost every aspect of life, consumers are looking for better services and eager to get the same from banking institutions. At the same time, banks that can provide more security and a personalized experience would attract more clients. Customers want digital banking products that are easy to use. One way in which ML improves the overall experience and services is by reducing the time it takes to make credit decisions and banking operations. Loan application, which used to take weeks, can now be made within days. Machine learning can make an unbiased analysis based on several credit factors.

**5. Internal operational solutions**

Machine learning in the banking sector has greatly changed internal operations for the better. Automation reduces the time staff spends on redundant tasks. Therefore, resources can be allocated towards improving the overall experience. Robots perform routine tasks with minimal risk of errors. So a bank can provide efficient solutions while automation gives employees the chance to pay more attention to the most important tasks.

Using ML has so many advantages, with the most important one being internal operational solutions today. Robots can go through a customer database at record time, thus reducing the need for employees to do this manually.

**6.**  **Marketing and lending solutions**

ML and AI in fintech collect data and also search for specific patterns that help banks make better marketing predictions. Examples of predictions that ML can make include:

• Changes in currencies,

• The best investment ideas,

• Credit risks,

• The optimum loan agreement for a client.

This data assists a bank in deciding where to invest, thus increasing their revenue. It also provides more accurate information on how to attract new clients.

**7.** **More personalization**

Banks can benefit from ML since it helps adopt great management in an organization, enhancing customer satisfaction, and providing more personalized and simplified operations and support.

# **1.3 About Industry**

The [National Institutes of Technology](https://en.m.wikipedia.org/wiki/National_Institutes_of_Technology) (NITs) are the autonomous public Universities which are controlled by government of India, there are 31 universities and all of them are autonomous yet they are linking with each other through a common council known as council of NITSER, this council will basically look after the administration and other important things of this universities and as I mentioned earlier the university is funded and run by the Central Government.

### **AI / ML Role in NIT’S and IIT’S**

Machine Learning is a sub-set of artificial intelligence where computer algorithms are used to autonomously learn from data. Machine learning (ML) is getting more and more attention and is becoming increasingly popular in many other industries. Within the insurance industry, there is more application of ML regarding the claims.

Product managers, along with insights from analysts and other business stakeholders decide the high-impact features that could benefit from Machine Learning. Any business requirement needs to be translated to a data science problem. For example, consider an online fashion website that is seeing a large number of clothing items being returned. The analysts found that this is because they tend to choose a smaller/bigger size while ordering. The data science approach would include training a model using user and product information and recommend a correct size.

However, just because the project makes sense, it doesn’t mean that the company should pursue it straight away. The business team first estimates the amount of savings in shipping an ML model would generate. Finally, since the company has limited human resources, it has to decide how feasible this project would be, at this point in time, as compared to other projects. These decisions are based on the company’s strategy, vision, and the calculated RoIs.

Generally, senior and lead data scientists in a company work with business managers to define a clear data science problem.

# **2.0 Claims**

It is very difficult for students to choose a college of their choice after clearing the board exams because of the high level of competition among IITs vs NITs. Just a few years ago, the competition was very low compared to the present. Around a decade ago, the number of top engineering institutions of India was limited and the level of education was low. But now, with the improvement in education and infrastructure, the number of IITs has increased, with a sharp increase in the number of students aiming for top-tier IITs. This has raised the bar for competition among IITs vs NITs.

Almost, each and every IIT is considered top-notch just because of the name ‘IIT’ associated with it. This is the reason why a lot of students do not even consider NITs. Moreover, old NITs have lost their glory with the newly established IITs. But, if considered thoroughly, NITs are now the hub of talent for industries. Let us compare in detail the difference between IIT and NIT, what should a student choose after JEE Main and how the procedure of admission through [**JEE Main**](https://zollege.in/exams/jee-main) follows.

The main factors for nit’s and iit’s are age, job, marital, education, default, housing, loan, contact, month, day\_of\_week, campaign, pdays, previous, poutcome, emp.var.rate , cons.price, cons.conf, euribor3m, employed and y(claim).

# **2.1 Intership Project-Data-Link**

The internship project data has taken from Kaggle and the link is https://www.kaggle.com/datasets/rumbleftw/iit-nit-data

# **3.0 AI / ML Modelling and** **Results**

# 

## **3.1 Your Problem of Statement**

Predictive models are most effective when they are constructed using acompany’s own historical claims data since this allows the model to recognize the specific nature of a company’s exposure as well as its claims practices. The construction of the model also involves input from the company throughout the process, as well as consideration of industry leading claims practices and benchmarks.

Predictive modelling can be used to quantify the impact to the claims department resulting from the failure to meet or exceed claim service leading practices. It can also be used to identify the root cause of claim leakage. Proper use of predictive modelling will allow for potential savings across two dimensions:

* Early identification of claims with the potential for high leakage, thereby allowing for the proactive management of the claim
* Recognition of practices that are unnecessarily increasing claims settlement payments.

## **3.2** **Data Science Project Life Cycle**

Data Science is a multidisciplinary field of study that combines programming skills, domain expertise and knowledge of statistics and mathematics to extract useful insights and knowledge from data

### TDSP lifecycle

### **3.2.1 Data Exploratory Analysis**

Exploratory data analysis has been done on the data to look for relationship and correlation between different variables and to understand how they impact or target variable.

The exploratory analysis is done for Auto Quote / Policy Conversionwith different parameters and all the charts are presented in **Appendices 6.2 - List of charts (6.2.1 to 6.2.9)**

### **3.2.2 Data Pre-processing**

We removed variables which does not affect our target variable(Claimed Target)as they may add noise and also increase our computation time, we checked the data for anomalous data points and outliers. We did principal component analysis on the data set to filter out unnecessary variables and to select only the important variables which have greater correlation with our target variable.

### **Check the Duplicate and low variation data**

As companies collect more and more data about their customers, an increased amount of duplicate information starts appearing in the data as well, causing a lot of confusion among internal teams. Having duplicate data means that the model will be trained more on that type of data and thus will be biased.  These duplicates add weight to the fit of those specific observations (cases). Whether this effect is big or small is hard to tell from the information we have provided. So that we checked the duplicates in our dataset as they are not required and also increase our computation time.

### 3.2.2.2 **Identify and address the missing variables**

### 

The real-world data often has a lot of missing values. Missing values are representative of the messiness of real world data. There can be a multitude of reasons why they occur — ranging from human errors during data entry, incorrect sensor readings, to software bugs in the data processing pipeline. The cause of missing values can be data corruption or failure to record data. The handling of missing data is very important during the pre-processing of the dataset as many machine learning algorithms do not support missing values. We usually use the simple imputer and KNN imputer to identify the missing values.

### 3.2.2.3 **Handling of Outliers**

An outlier is an object that deviates significantly from the rest of the objects. They can be caused by measurement or execution error. The analysis of outlier data is referred to as outlier analysis or outlier mining.  There is no one method to detect outliers because of the facts at the center of each dataset. One dataset is different from the other. A rule-of-the-thumb could be that you, the domain expert, can inspect the unfiltered, basic observations and decide whether a value is an outlier or not. These are different outlier methods for outlier analysis: Box plot, Scatter plot, Mathematical Function. There are different types of outliers:  Global Outliers, Contextual Outliers, Collective Outliers.

### **Categorical data and Encoding Techniques**

Machine learning models require all input and output variables to be numeric. This means that if your data contains categorical data, you must encode it to numbers before you can fit and evaluate a model. The two most popular techniques are an**Ordinal Encoding and a One-Hot Encoding**. They are 2 types of techniques we are using in this dataset one is label encoder and other is label Binarizer. Label encoder is used for ordinal data and label Binarizer is used for numerical data. The other technique is manual which was used sometimes.

### **Feature Scaling**

### Feature scaling is the process of normalising the range of features in a dataset. There are three different types of feature scaling : Min Max Scaler , Standard Scaler, Robust Scaler.

Min Max scaler : Normalisation, also known as min-max scaling, is a scaling technique whereby the values in a column are shifted so that they are bounded between a fixed range of 0 and 1.

Standardization : On the other hand, standardisation or Z-score normalisation is another scaling technique whereby the values in a column are rescaled so that they demonstrate the properties of a standard Gaussian distribution, that is mean = 0 and variance = 1.

Robust scaler :  [Robust scaler](https://scikit-learn.org/stable/modules/generated/sklearn.preprocessing.RobustScaler.html) scales features using statistics that are robust to outliers. More specifically, Robust scaler removes the median and scales the data according to the interquartile range, thus making it less susceptible to outliers in the data.

### **3.2.3** **Selection of Dependent and Independent variables**

The variable that are not affected by the other variables are called independent variables. The variables which depend on other variables or factors. We expect these variables to change when the independent variables, upon whom they depend, undergo a change. The dependent or target variable here isClaimed Target which tells us a particular policy holder has filed a claim or not the target variable is selected based on our business problem and what we are trying to predict.

In our data set ‘y’ which is nothing but a claim is the dependent variable and remaining all are independent variables.

### **3.2.4** **Data Sampling Methods**

The data we have is highly unbalanced data so we used some sampling methods which are used to balance the target variable so we our model will be developed with good accuracy and precision. We used three Sampling methods.

### **3.2.4.1** **Stratified sampling**

Stratified sampling randomly selects data points from majority class so they will be equal to the data points in the minority class. So, after the sampling both the class will have same no of observations.

It can be performed using strata function from the library sampling.

### **3.2.4.2** **Simple random sampling**

Simple random sampling is a sampling technique where a set percentage of the data is selected randomly. It is generally done to reduce bias in the dataset which can occur if data is selected manually without randomizing the dataset.

We used this method to split the dataset into train dataset which contains 70% of the total data and test dataset with the remaining 30% of the data.

### **3.2.5 Models Used for Development**

We built our predictive models by using the following ten algorithms.

### **Model 01**

Logistic uses logit link function to convert the likelihood values to probabilities so we can get a good estimate on the probability of a particular observation to be positive class or negative class. The also gives us p-value of the variables which tells us about significance of each independent variable.

### **Model 02**

Random forest is an algorithm that consists of many decision trees. It was first developed by Leo Breiman and Adele Cutler. The idea behind it is to build several trees, to have the instance classified by each tree, and to give a "vote" at each class. The model uses a "bagging" approach and the random selection of features to build a collection of decision trees with controlled variance. The instance's class is to the class with the highest number of votes, the class that occurs the most within the leaf in which the instance is placed.

The error of the forest depends on:

* Trees correlation: the higher the correlation, the higher the forest error rate.
* The strength of each tree in the forest. A strong tree is a tree with low error. By

using trees that classify the instances with low error the error rate of the forest decreases.

### **Model 03**

Artificial neural networks can theoretically solve any problem. ANNs can identify hidden patterns between the variables and can find how different combinations of variables can affect the target variable. The error correction is done by gradient descent algorithm which can reduce the error rate as much as possible for the given data.

### **Model 04**

A decision tree is drawn upside down with its root at the top. In the image on the left, the bold text in black represents a condition/**internal node**, based on which the tree splits into branches/ **edges**. The end of the branch that doesn’t split anymore is the decision/**leaf**, in this case, whether the passenger died or survived, represented as red and green text respectively.

### **Model 05**

**Extremely Randomized Trees Classifier(Extra Trees Classifier)** is a type of ensemble learning technique which aggregates the results of multiple de-correlated decision trees collected in a “forest” to output it’s classification result. In concept, it is very similar to a Random Forest Classifier and only differs from it in the manner of construction of the decision trees in the forest.

### **Model 06**

K-nearest neighbors (KNN) is a type of supervised learning algorithm used for both regression and classification. KNN tries to predict the correct class for the test data by calculating the distance between the test data and all the training points. Then select the K number of points which is closet to the test data. The KNN algorithm calculates the probability of the test data belonging to the classes of ‘K’ training data and class holds the highest probability will be selected. In the case of regression, the value is the mean of the ‘K’ selected training points.

### **Model 07**

Gaussian Naive Bayes classifier In Gaussian Naive Bayes, continuous values associated with each feature are assumed to be distributed according to a Gaussian distribution. A Gaussian distribution is also called Normal distribution.

### **Model 08**

Support Vector Machine(SVM) is a supervised machine learning algorithm used for both classification and regression. Though we say regression problems as well its best suited for classification. The objective of SVM algorithm is to find a hyperplane in an N-dimensional space that distinctly classifies the data points. The dimension of the hyperplane depends upon the number of features. If the number of input features is two, then the hyperplane is just a line. If the number of input features is three, then the hyperplane becomes a 2-D plane. It becomes difficult to imagine when the number of features exceeds three.

### **Model 09**

Bagging classifier is meta- estimator which can be utilized for predictions in classification and regression problems by means of Bagging Classifier and Bagging Regressor, which are available in scikit learning library. This method involves the following steps: The original data set is used to create some random subsets which is nothing but bagging or Bootstrapping.

### **Model 10**

XGboost is the most widely used algorithm in machine learning, whether the problem is a classification or a regression problem. It is known for its good performance as compared to all other [machine learning algorithms](https://www.analyticssteps.com/blogs/top-10-machine-learning-algorithms). **It is the execution of gradient boosted decision trees that is designed for high speed and performance.**

## **3.3 AI / ML Models Analysis and Final Results**

We used our train dataset to build the above models and used our test data to check the accuracy and performance of our models.

We used confusion matrix to check accuracy, Precision, Recall and F1 score of our models and compare and select the best model for given auto dataset of size ~ 272252 policies.

### **3.3.1 Different Model codes**

* The Python code for models with stratified sampling technique as follows:
* The Python code for models with simple random sampling technique as follows:

### **3.3.2 Random Forest Python Code**

* The Python code for models with stratified sampling technique as follows:

# To build the 'Multinominal Decision Tree' model with random sampling

from sklearn.ensemble import RandomForestClassifier

ModelRF=RandomForestClassifier(n\_estimators=100,criterion='gini', max\_depth=None,min\_samples\_split=2,min\_samples\_leaf=1, min\_weight\_fraction\_leaf=0.0,max\_features='sqrt',max\_leaf\_nodes=None, min\_impurity\_decrease=0.0,bootstrap=True,oob\_score=False,n\_jobs=None, random\_state=None,verbose=0,warm\_start=False,class\_weight=None, ccp\_alpha=0.0, max\_samples=None)

# Train the model with train data

ModelRF.fit(x\_train,y\_train)

# Predict the model with test data set

y\_pred = ModelRF.predict(x\_test)

y\_pred\_prob = ModelRF.predict\_proba(x\_test)

# Confusion matrix in sklearn

from sklearn.metrics import confusion\_matrix

from sklearn.metrics import classification\_report

# actual values

actual = y\_test

# predicted values

predicted = y\_pred

# confusion matrix

matrix = confusion\_matrix(actual,predicted, labels=[1,0],sample\_weight=None, normalize=None)

print('Confusion matrix : \n', matrix)

# outcome values order in sklearn

tp, fn, fp, tn = confusion\_matrix(actual,predicted,labels=[1,0]).reshape(-1)

print('Outcome values : \n', tp, fn, fp, tn)

# classification report for precision, recall f1-score and accuracy

C\_Report = classification\_report(actual,predicted,labels=[1,0])

print('Classification report : \n', C\_Report)

# calculating the metrics

sensitivity = round(tp/(tp+fn), 3);

specificity = round(tn/(tn+fp), 3);

accuracy = round((tp+tn)/(tp+fp+tn+fn), 3);

balanced\_accuracy = round((sensitivity+specificity)/2, 3);

precision = round(tp/(tp+fp), 3);

f1Score = round((2\*tp/(2\*tp + fp + fn)), 3);

# Matthews Correlation Coefficient (MCC). Range of values of MCC lie between -1 to +1.

# A model with a score of +1 is a perfect model and -1 is a poor model

from math import sqrt

mx = (tp+fp) \* (tp+fn) \* (tn+fp) \* (tn+fn)

MCC = round(((tp \* tn) - (fp \* fn)) / sqrt(mx), 3)

print('Accuracy :', round(accuracy\*100, 2),'%')

print('Precision :', round(precision\*100, 2),'%')

print('Recall :', round(sensitivity\*100,2), '%')

print('F1 Score :', f1Score)

print('Specificity or True Negative Rate :', round(specificity\*100,2), '%' )

print('Balanced Accuracy :', round(balanced\_accuracy\*100, 2),'%')

print('MCC :', MCC)

# Area under ROC curve

from sklearn.metrics import roc\_curve, roc\_auc\_score

print('roc\_auc\_score:', round(roc\_auc\_score(y\_test, y\_pred), 3))

# ROC Curve

from sklearn.metrics import roc\_auc\_score

from sklearn.metrics import roc\_curve

logit\_roc\_auc = roc\_auc\_score(y\_test, y\_pred)

fpr, tpr, thresholds = roc\_curve(y\_test,ModelRF.predict\_proba(x\_test)[:,1])

plt.figure()

#--------------------------------------------------------------------

plt.plot(fpr, tpr, label= 'Classification Model' % logit\_roc\_auc)

plt.plot([0, 1], [0, 1],'r--')

plt.xlim([0.0, 1.0])

plt.ylim([0.0, 1.05])

plt.xlabel('False Positive Rate')

plt.ylabel('True Positive Rate')

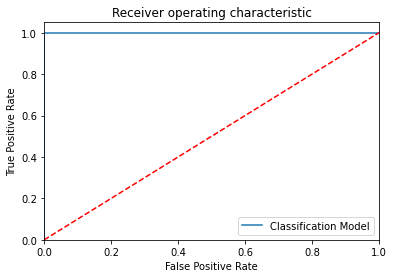
plt.title('Receiver operating characteristic')

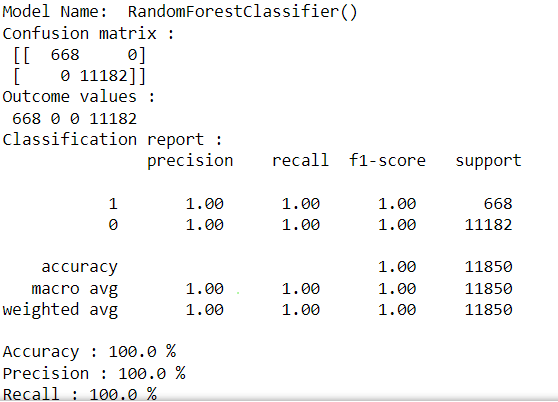
plt.legend(loc="lower right")

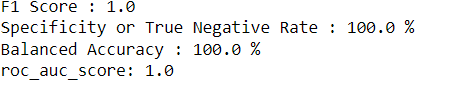
plt.savefig('Log\_ROC')

plt.show()

print('-----------------------------------------------------------------------------------------------------')







### **3.3.3 Extra Trees Python code**

* The Python code for models with stratified sampling technique as follows:

**from sklearn.ensemble import ExtraTreesClassifier**

**ModelET = ExtraTreesClassifier()**

models.fit(x\_train, y\_train)

# Prediction the model with test dataset

y\_pred = models.predict(x\_test)

y\_pred\_prob = models.predict\_proba(x\_test)

# Print the model name

print('Model Name: ', models)

# confusion matrix in sklearn

from sklearn.metrics import confusion\_matrix

from sklearn.metrics import classification\_report

# actual values

actual = y\_test

# predicted values

predicted = y\_pred

# confusion matrix

matrix = confusion\_matrix(actual,predicted, labels=[1,0],sample\_weight=None, normalize=None)

print('Confusion matrix : \n', matrix)

# outcome values order in sklearn

tp, fn, fp, tn = confusion\_matrix(actual,predicted,labels=[1,0]).reshape(-1)

print('Outcome values : \n', tp, fn, fp, tn)

# classification report for precision, recall f1-score and accuracy

C\_Report = classification\_report(actual,predicted,labels=[1,0])

print('Classification report : \n', C\_Report)

# calculating the metrics

sensitivity = round(tp/(tp+fn), 3);

specificity = round(tn/(tn+fp), 3);

accuracy = round((tp+tn)/(tp+fp+tn+fn), 3);

balanced\_accuracy = round((sensitivity+specificity)/2, 3);

precision = round(tp/(tp+fp), 3);

f1Score = round((2\*tp/(2\*tp + fp + fn)), 3);

# Matthews Correlation Coefficient (MCC). Range of values of MCC lie between -1 to +1.

# A model with a score of +1 is a perfect model and -1 is a poor model

#from math import sqrt

#mx = (tp+fp) \* (tp+fn) \* (tn+fp) \* (tn+fn)

#MCC = round(((tp \* tn) - (fp \* fn)) / sqrt(mx), 3)

print('Accuracy :', round(accuracy\*100, 2),'%')

print('Precision :', round(precision\*100, 2),'%')

print('Recall :', round(sensitivity\*100,2), '%')

print('F1 Score :', f1Score)

print('Specificity or True Negative Rate :', round(specificity\*100,2), '%' )

print('Balanced Accuracy :', round(balanced\_accuracy\*100, 2),'%')

#print('MCC :', MCC)

# Area under ROC curve

from sklearn.metrics import roc\_curve, roc\_auc\_score

print('roc\_auc\_score:', round(roc\_auc\_score(actual, y\_pred), 3))

# ROC Curve

from sklearn.metrics import roc\_auc\_score

from sklearn.metrics import roc\_curve

Model\_roc\_auc = roc\_auc\_score(actual, y\_pred)

fpr, tpr, thresholds = roc\_curve(actual, models.predict\_proba(x\_test)[:,1])

plt.figure()

#

plt.plot(fpr, tpr, label= 'Classification Mode Model' % Model\_roc\_auc)

plt.plot([0, 1], [0, 1],'r--')

plt.xlim([0.0, 1.0])

plt.ylim([0.0, 1.05])

plt.xlabel('False Positive Rate')

plt.ylabel('True Positive Rate')

plt.title('Receiver operating characteristic')

plt.legend(loc="lower right")

plt.savefig('Log\_ROC')

plt.show()

print('-----------------------------------------------------------------------------------------------------')

#

new\_row = {'Model Name' : models,

'True Positive': tp,

'False Negative': fn,

'False Positive': fp,

'True Negative': tn,

'Accuracy' : accuracy,

'Precision' : precision,

'Recall' : sensitivity,

'F1 Score' : f1Score,

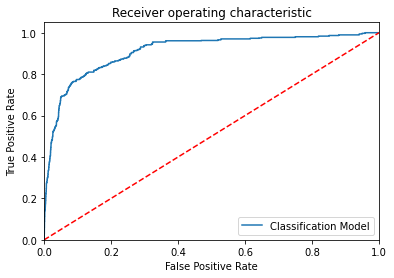
'Specificity' : specificity,

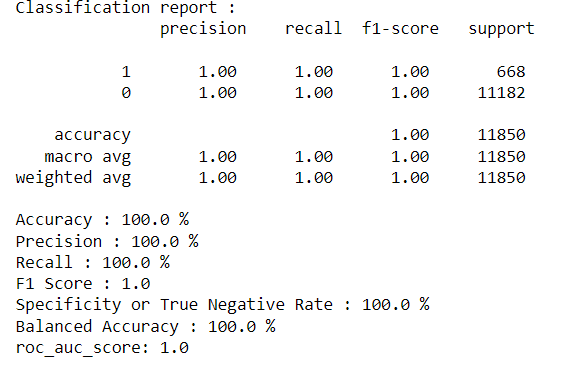
'MCC': 'MCC',

'ROC\_AUC\_Score':roc\_auc\_score(actual, y\_pred),

'Balanced Accuracy':balanced\_accuracy}

CSResults = CSResults.append(new\_row, ignore\_index=True)





**Stratified Sampling**: Random Forest model performance is good, by considering the confused matrix, highest accuracy (1.0) &good F1 score (1.0).This is because random forest uses bootstrap aggregation which can reduce bias and variance in the data and can leadsto good predictions withclaims dataset.

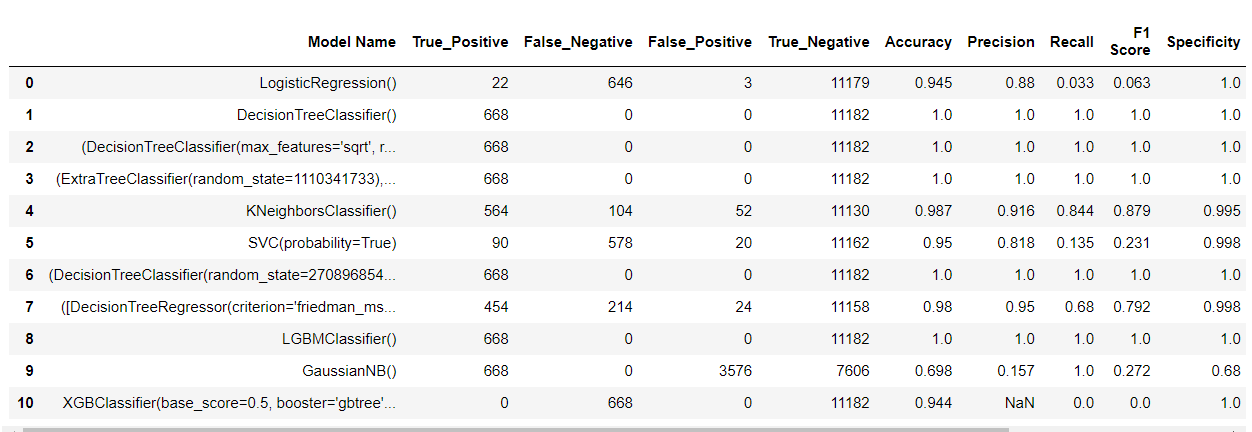
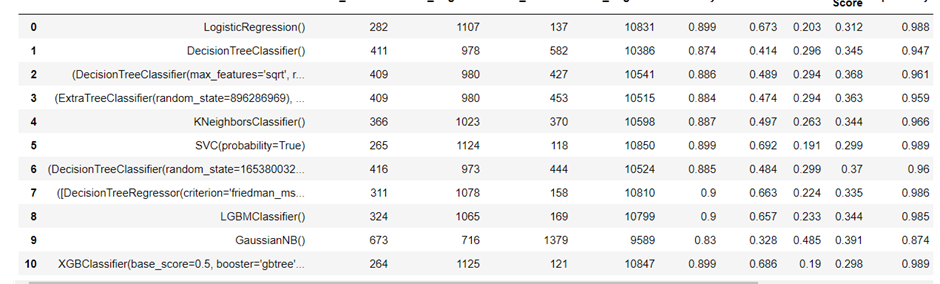
**Simple Random Sampling**: Artificial Neural Networks / Random Forest are out performed Logistic Regression model, by considering the confused matrix, highest accuracy (1.0) &good F1 score (1.0).This is because Artificial Neural Networks havehidden and complex patterns between different variables and can leads to good predictions with claims dataset.

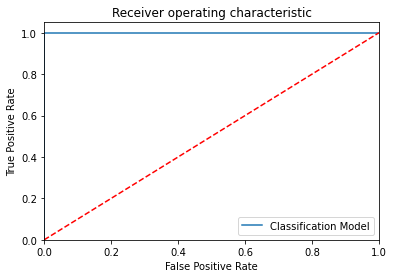
# **4.0 Conclusions and Future work**

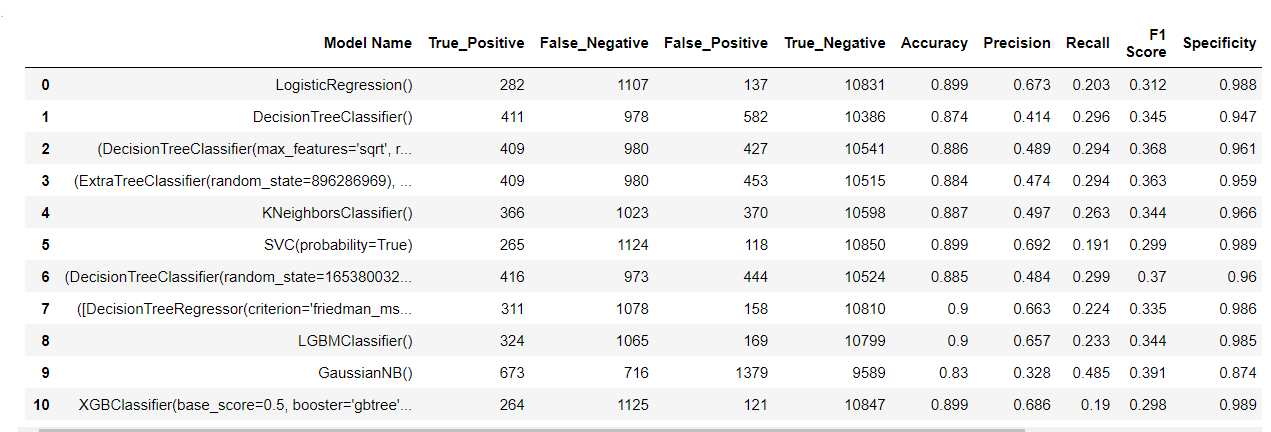
The model results in the following order by considering the model accuracy, F1 score and RoC AUC score.

1. **Decision Tree**with Stratified and Random Sampling
2. **RandomForestClassifier** with Simple Random Sampling
3. **ExtraTreeClassifier** with Simple Random Sampling

We recommend model - **Decision Tree** with Stratified and Random Sampling technique as a best fit for the give n BI claims dataset. We considered **Decision Tree**because it uses bootstrap aggregation which can reduce bias and variance in the data and can leads to good predictions with claims dataset.







### **5.0 Reference**

Kaggle link : <https://www.kaggle.com>

Git hub link : <https://github.com>