REPORT ON MUSIC GENRE

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

**BACHELOR OF TECHNOLOGY**

**in**

**COMPUTER SCIENCE AND ENGINEERING**

**By**

**VEERNALA DHANUNJAYA RAO**

**Regd. No.: 21B95A0526**

### Under Supervision of Mr. GundalaNagarajuHenotic Technology Pvt Ltd, Hyderabad (Duration: 5th July, 2022 to 4th September, 2022)



**DEPARTMENT OF COMPUTER SCIENCE AND ENGINEERING SAGI RAMA KRISHNAM RAJU ENGINEERING COLLEGE**

(An Autonomous Institution)

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

CHINNA AMIRAM, BHIMAVARAM, ANDHRA PRADESH

SAGI RAMA KRISHNAM RAJU ENGINEERING COLLEGE

(Autonomous)

ChinnaAmiram, Bhimavaram

DEPARTMENT OF COMPUTER SCIENCE AND ENGINEERING



CERTIFICATE

This is to certify that the “**Summer Internship Report”** submitted by **(VEERNALA DHANUNJAYARAO), (21B95A0526)** is work done by him/her and submitted during 2021 - 2022 academic year, in partial fulfillment of the requirements for the award of the Summer Internship Program for **Bachelor of Technology in (COMPUTER SCIENCE AND ENGINEERING),** at **(HENOTIC TECHNOLOGYPvt**

**Ltd)** from *05.07.2022 to04.09.2022 for AIML*

**DepartmentInternshipCoordinato Dean -T &PCell Head of the Department**

## Table of Contents

[1.0 Introduction 6](#_bookmark0)

* 1. [What are the different types ofMachine Learning? 7](#_bookmark1)
  2. Benefits of Using Machine Learningin PHD 9
  3. About Industry (examplePHD dataset) 10
     1. AI / ML Rolein PHD 11

2.0 PHDDataset 12

* 1. Main Drivers for PHD ofaPerson 13
  2. Internship Project -DataLink 13

[3.0 AI / ML Modellingand Results 14](#_bookmark2)

* 1. [Your Problemof Statement 14](#_bookmark3)
  2. [Data Science Project LifeCycle 14](#_bookmark4)
     1. [DataExploratoryAnalysis 15](#_bookmark5)
     2. [DataPre-processing 15](#_bookmark6)
        1. [Check the Duplicate and lowvariationdata 16](#_bookmark7)
        2. [Identify and address themissingvariables 16](#_bookmark8)
        3. [Handlingof Outliers 17](#_bookmark9)
        4. Categorical data andEncoding Techniques 18
        5. [FeatureScaling 18](#_bookmark10)
     3. [Selection of Dependent andIndependent variables 19](#_bookmark11)
     4. [DataSamplingMethods 19](#_bookmark12)
        1. [Stratifiedsampling 19](#_bookmark13)
        2. [Simplerandom sampling 19](#_bookmark14)
     5. [Models Usedfor Development 19](#_bookmark15)
        1. [Model01(Logistic Regression) 19](#_bookmark16)
        2. [Model 02(DecisionTreeClassifier) 19](#_bookmark17)
        3. Model 03(RandomForest Classifier) 20
        4. [Model 04(ExtraTreeClassifier) 20](#_bookmark18)
        5. [Model05(KNNClassifier) 20](#_bookmark19)
        6. [Model06(GaussianNB) 21](#_bookmark20)
        7. Model07(SVC) 21
        8. [Model08(XGB Classifier) 21](#_bookmark21)
        9. [Model09(LGBMClassifier) 22](#_bookmark22)
        10. [Model10(BaggingClassifier) 22](#_bookmark23)
        11. [Model 11(GradientBoostingClassifier) 22](#_bookmark24)
  3. [AI / ML Models Analysis andFinal Results 23](#_bookmark25)
     1. Modelcodes 23

[4.0 Conclusions andFuture work 27](#_bookmark26)

[5.0 References 28](#_bookmark27)

* 1. [Appendices 29](#_bookmark27)
  2. Pythonresult… 30
  3. [Listof charts… 30](#_TOC_250000)
     1. Chart01:Loanstatus 30
     2. Chart 02:Loam amount W.R.TLoanstatus 31
     3. Chart 03: Loan Amount Passed ForEach Purpose 31

**Abstract**

Automatic music genre classification is a fundamental component of music information retrieval systems and has been gaining importance and enjoying a growing amount of attention with the emergence of digital music on the Internet. Although considerable research has been conducted in automatic music genre classification, little has been done on hierarchical classification with taxonomies.

The underlying hierarchical taxonomy identifies the relationships of dependence between different genres and provides valuable sources of information for genre classification. This paper investigates the use of taxonomy for music genre classification. Our empirical experiments on two datasets show that using taxonomy improves the classification performance. We also propose an approach for automatically generating genre taxonomies based on the confusion matrix via linear discriminant projection. Our work also provides some insights for future research. The given dataset falls into classification and here we are predicting the types of Music Genre. The given dataset is a multinominal dataset, and it has multiple categories like Jazz, Electronic etc.

# 1.0Introduction

Wikipedia states that “music genre is a conventional category that identifies pieces of music as belonging to a shared tradition or set of conventions.” The term “genre” is a subject to interpretation, and it is often the case that genres may be very fuzzy in their definition. Further, genres do not always have sound music theoretic foundations, e.g. - Indian genres are geographically defined, Baroque is classical music genre based on time period.

Despite the lack of a standard criteria for defining genres, the classification of music based on genres is one of the broadest and most widely used. Genre usually assumes high weight in music recommender systems. Genre classification, till now, had been done manually by appending it to metadata of audio files or including it in album info.

This project however aims at content-based classification, focusing on information within the audio rather than extraneously appended information. The traditional machine learning approach for classification is used - find suitable features of data, train classifier on feature data, make predictions.

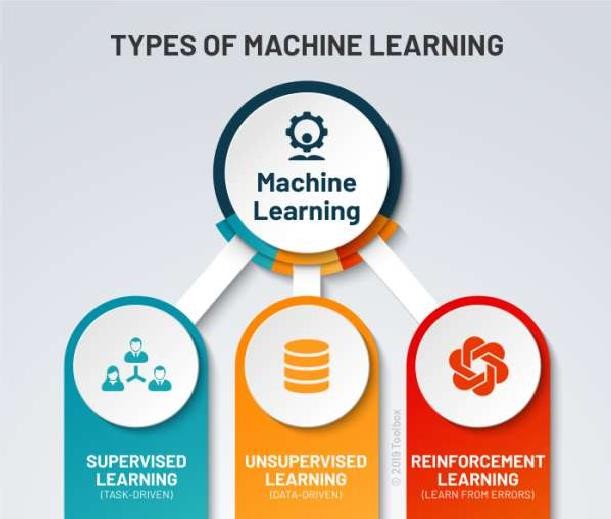
The novel thing that we have tried is the use of ensemble classifier on fundamentally different classifiers to achieve our end goal.

**music**, art concerned with combining vocal or instrumental [sounds](https://www.britannica.com/science/sound-physics) for beauty of form or emotional expression, usually according to cultural standards of [rhythm](https://www.britannica.com/art/rhythm-music), [melody](https://www.britannica.com/art/melody), and, in most Western music, [harmony](https://www.britannica.com/art/harmony-music). Both the simple [folk song](https://www.britannica.com/art/folk-music) and the complex [electronic composition](https://www.britannica.com/art/electronic-music) belong to the same activity, music. Both are humanly engineered; both are [conceptual](https://www.merriam-webster.com/dictionary/conceptual) and auditory, and these factors have been present in music of all styles and in all periods of history, throughout the world.

## What are the different types of MachineLearning?

### How Machine Learning Works

Machine learning uses two types of techniques: **Supervised learning**, which trains a model on known input and output data so that it can predict future outputs, and **Unsupervised learning**, which finds hidden patterns or intrinsic structures in input data.



## Supervised Learning

Supervised machine learning builds a model that makes predictions based on evidence in the presence of uncertainty. A supervised learning algorithm takes a known set of input data and

known responses to the data (output) and trains a model to generate reasonable predictions for the response to new data. Use supervised learning if you have known data for the output you are trying topredict.

**Supervised learning**uses**Regression** and **Classification** techniques models.

to develop predictive

**Regression techniques**predict continuous responses - for example, changes in temperature or

fluctuations in power demand. Typical applications include electricity algorithmic trading.

load forecasting and

Use regression techniques if you are working with a data range or if the nature of your response is a real number, such as temperature or the time until failure for a piece ofequipment.

Common regression algorithms include linear model, nonlinear model, stepwise regression, Gradient Descent Regression, Support Vector Regression, Ridge and Lasso Regressions

**Classification techniques**predict discrete responses - for example, whether an email is genuine or spam, or whether a tumour is cancerous or benign. Classification models classify input data into categories. Typical applications include medical imaging, speech recognition, and credit scoring.

Use classification if your data can be tagged, categorized, or separated into specific groups or classes. For example, applications for hand-writing recognition use classification to recognize letters and numbers. In image processing and computer vision, unsupervised pattern recognition techniques are used for object detection and imagesegmentation.

Common algorithms for performing classification include support vector machine (SVM), boosted and bagged decision trees, k-nearest neighbour, Naïve Bayes, discriminant analysis, logistic regression, and neural networks.

**Using Supervised Learning to Predict Heart Attacks:** Suppose clinicians want to predict whether someone will have a heart attack within a year. They have data on previous patients, including age, weight, height, and blood pressure. They know whether the previous patients had heart attacks within a year. So, the problem is combining the existing data into a model that can predict whether a new person will have a heart attack within ayear.

## Unsupervised Learning

Unsupervised learning finds hidden patterns or intrinsic structures in data. It is used to draw inferences from datasets consisting of input data without labelledresponses.

**Clustering** is the most common unsupervised learning technique. It is used for exploratory data analysis to find hidden patterns or groupings in data. Applications for cluster analysis include gene sequence analysis and market research.

For example, if a cell phone company wants optimize the locations where they build cell phone towers, they can use machine learning to estimate the number of clusters of people relying on their towers. A phone can only talk to one tower at a time, so the team uses clustering algorithms to design the best placement of cell towers to optimize signal reception for groups, or clusters, of theircustomers.Common algorithms for performing clustering include k-means and k-medoids, Apriori algorithms, hierarchical clustering, Gaussian mixture models and hidden Markov models

## Benefits of Using Machine Learning in Music Genre

Music Genre Classification using Machine Learning is what we will be discussing in this segment. Before understanding music genre classification, we shall first look into the definition of machine learning. A technology of Artificial Intelligence, Machine Learning is a concept wherein computers or machines learn from information (data) fed to them. Based on data that is entered into the machines, Machine Learning helps computers to build interpretive patterns and build analytical models automatically.

Music Genre Classification using Machine Learning is acomparatively newer concept that has emerged on the surface in recent times. While music genres have been known to the world for decades, machines have been able to work along the lines of music genre classification in the contemporary world where every other person is listening to music. As computers and smart phones have become the new music equipment, Machine Learning has facilitated the classification of music genres using several techniques.

As computers and smart phones have become the new music equipment, Machine Learning has facilitated the classification of music genres using several techniques.

## About Industry music

**Industrial Music** is a difficult genre to define, you have many hardcore critics who become very outspoken concerning their opinions of the genre , many believe that if a band isn’t compromised of 100% electronic components , that is 100% synthesized, and makes use of real musical instruments it isn’t true industrial and is industrial[metal fusion.](https://www.musicgenreslist.com/metal-music-genre/)

## In the beginning industrial music was mainly composed of highly distorted music and keyboards as well as very controversial lyrics and remained mainly underground, consisting of bands like Throbbing Gristle and Cabaret Voltaire, bands of the like are considered to be of the “first wave” of industrial music between the 70’s to 80’s. The “Second Wave” is when Industrial Music started to come into it’sown and became more well known, this mainly happened during the 80’s and 90’s, this is also where industrial becomes hard to pin down because people started to break it off into many different sub genre’s which I will list later on. During the second wave there was a flood of great bands that are still around today and some that have unfortunately fallen to the wayside, Skinny Puppy, Frontline Assembly, Front 242, Leather Strip, (now widely considered Dark wave), and many others.

### Industrial Sub Genres

* Aggrotech – Ambient Industrial – Cyber grind – Dark Ambient – Dark[Electro](https://www.musicgenreslist.com/dance-music-dancegenre/)– Dark Wave – Death Industrial
* Electronic Body Music Electro-Industrial – Electro-Pop – Future pop – Industrial

DnB – Industrial Metal – Industrial Rock – Industrial

* Techno – Japanoise – Martial Industrial Neofolk
* Noise – Power Noise – Synth- Pop, and others

## 1.3.1 AI / ML Role in Prediction Of Music Industry

The days of arguing whether artificial intelligence (AI) would influence the music industry are long gone. Artificial intelligence is already being employed in a variety of applications. Now it’s time to think about how it will affect the way we make and consume music. AI automates services, identifies patterns and insights in massive data sets, and helps generate efficiency in the music business, just like it does in other sectors. Companies in the music industry must acknowledge and prepare for the impact of artificial intelligence on their operations; those that do not will be left behind.

### AI In Music Industry: Use Cases

The following are some of the most common artificial intelligence applications in music apps:

### ● Music Composition

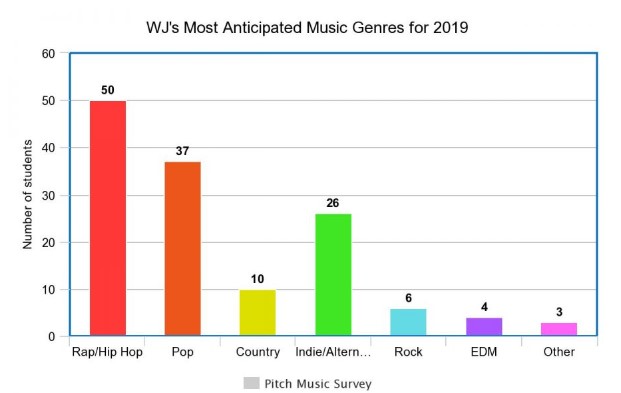
Consumers may utilize artificial intelligence to access various music composition and remix tools, allowing them to create their rhythms and songs. The popularity of artificial intelligence-generated music is growing, with numerous well-known performers releasing AI-generated albums. Other[**AItechnologies**](https://onpassive.com/blog/next-gen-onpassive-artificial-intelligence/)are available to help both new and seasoned artists create their unique music. Users may perfect their music with the LANDR audio mastering program, which uses artificial intelligence. By matching production components to album-quality music, you may create outputfilters.

**●Music Streaming**

This method evaluates the listening habits and music preferences of millions of people who utilize broadcasting services. It’s particularly good at spotting patterns in tens of thousands of pieces of music. Spotify’s Discover Weekly, an AI-powered feature, curates a weekly playlist for each user based on their artist preferences and listening habits. Spotify recommends music based on how you interact with different artists and songs, how often you play specific tracks on loop and other actions. The significant part is that millions of people are examined simultaneously, allowing you to see who else is listening to similar playlists.

### ● Mastering Music

Audio mastering has traditionally been done in a studio with specific acoustics, allowing humans to detect sound balance and spectral range issues. It’s quality control, which aids in detecting the problems and optimizing music playback on any device. AI may be integrated into the mastering process to minimize the manual work and human error involved in mastering an entire album, allowing for smoother track flow. Artificial intelligence for mastering has increased in favour among newer, less experienced musicians. One of the key reasons for its popularity is that AI mastering is high-quality without taking time and effort.



**2.1. Internship Project - DataLink**

The internship project data has taken from Kaggle and the link is

­­­­­­­­­­­

**(18 Columns& 50000 Rows)**

* 1. **AI / ML Modelling andResults**

## Your Problem ofStatement

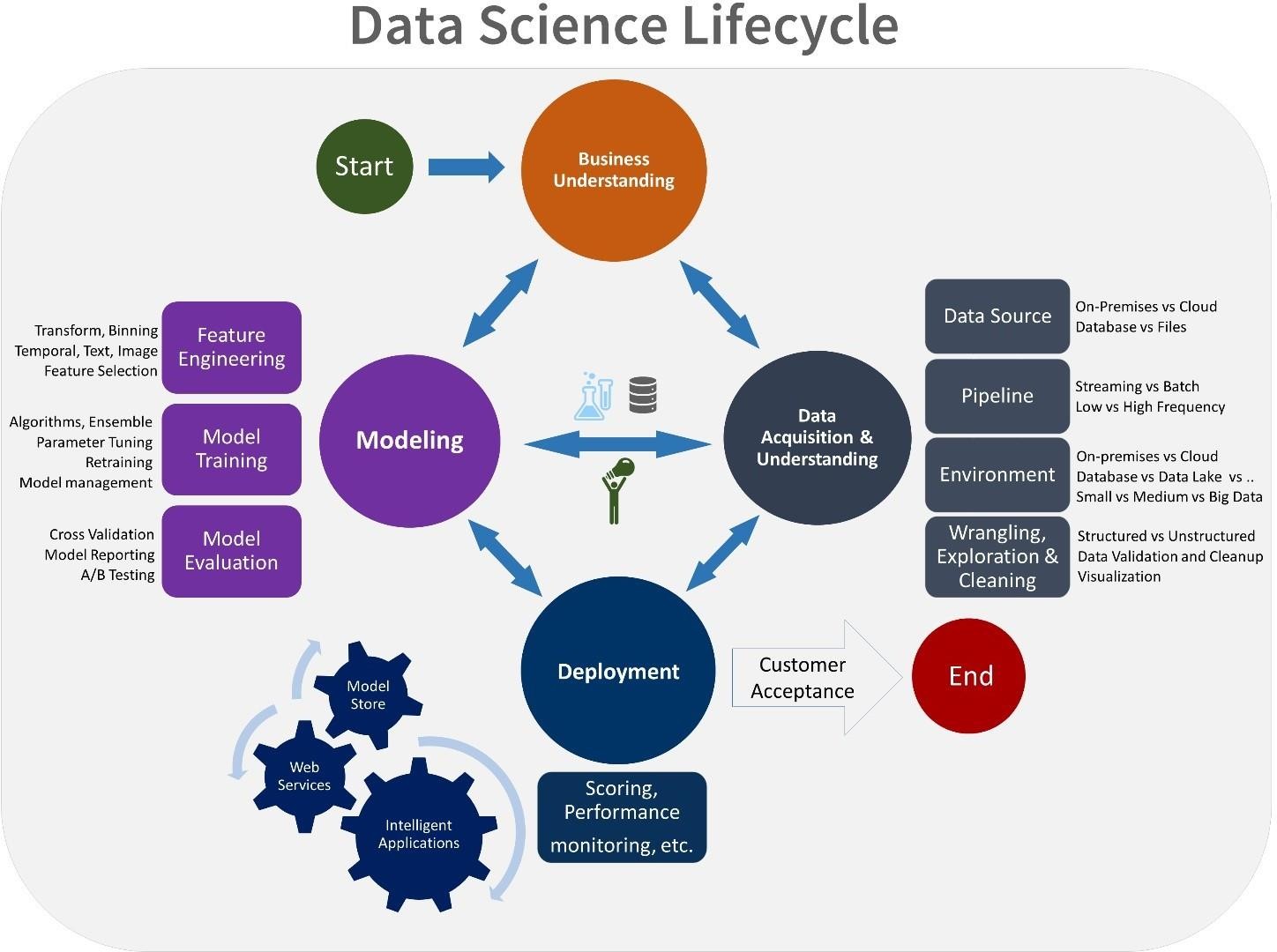
A Music Genre can be defined as a conventional category that groups together pieces of music sharing a set of conventions or traditions. Music is divided into different genres using a range of criteria. However, due to music’s artistic nature, classifications are almost always controversial and subjective, with many genres overlapping. This dataset outlines some of the top music genres in terms of their widespread popularity. This dataset has a list of music artists by genre and here we are supposed to predict the type of Music Genre based on the track name, popularity, acousticness, danceability, duration\_ms, energy, instrumentalness, key etc.

## Data Science Project LifeCycle

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.

In simple terms, a data science life cycle is nothing but a repetitive set of steps that you need to take to complete and deliver a project/product to your client. Although the data science projects and the teams involved in deploying and developing the model will be different, every data science life cycle will be slightly different in every other company. However, most of the data science projects happen to follow a somewhat similar process.

In order to start and complete a data science-based project, we need to understand the various roles and responsibilities of the people involved in building, developing the project.Let us take a look at those employees who are involved in a typical data science project:



### Data ExploratoryAnalysis

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.Exploratory Data Analysis refers to the critical process of performing initial investigations on data so as to discover patterns, to spot anomalies, to test hypothesis and to check assumptions with the help of summary statistics and graphicalrepresentations.

### DataPre-processing

We removed variables which does not affect our target variable(income\_>50k)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.

Data preprocessing transforms the data into a format that is more easily and effectively processed in data mining, machine learning and other data science tasks. The techniques are generally used at the earliest stages of the machine learning and AI development pipeline to ensure accurate results.

### Check the Duplicate and low variationdata

Duplicate Values: When two features have the same set of values

Duplicate Index: When the value of two features are different, but they occur at the same index

Steps to delete duplicates:

1. Usethegetduplicatefeaturesfunctionsto getalltheconstantfeatures.
2. Store all the duplicate features as a list for removing from the dataset. 3.Drop all such features from thedataset.

## Identify and address the missingvariables

How to Identify Missing Values?

We can check for null values in a derived dataset. But, sometimes, it might not be this simple to identify missing values. One needs to use the domain knowledge and look at the data description to understand the variables. There are variables that have a minimum value of zero. On some columns, a value of zero does not make sense and indicates an invalid or missingvalue.

Quick Classification of Missing Data

There are three types of missing data as below:

**Missing CompletelyAt Random (MCAR):** It is the highest level of randomness. This means that the missing values in any features are not dependent on any other feature’s values. This is the desirable scenario in case of missingdata.

**Missing At Random (MAR):** This means that the missing values in any feature are dependent on the values of other features.

**Missing Not At Random (MNAR):** Missing not at random data is a more serious issue and,in this case, it might be wise to check the data gathering process further and try to understand why the information is missing.

**What to Do with the Missing Values?**

We identified the missing values in a derived dataset, next we should decide the further Course of action.

Ignore the missing values

Drop the missing values Case Deletion Imputation

**None**: None is a Python singleton object that is often used for missing data in Python code.

**NaN**: Nan (an acronym for Not a Number), is a special floating-point value recognized by all systems that use the standard IEEE floating-point representation

## Handling ofOutliers

An outlier is a data point in a data set that is distant from all other observations. A data point that lies outside the overall distribution of dataset (95% of customer behaviour /claims

/ spending nature of customer)

Detecting outliers or anomalies is one of the core problems in data mining. The emerging expansion and continued growth of data and the spread of IoT devices, make us rethink the way we approach anomalies and the use cases that can be built by looking at those anomalies.

We now have smart watches and wristbands that can detect our heartbeats every few minutes. Detecting anomalies in the heartbeat data can help in predicting heart diseases. Anomalies in traffic patterns can help in predicting accidents. It can also be used to identify bottlenecks in network infrastructure and traffic between servers. Hence, the use cases and solution built on top of detecting anomalies arelimitless.

Another reason why we need to detect anomalies is that when preparing datasets for machine learning models, it is really important to detect all the outliers and either get rid of them or analyse them to know why you had them there in the firstplace.

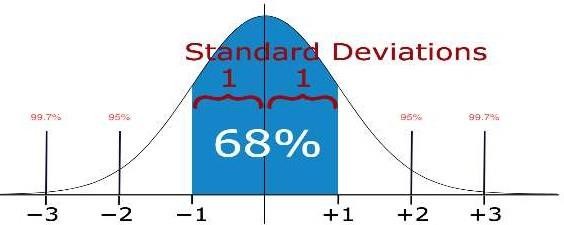
**Method 1 - Standard Deviation:**

In statistics, if a data distribution is approximately normal then about 68% of the data values lie within one standard deviation of the mean and about 95% are within two standard deviations, and about 99.7% lie within three standard deviations.Therefore, if you have any data point that is more than 3 times the standard deviation, then those points are very likely to be anomalous or outliers.

**Z score**

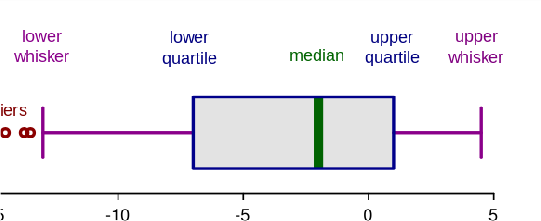
Z score indicates how many standard deviation away a data point

Calculate the **Z score = (X - m)/Sigma**, where m = mean, Sigma = standard deviation



**Method 2 - Boxplots:**

Box plots are a graphical depiction of numerical data through their quantiles. It is a very simple but effective way to visualize outliers. Think about the lower and upper whiskers as the boundaries of the data distribution. Any data points that show above or below the whiskers, can be considered outliers or anomalous.



### 3.2.2.3. Categorical data and EncodingTechniques

A categorical variable is one that has two or more categories (values). There are two types of categorical variable, **nominal** and **ordinal**. A nominal variable has no intrinsic ordering toits categories. For example, gender is a categorical variable having two categories (male and female) with no intrinsic ordering to the categories. An ordinal variable has a clear ordering.

Many ML algorithms are unable to operate on categorical or label data directly. However, Decision tree can directly learn from such data. Hence, they require all input variables and output variables to be numeric. This means that categorical data must be converted to a numerical form. Few types of categorical variable encoding are:

1. **One hot encoding**: Encoding each categorical variable with different Boolean variables (also called dummy variables) which take values 0 or 1, indicating if a category is present in an observation.
2. **Integer Encoding / Label Encoding**: Replace the categories by a number from 1 to n (or 0to n-1, depending the implementation), where n is the number of distinct categories of thevariable.
3. **Count or frequency encoding**: Replace the categories by the count of the observations that show that category in the dataset. Similarly, we can replace the category by the frequency -or percentage- of observations in the dataset. That is, if 10 of our 100 observations show the colour blue, we would replace blue by 10 if doing count encoding, or by 0.1 if replacing by the frequency.
4. **Ordered Integer Encoding**: Categories are replaced by integer 1 to k, where k is thedistinct categories in variable, but this numbering is decided by mean of target of eachcategory

### 3.2.2.3. FeatureScaling

Feature Scaling is done on the dataset to bring all the different types of data to a ***Single Format***. Done on **Independent Variable**.Types of Feature Scaling: Min Max Scaler, Standard Scaler.

Min Max Scalar:

It scales and transforms the data in-between 0 and 1.

ANN performs well when do scale the data using Min-max-Scalar.

**Standard Scalar:**

It scales and transform the data with respect to *Mean = 0* and *Standard Deviation =****1.***

### Selection of Dependent andIndependent variables

The dependent or target variable here is income\_>50k which tells us a particular personhas income greater than 50k or not the target variable is selected based on our requirement and what we are trying topredict.

The independent variables are selected after doing exploratory data analysis and we used Boruta to select which variables are most affecting our target variable.

### Data SamplingMethods

The data we have is a balanced data. In caseof unbalanced data we usedsome 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

### Stratifiedsampling

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.

### Simple randomsampling

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 datais 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.

### Models Used forDevelopment

We built our predictive models by using the following ten algorithms

* + - 1. **Model 01(LogisticRegression)**

Logistic uses logit link function to convert the likelihood values to probabilities so we cangeta 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 independentvariable.

* + - 1. **Model 02(Decision TreeClassifier)**

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 a graphical representation for getting all the possible solutions to a problem/decision based on given conditions.

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.

Decision Trees (DTs) are a non-parametric supervised learning method used for classification and regression. The goal is to create a model that predicts the value of a target variable by learning simple decision rules inferred from the data features. A tree can be seen as a piecewise constantapproximation.

* + - 1. **Model 03(Random ForestClassifier)**

Random forest is an algorithm that consists of many decision trees. It was first developed by Leo Bierman 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 isplaced.

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.

* + - 1. **Model 04(Extra TreeClassifier)**

**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 its 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.

* + - 1. **Model 05(KNNClassifier)**

K-Nearest Neighbour is one of the simplest Machine Learning algorithms based on Supervised Learning technique.

K-NN algorithm assumes the similarity between the new case/data and available cases and put the new case into the category that is most similar to the available categories.

K-NN algorithm stores all the available data and classifies a new data point based on the similarity. This means when new data appears then it can be easily classified into a well suite category by using K- NN algorithm. K-NN algorithm can be used for Regression as well as for Classification but mostly it is used for the Classification problems. K-NN is a non-parametric algorithm, which means it does not make any assumption on underlying data. It is also called a lazy learner algorithm because it does not learn from the training set immediately instead it stores the dataset and at the time of classification, it performs an action on the dataset. KNN algorithm at the training phase just stores the dataset and whenit gets new data, then it classifies that data into a category that is much similar to the newdata.

### Model06(GaussianNB)

Naïve Bayes is a probabilistic machine learning algorithm used for many classification functions and is based on the Bayes theorem. Gaussian Naïve Bayes is the extension of naïve Bayes. While other functions are used to estimate data distribution, Gaussian or normal distribution is the simplest to implement as you will need to calculate the mean and standard deviation for the training data.

* + - 1. **Model07(SVC)**

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 suitedfor 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 exceedsthree.

* + - 1. **Model 08(XGBClassifier)**

XGB is an implementation of Gradient Boosted decision trees. This library was written in C++. It is a type of Software library that was designed basically to improve speed and model performance. It has recently been dominating in applied machine learning. XGB models majorly dominate in many Kaggle Competitions. In this algorithm, decision trees are created in sequential form. Weights play an important role in XGB. Weights are assigned to all the independent variables which are then fed into the decision tree which predicts results. The weight of variables predicted wrong by the tree is increased and the variables are then fed to the second decision tree. These individual classifiers/predictors then ensemble to give a strong and more precise model. It can work on regression, classification, ranking, and user-defined predictionproblems.

* + - 1. **Model 09(LGBMClassifier)**

LGBM is a gradient boosting framework based on decision trees to increases the efficiency of the model and reduces memory usage.

It uses two novel techniques: Gradient-based One Side Sampling and Exclusive Feature Bundling (EFB) which fulfills the limitations of histogram-based algorithm that is primarily used in all GBDT (Gradient Boosting Decision Tree) frameworks. The twotechniques of GOSS and EFB described below form the characteristics of LGBM Algorithm. They comprise together to make the model work efficiently and provide it a cutting edge over other GBDT frameworks Gradient-based One Side Sampling Technique for LGBM: Different data instances have varied roles in the computation of information gain. The instances with larger gradients(i.e., under-trained instances) will contribute more to the

information gain. GOSS keeps those instances with large gradients (e.g., larger than a predefined threshold, or among the top percentiles), and only randomly drop those instances with small gradients to retain the accuracy of information gain estimation. This treatment can lead to a more accurate gain estimation than uniformly randomsampling, withthesame target sampling rate, especially when the value of information gain has a largerange.

### Model 10(Bagging Classifier)

A Bagging classifier is an ensemble meta-estimator that fits base classifiers each on random subsets of the original dataset and then aggregate their individual predictions (either by voting or by averaging) to form a final prediction. Such a meta-estimator can typically be used as a way to reduce the variance of a black-box estimator (e.g., a decision tree), by introducing randomization into its construction procedure and then making an ensemble out of it. Each base classifier is trained in parallel with a training set which is generated by randomly drawing, with replacement, N examples(or data) from the original training dataset – where N is the size of the original training set. Training set for each of thebase classifiers is independent of each other. Many of the original data may be repeated inthe resulting training set while others may be leftout.

Bagging reduces overfitting (variance) by averaging or voting, however, this leads to an increase in bias, which is compensated by the reduction in variance though.

### Model 11(Gradient BoostingClassifier)

**Gradient boosting classifier** is a set of machine learning algorithms that include several weaker models to combine them into a strong big one with highly predictive output. Models of a kind are popular due to their ability to classify datasets effectively.

Using **gradient boost for classification** we discover the initial prediction for every patient in the **log (odds)**.

## AI / ML ModelsAnalysis 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 dataset of predicting heartdisease.

### 3.1.1 Modelcode

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

from sklearn.linear\_model import LogisticRegression from sklearn.tree import DecisionTreeClassifier

from sklearn. ensemble import RandomForestClassifierfromsklearn. ensemble import ExtraTreesClassifier from sklearn. neighbors import KNeighborsClassifier from sklearn. naive\_bayes import GaussianNB

from sklearn.svm import SVC

from sklearn. ensemble import BaggingClassifier

from sklearn. ensemble import GradientBoostingClassifier import lightgbm as lgb

#Classification algorithms ModelLR

= LogisticRegression ()

ModelDC = DecisionTreeClassifier () ModelRF = RandomForestClassifier () ModelET = ExtraTreesClassifier () ModelKNN = KNeighborsClassifier () ModelSVM = SVC ()

modelGNB = GaussianNB () modelBAG = BaggingClassifier () ModelLGB = lgb. LGBMClassifier () ModelGNB = GaussianNB ()

# Evaluation matrix for all the algorithms

MM=[ModelLR,ModelDC,ModelRF,ModelET, ModelKNN,ModelSVM,modelGNB,modelBAG,ModelLGB,ModelGNB]

#for models in MM:

# Fit the model models.fit(x\_train, y\_train)#

Prediction

y\_pred = models. predict(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 value: \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 modelfrom 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\_scoreprint('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=round(tp/(tp+fn),3);

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

plt. plot([fpr, tpr], label= 'Classification 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**: Gradient Boosting Classifier model performance is good, by consideringthe confused matrix, highest accuracy (0.858) &good F1 score (0.663).

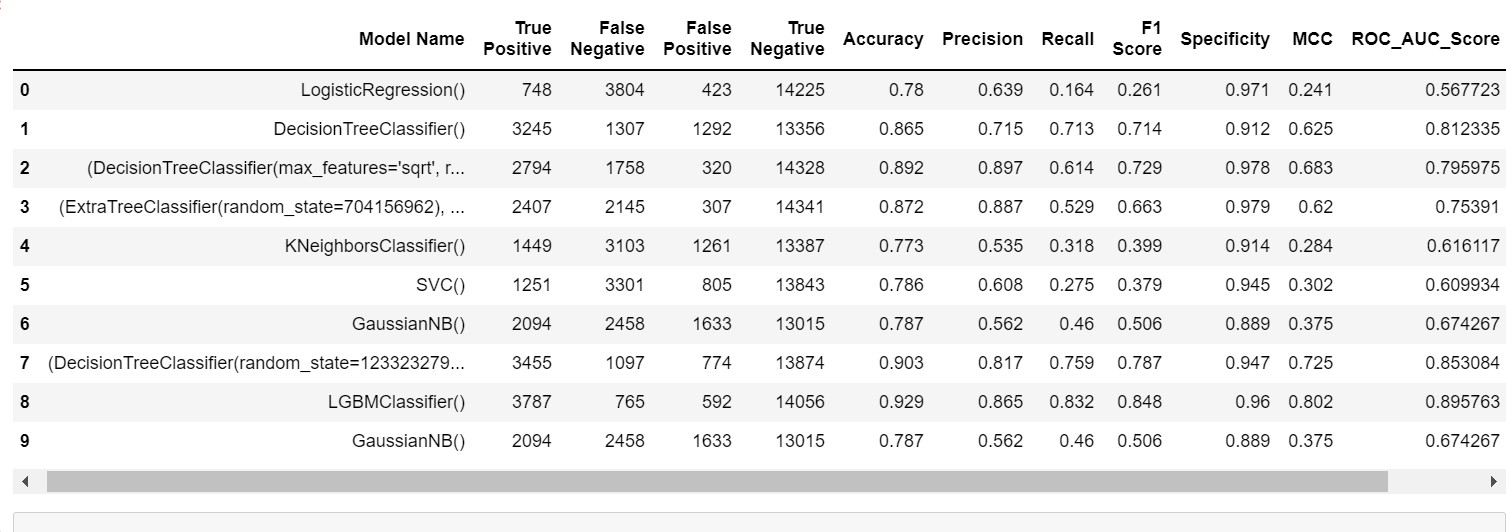
# 

# 4.0 Conclusions and Futurework

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

### LGBMClassifier

* + 1. **Decision treeclassifier**
    2. **Extra treeclassifier**



# References

The dataset has taken from Kaggle and the links

* <https://seaborn.pydata.org/generated/seaborn.kdeplot.html>
* <https://www.geeksforgeeks.org/>
* <https://www.w3schools.com/>
* <https://www.kaggle.com/>
* <https://matplotlib.org/>
* <https://www.python.org/>

**6.0**

**Appendices**

**6.1**

**. Python code Results**

**6.2**

**. List Of**

**Charts**

**6.2.1**

**Chart**

**01:**

**Pie**

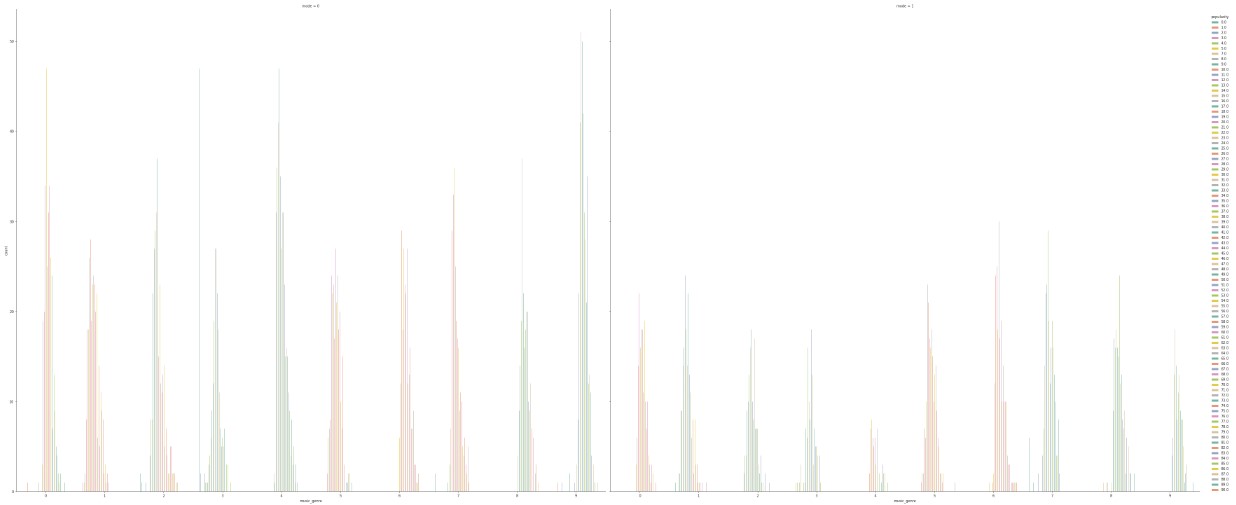
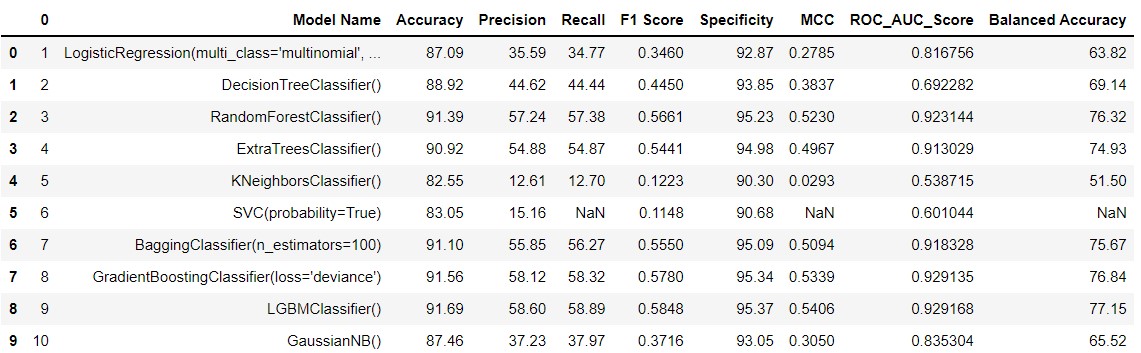
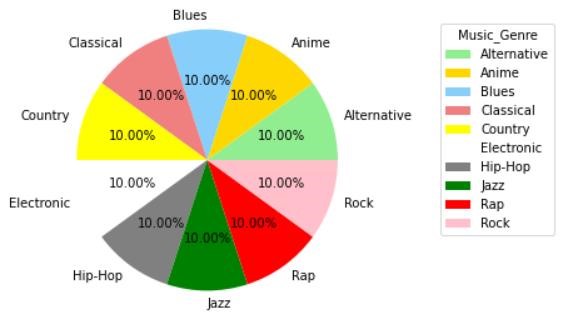
**Chart representing**

**Music Genre Fields**

**6.2.2.**

**Chart 02:**

**Count Plot for Music\_Genre and Mode**



**6.2.**

**3**

**. Chart**

**3:**

**Kdeplot**

**for**

**Acousticness, Instrumentalness and**

**Speechiness**

**6.2.4.**

**Chart**

**4:**

**Heat Map**

**for Music\_Genre**

**D**

**ataset**

