# Machine Learning Classification Models

## Logistic regression

Logistic regression (LR) is a supervised learning algorithm that can be used to solve classification problems. LR is one of the most popular methods that have been used in classification tasks (Ajoodha, et al., 2015) which used logistic function to predict the categorical dependent variable. Where x represents the features, g(z) is the activation function and h(x) is the hypothesis function. Then we can illustrate the definition of the logistic model as:

### Advantages

* It performs well even when the dataset size is small.
* Because logistic regression calculates coefficients for each factor, it is simpler to understand how each feature influences the result.
* It has a low resource requirement and is computationally effective.
* Provides probabilities for every category that can be used to gauge the model's level of competence.

### Disadvantages

* Considers a linear decision boundary, therefore it is inappropriate for interactions that are complex and non-linear.
* Multicollinearity: If features are significantly linked, efficiency decreases.
* Robust to outliers that could have a negative impact on performance.
* Limited Complexity: For sophisticated data sets with intricate interactions between features, less effective than other methods.

## Naïve Bayes

Naïve Bayes algorithm (NB) is a supervised learning algorithm, depending on Bayes theorem which is used for solving classification problems like text classification (Bužić & Dobša, 2018) with a high-dimensional training dataset. Naïve Bayes algorithm helps us to build a fast machine learning model that can make quick predictions depending on the probability of an object.

Naive Bayes predicts depending on past experience. Below mentioned equation determined the working of NB.

Where:

* *X: Unknown class’s data*
* *c: Specific class*
* *P(c | x): posterior probability*
* *P (c): prior probability*
* *P(X | c): Probability that is based on conditions on the hypothesis*
* *P (x): Probability A*

### Advantages

### It is simple and straightforward to grasp because it is built on the Bayes theorem and implies prediction independence.

### Because of its simplicity, it is able to deal with big datasets efficiently. Computationally efficient, it takes lesser time to train large dataset.

### Easily tackles challenges with multiclass classification.

### Excellent in text data classification.

### Disadvantages

* It have naive assumption that all features are independent but it’s not possible in real-time dataset.
* Probability estimations are frequently unreliable, despite the fact that classification may be precise.
* The algorithm gives a class zero probability when it does not appear in the training data
* It is a simple algorithm and not appropriate for really complex data.

## Random Forest Classifier

A Random Forest Classifier is an ensemble learning algorithm that constructs multiple decision trees and combines their predictions using a majority voting mechanism. For multi-class problems (Bahuleyan, 2018), each decision tree in the random forest is trained on a bootstrapped sample of the dataset and uses a random subset of features at each split. This strategy introduces diversity among the trees, reducing overfitting and improving generalization. The predicted class for an instance is the one with the majority vote among all trees. Random forests are robust to overfitting, handle non-linear relationships well, and often achieve better performance compared to individual decision trees.

### Advantages

### Proficient of detecting associations among features that are not linear.

### It reduces probability of overfitting by averaging several decision trees,

### It specifies information about the significance of a feature, assisting in feature selection.

### Possibility of employing proximity-weighted imputation or median values to address missing data.

### Disadvantages

* Random forests can be complicated, requiring more memory and processing power.
* Due to the construction of numerous trees, this model requires more time to train than simpler models.
* Compared to logistic regression, it is a black-box model that is more difficult to comprehend.
* When dealing with noisy data, it may be possible to biased to the noise.

# LDA Topic Model

A particular type of text mining is called topic modelling which extracts recurring patterns from text-based information. It involves applying models to the collection of text data in order to create a representation of the subjects covered there. An NLP method known as topic modelling uses statistical connections of words in text data to construct latent topics. Topic modelling is a method of information extraction and classification of enormous collections of texts. Without using any predetermined dictionaries or interpretive rules, these latent themes are constructed. A set of texts' fundamental subjects are discovered using topic modelling, which also provides quantitative measurements that identify and characterise the topics of specific texts among a set of texts and enable thematic comparisons both inside and between texts. Gensim library has been used to build model. To build LDA on the dataset we have extracted unigram features from the text. We have bag of words (Bow) features from the dataset and apply LDA model on extracted features. To chose optimal topic number for our dataset we have evaluated LDA model on topics numbers in range 5 to 30. To evaluate LDA model’s performance, we have used the Coherence evaluation metric the highest score determines the best number of topics. LDA model extracts 7 number of topics over the entire dataset having coherence score 0.401.

# Experimental Setup

This segment designates the experimental setup used for developing machine learning techniques for genre and topic prediction tasks and topic modeling using music lyrics dataset. This module illustrates experimental settings and evaluation methods.

## Evaluation Methodology

The problems of genre prediction and topic prediction are treated as supervised multiclass classification (Prajwal, et al., 2021) tasks. For genre prediction task input will be lyrics and other supporting features according to research question and output attribute will be genre. For topic prediction task input will be lyrics and other supporting features according to research question and output attribute will be topic. The problem of theme extraction from text is treated as topic modeling task.

## Models

For the multiclass classification task we have built Logistic Regression, Naïve Bayes and Random Forest classifier. We have extracted features using tfidf vectorizer. As it can be seen from above exploratory data analysis dataset is highly imbalanced to balance the dataset we have used SMOTE oversampling technique. Synthetic Minority Over-sampling Technique (SMOTE) is an advanced oversampling method that generates synthetic samples for the minority class to balance the class distribution. SMOTE works by selecting samples from the minority class and generating new synthetic instances based on the feature space similarities between the selected samples and their k-nearest neighbors. After applying SMOTE we passed those features to the models. After the model is trained on the dataset we have used the test data to evaluate models performance using multiple evaluation measures.

## Evaluation Measures

For experimentations evaluated in this project, we have used the following mentioned evaluation metrics. To compare different models, we used five metrics to evaluate the model’s performance. Starting with Accuracy, that is the most common criteria for measuring classification models performance, but if working with a multiclass classification dataset, this criterion is not appropriate because the minority class will have a small contribution to the accuracy criteria. So, we will use six evaluation measures including Accuracy, Recall, Precision, F1 Score and Confusion Matrix. The formulas are defined as follows:

**Accuracy**

Accuracy provides the ratio among true predictions and overall predictions of the model.

**Recall**

It depicts that how many classes from positive class are actually model classified as positive.

**Precision**

It depicts that how many classes predicted by model as positive are actually positive.

**F1 score**

It provides the average score between precision and recall.

**Confusion Matrix**

It provides the complete frequency of TP, FN, TN, and FP in overall model predictions.

Where TP, FN, TN, and FP given in equations represent the true positive prediction, false negative prediction, true negative prediction, false positive prediction, respectively.