**MACHINE LEARNING REPORT**

**Comparing different classification models for**

**Taylor’s Manifest Anxiety Scale (TMAS)**

Team members: Manasi Murarka – J038, Khyaati Nayak – J044, Bharati Panigrahi – J045

ABSTRACT

We all experience anxiety; it is a natural human state and a vital part of our lives. Anxiety is often regarded as an artefact of modern societies, one that is increasingly represented in visual arts, music, literature, and social media. Anxiety helps us to identify and respond to danger in ‘fight or flight’ mode. It can motivate to us face up to dealing with difficult challenges. The ‘right’ amount of anxiety can help us perform better and stimulate action and creativity. But there is another side to anxiety. Persistent anxiety causes real emotional distress and can lead to us becoming unwell and, at worst, developing anxiety disorders such as panic attacks, phobias and obsessional behaviours. Anxiety at this level can have a truly distressing and debilitating impact on our lives and impact on our physical as well as our mental health.

Machine learning, a fast-growing approach that’s spreading out and helping every sector in making viable decisions to create the foremost of its applications. Our main objective is to classify and find out the levels of anxiety among different age groups. Various Machine Learning algorithms were applied to predict the level of anxiety. GridSearchCV, a tuning technique was applied compute the optimum values of hyperparameters. Amongst all, Voting Classifier works the best with 96.95% accuracy followed by Gradient Boosting with Support Vector Classifier as the base estimator with 96.67% accuracy.

Keywords: anxiety, obsessional behaviours, GridSearchCV, Gradient Boosting, Voting Classifier, SVM

**INTODUCTION - Taylor Manifest Anxiety Scale**

**Description:**The **Taylor Manifest Anxiety Scale**, often shortened to TMAS, is a test of [anxiety](https://en.wikipedia.org/wiki/Anxiety) as a [personality trait](https://en.wikipedia.org/wiki/Personality_trait), and was created by [Janet Taylor](https://en.wikipedia.org/wiki/Janet_Taylor_Spence) in 1953 to identify subjects who would be useful in the study of anxiety disorders. It was originally developed as a device for selecting subjects for inclusion in psychological experiments on stress, motivation, and human performance. It has subsequently been used as a general indicator of anxiety as a personality trait, it is not intended as a specific measure of anxiety as a clinical entity. Taylor thought that personality drive level would be reflected in the intensity of “manifested anxiety” and measured it using true/false responses. Items judged by clinicians as being indicative of manifest anxiety were selected from the Minnesota Multiphasic Personality Inventory.

**Scoring/Interpretation:**True-false responses are used for each item, and the replies indicating anxiety are counted, giving a score from 0 to 50 with the higher the score representing a higher level of anxiety. It is up to the discretion of the psychiatrist to decide where they fit in the “manifest anxiety” interpretation.

**MATERIALS AND METHODOLOGY**

This research focused on detecting anxiety using the **Taylor Manifest Anxiety Scale** questionnaire (TMAS). We used datasets which had information about the person's age, gender, and anxiety score based on the response to the 50 questions. We performed Principal Component Analysis, a dimension reduction technique and subsequently classified using six machine learning algorithms – namely Naïve Bayes, KNN, Logistic Regression, Support Vector Machine, Decision Tree, Random Forest Tree. We also applied GridSearchCV, K-fold cross validation, two techniques of boosting – Adaboost and Gradient Boosting and Voting Classifier.

1. DATA PRE-PROCESSING:

The dataset has 53 columns – anxiety score, gender, age and 50 questions few of which are listed below.

Q1. I do not tire quickly.

Q2. I am troubled by attacks of nausea.

Q3. I believe I am no more nervous than most others.

Q4. I have very few headaches.

Q5. I work under a great deal of tension.

We saw there are outliers in the “age” variable, so we removed them. We then converted the continuous variables “age” and “score” into categorical variables:

For score:

0 to 5 => Normal

6 to 20 => Mild

21 to 35 => Moderate

36 to 50 => Severe

For age:

14 to 25 => Student

26 to 45 => Working Class

46 to 60 => Middle Age

Above 60 => Senior Citizen

After this we performed Exploratory Data Analysis. Few of the visualizations are as follows:

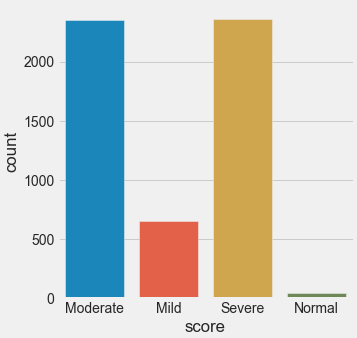


Fig- 1.A

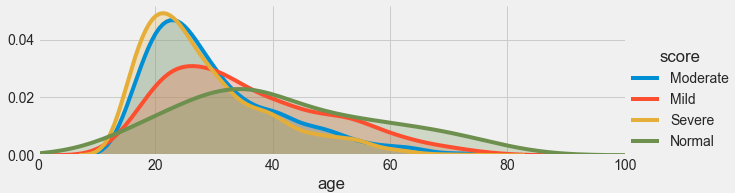


Fig – 1.B

The graph 1.A depicts the count of people suffering from different levels of anxiety. It is quite evident that most of them suffer from Moderate to Severe levels of anxiety while very few people suffer from normal levels of anxiety. Although the number of people suffering from mild anxiety are less than those suffering from moderate or severe anxiety it is still a significant number of people. The graph 1.B showing the levels of anxiety across different age groups in the form of KDE plot indicates similar pattern where moderate and severe anxiety are most frequent followed by mild anxiety with normal anxiety being the least frequent.

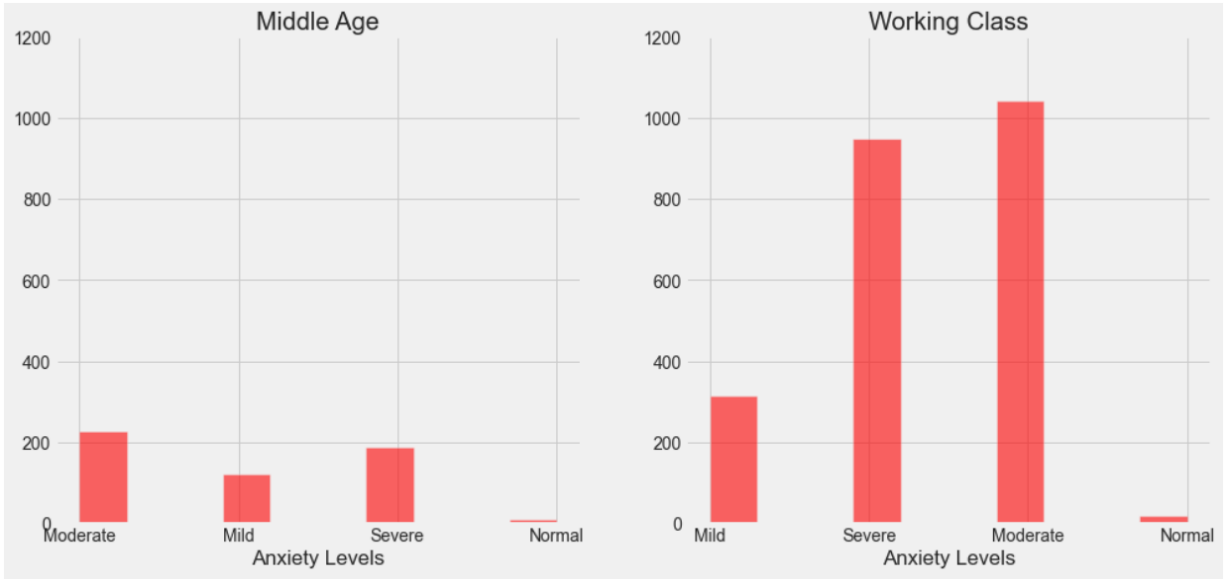


Fig – 2.A

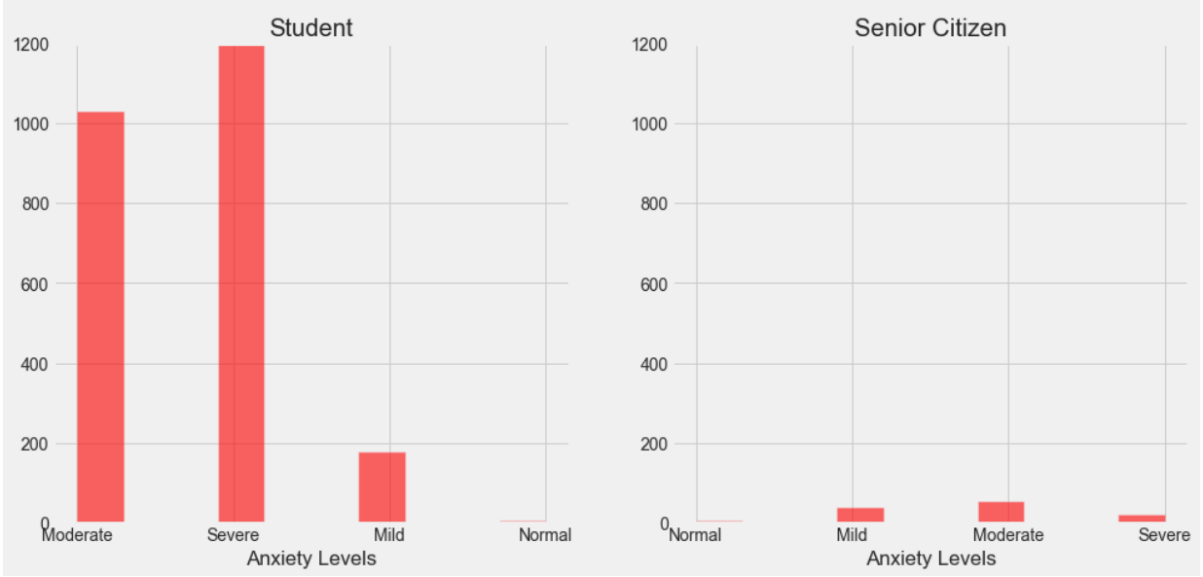


Fig – 2.B

These graphs indicate the frequency of different levels of anxiety across various age groups, viz., Student, Working class(25-45yrs), Middle Age(45-60yrs) and senior citizens (above 60yrs). It is a comparison of individual anxiety forms across various ages. It can be observed from the above graphs that students have a high frequency of severe anxiety followed by moderate anxiety, but the reverse is true for the working class as they have higher frequency of moderate anxiety followed by severe anxiety. Mild and normal forms of anxiety are higher in the working class compared to students. As against this the Middle-aged population have significantly lower levels of anxiety while senior citizens have even lower frequency of the different types of anxiety.

Using ordinal encoder, we encode our categorical features as an integer array. Using pipeline, we replaced the missing values with the median of the respective columns. We also performed PCA to find which factor amongst the ones we were taking into consideration had the highest impact on anxiety. We split our dataset into train and test set with a ratio of 80 - 20.

1. CLASSIFICATION:
   * 1. Naïve Bayes Classification

The Naïve Bayes classifier calculates conditional probability using Bayes theorem to divide into different classes. This theorem depends on the naïve assumption, in which input factors are independent of each other.

* + 1. K-Nearest Neighbour Classification

It finds similarity between predefined classes and the classes to be classified using Euclidean distance. Another algorithm used in this category is K-star, uses similarity measure as entropy distance.

* + 1. Logistic Regression

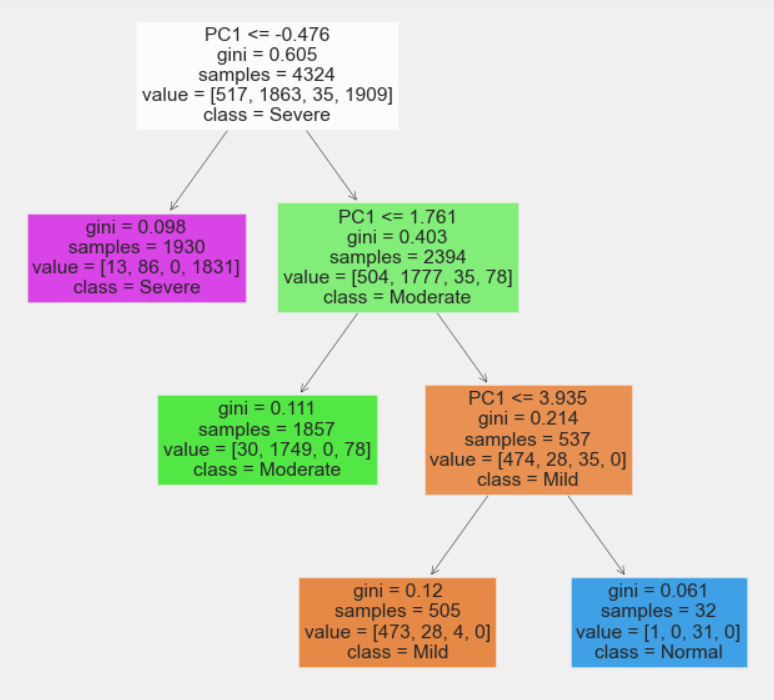
It is a statistical analysis method used to predict a data value based on prior observations of a [data set](https://whatis.techtarget.com/definition/data-set).

* + 1. Support Vector Machine

The algorithm creates a line or a hyperplane (a subspace whose dimension is one less than that of its ambient space) which separates the data into classes. It works for both regression and classification tasks but is mainly used in classification.

* + 1. Decision Tree Classifier

The decision tree method of machine learning makes decisions at different levels using tree data structure – this is suitable for predictive problems because they are easy to interpret, and the structure is stable. It covers both classification as well as regression.

****

* + 1. Random Forest

The Random Forest classifier creates multiple decision trees from randomly selected subset of training dataset. Then it aggregates the votes from different decision trees to decide the final class of test objects.

* + 1. Voting Classifier

A Voting Classifier is a machine learning model that trains on an ensemble of numerous models and predicts an output (class) based on their highest probability of chosen class as the output. It simply aggregates the findings of each classifier passed into Voting Classifier and predicts the output class based on most of the voting.

1. BOOSTING TECHNIQUES:
   * 1. Adaboost:

AdaBoost algorithm, short for Adaptive Boosting, is a Boosting technique used as an Ensemble Method in Machine Learning. It is called Adaptive Boosting as the weights are re-assigned to each instance, with higher weights assigned to incorrectly classified instances.

* + 1. Gradient Boosting:

Gradient boosting is a type of machine learning boosting. It relies on the intuition that the best possible next model, when combined with previous models, minimizes the overall prediction error. The key idea is to set the target outcomes for this next model in order to minimize the error.

1. GridSearchCV:

GridSearchCV tries all the combinations of the values passed in the dictionary and evaluates the model for each combination using the Cross-Validation method. Hence after using this function, we get accuracy/loss for every combination of hyperparameters, and we can choose the one with the best performance.

1. K-FOLD CROSS VALIDATION:

Cross-validation is a resampling procedure used to evaluate machine learning models on a limited data sample. The procedure has a single parameter called K that refers to the number of groups that a given data sample is to be split into. As such, the procedure is often called K-fold cross-validation. When a specific value for K is chosen, it may be used in place of K in the reference to the model, such as K=10 becoming 10-fold cross-validation.

**RESULTS**

1. CONFUSION MATRIX:

TP (True positive) = Diagonals of matrix

FN (False Negative) = Sum of the corresponding row for class (excluding TP of that class)

FP (False Positive) = Sum of the corresponding column for class (excluding TP of that class)

TN (True Negative) = Sum of all the row and column (excluding row and column of that class)

Calendar

Description automatically generated with medium confidence A picture containing table

Description automatically generated

Table

Description automatically generated with medium confidence Table

Description automatically generated

A picture containing calendar

Description automatically generated A picture containing table

Description automatically generated

1. CLASSIFICATION REPORT:

A picture containing text, receipt

Description automatically generated

A picture containing text, receipt

Description automatically generated

A picture containing text, receipt

Description automatically generated

Table

Description automatically generated

A picture containing text, receipt

Description automatically generated

A picture containing text, receipt

Description automatically generated

1. ACCURACY vs AUC – K-FOLD:

NAÏVE BAYES

Chart, line chart

Description automatically generated

K-NEAREST NEIGHBOUR

Chart, line chart

Description automatically generated

LOGISTIC REGRESSIONChart, line chart

Description automatically generated

SVM

Chart, line chart

Description automatically generated

DECISION TREE

Chart, line chart

Description automatically generated

RANDOM FOREST

Chart, line chart

Description automatically generated

1. COMPARING DIFFERENT MODELS:

Table

Description automatically generated

Chart, bar chart

Description automatically generated

It is quite evident that Random Forest, Decision Tress, SVM and Logistic Regression perform best on test dataset. So, we apply boosting techniques on these classifiers to improve their performance.

1. MODEL ACCURACY AFTER BOOSTING:

Table

Description automatically generated

Chart

Description automatically generated

**CONCLUSION**

Various machine learning algorithms were applied to determine four different severity levels of anxiety. Data was collected using a standard questionnaire i.e., TMAS. Subsequently six different classification techniques were applied – namely Naïve Bayes, KNN, Logistic Regression, Support Vector Machine, Decision Tree, Random Forest Tree. Also, GridSearchCV, K-fold cross validation, two techniques of boosting – Adaboost and Gradient Boosting and Voting Classifier were applied. Amongst all, Voting Classifier works the best with 96.95% accuracy followed by Gradient Boosting with Support Vector Classifier as the base estimator with 96.67% accuracy. The important variable found were:

Q3. I believe I am no more nervous than most others.

Q24. I dream frequently about things that are best kept to myself.

Q30. I cry easily.

Q11. I worry quite a bit over possible misfortune.

As such these variables were considered most important in detecting anxiety.

**REFERENCES**

1. *https://instruct.uwo.ca/kinesiology/9641/Assessments/Psychological/TMAS.html*
2. *https://en.wikipedia.org/wiki/Taylor\_Manifest\_Anxiety\_Scale*
3. *https://www.kaggle.com/lucasgreenwell/manifest-anxiety-scale-responses?select=data.csv*
4. *https://www.sciencedirect.com/science/article/pii/S1877050920311984*
5. *https://scholars.direct/Articles/biomedical-research/ibr-4-020.php?jid=biomedical-research*
6. *https://arxiv.org/ftp/arxiv/papers/1903/1903.10222.pdf*
7. *https://www.mdpi.com/1660-4601/18/14/7625*
8. *https://www.sciencedirect.com/science/article/pii/S1877050920309091*