## **Stress Detection using Naive Bayes and Decision Tree Classifiers**

**N Krishna Chowdary Pathuri**

Student, KL University, Vijayawada, 522502, Vijayawada, India

**Abstract.**

A large percentage of people deal with the ubiquitous problem of stress in our ever-busier lives, which includes both emotional and physical strains. Regrettably, a lack of awareness causes many people to fail to identify the symptoms of stress. Early detection of stress is essential because, if left unchecked, its effects can lead to serious health issues that progressively erode one's physical and mental health.

Automated methods help to lower the risks related to stress that go untreated in addition to making stress detection easier. Although a number of models, including the multi-attention model, the Factor Graph model, and the Personal Knowledge model, have been developed for social media blogs, their efficacy has been limited because they have mainly examined single-line text.

To overcome this constraint, a new Decision Tree algorithm classifier has been presented, with the goal of identifying stress from multi-line text data in the dataset that includes user-generated content that is both stressed and unstressed. The method has practical potential as evidenced by the promising accuracy of the experimental results.

## **Introduction**

The identification and classification of stress has become a significant difficulty in recent times. There are three types of stress: acute, episodic acute, and chronic stress. Stress is described as a condition of mental or emotional strain. It affects several body systems, including as the immune system, metabolism, and memory. Unfortunately, people who experience stress are frequently ignorant of their situation due to the lack of trustworthy stress-detecting technologies. Effective treatment of stress requires early identification. The importance of stress detection in modern culture is shown by the reliance of stress detection techniques on textual, visual, and social indicators. Early detection of stress has a major influence on treatment plans and patient outcomes.

In addition to being difficult and time-consuming, manual stress detection is also error prone. Therefore, it is imperative to have automated stress detection techniques. This chapter provides an overview, a problem description, objectives, a study of current systems, importance, and limitations of the project. Stress, which is basically a sensation of physical or emotional pressure, can have many causes, such as anxiety, rage, or irritation. Stress is a frequent problem, but if ignored, it can have serious consequences that could result in fatal illnesses including diabetes, cancer, heart disease, and depression. Direct questioning is a key component of traditional stress detection techniques, and if it is ignored, it can lead to depression. Since stress is not always obvious, it is critical to comprehend the causes that lead to it. Sleep issues, light-headedness, worry, exhaustion, mood fluctuations, sickness, tense muscles, and shaking are common indicators of stress. Stress may come from a variety of places, including the workplace, education, family, and social relationships. When under pressure, the degree of stress can quickly increase. Chronic mood disorders or addiction are examples of long-lasting stress that can have severe effects over a lengthy period of time. Short-term stress, on the other hand, can typically be resolved quickly and arises from everyday obstacles like traffic or interpersonal issues. Because social media is so widely used as a platform for emotional expression, a lot of study has been done in this area, yet the findings are frequently disappointing. Enhancing efficiency requires identifying and overcoming the restrictions in stress detection via study. Furthermore, it is imperative to investigate various stress indicators and their impact on individuals. In order to better understand this intricate phenomenon, a text segment is used as an input for stress detection in this study.

**Background Work**

The ability of a digital system or device under system control to carry out tasks that are frequently performed by intelligent species is known as artificial intelligence (AI). The phrase describes the endeavour to develop artificial intelligence (AI) systems that possess cognitive abilities like those of humans, including knowledge-based learning, reasoning, significance, and generalisation. Nevertheless, no service can match human adaptability to a greater variety of tasks or those requiring a significant amount of prior knowledge, even with the continuous advancements in computer processor speed and memory capacity.

Learning, reasoning, and personality are the three main psychological skills in AI programming. Techniques of learning: The area of AI programming is focused on data collection and the development of rules that guarantee that the data can be changed into useful information. The rules, sometimes referred to as methodologies, give computing devices exact instructions on how to carry out a specific task.

A machine learning approach uses a forecasting model that is built using historical data to predict the overall outcome when it receives user input. How well the output is estimated depends on the amount of information used, since a large dataset makes it possible to create a model that quantifies the result. Machine learning has gained importance. Because machine learning can accomplish tasks that would be too complicated for a human to perform directly, it is thought to be necessary. We require computer systems because we are unable to physically obtain such vast amounts of data; machine learning can help with this. By adding large amounts of data to machine learning algorithms, we can improve them. Training data is used to train machine learning algorithms. They can quickly forecast and make decisions based on past data when new data is received. supervised learning, unsupervised learning, and semi-supervised learning are the two categories of machine learning. The machine learns while it is being watched in supervised learning. One of its features is a model that can forecast using labelled data. A labelled dataset suggests that you are aware of the ideal viewpoint already. The deep learning methodology is divided into two groups. They are regression and classification, in that order.

Classification is used to describe a regression model that is categorical and has two or more classes. True or false, male or female, yes or no, and so forth are a few examples. Whether the output variable is imaginary or real A change in one variation is connected to a transformation in another because there is an interaction between the two or more variations. Pay according to work history, weight according to height, etc. are a few examples. When a device uses a training data set for unsupervised machine learning, it typically trains itself completely. In a big dataset, the machine must look for configurations and react to them correctly. Two categories of machine learning exist. They are association and clustering. An argument with a friend or a traffic jam are examples of everyday activities that can quickly relieve stress. These kinds of stress are referred to as acute or transient stress. Long-lasting stress, like that caused by alcoholism or mood disorders, is referred to as chronic stress. The scientific community should identify and address the implications in stress detection for ongoing advancements. Furthermore, a variety of results ought to be acquired in order to ascertain the individual's level of impact. A few lines of text were used as input to identify stress to ascertain all of this.

**System Design:**

In figure 1, the architecture of the Stress Detection Architecture describes the flow of the project where the dataset is loaded, and the data pre-processing step is carried out along with the feature selection. Then the dataset is spitted into test and train which carried out the machine learning techniques are carried out. Here we use two algorithms to calculate the accuracy of the system, namely Decision Tree and Naïve Bayes algorithms, where it finally results whether a person contains stress or not.

**A diagram of a data processing process

Description automatically generated**

Fig1: Stress Detection Architecture

Here fig (2), explains about the main dataset divided into testing and training datasets for simplification.

**A diagram of a full data set

Description automatically generated**

**Stress:**

In this busy world, stress has become an integral part in everyone’s life. Stress leads to both physical and mental illness. T shouldn’t be neglected, ad it leads to other serious problems like depression and cancer. Therefore, stress can be defined as the mental pressure and worry about the difficulties in life.

### **Cause of Stress**

You might be feeling under immense pressure, facing significant life changes, worrying about certain matters, lacking control over outcomes, dealing with overwhelming responsibilities, experiencing discrimination or abuse, and going through a period of uncertainty while also not having enough work, activities, or change in your life.

Stress, a common experience across various life forms, can be triggered by a multitude of factors. Among these, work-related stress stands as a prominent catalyst for stress in individuals. However, the spectrum is broader, encompassing a range of other stressors. Instances such as unemployment, health issues, and the associated uncertainties can heavily contribute to heightened stress levels. Additionally, the pressures emanating from work environments, compounded by long hours and insufficient support, can further exacerbate this psychological strain. Moreover, traumatic events, poor mental health, and the weight of financial commitments constitute additional stress-inducing elements, forming a complex interplay of factors that can significantly impact an individual's overall well-being.

### **Symptoms of Stress**

The symptoms of stress can vary from person to person, and it is not imperative for everyone to experience the same manifestations. However, if an individual's usual behavior shifts towards abnormal patterns, it becomes crucial to observe and take necessary steps to identify the underlying issue. Stress can manifest in both physical and mental symptoms within the body, with common physical indicators encompassing aches and pains, sleep disturbances, high blood pressure, and digestive issues.

### **Advantages of stress detection**

### Stress detection mainly benefits to protect the health and life of a person.

* Accurate than manual detection
* Decreases the risk of getting effected.
* Increases the lifespan of pupil.
* Saves money invested for medical alliances.

### **Challenges in stress detection**

* The classification of the different words into different categories, because all words are not similar.
* Identifying the level of the stress means at which level the person is having stress.
* Word embeddings have a significant impact on identifying the accuracy.

**Methodology**:

Detecting stress levels from social media blogs can be difficult since stress can appear in several ways and individuals may not always directly state their stress levels in their articles. The proposed Decision tree algorithm is a machine learning technique that attempts to determine whether a person is stressed based on the content of their social media blogs.

The decision tree algorithm is a form of supervised learning algorithm that makes predictions using a tree-like model. In this scenario, the algorithm would be trained on a dataset of social media blogs that had been labelled as stress-inducing or stress-free. The system would analyze numerous blog post features, such as the language used, the length of the post, and the presence of specific keywords, to determine whether a new post indicates stress or not.

To use this technique, you must first collect a collection of social media blogs that have been labelled as suggesting stress or not. This can be accomplished by manually labelling a selection of posts or by using an already labelled dataset. Once you have a labelled dataset, you can train a model to predict if a new post signals stress or not using a decision tree approach. To assess the accuracy of your model, divide your dataset into training and testing sets.

One potential limitation of this approach is that people may use different language and keywords to indicate stress, making it difficult to develop a comprehensive set of features capable of accurately capturing all instances of stress. Furthermore, other factors such as a person's personality and overall mental health may influence whether they indicate stress in their social media blogs, which may be difficult to capture using only the content of their blog posts.

However, there are some possible methods for detecting stress levels based on social media blogs:

**Decision Tree:**

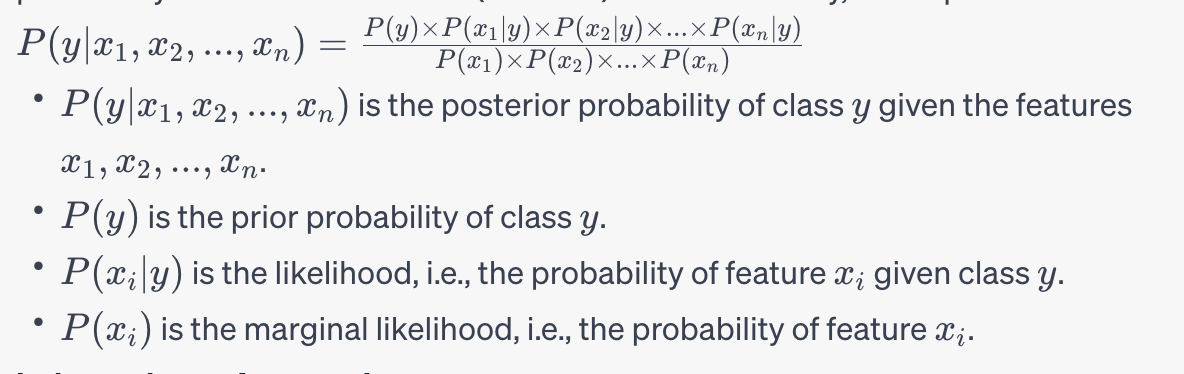
For both classification and regression tasks[4], a well-liked machine learning algorithm is the decision tree classifier. It functions with both continuous and categorical input and output variables. It forecasts the class label for a specific input sample in the context of classification. Here's how it functions:

* Decision trees are one of the most effective methods for data mining; they have been widely used in several disciplines[2] because they are easy to be used, free of ambiguity, and robust even in the presence of missing values.[3]A common method of selecting the best possible sub-tree from several candidates is to consider the proportion of records with error prediction .
* Recursively dividing the data into subsets according to the input feature values creates the tree. At each stage, the algorithm determines which feature is best to divide the data into. The attribute that best distinguishes classes is the "best" attribute. This procedure carries on until the algorithm decides, based on predetermined parameters (such as purity levels or a depth limit), that additional splitting is not required.
* The various algorithms have their own merits on execution speed, scalability, comprehensibility of output result and accuracy of the Classification forecasting [5]
* The choice of which feature to split on is made at each node in the tree. Metrics such as mean squared error (for regression problems) or Gini impurity (for classification problems) are used to select the feature.
* The algorithm keeps splitting until it decides to stop for various reasons, such as reaching a maximum depth, having fewer samples in a node than a threshold, or meeting other stopping requirements. The final nodes, referred to as leaf nodes, represent the classes or class probabilities in classification problems after the splitting process is complete.
* A new data point is sent from the root to a leaf node of the tree in order to generate predictions. The input sample is then given that leaf node's class label.

**Naive Bayes Classifier:**

Based on Bayes' theorem, the Naive Bayes Classifier is a probabilistic machine learning algorithm. It works especially well for text classification issues, like sentiment analysis and spam detection, but it can also be applied to other kinds of data. In many real-world scenarios, Naive Bayes classifiers—hence the name "naive"—perform surprisingly well for such a simple machine, especially when the assumption of feature independence holds true. This is how it operates:

* Based on the likelihood of the observed data (features), the Bayes theorem determines the probability of a hypothesis (class label). In terms of math, it is expressed as:



* Given the class label, the "naive" assumption of Naive Bayes is that features are conditionally independent. Put differently, the existence or lack of a specific feature has no bearing on the existence or lack of any other feature. This assumption greatly streamlines the calculations.
* It was easy to understand, required training data to parameters estimate, Unresponsive to unrelated features, handled real and distinct data well [6,7].

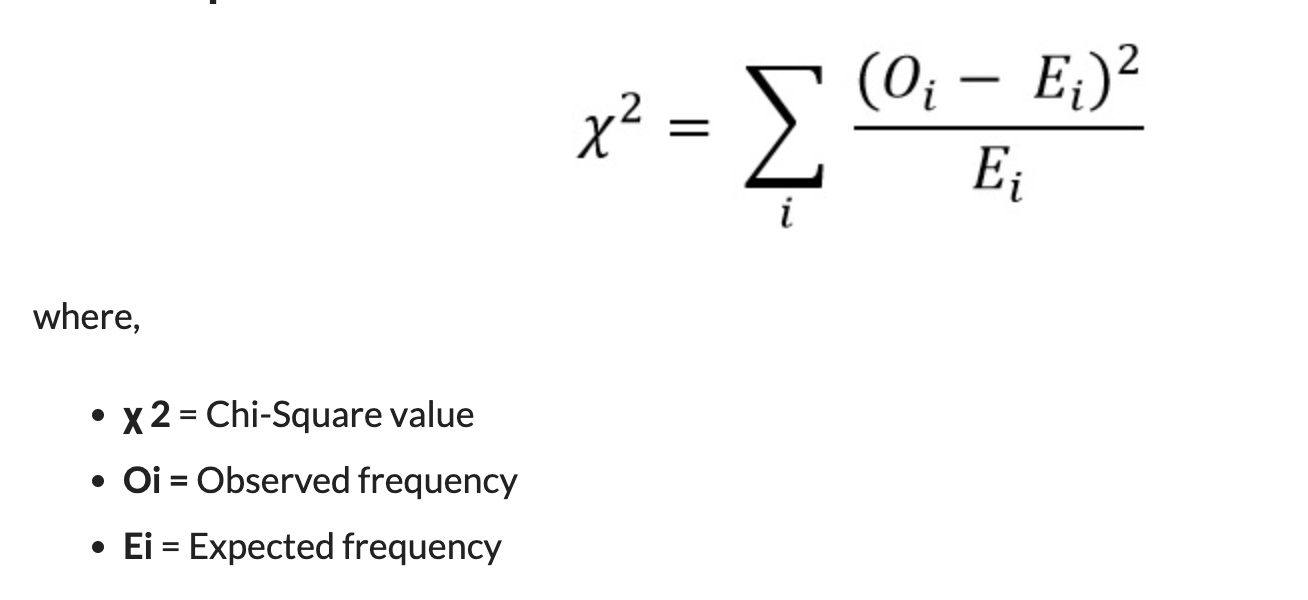
**Data Collection:**

* The process of collecting data for stress detection entails compiling multiple lines of text from different sources. This information is then utilized to train a machine learning algorithm to identify and categorize stress detections.
* Train and test data are both included in the data set. There are 2838 multi-line stressed words in the training dataset and 715 multi-line stressed words in the testing dataset.

**Data Preprocessing:**

* There are sometimes missing values in the sensor data collected for equipment failure, maintenance, and the repair of equipment [9], and missing values in the data obtained in real time affect the final performance of machine learning models [10].
* One crucial step in getting data ready for machine learning algorithms is data pre-processing. Pre-processing in the context of stress detection is converting the unprocessed picture data into a format suitable for being supplied into a machine learning algorithm.
* Here in this project, we used chi-square test as a data preprocessing technique to increase accuracy score and for comparison of the results.
* Raw datasets are usually characterized by incompleteness, inconsistencies, lacking in behaviour, and trends while containing errors [8]. The pre-processing is essential to handle the missing values and address inconsistencies.
* **Chi-square test :**

A statistical technique used in machine learning to choose features with the strongest correlation with the target variable is called Chi-Square (χ²) Feature Selection. When working with categorical features and target variables, it is especially helpful. [1] CHAID (Chi-Squared Automatic Interaction Detection).Chi-square tests are useful in establishing whether the features and the target variable significantly correlate. This is how it operates:



**Data Splitting:**

* In machine learning, data splitting is the process of separating a dataset into two or more subsets in order to make model testing, evaluation, and training easier. Partitioning the data is primarily done to evaluate a machine learning model's performance on novel, unseen data, which aids in determining how well the model will generalize to novel, unseen data in the real world.

Here are the common types of data splits in machine learning:

* **Training Dataset:**

The training set receives the lion's share of the dataset. This data is used to teach the model during the training process. To guarantee that the model learns significant patterns, it is imperative that the training set accurately represents the distribution of all the data.

* **Testing Dataset:**

The test set is a wholly unique dataset that is unfamiliar to the model. It is only used once to assess the model's performance following training and fine-tuning. An objective assessment of the final model's capacity to generalize to fresh, untested data is given by the test set.

**Training the model:**

* In figure 3, we discuss the process of teaching a machine learning algorithm to identify patterns in data, which is known as model training. The algorithm gains knowledge of the connections between input features and target outputs from a labelled dataset during training. After being trained, this model can be applied to new, unseen data to make predictions or decisions.

The below fig(3), describes about the process involved in the evaluating the model .Simply, collect data means searching(or) retrieving the dataset of model training. Prepare data refers for data preprocessing and the testing , training of dataset to deploy a model.

A black line with text

Description automatically generated

Fig 3: Steps involved in model training.

**Results Discussion:**

* **Without Applying Feature Selection:**

**A screen shot of a computer code

Description automatically generated**

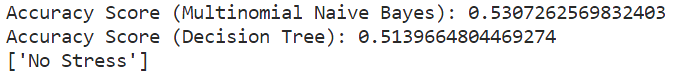
**A screenshot of a computer

Description automatically generated**

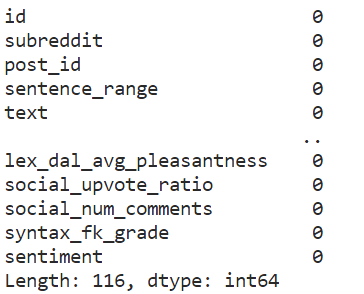
**A screenshot of a computer

Description automatically generated**

**Accuracy Score:**

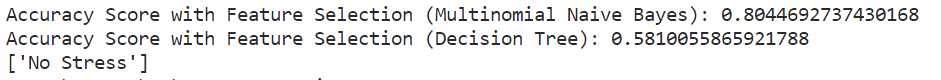
****

* **With Applying Feature Selection:**

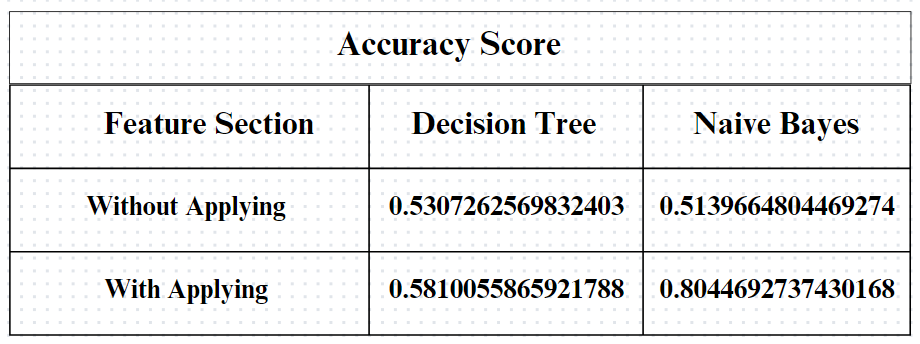
**A screenshot of a computer screen

Description automatically generated**

**Accuracy Score:**

****

**Final Table of the Accuracy Scores:**

****

**Conclusions**

By using the two methodologies, named Naïve Bayes Classifier and Decision Tree Algorithm, we get the accuracy score with less accuracy. Later by applying Feature Selection and using Chi–Square test we increase the accuracy of our model, as the Chi–Square test is used for calculating Chi-square between each feature and the target and select the desired number of features with best Chi-square score.

**Acknowledgement.**

I would like to thank my supervisor for their guidance and support during this project. I am grateful to the participants for their time and contributions. My heartfelt appreciation goes to my friends and family for their unwavering support. Lastly, I acknowledge the research community for laying the groundwork for this study.

**References.**

1. Quinlan RJ. *C4.5: Programs for Machine Learning.* San Mateo California: Morgan Kaufmann Publishers, Inc.; 1993.
2. Hastie TJ, Tibshirani RJ, Friedman JH. *The Elements of Statistical Learning: Data Mining Inference and Prediction. Second Edition.* Springer; 2009. ISBN 978-0-387-84857-0
3. Hastie T, Tibshirani R, Friedman J.*The Elements of Statistical Learning.* Springer; 2001. p. 269-272 G. Saranya, G. Geetha, C. K, M. K and S. Karpagaselvi, "Sentiment analysis of healthcare Tweets using SVM Classifier," 2020 International Conference on Power, Energy, Control and Transmission Systems (ICPECTS), 2020, pp. 1-3, doi:

10.1109/ICPECTS49113.2020.9336981.

1. Breiman L, Friedman JH, Olshen RA, Stone CJ. *Classification and Regression Trees.* Belmont California: Wadsworth, Inc.; 1984.
2. N. Bissantz and J. Hagedorn, “Data mining”, Business and Information Systems Engineering, vol. 1, (2009), pp. 118-122.
3. UmadeviD.Sundar, Dr.P.Alli,”An Optimized Approach to Predict the Stock Market Behavior and Investment Decision Making using Benchmark Algorithms for Naive Investors”, Computational Intelligence and Computing Research (ICCIC), 2013 IEEE
4. Gholap, J. (2012). Performance Tuning Of J48 Algorithm For Prediction Of Soil Fertility. Asian Journal of Computer Science and Information Technology, 2(8).
5. Zhang, Limao,Huazhong University of Science and Technology, Wuhan, China
6. Kim, J.; Han, Y.; Lee, J. Data imbalance problem solving for smote based oversampling: Study on fault detection prediction model in semiconductor manufacturing process. *Adv. Sci. Technol. Lett.* **2016**, *133*, 79–84.
7. García-Laencina, P.J.; Sancho-Gómez, J.L.; Figueiras-Vidal, A.R.; Verleysen, M. K nearest neighbours with mutual information for simultaneous classification and missing data imputation. *Neurocomputing* **2009**, *72*, 1483–1493.
8. M. Trotzek, S. Koitka and C. M. Friedrich, "Utilizing Neural Networks and Linguistic Metadata for Early Detection of Depression Indications in Text Sequences," in IEEE Transactions on Knowledge and Data Engineering, vol. 32, no. 3, pp. 588-601, 1 March 2020, doi: 10.1109/TKDE.2018.2885515.
9. David William, Derwin Suhartono,”Text-based Depression Detection on Social Media Posts: A Systematic Literature Review”,Procedia Computer Science,Volume

179,2021, Pages 582-

589,ISSN18770509,https://doi.org/10.1016/j.procs.2021.01.043.

1. Arfan Ahmed, Sarah Aziz, Carla T. Toro, Mahmood Alzubaidi, Sara Irshaidat, Hashem Abu Serhan, Alaa A. Abd-alrazaq, Mowafa Househ,”Machine learning models to detect anxiety and depression through social media: A scoping review”,Computer Methods and

Programs in Biomedicine Update,Volume 2,2022,100066,ISSN

26669900,https://do.org/10.1016/j.cmpbup.2022.100066.

1. Kailai Yang, Tianlin Zhang,Sophia Ananiadou,”A mental state Knowledge–aware and Contrastive Network for early stress and depression detection on social media”,Information

Processing & Management,Volume 59,Issue 4,2022,102961,ISSN 0306-

4573,https://doi.org/10.1016/j.ipm.2022.102961.

1. Ana-Sabina Uban, Berta Chulvi, Paolo Rosso,”An emotion and cognitive based analysis of mental health disorders from social media data”,Future Generation Computer

Systems,Volume 124,2021,Pages 480-494,ISSN 0167-

739X,https://doi.org/10.1016/j.future.2021.05.032.