iJRASET
International Journal For Research in Applied Science and Engineering Technology

INTERNATIONAL JOURNAL FOR RESEARCH

IN APPLIED SCIENCE & ENGINEERING TECHNOLOGY

Volume: 12 **Issue:** III **Month of publication:** March 2024

DOI: https://doi.org/10.22214/ijraset.2024.59219

www.ijraset.com

Call: 08813907089 | E-mail ID: ijraset@gmail.com

International Journal for Research in Applied Science & Engineering Technology (IJRASET)
ISSN: 2321-9653; IC Value: 45.98; SJ Impact Factor: 7.538
Volume 12 Issue III Mar 2024- Available at www.ijraset.com

Mental Health Prediction Using Catboost Algorithm

Mr. M. Jeevan Babu[1], Sirisha Kamunuri[2], Bhavana Sri Keerthi Jarugu[3], Srisaiteja Guttikonda[4], Venkata Srinivasa Rao Kesanam[5]

[1]*Assistant Professor,* [2, 3, 4, 5]*Student, Dept. of Computer Science and Engineering, Vasireddy Venkatadri Institute of Technology, Nambur, Guntur Dt., Andhra Pradesh.*

Abstract: ***This study investigates the application of the CatBoost algorithm in predicting mental health outcomes using Python programming language. Mental health prediction is a critical area of research due to its significant impact on individuals and society. Traditional predictive modeling techniques often encounter challenges in handling complex and high-dimensional data inherent in mental health datasets. CatBoost , a state- of-the-art gradient boosting algorithm, has shown promise in effectively addressing these challenges by handling categorical variables seamlessly and exhibiting robust performance in various domains. Leveraging its powerful capabilities, this study aims to develop predictive models for mental health outcomes utilizing a comprehensive dataset encompassing diverse socio- demographic, behavioural , and clinical factors. The predictive performance of the CatBoost algorithm will be evaluated and compared against other commonly used machine learning algorithms, demonstrating its effectiveness in accurately predicting mental health outcomes. This research contributes to the advancement of predictive modeling in mental health research and holds potential implications for personalized interventions and resource allocation in mental healthcare systems.***

Keywords: CatBoost Algorithm , Machine Learning ,Mental Health Prediction, College Students, psychology.

I. INTRODUCTION

Predicting mental health outcomes using machine learning algorithms has gained significant attention in recent years due to its potential to enhance early intervention and treatment planning. Among various algorithms, CatBoost stands out for its effectiveness in handling categorical variables and dealing with imbalanced datasets, which are common in mental health research. In this study, we aim to leverage the power of CatBoost algorithm implemented in Python to develop a predictive model for mental health outcomes.

By analyzing diverse socio-demographic, behavioural, and clinical factors, we seek to accurately predict the likelihood of individuals developing mental health issues. Such predictive models hold promise in informing targeted interventions and resource allocation, ultimately improving mental health outcomes and quality of life for individuals at risk.

In this project, we leverage the power of the CatBoost algorithm within the Python framework Flask to develop a predictive model for mental health disorders. CatBoost, known for its robust handling of categorical features and excellent performance in classification tasks, is an ideal choice for this endeavor. Through the integration with Flask an intuitive web application framework, we not only build a powerful predictive model but also create an accessible interface for users to interact with and understand the predictions. This project holds the promise of enhancing early detection and intervention efforts, ultimately contributing to improved mental health outcomes.

II. LITERATURE SURVEY

1) *Jinping Liu, Fang Xia, Yanyin Cui, Zixu Hao: The promotion effect of innovation and entrepreneurship education on medical student's ,mental health based on stepwise regression.*

Paper explores the impact of Advantages innovation and entrepreneurship. Improved. education on the mental health of medical students, and provides a basis for enhancing their mental health status.

The regression Mechanism Stress Caping Reduction through Creativity results showed that interpersonal Disadvantages communication, social support and, emotional Intelligence were the main factors influencing mental health before mass entrepreneurship and innovation. education, and interpersonal communication, social support and self-efficacy were the main factors influencing mental health after mass entrepreneurship innovation education.

2) *Youngji Koh;Chanhee Lee; Yunhee Ku; Uichin Lee: Data Visualization for Mental Health Monitoring in Smart Home Environment: A Case Study*

Mental health care and monitoring are important. Advancements in smart home sensing technology also make tracking people's activities easy in the home, enabling the monitoring of mental health more effectively. Some related works have demonstrated the possibilities of mental health monitoring using sensor data collected in smart home. However there is a lack of prior research on how to effectively utilize smart home data visualization to help people understand how their everyday behaviour are related to their mental health status.

III. METHODOLOGY

A. Dataset

Our dataset was gathered via Kaggle, a platform that lets users search for datasets to utilize in the construction of artificial intelligence models. 7023 rows and 19 columns make up the dataset. The dataset includes both categorical and numerical variables. To encode the categorical data, techniques such as label encoding are employed. To preprocess and analyze data, utilize the Pandas package and the Python programming language. The dataset is split in a 20-to-80 ratio.

Our dataset consists of information about students and the questions they answered. There are 19 features in the dataset.

B. Model Algorithm Selection

We focus on the CatBoost algorithm, a gradient boosting algorithm specifically designed for handling categorical features efficiently.

CatBoost is a machine learning algorithm specifically designed for gradient boosting on decision trees. It's known for its ability to handle categorical features seamlessly without the need for pre-processing, making it particularly useful for real-world datasets with a mix of categorical and numerical features. Key features of CatBoost include: Gradient Boosting, Categorical Feature Handling, Regularization. Optimized Training Process, Scalability

Overall, CatBoost is a powerful and versatile algorithm that excels in handling categorical data, providing high accuracy, and being relatively easy to use compared to other gradient boosting implementations.

Among various algorithms, CatBoost stands out for its effectiveness in handling categorical variables and dealing with imbalanced datasets, which are common in mental health research. In this study, we aim to leverage the power of CatBoost algorithm implemented in Python to develop a predictive model for mental health outcomes

C. Analysis of Model Construction Process

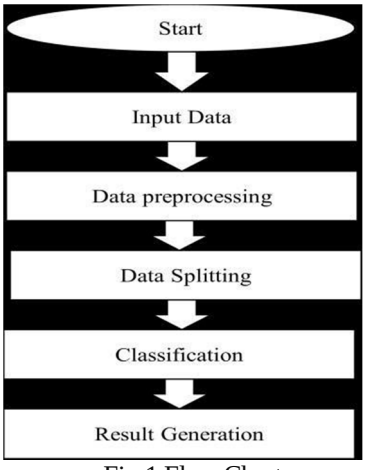


Fig.1.Flow Chart

In the proposed system for mental health prediction utilizing the CatBoost algorithm within the Python framework Flask, we aim to develop a user-friendly and efficient tool for early detection and intervention in mental health disorders. Leveraging the robustness of CatBoost, a powerful gradient boosting algorithm, we intend to analyze diverse datasets encompassing psychological, demographic, and behavioral attributes.

ISSN: 2321-9653; IC Value: 45.98; SJ Impact Factor: 7.538
Volume 12 Issue III Mar 2024- Available at www.ijraset.com

1) Data Collection

We have taken our data set from Kaggle. The dataset consists of 7028 data points about students. After that, we should decide on the method for data collection based on our question and the available resources. Common methods include surveys, interviews, observations, and experiments.

2) Data preprocessing

Data preprocessing is a critical step in preparing data for training with the Cat Boost algorithm. It involves handling missing values, encoding categorical features, and potentially scaling features. Cat Boost can handle missing values internally, but preprocessing them beforehand can be beneficial. Categorical features need not be one-hot encoded; instead, Cat Boost automatically encodes them using target encoding. Feature scaling is not necessary for Cat Boost due to its use of gradient boosting, which is not sensitive to feature scales. However, outliers should be handled appropriately. Splitting the data into training, validation, and test sets is essential for evaluating the model's performance.

3) Feature Extraction

Feature extraction is the process of deriving new features from existing ones to improve a model's performance. In the context of the Cat Boost algorithm, feature extraction is not explicitly performed as it is in traditional machine learning models. This means that Cat Boost can handle complex, high-dimensional datasets without the need for manual feature extraction. Additionally, Cat Boost can automatically handle categorical features and missing values, further simplifying the data preprocessing step. Overall, while traditional feature extraction techniques may not be required, thoughtful feature engineering and selection can still play a crucial role in optimizing model performance when using Cat Boost.

4) Data Splitting

It involves dividing the dataset into separate subsets for training, validation, and testing. The training set is used to train the model; the validation set is used to tune hyperparameters and evaluate model performance during training; and the test set is used to evaluate the final model's performance on unseen data. Proper data splitting is essential to ensure that the model generalizes well to new, unseen data.

In Cat Boost, data splitting can be done using the train_test_split function from the sklearn.model_selection module. This function randomly splits the dataset into training and testing sets based on a specified test size or a specified number of samples. For example, to split a dataset into 70% training and 30% testing sets, you can use(X_train, X_test, y_train, y_test) = train_test_split(X, y, test_size=0.2, random_state=42). This helps prevent bias in the model and ensures that it performs well on new data. Additionally, you should use a random seed (random_state parameter) to ensure the reproducibility of the results. By properly splitting the data, you can train and evaluate the Cat Boost model effectively, leading to better performance and generalization.

5) Hyperparameter Tuning

Hyperparameter tuning in the Cat Boost algorithm involves optimizing parameters like learning rate, tree depth, and number of iterations. Techniques like grid search and random search can be used to find the best combination of hyperparameters for improved model performance.

6) Model Training

Training a model with the cat Boost algorithm involves several key steps. First, the dataset is prepared by encoding categorical features and handling missing values. Then, the data is split into training, validation, and test sets. The model is trained on the training set using the Cat Boost Classifier class, specifying hyperparameters such as the learning rate and tree depth. During training, the model's performance is evaluated on the validation set. Once training is complete, the final model is evaluated on the test set to assess its performance on unseen data.

7) Model Evaluation

Model evaluation in the Cat Boost algorithm involves assessing the model's performance using metrics such as accuracy, precision, recall, and F1 score for classification tasks. Cross-validation can be used to get a more robust estimate of the model's performance.

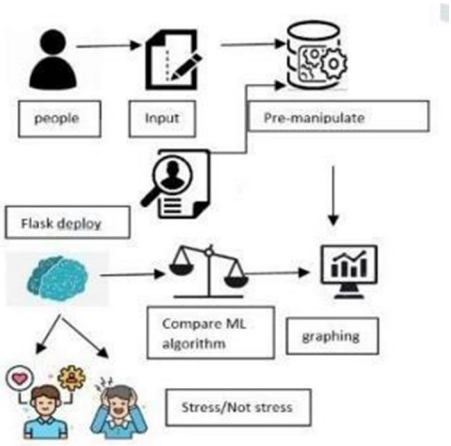


Fig.2.Architecture

IV. RESULTS AND DISCUSSIONS

The Final Result will get generated based on the overall classification and prediction. The mental health is classified into 3 different categories which includes Low stressed, Moderately Stressed ,Highly stressed.
The performance of this proposed approach is evaluated using some measures like,

A. Accuracy

Accuracy of classifier refers to the ability of classifier. It predicts the class label correctly and the accuracy of the predictor refers to how well a given predictor can guess the value of predicted attribute for a new data.
Here's the formula for accuracy:
Accuracy= TP+TN/TP+TN+FP+FN Where:
TP (True Positives) are the instances that are actually positive and are predicted as positive. TN (True Negatives) are the instances that are actually negative and are predicted as negative. FP (False Positives) are the instances that are actually negative but are predicted as positive. FN (False Negatives) are the instances that are actually positive but are predicted as negative.

V. FUTURE ENHANCEMENT

In future enhancements for mental health prediction using the CatBoost algorithm within the Python framework Flask, several avenues can be explored to refine and improve the predictive model's performance and usability. Firstly, incorporating more diverse and comprehensive datasets from various demographic groups and geographic regions can enhance the model's accuracy and generalizability, ensuring it caters to a broader population. Additionally, integrating advanced feature engineering techniques such as text sentiment analysis for social media data or incorporating wearable device data for physiological markers can provide richer inputs for the prediction model, capturing nuanced aspects of an individual's mental well-being. Furthermore, implementing interpretability techniques such as SHAP (SHapley Additive exPlanations) values or LIME (Local Interpretable Model-agnostic Explanations) can enhance the transparency of the model's decision-making process, fostering trust and understanding among users.

VI. CONCLUSION

In summary, our integration of the CatBoost algorithm into the flask framework for mental health prediction exhibits encouraging outcomes and potential applications within the realm of mental healthcare. By employing robust machine learning techniques and flask's user-friendly interface, we have developed a tool proficient in predicting mental health conditions based on pertinent input features.
The CatBoost algorithm, renowned for its adeptness in handling categorical features and robust performance in predictive tasks, forms a dependable foundation for our prediction model. Through its utilization, we have attained precise predictions while mitigating the risk of overfitting and enhancing interpretability.

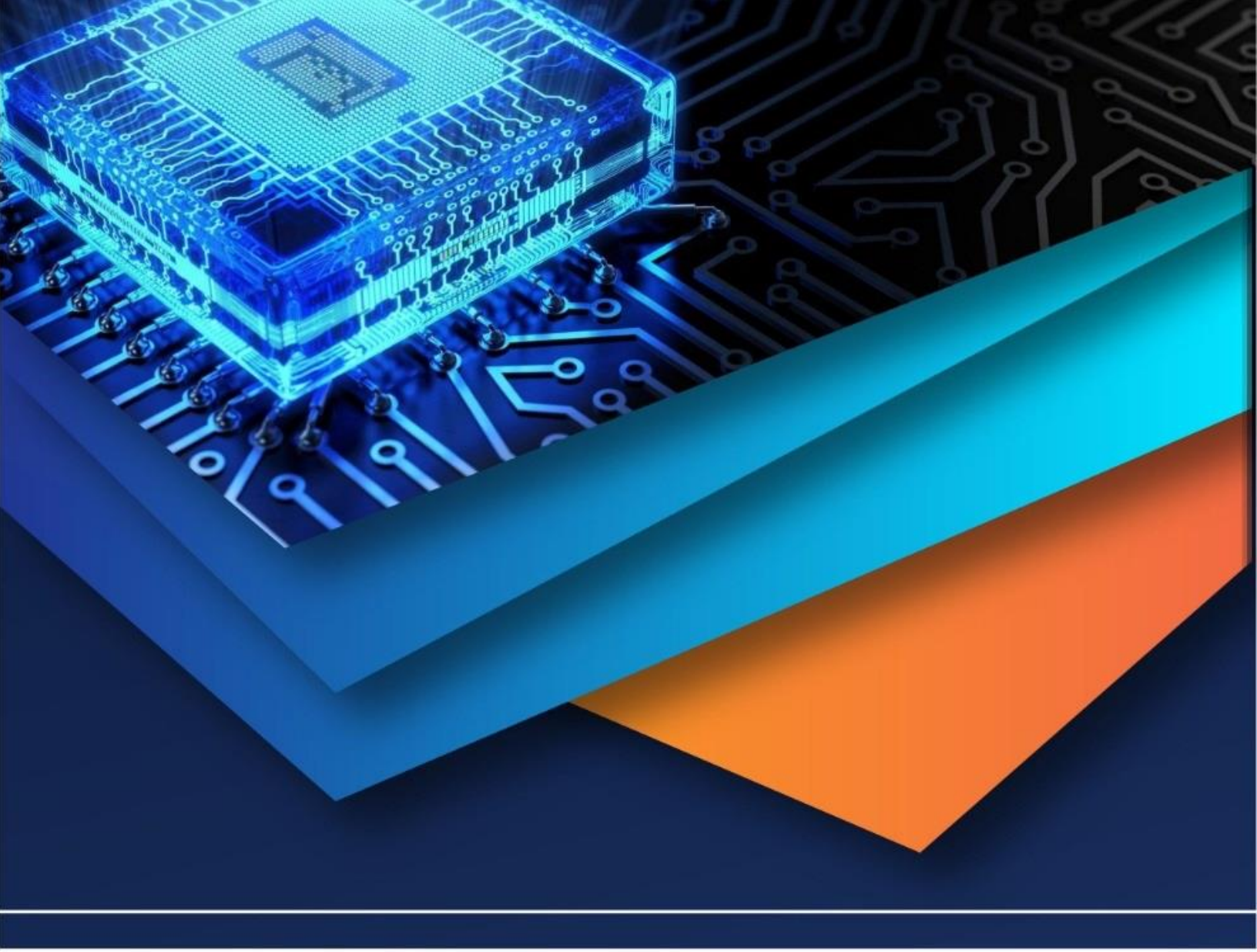
Conversely, flask furnishes an instinctive and interactive platform for users to input their data and acquire real-time predictions. Its simplicity and customizable attributes render it an optimal choice for deploying machine learning models and presenting their outcomes to diverse audiences, including healthcare professionals and individuals in need of assistance.

VII. ACKNOWLEDGEMENT

It is nature and inevitable that the thoughts and ideas of other people tend to drift in to the subconscious due to various human parameters, where one feels acknowledge the help and guidance derived from others. We acknowledge each of those who have contributed for the fulfillment of this project. We take the opportunity to express our sincere gratitude to our guide, Mr. M. Jeevan Babu, and project Coordinator Mr. P.R. Krishna Prasad, whose guidance from time to time helped us to complete this project successfully.

REFERENCES

[1] Jung Yuchae, Yong Ik Multimedia Tools and Applications, 76 (9) (2020), pp. 11305-11317 View PDF CrossRefView Record in ScopusGoogle Scholar
[2] Norizam, Sulaiman. Determination and classification of human stress index using the nonparametric analysis of EEG signals. Diss. UniversitiTeknologi MARA, 2020. Google Scholar
[3] Lawrence, O, Hall. A Primer on Cluster Analysis by James C. Bezdek [By the Book]. IEEE Systems, Man, and Cybernetics Magazine, 2018, 4(1):48-50.
[4] Ezugwu E S, Agbaje M B, Aljojo N, et al. A Comparative Performance Study of Hybrid Firefly Algorithm for Automatic Data Clustering. IEEE Access, 2020, 8(2020):121089-121118.
[5] Pastor K, Aanski M, Vujic D, et al. A rapid dicrimination of wheat, walnut and hazelnut flour samples using chemometric algorithms on GC/MS data. Journal of Food Measurement and Characterization, 2019, 13(3):2961-2969.
[6] Elizabeth S, Rebecca G, Margarita M B, et al. Preconception prediction of expectant fathers' mental health: 20-year cohort study from adolescence. Bjpsych Open, 2018, 4(02):58-60.
[7] Singh H, Kumar Y. Hybrid Artificial Chemical Reaction Optimization Algorithm for Cluster Analysis. Procedia Computer Science, 2020, 167(4):531-540.
[8] Urban M, Klemm M, Ploetner K O, et al. Airline categorization by applying the business model canvas and clustering algorithms. Journal of Air Transport Management, 2018, 71(AUG.):175-192.
[9] Yoganathan D, Kondepudi S, Kalluri B, et al. Optimal sensor placement strategy for office buildings using clustering algorithms. Energy and Buildings, 2018, 158(PT.2):1206-1225.
[10] Munshi A. Clustering of Wind Power Patterns Based on Partitional and Swarm Algorithms. IEEE Access, 2020, PP (99):1-1.
[11] Mccloskey S, Jeffries B, Koprinska I, et al. Data-driven cluster analysis of insomnia disorder with physiology-based qEEG variables. Knowledge-Based Systems, 2019, 183(Nov.1):104863.1-104863.11.
[12] Fritz M, Behringer M, Schwarz H. Quality-driven early stopping for explorative cluster analysis for big data. Computer Science, 2019, 34(2-3):129-140
[13] Kuo R J, Lin J Y, Nguyen T. An application of sine cosine algorithm-based fuzzy possibilistic c-ordered means algorithm to cluster analysis. Soft Computing, 2021, 25(11):1-16.
[14] Whiting D, Fazel S. How accurate are suicide risk prediction models? Asking the right questions for clinical practice. Evidence-Based Mental Health, 2019, 22(3): ebmental-2019-300102
[15] Pereira A , Borim F, Aprahamian I, et al. Comparison of Two Models of Frailty for the Prediction of Mortality in Brazilian Community-Dwelling Older Adults: The FIBRA Study. The journal of nutrition, health & aging, 2019, 23(10):1004-1010
[16] Jeevan Babu Maddala, M. Vanaja, P. Satya, N. Harika, N. Dinesh, Chronic Kidney Disease Prediction. Date of Publication: December 2022.Volume:11



10.22214/IJRASET

45.98

IMPACT FACTOR: 7.129

IMPACT FACTOR: 7.429

INTERNATIONAL JOURNAL FOR RESEARCH

IN APPLIED SCIENCE & ENGINEERING TECHNOLOGY