Comparison of Logistic Regression and Random Forest On Student Stress Prediction

Name : Mohammad Talib Akhtar

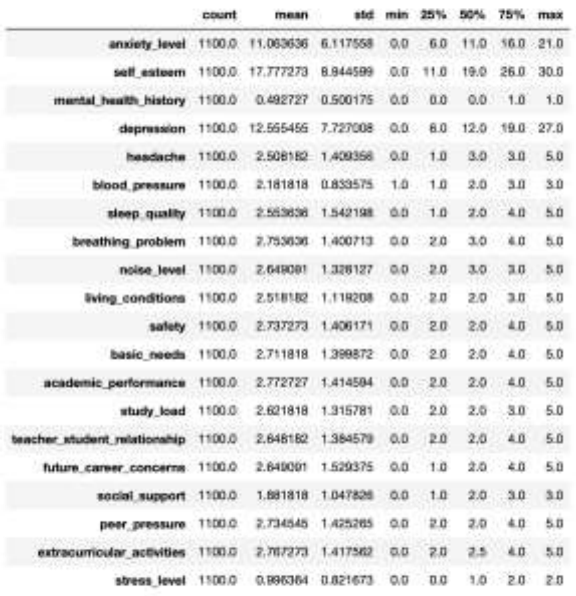
Coursework: INM431 Machine Learning

Description and Motivation

- Using Random Forest and Logistic Regression we will solve multi classification problem of Student Stress Level Prediction.
- We'll evaluate and compare these two supervised model’s performances by predicting student stress level (0: Low stress, 1: Average Stress, 2: High Stress).
- We will also compare our work with previous similar implementations done by S. Monisha and R. Meera (2020)

Initial analysis of the data set including basic statistics

- The dataset is Student Stress Factors from Kaggle.
- The dataset contains 1100 rows and 21 columns, where 20 are independent features and 1 is our target variable.
- Table 1 shows the description of the dataset, which is number of data, mean, standard deviation, min, max & data distribution.
- Pie chart depicts that dataset is balanced because 33.9 % data is of low stress, 32.5 % is about average stress level and 33.5 % is of high stress.
- In the self-esteem histogram it shows that the students having low self-esteem have higher stress level.
- Lastly the data correlation heatmap depicts the correlation between the features. For example, Academic performance is negatively correlated to stress level.



	count	mean	std	min	25%	50%	75%	max
anxiety_level	1100.0	11.063636	6.117558	0.0	6.0	11.0	16.0	21.0
self_esteem	1100.0	17.777273	8.944599	0.0	11.0	19.0	26.0	30.0
mental_health_history	1100.0	0.492727	0.500175	0.0	0.0	0.0	1.0	1.0
depression	1100.0	12.555455	7.727008	0.0	6.0	12.0	19.0	27.0
headache	1100.0	2.508182	1.409356	0.0	1.0	3.0	3.0	5.0
blood_pressure	1100.0	2.181818	0.833575	1.0	1.0	2.0	3.0	3.0
sleep_quality	1100.0	2.553636	1.542198	0.0	1.0	2.0	4.0	5.0
breathing_problem	1100.0	2.753636	1.400713	0.0	2.0	3.0	4.0	5.0
noise_level	1100.0	2.649091	1.328127	0.0	2.0	3.0	3.0	5.0
living_conditions	1100.0	2.518182	1.119208	0.0	2.0	2.0	3.0	5.0
safety	1100.0	2.737273	1.406171	0.0	2.0	2.0	4.0	5.0
basic_needs	1100.0	2.711818	1.399872	0.0	2.0	2.0	4.0	5.0
academic_performance	1100.0	2.772727	1.414594	0.0	2.0	2.0	4.0	5.0
study_load	1100.0	2.621818	1.315781	0.0	2.0	2.0	3.0	5.0
teacher_student_relationship	1100.0	2.648182	1.384579	0.0	2.0	2.0	4.0	5.0
future_career_concerns	1100.0	2.649091	1.529375	0.0	1.0	2.0	4.0	5.0
social_support	1100.0	1.881818	1.047826	0.0	1.0	2.0	3.0	3.0
peer_pressure	1100.0	2.734545	1.425265	0.0	2.0	2.0	4.0	5.0
extracurricular_activities	1100.0	2.767273	1.417562	0.0	2.0	2.5	4.0	5.0
stress_level	1100.0	0.996364	0.821673	0.0	0.0	1.0	2.0	2.0

Figure 1 Data Description

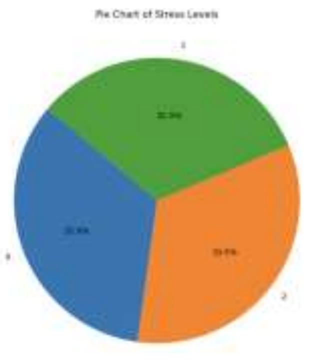


Figure 3. Pie Chart

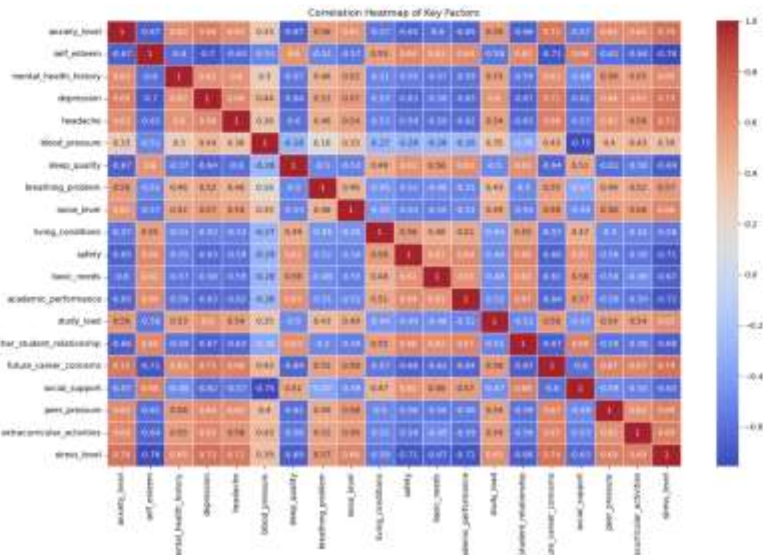


Figure 3. Heatmap

Hypothesis statement

- LR is faster to train and requires less computational resources while on the other hand RF is more computationally intensive, especially with a large number of trees.
- It is predicted that in terms of predicting student stress levels, both Random Forest and Logistic Regression will perform noticeably better than a random chance model.
- Random Forest is anticipated to produce a more intricate, potentially multi-dimensional decision boundary, whereas Logistic Regression will model the boundary linearly.
- In this particular context, it is anticipated that the Random Forest model will produce higher accuracy than Logistic Regression due to its capacity to capture complex, non-linear relationships in data.

Explanation of the methodology selected for training and evaluation

- Clean the dataset, Since variables like sleep quality are ordinal, it is best to replace missing values with the mode and eliminate ‘bullying’ column as it contains high percentage of missing values.
- Divide the dataset into subsets for testing and training, reserving 25% for testing and 75% for model training.
- Create predictive models by applying the algorithms for Logistic Regression (LR) and Random Forest (RF). These models will serve as the foundation for our knowledge of the variables influencing the stress levels of students.
- Use Grid Search Cross-Validation (GridSearchCV) for hyperparameter tuning to optimize both models. This step aims to improve performance by carefully iterating through different combinations of parameters to fine-tune the models.
- Get reliable estimates of the model’s accuracy, error rate, recall, and other relevant performance metrics by thoroughly evaluating them.

Results & Choice of Parameters

Logistic Regression

- Many of the learning techniques have faced issues in determining the optimized subset of features in order to improve the classification [2].
- The large feature set we used allowed our multiclass classification model to show a high degree of complexity. Even though this complexity was extensive, it presented problems for the model's interpretability and performance.

The choice of parameters

- The hyperparameter grid examined two types of regularization penalties ('l1' and 'l2'), a range of regularization strengths ('clf__C' values from 0.01 to 10), and It was determined that {'clf__C’: 10, 'clf__penalty': 'l2'} was the best configuration in the model.

Random Forest

- To ensure diversity and lessen overfitting, the model trains an ensemble of 100 decision trees using a bagging algorithm that uses entropy as the splitting criterion at each node [2].
- During optimization the Random Forest model has critical parameters that can be optimized to improve generalization and accuracy. 'n_estimators' (100–300) modifies the ensemble size to optimize performance and efficiency.

The choice of parameters

- The log scale range for hyperparameter tuning was 4 to 100.

Stress Prediction results		
LR	Model	**RF**
88.73	Accuracy	89.45
88.66	F1 Score	88.77
88.89	Precision	88.36
88.26	Recall	88.39

Table 1. Results

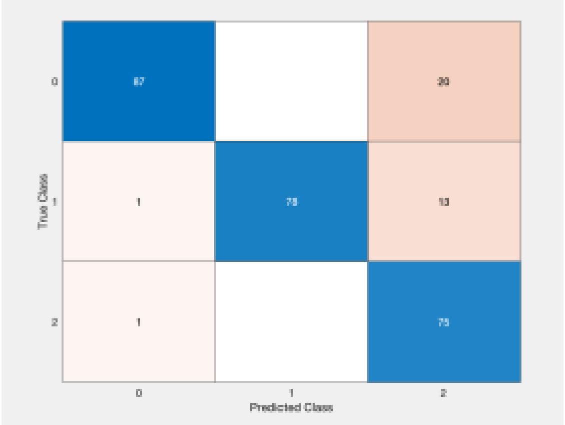


Figure 4. Confusion Matrix of Optimized LR

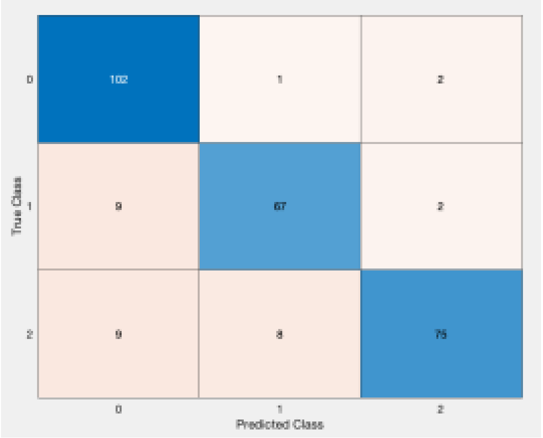


Figure 5. Confusion Matrix of RF

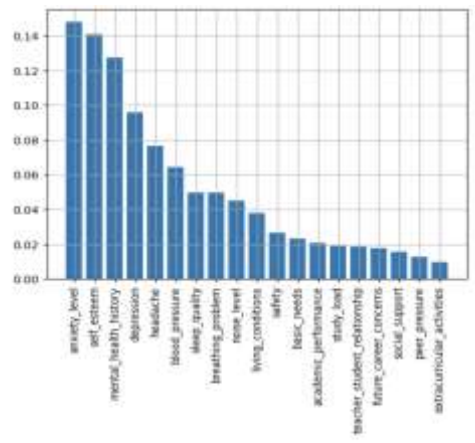


Figure 3. Bar graph of feature importance

Lessons learned and future work

Lessons learned

- A crucial stage in guaranteeing the quality and appropriateness of data for building precise and effective predictive models is data preprocessing.
- When interpreting, one must be aware of imbalance labels because precise, recall, and F1 score are additional outcomes that should be considered in addition to accuracy.

Future work

- More individualized interventions might be possible by creating customized models that take into consideration each person's unique stress thresholds and baselines.
- Integrating user input to improve the model and its forecasts over time.
- To optimize the model more further we can include k fold cross validation and use maximum depths, Bootstrap sampling, criterion, etc.

The summary of implementing model with their pros and cons

Logistic Regression (LR)

- Logistic regression is a statistical technique used to analyze datasets where one or more independent variables influence the outcome. The method helps in solving classification problems.
- A logistic regression model is very interpretable because its coefficients can be used to understand the relationship between the independent variables and the dependent variable's log odds.
- There is no presumption of linearity between the dependent and independent variables, in contrast to linear regression.

Pros

- Logistic Regression is more effective when data is linearly separated.
- It's known for being easy to understand and doesn't need high processing power.
- An understanding of the impact of features can be gained from analyzing the highly interpretable output of a logistic regression model.
- In order to categorize observations, it offers probabilities for outcomes that can be thresholded at various levels.

Cons

- It does not assume that the dependent or independent variables are linear, but it does assume that the log odds are linear.
- Alternative models may outperform Logistic Regression when it comes to complex relationships in data.
- Because of the significant influence of outliers on the result, the results may become skewed.
- It may underfit when dealing with intricate and non-linear relationships present in high-dimensional data.

Random Forest (RF)

- Random forest is a commonly-used machine learning algorithm which combines the output of multiple decision trees to reach a single result [3].
- In order to create diversity among the trees, Random Forest adds extra randomness to the tree-building process by splitting a node based on the best feature chosen from a random subset of features.
- Random Forest performs classification by generating trees and iteratively eliminates irrelevant features [4].

Pros

- To produce a prediction that is more reliable and accurate, it constructs numerous decision trees and combines them.
- To create the trees, Random Forest employs the bagging technique (Bootstrap Aggregating), which selects samples at random from the dataset and replaces them.
- The Random Forest algorithm achieves its high accuracy by averaging uncorrelated decision trees.

Cons

- Even though Random Forest is robust, biased decision trees especially if the data is unbalanced can still cause it to display bias.
- The model can be memory-intensive, which makes it difficult to run on computers with low memory.

Analysis and evaluation of results

- The accuracy Random Forest model's stress prediction on unseen data should be similar for each new set of samples. Estimates of a model's reliability can be gathered [5].
- There is a marginal performance difference between Random Forest and Logistic Regression, with Random Forest outperforming Logistic Regression by about 1%. This advantage is due to Random Forest's built-in feature ranking ability, which makes it easier to select features efficiently than Logistic Regression, which requires a different feature selection process.
- With an accuracy of 88.73%, precision of 88.89%, recall of 88.26%, and F1 score of 88.66%, Logistic Regression demonstrated excellent performance metrics. Its effectiveness as a baseline model is noteworthy; it implies that the decision boundaries between classes are probably linear, which may eliminate the need for more sophisticated models in some circumstances.
- The algorithmic performance was improved by using Grid Search Cross-Validation optimization. We thoroughly investigated the hyperparameter space, which includes regularization penalties ('l1' and 'l2') and a regularization strength continuum ('clf__C' from 0.01 to 10.
- A moderate regularization strength combined with an L2 penalty results in a balanced model complexity and peak performance. This is demonstrated by the identification of the optimal model configuration, which was {'clf__C’: 10, 'clf__penalty': 'l2}.
- Random Forest is a flexible ensemble tree model that can naturally model non-linear relationships and capture intricate feature interactions. It obtained 89.45% accuracy, 88.36% precision, 88.39% recall, and 88% F1 score.
- An illustrative analysis has shown that the predictive power of the model is significantly influenced more by factors like depression, anxiety, headaches, mental health, and self-esteem than by other features.
- Critical hyperparameters for Random Forest, like 'n_estimators is essential for fine-tuning the model to improve accuracy and generalization 'n_estimators', which range from 100 to 300, affects ensemble size and performance.
- As seen in the figure 3, we can see out of all the features anxiety level is the most important feature which affect the stress level prediction in random forest model.
- Unlike Logistic Regression, which has hyperparameters mainly related to optimization and regularization, Random Forest has a wide range of hyperparameters related to its structural design. This results in a much higher computational intensity for Random Forest tuning, especially when varying the number of trees and their depth. Nevertheless, the possible improvements in predictive performance frequently outweigh the computational burden.
- Confusion matrix (in the figure 4 & 5) analysis using Random Forest and Logistic Regression models for a multiclass classification dataset reveals interesting differences. The first matrix's representation of the Random Forest model shows that it performs well at all stress levels, and class-specific accuracy values imply that its predictive power is fairly equal. Conversely, the second matrix illustrates the slight variation in the accuracy of the Logistic Regression model's ability to identify the various stress levels, with a slightly lower precision in classifying the 'high stress' (class 2) category.
- For the 'medium stress' (class 1) instances, both models perform admirably. This is probably because the training data has a higher representation of this class, which results in a more confident classification outcome for this category. It appears that the Logistic Regression model has greater difficulty classifying the 'low stress' (class 0) and 'high stress' (class 2) levels, which may be a sign of the limitations of the linear decision boundaries in capturing the complexities of the data. Given its more consistent classification performance across all stress levels, one would expect the Random Forest model to have a higher F1-score a crucial metric that combines precision and recall. This measure would support Random Forest's resilience in managing multiclass classifications in an unbalanced dataset, surpassing Logistic Regression, which could have trouble with non-linearly separable class boundaries. The two models' different F1-scores are probably a reflection of Random Forest's inherent advantages in controlling class imbalance and capturing non-linear relationships within the feature set.

References

T. Mishra, D. Kumar and S. Gupta, "Mining Students' Data for Prediction Performance," *2014 Fourth International Conference on Advanced Computing & Communication Technologies*, Rohtak, India, 2014, pp. 255-262, doi: 10.1109/ACCT.2014.105. https://ieeexplore.ieee.org/stamp/stamp.jsp?tp=&arnumber=6783461&isnumber=6783406

P. B. Pankajavalli, G. S. Karthick and R. Sakthivel, "An Efficient Machine Learning Framework for Stress Prediction via Sensor Integrated Keyboard Data," in IEEE Access, vol. 9, pp. 95023-95035, 2021, doi: 10.1109/ACCESS.2021.3094334. https://ieeexplore.ieee.org/abstract/document/9475069

References

B. M. Booth, H. Vrzakova, S. M. Mattingly, G. J. Martinez, L. Faust and S. K. D’Mello, "Toward Robust Stress Prediction in the Age of Wearables: Modeling Perceived Stress in a Longitudinal Study With Information Workers," in IEEE Transactions on Affective Computing, vol. 13, no. 4, pp. 2201-2217, 1 Oct.-Dec. 2022, doi: 10.1109/TAFFC.2022.3188006. https://ieeexplore.ieee.org/stamp/stamp.jsp?tp=&arnumber=9813544&isnumber=9964459

S. Monisha, R. Meera, V. Swaminath R. and A. Raj L., "Predictive Analysis of Student Stress Level Using Naïve Bayesian Classification Algorithm," 2020 International Conference on Computer Communication and Informatics (ICCCI), Coimbatore, India, 2020, pp. 1-7, doi: 10.1109/ICCCI48352.2020.9104113. https://ieeexplore.ieee.org/stamp/stamp.jsp?tp=&arnumber=9104113&isnumber=9104005

“What Is Random Forest?” *IBM*, https://www.ibm.com/topics/random-forest#:~:text=Random%20forest%20is%20a%20commonly,both%20classification%20and%20regression%20problems.