Principal Component Analysis and Multiclass Classification of Stress

Ajay Rawat - 40221356

GitHub link: https://github.com/AjayINSE6220/Ajay_INSE_6220_Project-/tree/main

***Abstract*— In the recent busy lifestyle, people have become careless and have been ignoring sleep, which is ultimately affecting their health, and they are becoming more prone to being significantly impacted due to the lack of sleep. It's a fact that if you neglect sound sleep, it will have an impact on your health and productivity. So, quality sleep is necessary to have a productive daytime, and it's good for your health too. It is crucial to make sure that getting enough sleep is a priority in maintaining a healthy lifestyle. To investigate the correlation between duration of sleep and stress levels, principal component analysis (PCA) was utilized. The stress levels range from low to medium-high, and there are five of them. The methods, which included linear discriminant analysis, naive bayes, K neighbor classifier, logistic regression, random forest classifier, and extra tree classifier, were used and achieved a perfect accuracy score of 1.0 with LR.**

Index terms – Principal component analysis (PCA), stress level, sleep, health, Machine learning

I. INTRODUCTION

A cutting-edge device called the Smart-Yoga Pillow (SaYoPillow) has been developed to explore how stress and sleep are linked, and to develop the idea of "Smart-Sleeping." This device employs a unique program along with a powerful computer chip that can examine the sleeping habits and physiological changes which happens during the time of your sleep. It is suggested to make a stress prediction for the following day based on these changes throughout sleep. It is implemented to safely transport the analyzed stress data and average physiological changes to the IoT cloud for storage.

It is also suggested to transfer data securely from the cloud to any applications run by other parties. The user interface makes it possible for them to manage the data visibility and accessibility. SaYoPillow is a revolutionary product with up to 96% accuracy, security features, and consideration of sleeping patterns for stress relief. SayoPillow.csv shows the correlation between the user's snoring range, heart rate, blood oxygen levels, respiration rate, body temperature, rate of limb movement, number of hours of sleep, and stress levels (0- low/normal, 1 - medium low, 2 - medium, 3 - medium high, 4 - high) derived from the literature review.

Figure 1: SaYoPillow purpose

II. PRINCIPAL COMPONENT ANALYSIS

In many real-world applications, the data sets can be very large and multifaceted, with many associated variables that can make it difficult to process, store, and visualize the data. The Principal Component Analysis technique prove to be in handy where the real-world high-dimensional data is transformed into the low dimensional data so that interpretation becomes easier and useful for researchers.

In order to minimize the dimensions of the data while retaining the most information or variance in the dataset, a dimensionality reduction approach called PCA is applied. It entails identifying principal components and applying them to the data to conduct a shift of foundation. A large number of related variables can be condensed into a manageable number of independent variables known as principle components (PCs) using the multivariate PCA technique. These PCs can be viewed as new data coordinates that streamline the analysis and capture as much variability in the original variables as is practical.

Each succeeding PC captures the remaining variation that the preceding ones did not capture, with the first PC capturing the biggest variance in the data. The PCs are also orthogonal to one another, i.e., they are independent of one another and do not overlap.

PCA maintains the data's variability by shortening the distance between the data and its projection onto the PCs. Researchers are now able to visualize the data and spot patterns that weren't necessarily obvious from the original data. This facilitates and improves understanding.

A. PCA Algorithm

1. Center data

Prior to data centering, we compute the average (sometimes referred to as the mean) for each column

in the dataset. With p being the number of columns in the dataset, this results in a vector for us of length p. Then, from each component in the relevant column, we deduct the associated mean value. As a result, a new dataset is created with a mean value of zero for each column. The centering data is given by

Y=HX

2. **Covariance matrix**

The aim of this stage is to identify the connections between the variables in a dataset. Some variables may be highly correlated, resulting in duplicate or redundant information. The covariance matrix is utilized to detect and quantify these relationships. It is given by,

S=1/n YtY

3. **Eigen decomposition**

The method known as Eigen decomposition can be used to calculate the Eigen values and Eigen vectors of the covariance matrix, denoted by S. Eigen vectors indicate the direction of the principal components, while Eigen values indicate the quantity of variance associated with each principal component. The equation provided below is utilized for this purpose,

S = AΛA T

4. **Principal Component**

Using the n x p matrix (Z), compute the transformed data. The rows of Z indicate the observations, while the columns of Z represent the PC. The number of PCs is the same as the dimension of the algorithm in the original data matrix.

Z = Y A

IV

III. CLASSIFICATION ALGORITHMS

Classification algorithms are machine learning approaches that use features to categorize data into pre-defined groups or classes. They are utilized in a wide range of applications, including image identification, audio recognition, natural language processing, and fraud detection.

There are various types of classification algorithms, such as:

1. **Logistics Regression**

It is the technique employs for binary classification problems where the aim is to predict an outcome that is binary in nature based on more than one predictor variables. It is a supervised learning algorithm that models the relationship between the predictor variables and the binary outcome by mapping the predictors to the probability of the outcome, utilizing the sigmoid function. With this, you can figure out what will happen in situations where there are more than two possible outcomes.

2. **K closest neighbors**

K-nearest neighbors (KNN) is a supervised classification technique in machine learning that seeks to forecast the output of a given input by analyzing the outputs of k nearby data points. KNN is regarded as a "lazy learner" because it only stores the training data, and during testing, it computes the distance between the input data point and every other training data point.

3. **Random forest**

It is also known as random choice forest, where a large number of decision trees are built during the training phase of an ensemble learning approach, which is used for classification, regression, and other tasks. The fundamental trees that make up Random Forest are completely distinct from one another. A subset of data points and a subset of features are used to train each base classifier.

4. **Naive Bayes**

An assumption made by a Naive Bayes classifier is that the existence of one feature in a class has no bearing on the presence of any other features. When dealing with high-dimensional and sparse data, where the number of features is significantly more than the number of samples, the Naive Bayes Classifier performs particularly well. Missing data and categorical features are also simply handled by it. It is frequently applied to sentiment analysis, spam filtering, text classification, and recommendation systems.

IV. DATASET DESCRIPTION

The population's stress levels are shown in the pie chart below; as can be seen, Level 1 has the largest count and Level 3 has the lowest count.

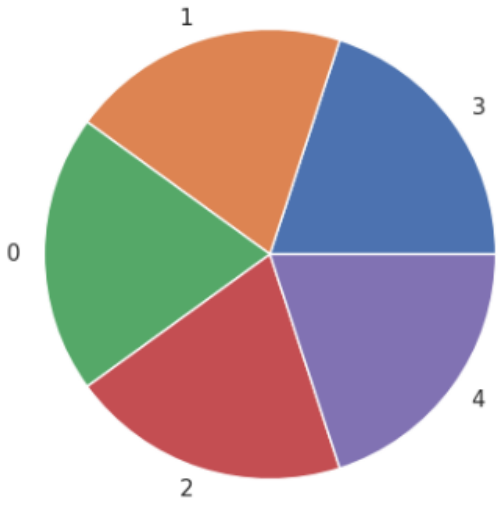

Figure 2: Stress level

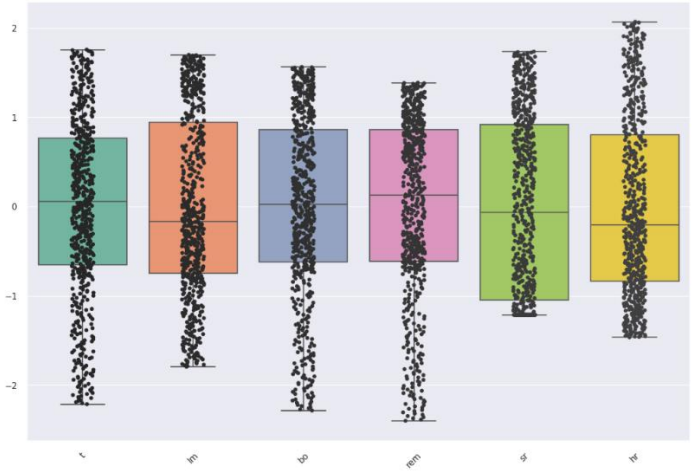

Figure 3: Strip plot

Figure 3 shows a strip plot, which is a box plot that has been expanded to include all of the normalized observations. It displays all of the data points as well as their distribution. This dataset has no outliers; every point falls inside the acceptable range."

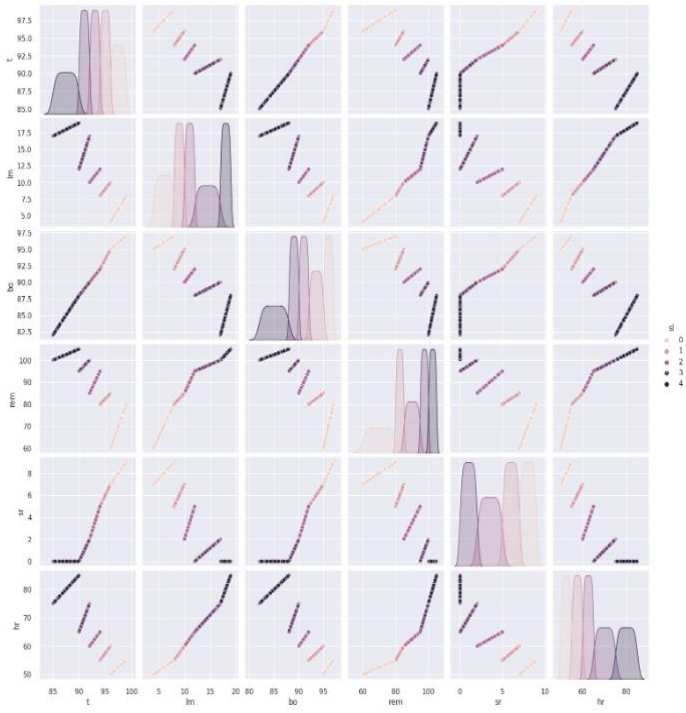

Figure 4: Pair plot

In figure 4, the pair plot shows the visual representation of the data each numerical attribute in the dataset. Pair plots are excellent for discovering unique or unexpected data points, as well as identifying patterns and relationships between variables.

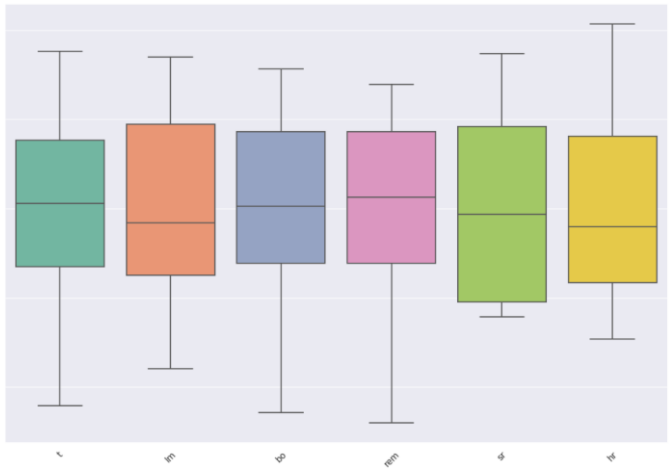

Figure 5: Box plot

The box plot illustrates the distributions of the attributes using lines and boxes. It uses lines and boxes to represent different values. The middle line shows the value in the middle of the data. The boxes show where most of the data is. Lines extend from the boxes to show the rest of the data. If there are any numbers that are far away from the rest of the data, they are marked with dots.

The correlation matrix is the matrix that shows relation between two elements in the random vector. The value is called highly correlated if it is near to 1 and least correlated if it falls near to 0.

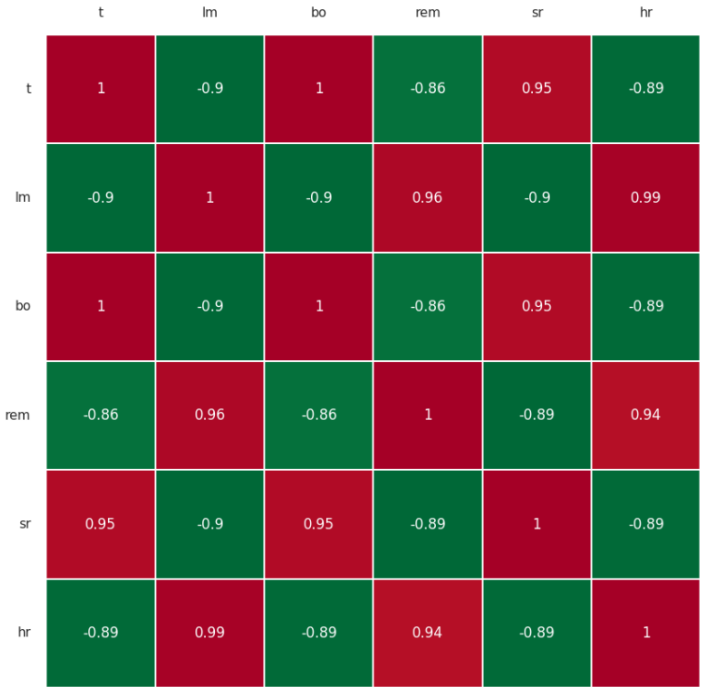

Figure 6: Pair plot

The lm i.e., limb movement is highly correlated with heart rate and eye movement.

V. PCA RESULTS

When the PCA library is used, the user has more flexibility and can accomplish a lot with just one line of code. Figures and charts from the PCA library implementation are shown in this report.

The 6 feature set can be reduced to fewer features by using the PCA stages. The eigen matrix A is used to condense the original n x p dataset. A PC is used to represent each column of the eigenvector matrix A. Each PC gathers a certain quantity of information, which establishes the dimension (r). The following information is provided for the stress level dataset's eigenvector matrix (A):

$$A = \begin{matrix} -0.408 & -0.467 & 0.19 & -0.222 & 0.085 & -0.721 \\ 0.412 & -0.385 & -0.237 & -0.021 & 0.788 & 0.053 \\ -0.408 & -0.451 & 0.196 & -0.349 & -0.003 & 0.684 \\ 0.402 & -0.451 & 0.568 & 0.497 & -0.249 & 0.032 \\ -0.408 & -0.290 & -0.554 & 0.656 & -0.081 & 0.058 \\ 0.408 & -0.375 & -0.486 & -0.385 & -0.550 & -0.064 \end{matrix}$$

The following equation is used to calculate the percentage of variation accounted for by the jth PC:

$$\ell_j = \frac{\lambda_j}{\sum_j^p \lambda_j} \times 100\%, \quad for\ j = 1, \ldots, p$$

The eigen values of each eigen vector from the eigen value matrix is given below:

5.60
2.76
8.25
4.48
2.97
1.50

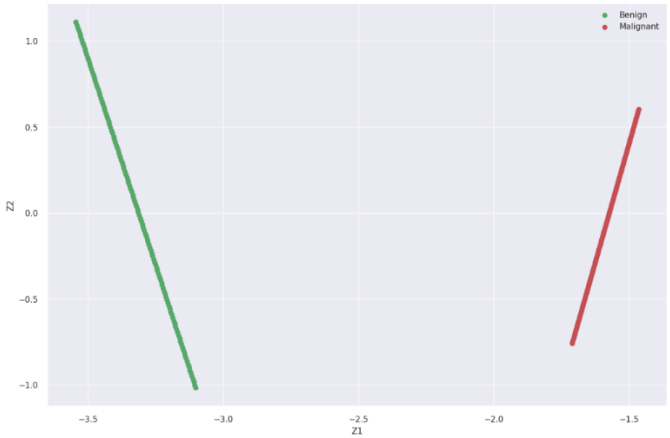

Figure 7: Scatter plot for components

The scatter plot in figure 7 displays the relationship between two components.

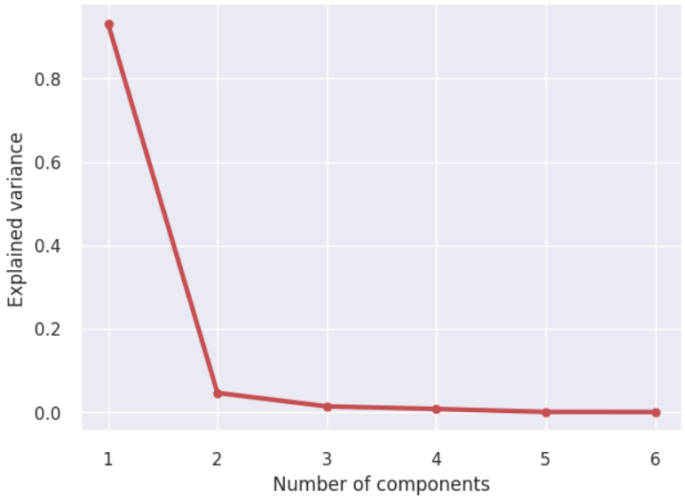

Figure 8: Scree plot

The scree plot shows that the first two principal components explain 99.17% of the total variance. This means that these two components contain most of the information in the original data set, and the remaining components contribute relatively little additional information. The elbow curve started at the second component, indicating that the first principal component explained most of the variance and the remaining components contributed relatively little additional information. Therefore, retaining the first two principal components is likely to be sufficient to capture most of the variation in the data set.

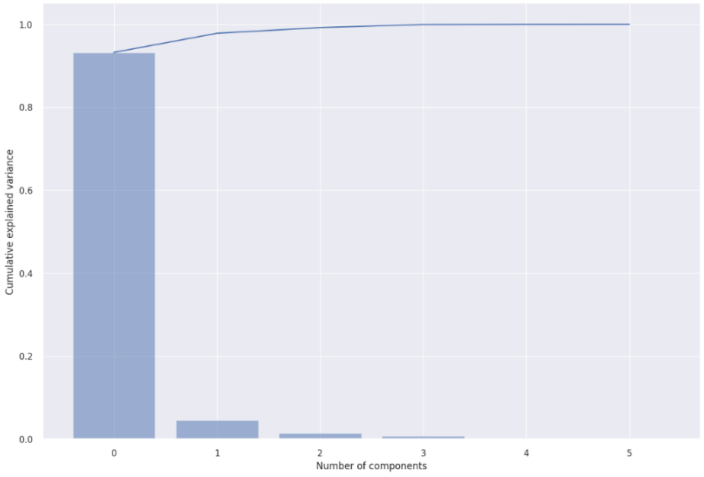

Figure 9: Pareto plot for components

The pareto plot shows the significant factors for sources of variation in the dataset.

The first principal component Z1 is given by:

Z1 = -0.408X1 + 0.412X2 - 0.408X3 + 0.402X4 - 0.408X5 + 0.408X6

The second principal component Z2 is given by:

Z2 = −0.467X1 - 0.385X2 - 0.451X3 - 0.451X4 - 0.290X5 – 0.375X6

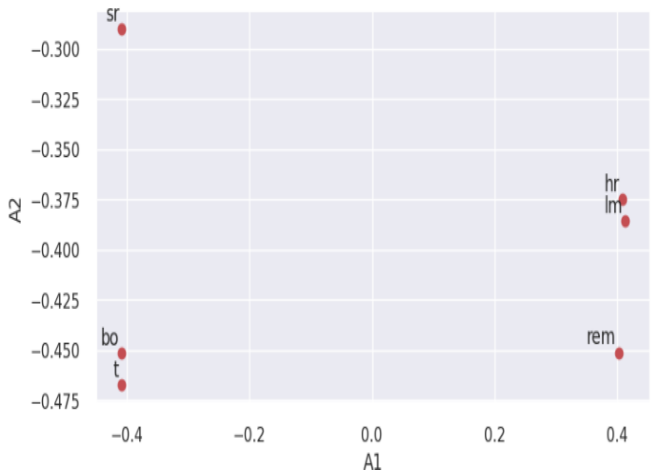

Figure 10: PC coefficient plot

Using a PC coefficient plot, above figure depicts each variable's contribution to the first two PCs. In other words, it displays which components have a comparable role in the first two PCs. Heart rate and limb movement has the highest contribution in first PC whereas ‘sr’ has has highest contribution in PC2. Temperature has the lowest contribution.

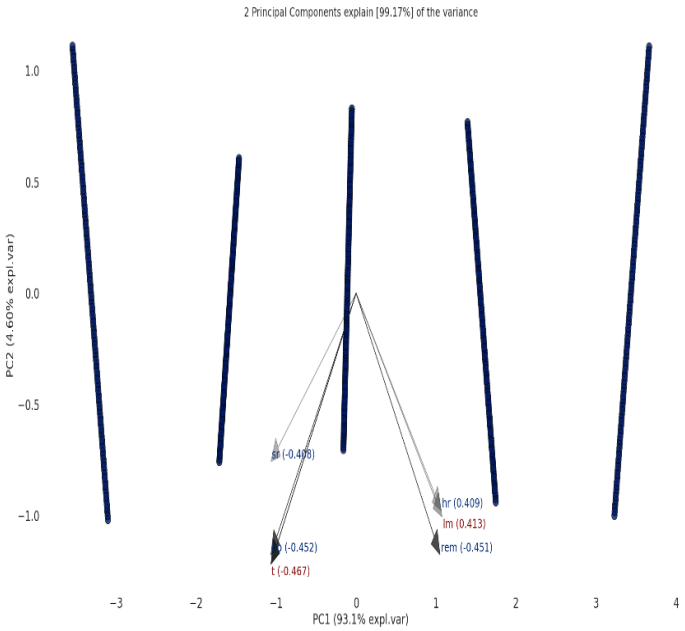

Figure 11: Biplot

Figure 11 is a biplot with PC1 and PC2 as the x and y axes, respectively. All of the observations are denoted by dots, while the rows of the eigenvector matrix are denoted by vectors. The angle formed by the vector and the axis indicates the variable's contribution. When the angle is modest, the contribution is large; as the angle increases, the contribution diminishes.

VI. CLASSIFICATION RESULT

In this part, machine learning classification algorithms is discussed. The objective of the analysis is to predict stress levels. The "compare model" function in pycaret, which compares the performance of multiple models in the library, was used to identify the best model. This function generated cross-validated performance metrics for each model, and then averaged the results. Based on the integrated approach, the model with the highest performance was determined to be the Logistic Regression model.

	Model	Accuracy	AUC	Recall	Prec.	F1	Kappa	MCC	TT (Sec)
lr	Logistic Regression	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000	1.2450
knn	K Neighbors Classifier	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000	0.1420
nb	Naive Bayes	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000	0.1120
lda	Linear Discriminant Analysis	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000	0.0750
et	Extra Trees Classifier	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000	0.6280
lightgbm	Light Gradient Boosting Machine	0.9925	0.9976	0.9925	0.9933	0.9923	0.9906	0.9909	0.3620
rf	Random Forest Classifier	0.9899	1.0000	0.9899	0.9910	0.9899	0.9874	0.9877	0.4200
xgboost	Extreme Gradient Boosting	0.9874	0.9973	0.9874	0.9887	0.9873	0.9842	0.9846	0.3690
dt	Decision Tree Classifier	0.9798	0.9872	0.9798	0.9824	0.9794	0.9747	0.9756	0.2160
gbc	Gradient Boosting Classifier	0.9773	0.9906	0.9773	0.9806	0.9771	0.9716	0.9726	1.3500
ridge	Ridge Classifier	0.8939	0.0000	0.8939	0.9094	0.8934	0.8673	0.8715	0.0950
ada	Ada Boost Classifier	0.7674	0.9312	0.7674	0.6636	0.6973	0.7080	0.7522	0.2650
svm	SVM - Linear Kernel	0.6390	0.0000	0.6390	0.5348	0.5526	0.5487	0.6080	0.1050
dummy	Dummy Classifier	0.2071	0.5000	0.2071	0.0430	0.0712	0.0000	0.0000	0.0750
qda	Quadratic Discriminant Analysis	0.1944	0.0000	0.1944	0.0379	0.0634	0.0000	0.0000	0.0680

Figure 12: PC coefficient plot

The above picture is the output of the comparison of all the models and their performance accuracy. Logistic regression has the highest performing accuracy of 1 along with K-neighbor classifier and Naïve bayes.

```
LogisticRegression
LogisticRegression(C=1.0, class_weight=None, dual=False, fit_intercept=True,
                   intercept_scaling=1, l1_ratio=None, max_iter=1000,
                   multi_class='auto', n_jobs=None, penalty='l2',
                   random_state=123, solver='lbfgs', tol=0.0001, verbose=0,
                   warm_start=False)
```

Figure 13: Top 3 algorithms

Fold	Accuracy	AUC	Recall	Prec.	F1	Kappa	MCC
0	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000
1	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000
2	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000
3	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000
4	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000
5	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000
6	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000
7	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000
8	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000
9	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000
Mean	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000
Std	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000

Figure 14: Tune logistic regression

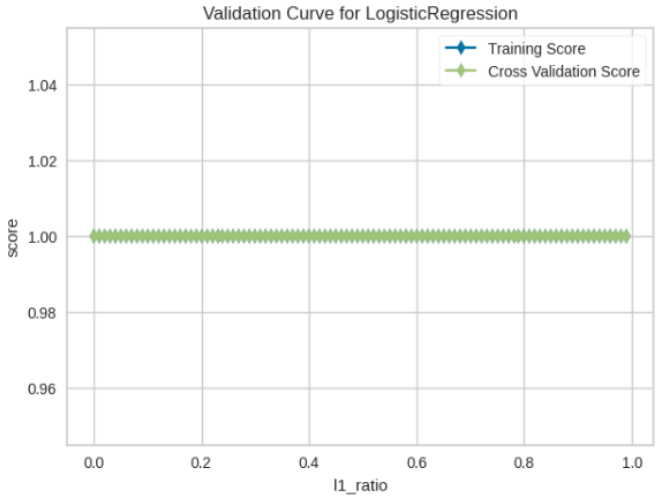

Figure 15: Logistic Regression validation curve

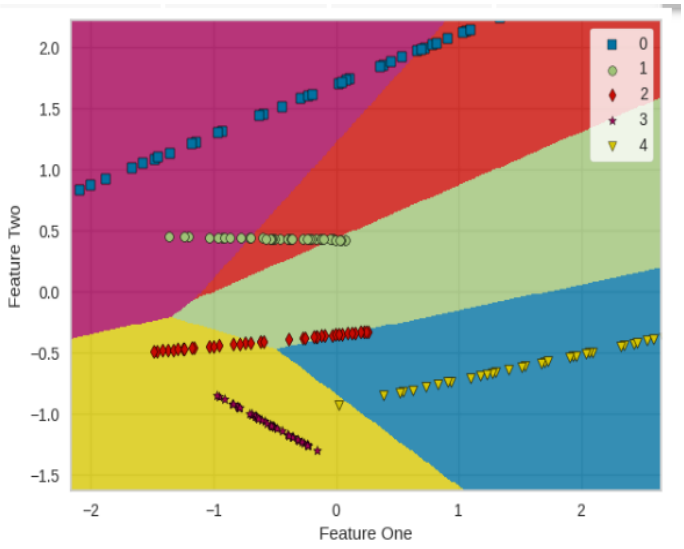

Figure 16: Decision curve

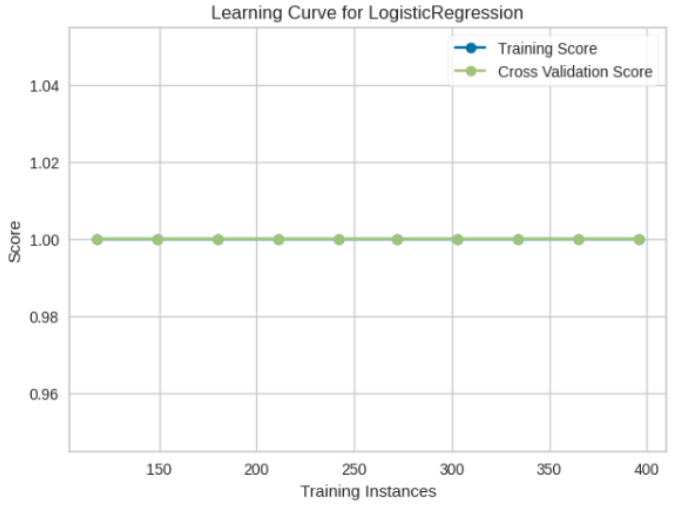

Figure 17: Logistic Regression leaning curve

The above curve describes the effectiveness of the machine learning model by charting the model's training and validation error as a function of training sample number.

We can assess whether a logistic regression model is under fitting, overfitting, or doing well by analyzing its learning curve and making suitable tweaks to optimize its performance.

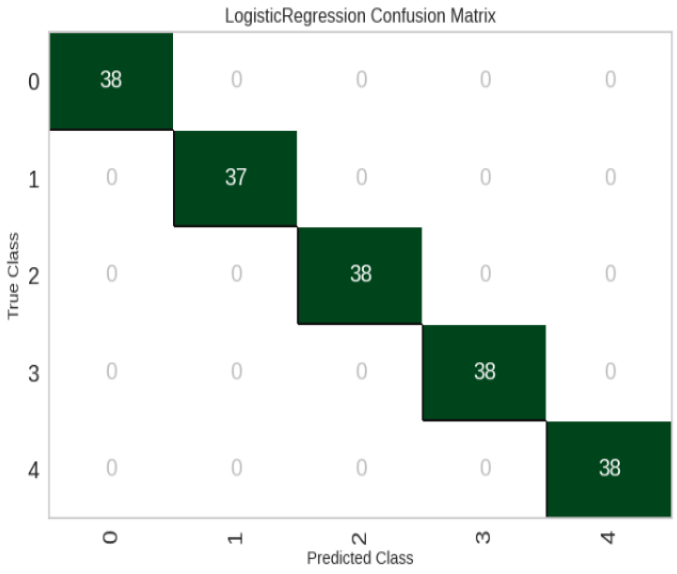

Figure 18: Confusion matrix of Logistic regression

A confusion matrix is a table used to assess the effectiveness of a classification model. In this matrix, the rows correspond to the actual class labels and the columns correspond to the predicted class labels. The number of right and unsuccessful predictions is summarized and separated by class. There are no inaccurate predictions in the dataset.

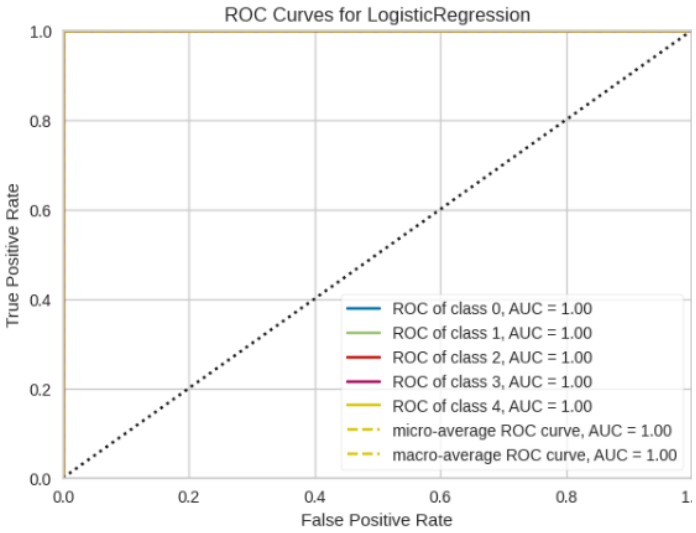

Figure 19: ROC curve for Logistic regression

In this graph, x-axis reflects the false positive rate while y-axis shows true positive rate. The curve displays the trade-off between false positive rate and true positive rate at different threshold values. The area under the ROC curve (AUC) is a popular metric for assessing the overall performance of the logistic regression model. AUC of 1 represents ideal performance. Hence, LR shows the best performance.

VII. CONCLUSION

In conclusion, Principal Component Analysis and classification algorithms have been applied on multiclass of stress dataset. After applying PCA, the first two principal components showed 99.1% of the variance. Only two major components were used. Logistic regression was found to be the highest performance model with accuracy of 1. It is observed that after tuning, performance metrics score of the algorithm has improved significantly. To summarize, the algorithm can successfully run on the multiclass of stress dataset.

REFERENCES

[1] L. Rachakonda, A. K. Bapatla, S. P. Mohanty, and E. Kougianos,"SaY-oPillow: Blockchain-Integrated Privacy-Assured IoMT Framework for Stress Management Considering Sleeping Habits", IEEE Transactions on Consumer Electronics (TCE), Vol. 67, No. 1,Feb 2021, pp. 20-29.

[2] L. Rachakonda, S. P. Mohanty, E. Kougianos, K. Karunakaran, and M. Ganapathiraju, "Smart-Pillow: An IoT based Device for Stress Detection Considering Sleeping Habits", in Proceedings of the 4th IEEE International Symposium on Smart Electronic Systems (iSES), 2018, pp. 161–166.

[3] C. Bishop. Patter Rocognition and Machine Learning 2007.

[4] S. F. Quan, B. V. Howard, C. Iber, J. P. Kiley, F. J. Nieto, G. T. O'Connor, D. M. Rapoport, S. Redline, J. Robbins, J. M. Samet, and P. W. Wahl, "The Sleep Heart Health Study: Design, Rationale, and Methods." Sleep., vol. 20, no. 12, pp. 1077–1085, Dec. 1997.

[5] G. Q. Zhang, L. Cui, R. Mueller, S. Tao, M. Kim, M. Rueschman, S. Mariani, D. Mobley, and S. Redline, "The National Sleep Research Resource: Towards a Sleep Data Commons." J Am Med Inform Assoc., vol. 25, no. (10, pp. 1351–1358., Oct. 2018.

[6] A. Ben Hamza, Advanced Statistical Approaches to Quality, unpublished