Project Sukoon: A Stress Predictor for Students

Apka Sukoon?

Group Leader: Muhammad Furqan Raza (460535)
Tamkeen Sara (474585)
Attiqa Bano (473781)
Rimsha Mahmood (455080)

CS-245 Machine Learning
Assignment 03: Proposed Approach & Results

1 Introduction

Mental health issues are increasingly prevalent, and early detection of stress is crucial in ensuring timely intervention. Our project, "Mental Health Stress Detector," uses machine learning techniques to classify individuals into stress categories based on questionnaire data. In Assignment 2, we implemented baseline models—Multiclass Logistic Regression and Random Forest. Although these models performed well, they exhibited limitations such as overfitting and sensitivity to high-dimensional data.

This report presents our enhanced approaches in Assignment 3, which incorporate regularization and dimensionality reduction techniques to improve model performance, generalization, and robustness.

2 Proposed Methods

To enhance the performance of our baseline models, we implemented four novel/improved methods:

2.1 L1 Regularization (Lasso)

L1 regularization adds a penalty equal to the absolute value of the magnitude of coefficients to the loss function. This helps in both regularization and feature selection. Unimportant feature weights are pushed to zero, reducing model complexity.

Motivation: Reduce overfitting and eliminate irrelevant features in high-dimensional space.

2.2 L2 Regularization (Ridge)

L2 regularization adds the squared magnitude of coefficients as a penalty to the loss function. It discourages large coefficients without necessarily eliminating them.

Motivation: Prevent overfitting and improve model stability, especially in presence of multicollinearity.

2.3 Principal Component Analysis (PCA)

PCA transforms the feature space into a set of orthogonal components that explain the most variance in the data. By reducing dimensionality, we aim to decrease noise and improve computational efficiency.

Motivation: Reduce feature space while preserving variance, mitigate noise, and enhance generalization.

2.4 PCA with L2 Regularization

We combined PCA with L2-regularized logistic regression to explore the synergy of dimensionality reduction and regularization.

Motivation: Combine benefits of lower-dimensional space with regularization to minimize overfitting while retaining most informative patterns.

3 Implementation

We used Python's scikit-learn library to implement all models. Below are key code snippets used in experimentation:

L1-Regularized Logistic Regression

```
from sklearn.linear_model import LogisticRegression
model = LogisticRegression(penalty='l1', solver='liblinear')
model.fit(X_train, y_train)
accuracy = model.score(X_test, y_test)
print("L1 Accuracy:", accuracy)
```

L2-Regularized Logistic Regression

```
model = LogisticRegression(penalty='l2', solver='liblinear')
model.fit(X_train, y_train)
accuracy = model.score(X_test, y_test)
print("L2 Accuracy:", accuracy)
```

PCA with Logistic Regression

```
from sklearn.decomposition import PCA
pca = PCA(n_components=10)
X_train_pca = pca.fit_transform(X_train)
X_test_pca = pca.transform(X_test)

model = LogisticRegression()
model.fit(X_train_pca, y_train)
accuracy = model.score(X_test_pca, y_test)
```

4 Results

Accuracy Comparison

Table 1: Model Performance Comparison

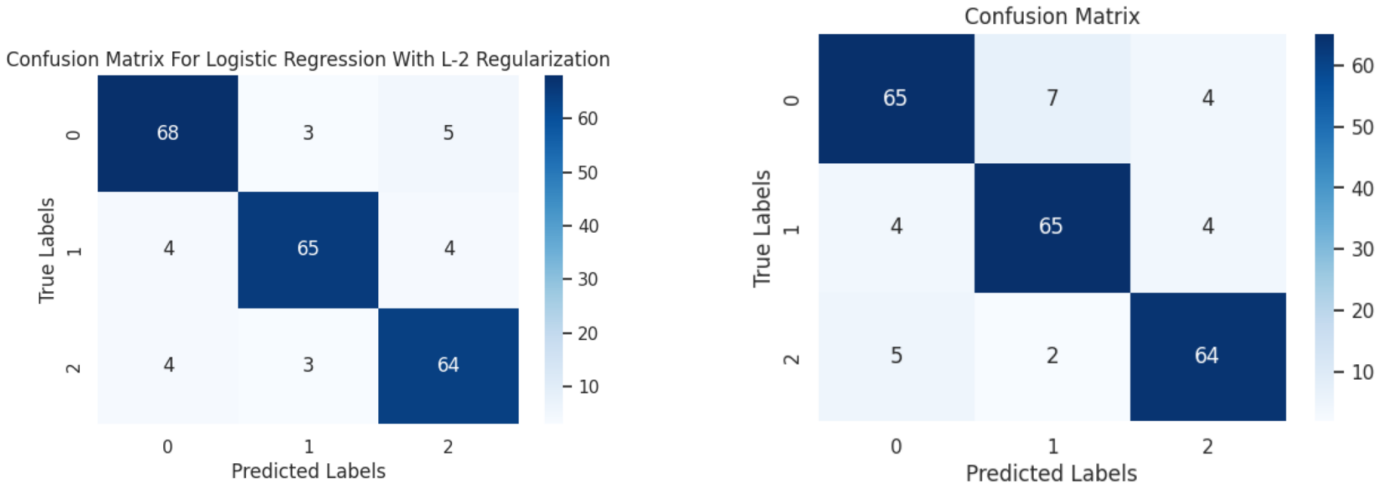
Model	Accuracy (%)
Multiclass Logistic Regression	86.81
Random Forest	88.79
L1-Regularized Logistic Regression	89.09
L2-Regularized Logistic Regression	**89.54**
PCA with Logistic Regression	88.18
PCA + L2-Regularized Logistic Regression	88.63

Observations

- L2 regularization offered the highest accuracy, likely due to better control over variance.
- L1 regularization gave a decent boost while also helping reduce the feature set.
- PCA slightly improved performance but alone wasn't enough to outperform regularization techniques.
- Combining PCA and L2 achieved a balance between dimensionality and generalization.

Output Snippets

The Confusion Matrices of Both Approaches are as follows:



(a) L-2 Regularization: Accuracy = 89.54%

(b) PCA & L-2 Regularization: Accuracy = 88.63%

Figure 1: Confusion Matrices for Logistic Regression With L-2 Regularization & PCA

5 Analysis

Compared to baseline models, all four improved approaches enhanced performance in varying degrees:

Why L2 Performed Best

L2 regularization avoids overfitting by penalizing large weights and is more robust in high-dimensional datasets with multicollinearity. It maintains all features but shrinks their influence, enabling the model to generalize better.

L1 vs L2

While both aim to prevent overfitting, L1 introduces sparsity which is beneficial when irrelevant or redundant features exist. L2, on the other hand, retains all features and stabilizes the model, making it slightly more accurate in our dataset.

Impact of PCA

PCA helped by reducing dimensionality and noise, but it may also lose interpretability and some information. This possibly explains the dip in performance in PCA-only models compared to L2.

Trade-offs in PCA+L2

While this combination was effective, dimensionality reduction via PCA may have suppressed some useful features which L2 could have leveraged better in its original form.

6 Conclusion

Through our enhanced methods, we demonstrated that careful application of regularization and dimensionality reduction techniques leads to better predictive performance in stress detection. Among all, L2 regularization proved the most effective, with an accuracy of 89.54%.

These improvements are practical, computationally efficient, and can be scaled to larger datasets. Future work may involve deeper models like Neural Networks or using feature selection with domain knowledge.

Notebook Links

- **L1 & L2 Implementation:** https://colab.research.google.com/drive/1bt2YmW3BG64smo0DcEKes-l1f9uB0va
usp=sharing
- **PCA + Logistic Regression Notebook:** https://colab.research.google.com/drive/1lXqxDozaXfpSN7KOPnUZ
usp=sharing