Assignment 2 - Cancer Type Prediction

Student Details:

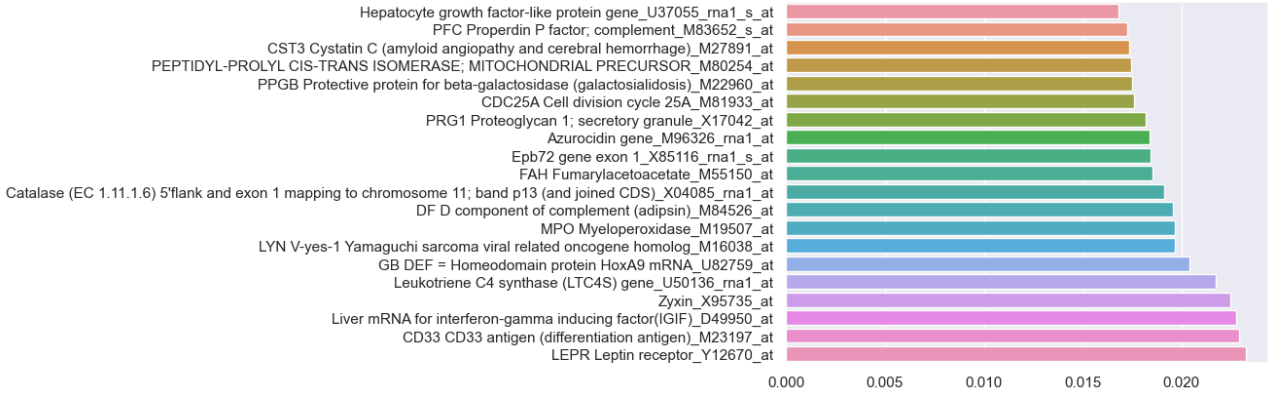
- **Name:** Yelisetty Karthikeya S M
- **Roll Number:** 21CS30060

Dataset:

The dataset focuses on classifying new cases of cancer based on gene expression monitoring via DNA microarray. It aids in identifying new cancer classes and assigning tumors to known classes, particularly patients with Acute myeloid leukemia (AML) and Acute lymphoblastic leukemia (ALL). There are two datasets: training (38 samples) and test (34 samples), containing measurements corresponding to ALL and AML samples.

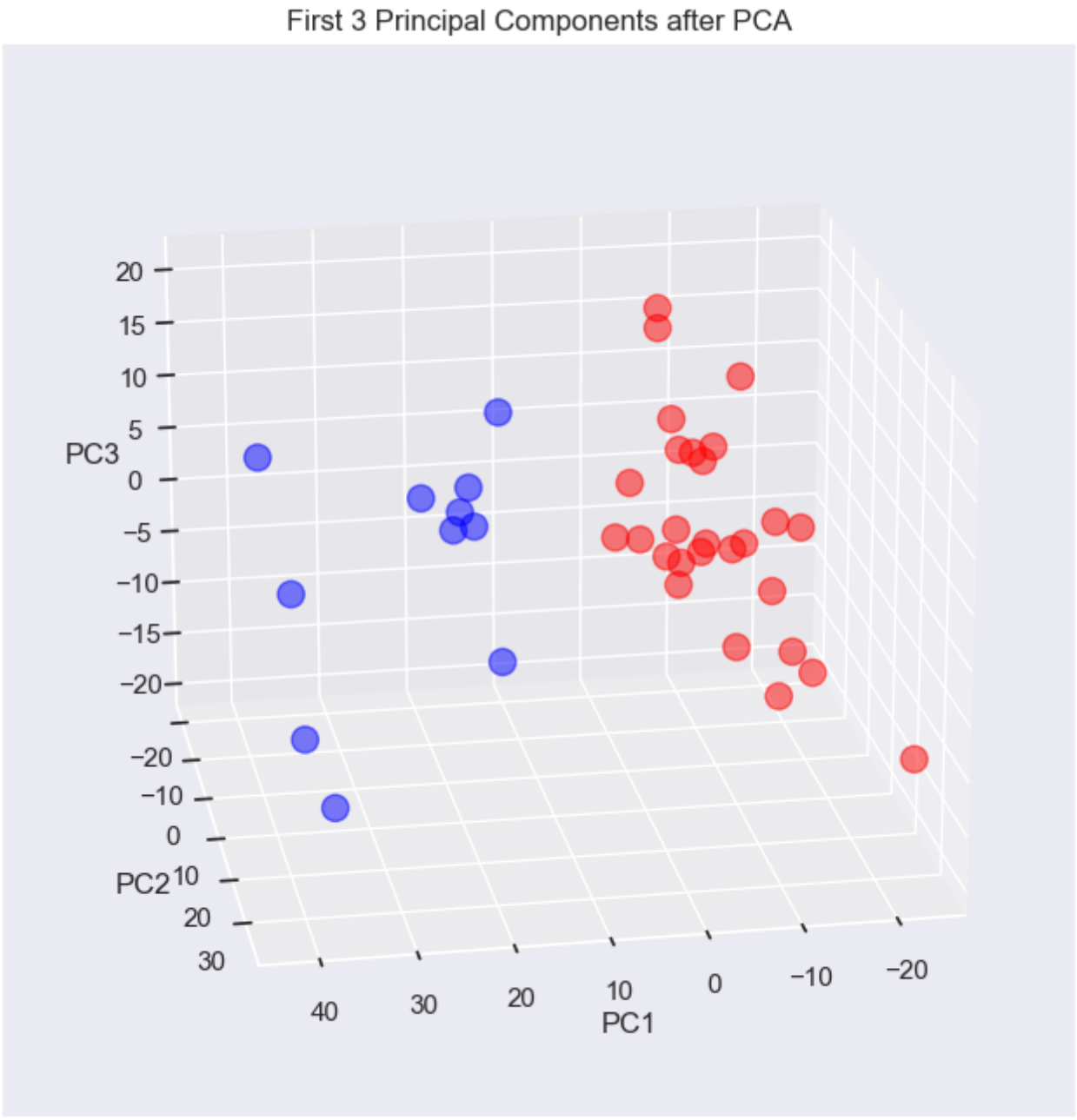
Data Pre-processing:

1. Combined both training and test datasets, removing duplicate columns, and scaling the data to the range [0, 1].
2. Utilized feature selection based on higher weights in a logistic regression model.



3. Performed Principal Component Analysis (PCA) to extract features, ultimately using 3 PCA components that showed a clear decision boundary

for classification.



Support Vector Machines (SVMs):

SVMs play a crucial role in cancer type prediction, classifying cancer patients into different types or subtypes based on various molecular and clinical features. Three kernels were used: Gaussian, Sigmoid, and Linear. The best SVM model was achieved with the following parameters:

- Parameters: {'C': 0.1, 'gamma': 1, 'kernel': 'linear'}
- Accuracy: 0.9411764705882353

Strengths:

- SVMs can handle high-dimensional data, making them suitable for tasks involving gene expression profiles, genetic mutations, and clinical features, common in cancer type prediction.
- SVMs provide relatively clear decision boundaries, making it easier to interpret the model's predictions, particularly for linear kernels.
- SVMs are less prone to overfitting, especially when using the appropriate regularization parameter (C), which is crucial when dealing with small or noisy datasets common in cancer research.

Weaknesses:

- SVMs can become computationally expensive and slow when dealing with large datasets or when using complex kernel functions.
- SVMs can struggle when faced with noisy data or when classes are heavily overlapping, as they strive to maximize the margin between classes.

Random Forest:

Random Forest is a versatile ensemble learning method often used in cancer type prediction. The best Random Forest model was achieved with the following parameters:

- Parameters: {'max_features': 0.6, 'min_samples_leaf': 8, 'min_samples_split': 3, 'n_estimators': 100}
- Accuracy: 0.9117647058823529

Neural Networks:

Neural networks can automatically learn relevant features from raw data, which is particularly valuable in cancer type prediction. The best Neural Network model was achieved with the following parameters:

- Parameters: {'activation': 'tanh', 'alpha': 0.0001, 'hidden_layer_sizes': (50, 50, 50), 'learning_rate': 'constant', 'solver': 'sgd'}

- Accuracy: 0.8823529411764706

Strengths:

- Neural Networks are capable of capturing complex, non-linear relationships within the data, which can be valuable when dealing with multifaceted factors influencing cancer types.

Weaknesses:

- Deep neural networks are susceptible to overfitting, especially when dealing with small or imbalanced datasets. Regularization techniques like dropout and early stopping are often needed to mitigate this.
- Training deep neural networks, particularly with large architectures, can be computationally expensive and require substantial computational resources.

Performance of the Three Models:

From both ROC curve and Precision-Recall curve, we can clearly see that SVMs with sigmoid and linear kernels are the best performing models for this task.

Model Selection:

Failing to select an appropriate model or neglecting parameter tuning can lead to underperforming models that are inaccurate and unreliable. This can be costly, especially in applications where the consequences of errors are significant, such as in medical diagnosis or autonomous vehicles. So, we need to do careful parameter tuning and select the most appropriate model.