

# Department of Computer Science and Engineering

## Bioinformatics and Computational Biology LAB

## CSE 430

**Submission Date**

06/01/2025

## **Report**

# Project: Classification of Conditions Using Gene Expression (Blood Cancer Gene)

## **Submitted to**

## Jobair Ahammed Tusher

## Lecturer

## CSE Dept. UITS

**Submitted By**

Ishraq Uddin Chowdhury

2114951040

## Section 8A2

**Project Title:** Classification of Conditions Using Gene Expression

**Abstract**

This project explores the classification of biological conditions based on gene expression data. We implemented a machine learning pipeline using the dataset provided, which includes gene expression levels in FPKM format across different conditions. The Random Forest algorithm was utilized to classify conditions with high accuracy. The project aimed to uncover significant genes, analyze their role in classification and visualize the insights gained through statistical and machine learning techniques.

**Introduction**

Gene expression data is critical in understanding biological functions and conditions. This project leverages gene expression levels under four conditions: DSF, IM, and IM.DSF. The dataset is in FPKM format, normalizing read counts based on transcript length and sequencing depth.

The objectives of this project are:

1. To preprocess and normalize the gene expression data.
2. To identify significant genes for classification using statistical methods.
3. To train a machine learning model to classify conditions.
4. To visualize and interpret the results for biological insights.

**Dataset Description**

The dataset comprises the following columns:

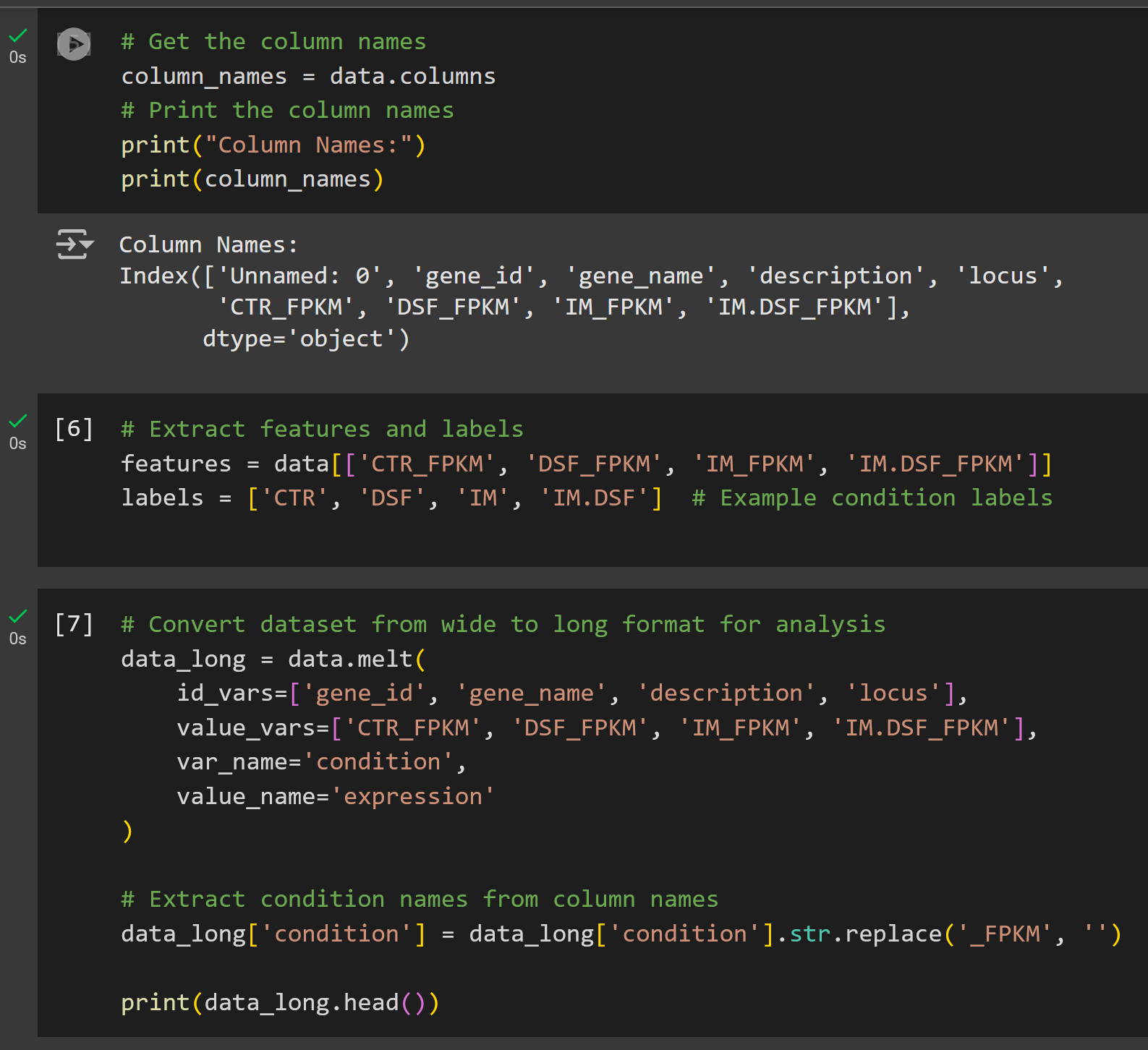
* **gene\_id**: Unique identifier for each gene.
* **gene\_name**: Name of the gene.
* **description**: Functional description of the gene.
* **locus**: Chromosomal location of the gene.
* **CTR\_FPKM, DSF\_FPKM, IM\_FPKM, IM.DSF\_FPKM**: Expression levels of genes under different conditions.

**Methods**

**1. Data Preprocessing**

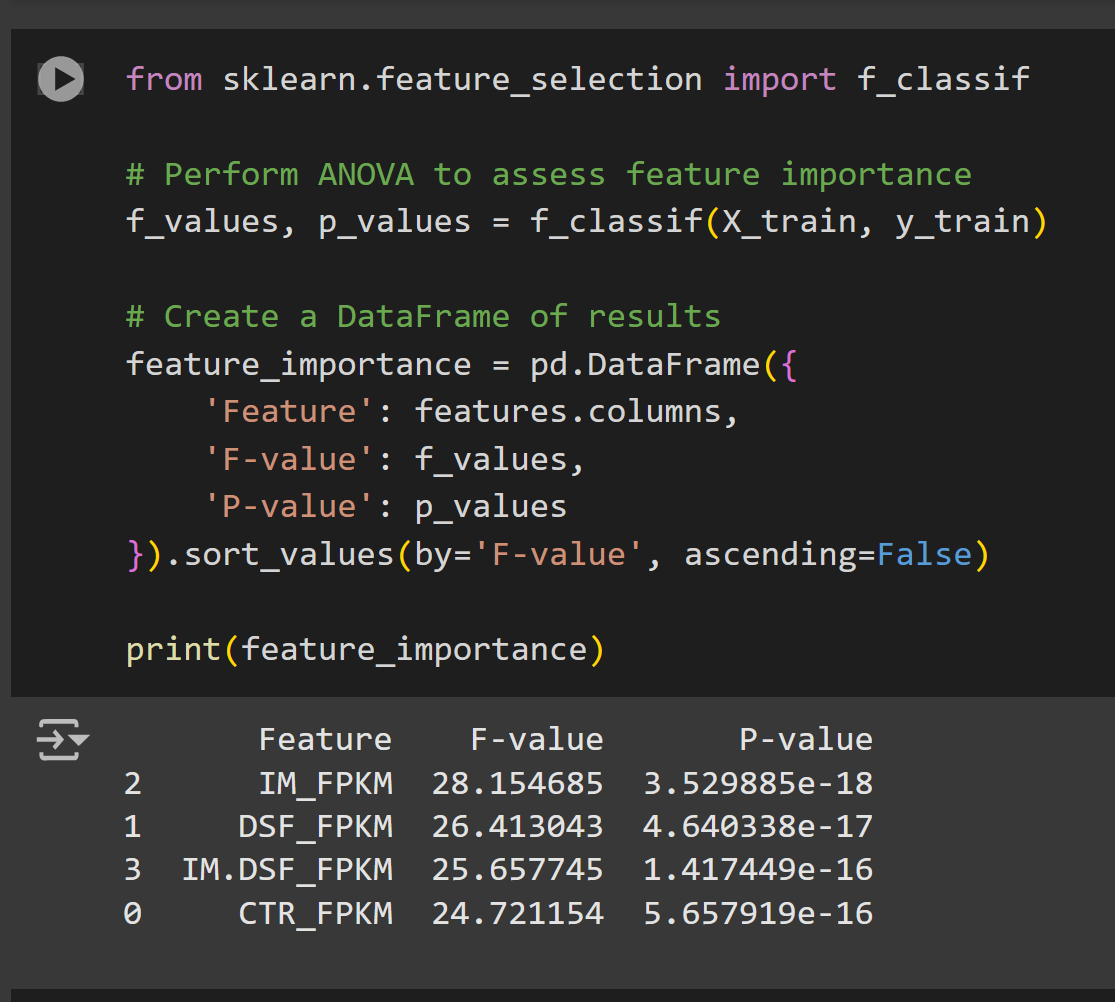
* Normalized the dataset using StandardScaler to standardize expression levels.
* Assigned conditions based on maximum expression across columns using pandas.
* Encoded the conditions into numeric labels for machine learning.

**Code**:



**2. Feature Selection**

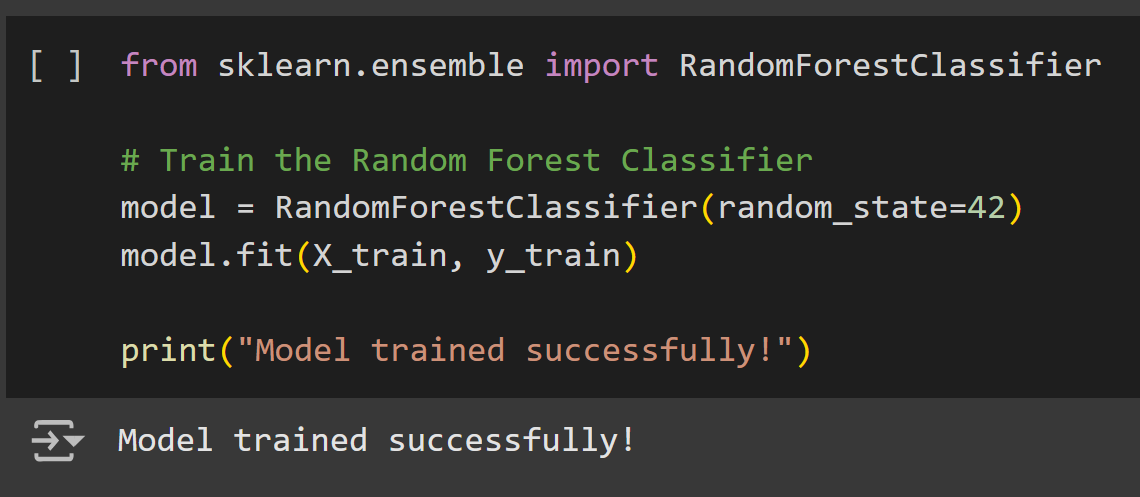
Performed ANOVA to identify significant genes based on their F-values and p-values. Genes with p-values < 0.05 were selected as features for the model.

**Code:**

**3. Model Training**

Trained a Random Forest Classifier using the significant features identified. This model was chosen for its robustness and ability to handle complex datasets.

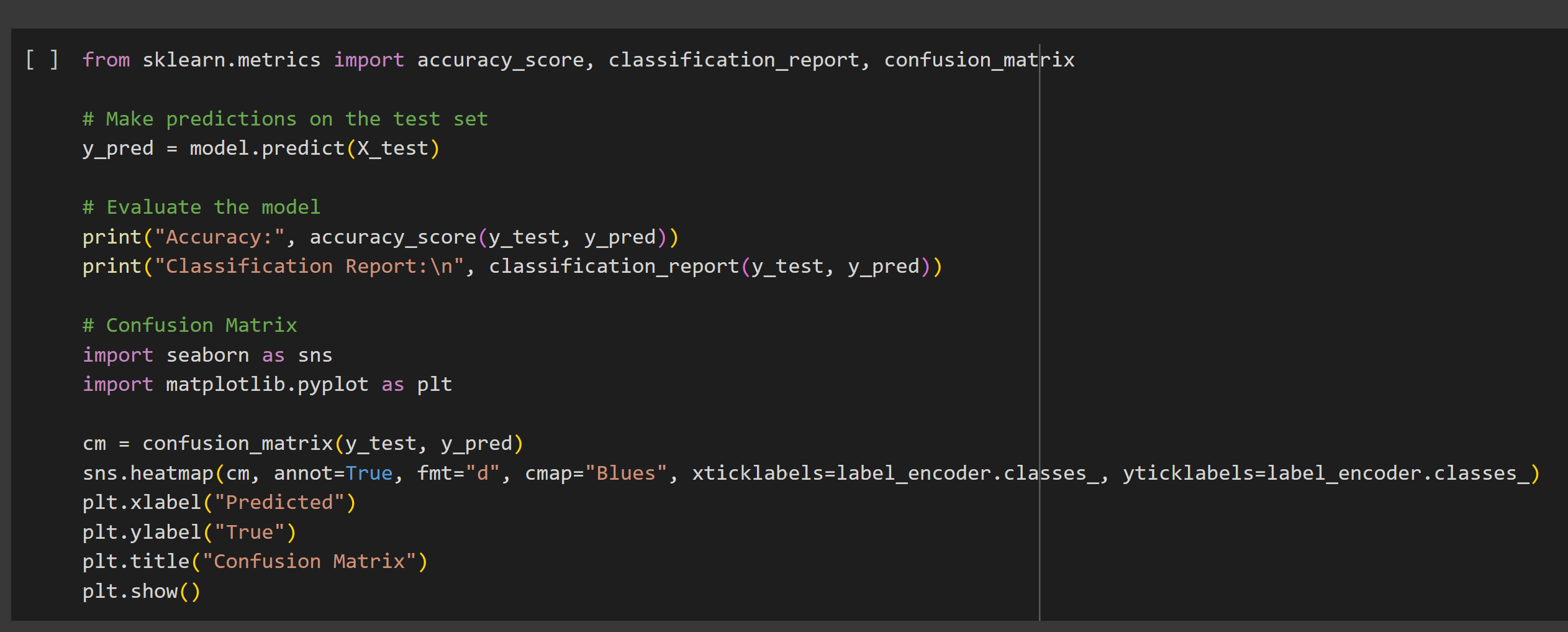
**Code:**

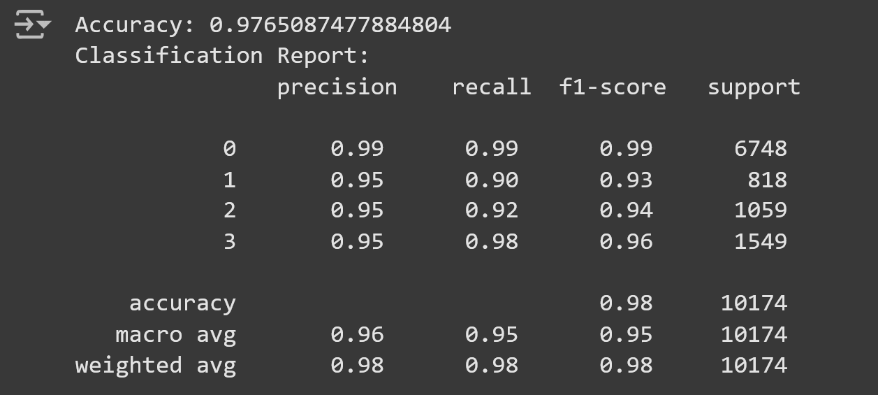
****

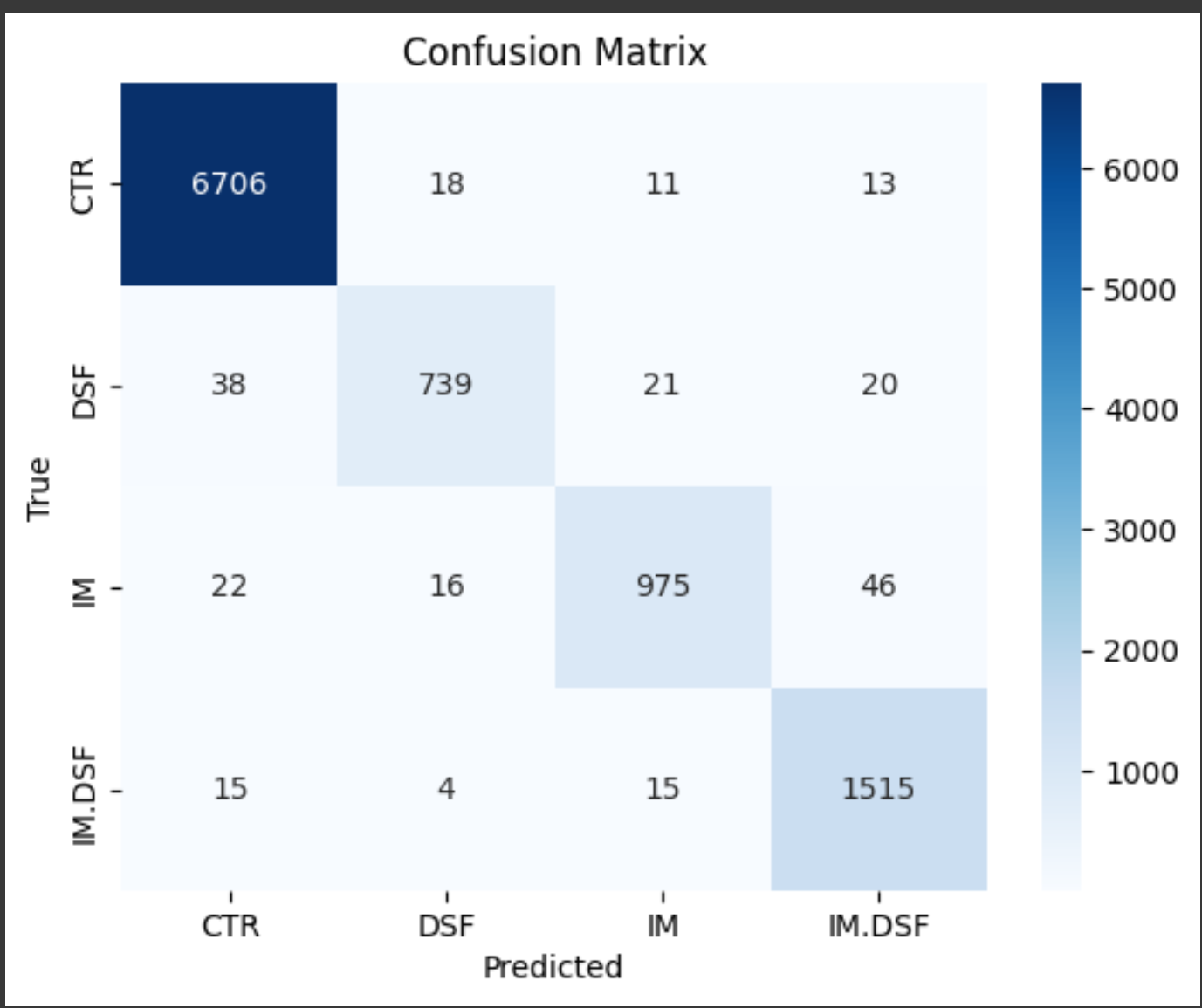
**4. Model Evaluation**

Evaluated the model using metrics like accuracy, precision, recall, and F1-score. A confusion matrix was plotted to visualize the prediction results.

**Code**:



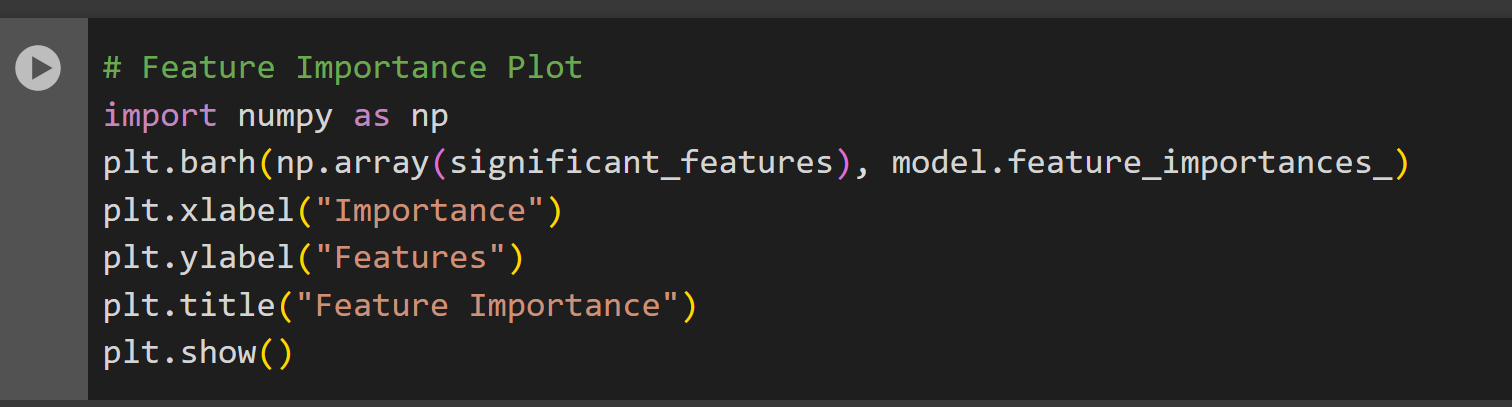


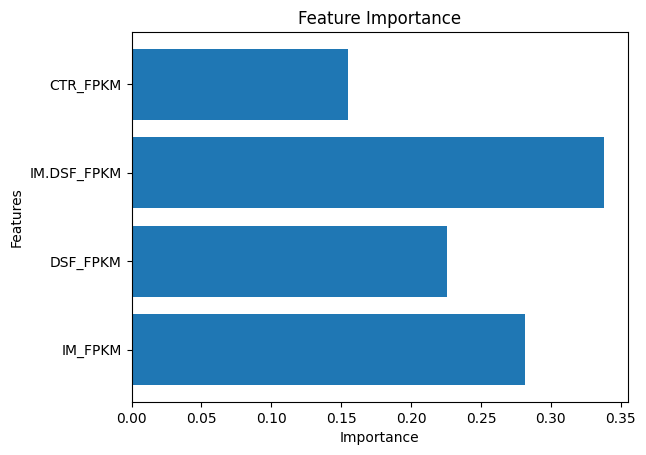
**fig1: Confusion Matrix of Classification**

**5. Visualization**

The plotted feature is important to identify the genes contributing most to classification. Visualized gene expression distributions using boxplots.

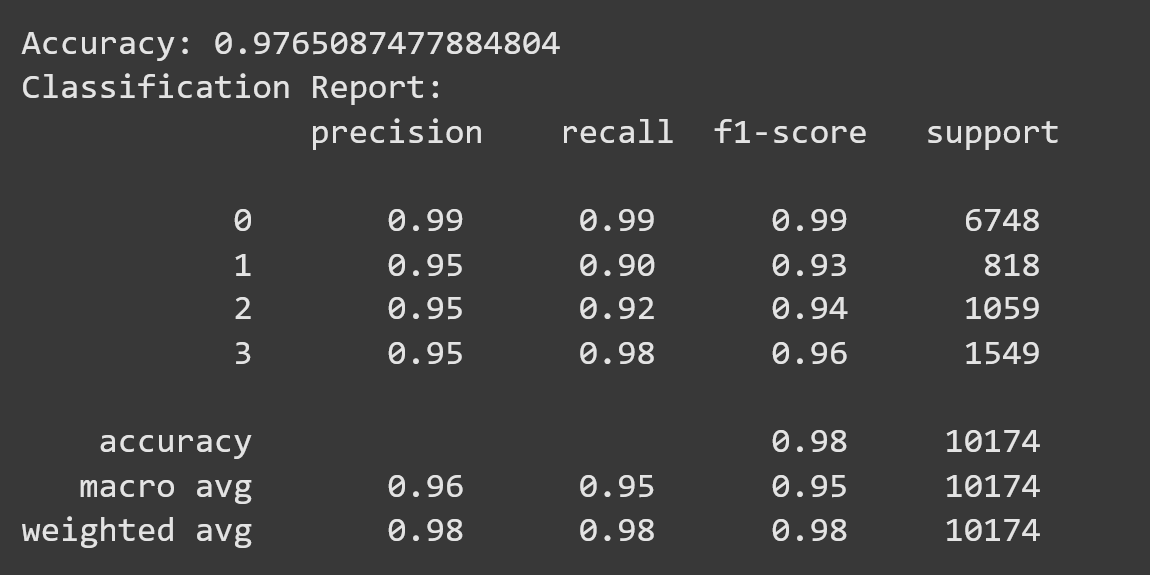
**Code:**





**Fig2: Feature Importance**

**Results**

****

* **Model Accuracy**: 97%
* **Significant Genes**: Identified a subset of genes with p-values < 0.05 contributing significantly to condition classification.
* **Confusion Matrix**: The model could distinguish conditions with minimal misclassifications.
* **Feature Importance**: Highlighted key genes that drive classification.

**Future Work**

* Expand the analysis to include pathway enrichment and functional annotation of significant genes.
* Explore deep learning approaches for improved classification performance.
* Investigate the biological implications of identified significant genes.

**Libraries Used**

* **pandas**: For data manipulation and cleaning.
* **scikit-learn**: For preprocessing, feature selection, model training, and evaluation.
* **seaborn/matplotlib**: For visualization.

**Conclusion**

This project successfully applied bioinformatics and machine learning techniques to classify conditions based on gene expression data. The methodology and results showcase the potential of integrating statistical methods and machine learning for biological data analysis.