CSE 4754 Bioinformatics

Tumor Sample Classification

TCGA: UCEC, KICH, LIHC, ESCA

Team Members

Tasnimul Hasnat
190041113

Nur Nasrum
1900411106

Khaja Abdus Sami
190041136

Introduction

Objective

Classification of UCEC, KICH, LIHC, ESCA tumor samples by tumor type using Logistic Regression, Decision Tree, KNN, Random Forest, Gradient Boosting and Support Vector Machine.

Dataset Size

Number of Samples: **1297**

Number of Genes: **60660**

Class-wise Samples (Total Count)

UCEC	588
KICH	91
LIHC	421
ESCA	197
Total	1297

Classification and Validation Process

Logistic Regression: Classification Outcomes

Confusion Matrix

$$\begin{bmatrix} 24 & 0 & 0 & 0 & 1 \\ 0 & 111 & 0 & 0 & 0 \\ 0 & 0 & 13 & 0 & 0 \\ 3 & 0 & 1 & 70 & 0 \\ 0 & 0 & 0 & 0 & 37 \end{bmatrix}$$

Logistic Regression: Statistical Validation

Accuracy: 98.08%
Macro Precision: 0.9582
Weighted Precision: 0.9820
Macro Recall: 0.9812
Weighted Recall: 0.9808
Macro F1-Score: 0.9690
Weighted F1-Score: 0.9809

Decision Tree: Classification Outcomes

Confusion Matrix

```
[ 20    1    0    3    1]
[  2  107    0    0    2]
[  0    0   13    0    0]
[  1    0    1   72    0]
[  1    0    0    0   36]
```

Decision Tree: Statistical Validation

Accuracy: **95.38%**
Macro Precision: **0.9271**
Weighted Precision: **0.9541**
Macro Recall: **0.9420**
Weighted Recall: **0.9538**
Macro F1-Score: **0.9341**
Weighted F1-Score: **0.9537**

Random Forest: Classification Outcomes

Confusion Matrix

```
[ 23    0    0    0    2]
[  0  111    0    0    0]
[  0    0   13    0    0]
[  2    0    1   71    0]
[  0    0    0    0   37]
```

Random Forest: Statistical Validation

Accuracy: **98.08%**
Macro Precision: **0.9595**
Weighted Precision: **0.9814**
Macro Recall: **0.9759**
Weighted Recall: **0.9808**
Macro F1-Score: **0.9672**
Weighted F1-Score: **0.9808**

KNN: Classification Outcomes

Confusion Matrix

$$\begin{bmatrix} 24 & 0 & 0 & 0 & 1 \\ 0 & 110 & 0 & 1 & 0 \\ 0 & 0 & 13 & 0 & 0 \\ 8 & 0 & 1 & 65 & 0 \\ 0 & 0 & 0 & 0 & 37 \end{bmatrix}$$

KNN: Statistical Validation

Accuracy: **95.77%**
Macro Precision: **0.9274**
Weighted Precision: **0.9643**
Macro Recall: **0.9659**
Weighted Recall: **0.9577**
Macro F1-Score: **0.9432**
Weighted F1-Score: **0.9588**

Gradient Boosting: Classification Outcomes

Confusion Matrix

```
[ 23    0    0    0    2]
[  0  111    0    0    0]
[  0    0   13    0    0]
[  1    0    1   72    0]
[  0    0    0    0   37]
```

Gradient Boosting: Statistical Validation

Accuracy: **98.46%**
Macro Precision: **0.961**
Weighted Precision: **0.9851**
Macro Recall: **0.9786**
Weighted Recall: **0.9846**
Macro F1-Score: **0.9729**
Weighted F1-Score: **0.9846**

Support Vector Machine: Classification Outcomes

Confusion Matrix

```
[ 20    1    0    3    1]
[  2  107    0    0    2]
[  0    0   13    0    0]
[  1    0    1   72    0]
[  1    0    0    0   36]
```

Support Vector Machine: Statistical Validation

Accuracy: **98.08%**
Macro Precision: **0.9582**
Weighted Precision: **0.9820**
Macro Recall: **0.9812**
Weighted Recall: **0.9808**
Macro F1-Score: **0.9690**
Weighted F1-Score: **0.9809**

In conclusion, Gradient Boosting gave the best accuracy of **98.46%**

Thank You!