**Assignment 2-21BT10004**

**Aishwary Yadav**

**Application of Machine Learning in Biological Systems**

**Problem Statement:**

**Objective**: This assignment focuses on understanding and comparing the performance of different machine learning models for predicting cancer types based on a given dataset. You will analyse Support Vector Machines (SVM), Random forest (RF), neural network (NN) regression and other relevant techniques.

**Dataset Description and Understanding:**

In the given problem, we have in total 72 patients and their respective gene expression values for different genes, and also we have the cancer status of each patient.

Out of the 72 patients, 38 are columns in training data and 34 are columns in testing data. We don't need to use '**call**' columns as they are not required in the analysis.

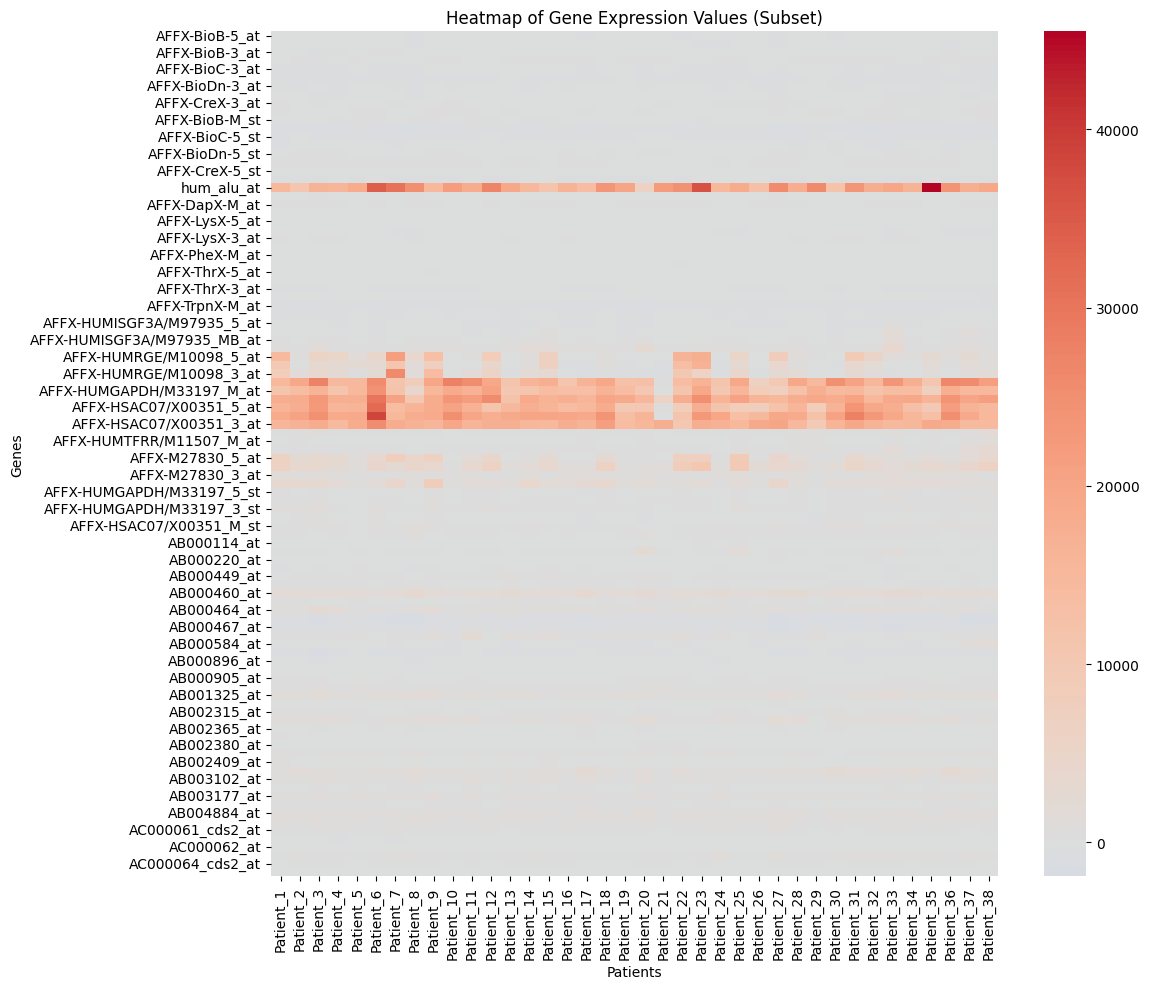
Our Target vector is the '**cancer**' column in the copy\_of\_actual dataset.

***Analysis of Gene Expression Data for 72 Patients for Predicting Cancer types***

1. **Heatmap of Gene Expression Values (Subset)**

**Findings:**

The heatmap provides a snapshot of gene expression patterns across patients.Blue regions indicate lower gene expression, while red regions indicate higher expression.



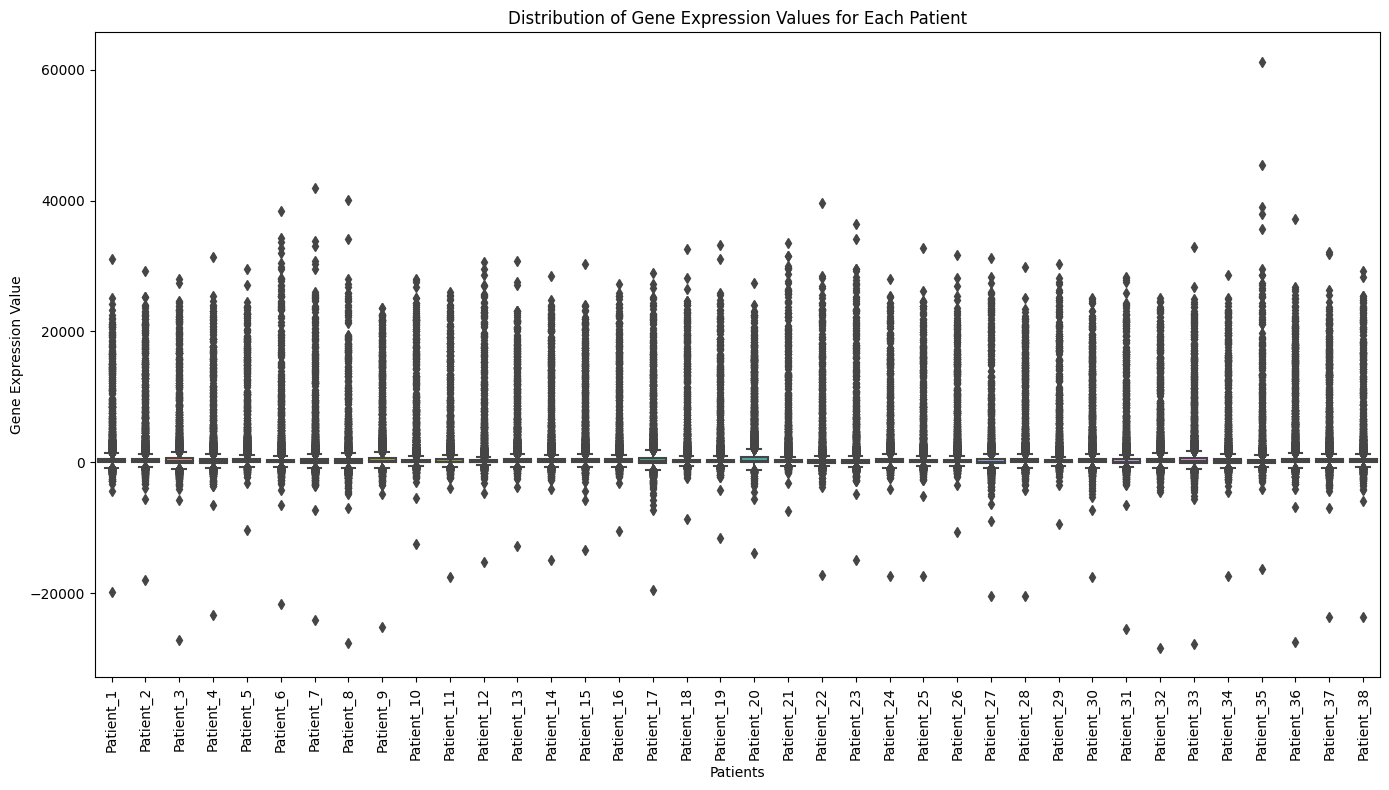
2. **Distribution of Gene Expression Values for Each Patient**

**Findings:**

The boxplot showcases the distribution of gene expression values for each patient.

Some patients, such as Patient\_8 and Patient\_33, have lower median expression values.

Variability in gene expression is observed across different patients, with some exhibiting more outliers than others.



**3.** **PCA of Gene Expression Data**

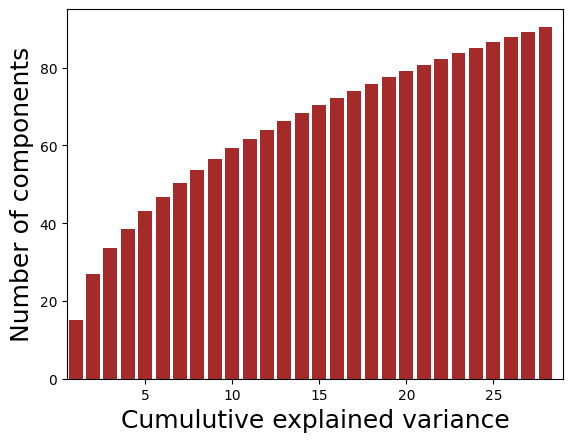
**Findings:**

The PCA scatterplot visualizes the variation between patients based on their gene expression values.

Some patients cluster together, suggesting similar gene expression patterns.

The spread of patients in the PCA space indicates diverse expression profiles.

28 features explain around 90% of the variance.



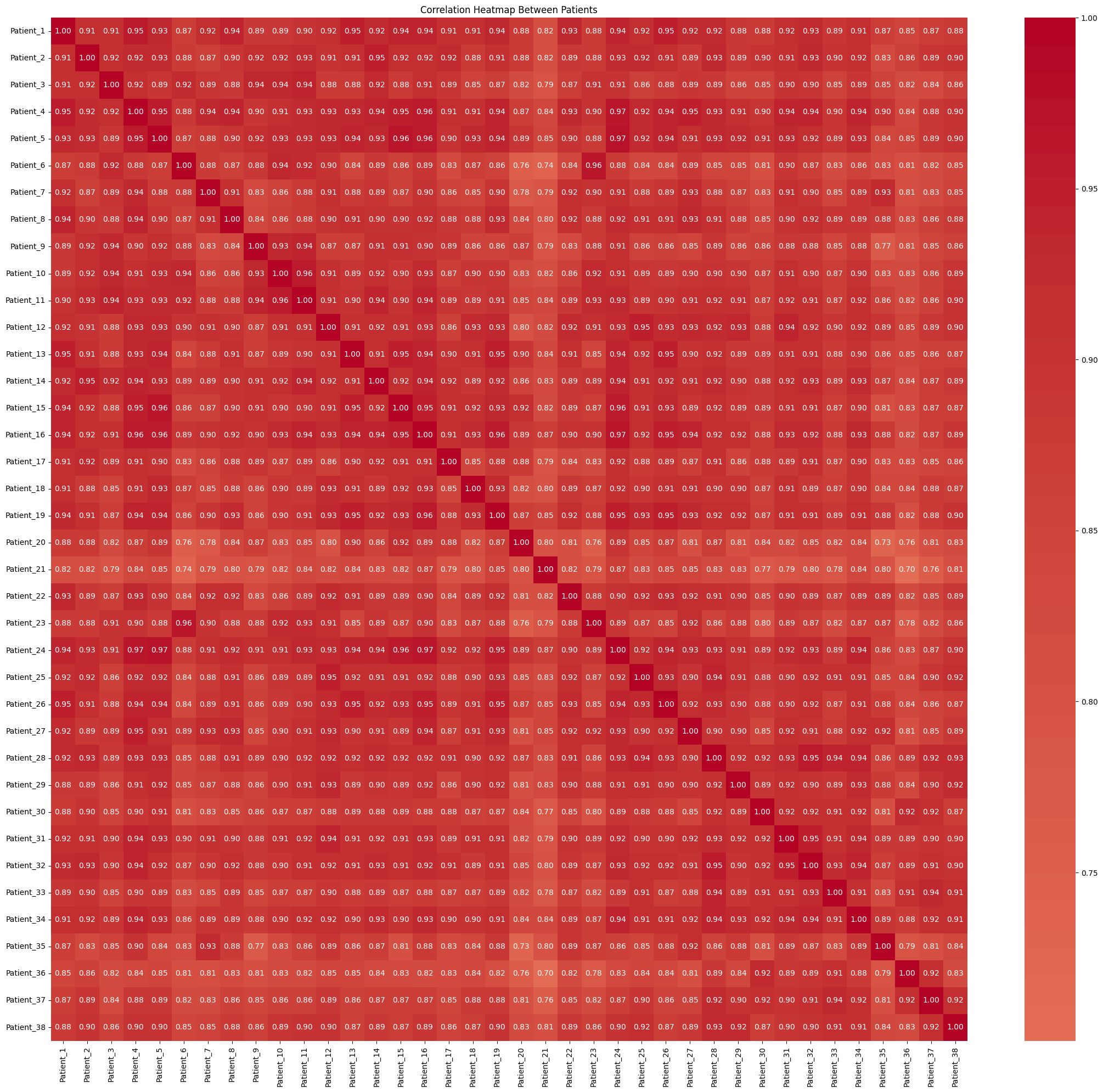
**4.Correlation Heatmap Between Patients**

**Findings:**

The heatmap visualizes pairwise correlations between patients based on their gene expression profiles.

Most patients have moderate to high positive correlations, indicating similar expression patterns.

Some pairs of patients exhibit lower correlations, suggesting distinct expression profiles.



Conclusion:

These visualizations provides a good overview about the given dataset and helpful for insights.

**SVM analysis:**

**Support Vector Machines** are widely used in cancer type prediction due to their ability to handle high-dimensional genomic data effectively. In the context of cancer research:

**Feature Selection**: Genomic data often involves thousands of genes or features. SVMs can identify the most relevant genes (features) for cancer prediction, aiding in feature selection and reducing the dimensionality of the data.

**Kernel Trick**: SVMs use kernel functions like radial basis function (RBF) to transform the data into higher-dimensional spaces, enabling the model to capture complex relationships among genes. This is crucial in understanding the intricate genetic patterns associated with different cancer types.

**Binary Classification**: SVMs are inherently binary classifiers. In cancer research, where distinguishing between cancer and non-cancer samples is vital, SVMs can effectively classify samples into two classes, making them valuable tools for tumor detection.

SVMs use kernels to transform data into ***higher-dimensional*** spaces. Different kernels serve various purposes:

**Linear Kernel**: Suitable for linearly separable data. It works well when the relationship between features and classes is approximately linear.

**Radial Basis Function (RBF) Kernel**: Also known as Gaussian kernel, it is effective for capturing complex, non-linear relationships. RBF kernel is commonly used in genomics due to its ability to handle intricate patterns in high-dimensional data.

**Polynomial Kernel**: Useful for capturing polynomial relationships in the data. It can be effective when the decision boundary is curved.

**Sigmoid Kernel**: Sigmoid kernel can be used when the data has a sigmoid (S-shaped) relationship. However, it's used less frequently compared to linear, RBF, and polynomial kernels.

Choosing the appropriate kernel is crucial. We need to experiment with different kernels to find the one that best fits the underlying patterns in the genomic data, leading to more accurate cancer type predictions.

***Results of SVM:***

On applying SVM on the dataset , we found out that linear kernel gives the best result and higher degree polynomial are performing well on the dataset as compared to rbf kernel.

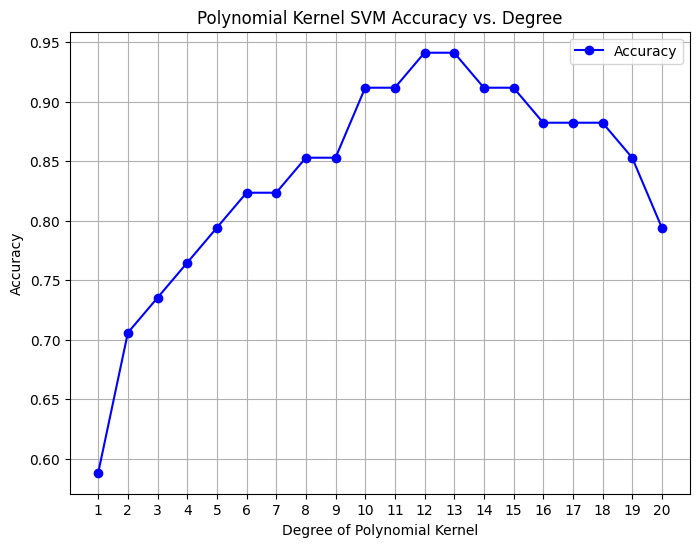
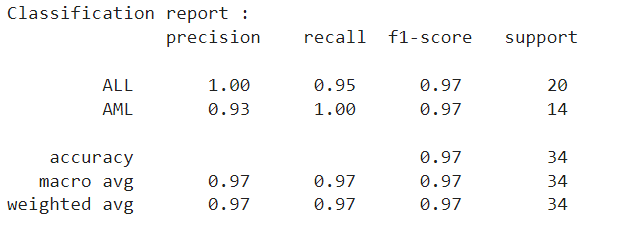
Results of these kernels are:

Linear Kernel SVM Accuracy: 0.9705882352941176

RBF Kernel SVM Accuracy: 0.6176470588235294

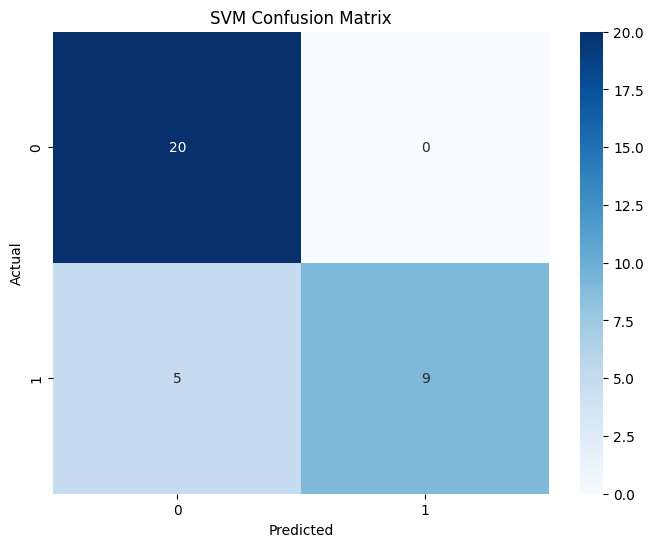
Polynomial Kernel SVM Accuracy: 0.9411764705882353

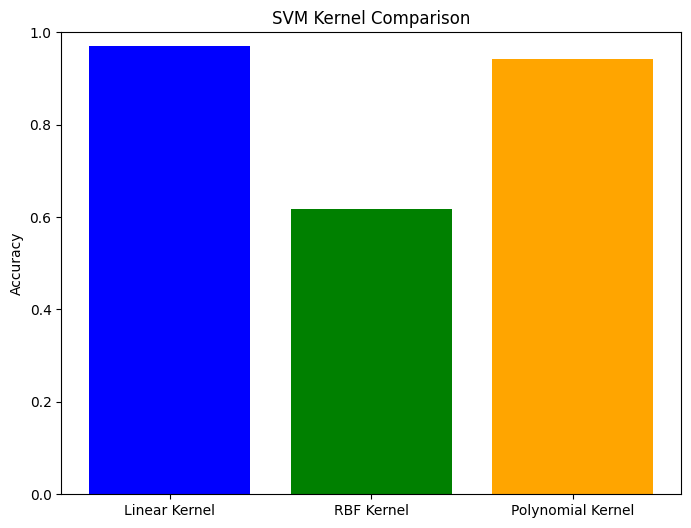
**Visualizations**



The Classification report of the best performing Linear SVM is given , from which we got the best F1-score of 0.97 , and also we plotted a comparative plot of degree of polynomial with accuracy and found out that polynomial with degree 12-13 is giving the accuracy .

Below is the confusion Matrix of the SVM.





**Random Forest Analysis:**

**Random Forest's Role in Cancer Type Prediction:**

Random Forests play a vital role in cancer type prediction due to their ability to analyze complex and high-dimensional genomic data. In the context of cancer research:

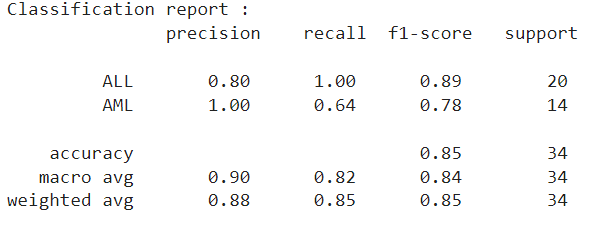
**Handling High-Dimensional Data**: Genomic data often consists of thousands of genes, which Random Forests handle well, identifying patterns and relationships that contribute to cancer types.

**Ensemble Learning**: By combining multiple decision trees, Random Forests improve accuracy and generalize better to unseen data, crucial when dealing with the diverse and evolving nature of cancer datasets.

**Feature Importance**: Random Forests rank genes based on their importance in predicting cancer types. Researchers can focus on these genes, potentially discovering novel biomarkers and therapeutic targets.

**Handling Noisy Data**: Genomic data can be noisy and contain outliers. Random Forests are robust to such noise, providing reliable predictions even with imperfect data.

**Results of Random Forest:**

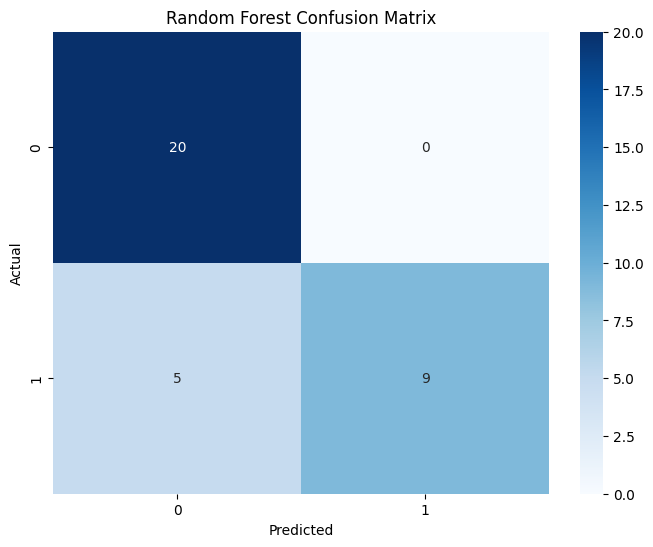
****

The classification report provided shows the performance of a binary classification model on two classes: ALL and AML. The model has an overall accuracy of 85%, meaning that it correctly predicted the class of 85% of the instances.

The model is performing well on both classes, with a precision of 1.0 for AML and 0.8 for ALL. This means that the model is correctly identifying almost all of the AML instances and a high proportion of the ALL instances.

However, the model has a higher recall for ALL (1.0) than for AML (0.64). This means that the model is missing some of the AML instances. This is a potential concern, as it is important to identify all of the AML cases in order to provide appropriate treatment.

The F1 score, which is a harmonic mean of precision and recall, is 0.89 for ALL and 0.78 for AML. This suggests that the model is performing better on ALL than on AML.



***Neural Network Regression:***

Neural networks, a subset of machine learning algorithms, have revolutionized various domains of research, including medical diagnosis and prognosis. Their application to cancer type prediction holds significant importance:

**High-dimensional Data Handling**: Cancer genomics and proteomics generate high-dimensional datasets. Neural networks can handle this high-dimensionality, extracting complex patterns that other algorithms might miss.

**Accuracy:** With the right architecture and tuning, neural networks can achieve high prediction accuracy. This is critical in medical applications, where the cost of misclassification can be extremely high.

**Feature Learning**: Traditional algorithms often require manual feature extraction. Deep neural networks, particularly convolutional neural networks (CNNs) for image data and recurrent neural networks (RNNs) for sequence data, can automatically learn relevant features from raw data.

**Integration of Multiple Data Types**: Neural networks can integrate diverse data types, such as genomic, proteomic, and imaging data, to make a unified prediction. This holistic approach can improve the accuracy of predictions.

**Personalized Medicine**: With accurate cancer type prediction, personalized treatment plans can be formulated for individual patients, improving outcomes and reducing side effects.

**Rapid Predictions:** Once trained, neural networks can make predictions in real-time, aiding in quicker clinical decisions.

**Grid Search for Neural Network Parameters**

Training a neural network involves determining the best values for several parameters. The selection of the right parameters, or hyperparameters, can significantly affect the model's performance. Grid search is a traditional method used to search for the optimal hyperparameters.

**Working:**

**Define the Grid**: List possible values for each hyperparameter you want to tune. For example, if you're tuning the learning rate, you might define a grid like

**Cartesian Product**: The grid search algorithm will consider every possible combination of the hyperparameters. So, if you have three values for the learning rate and two for batch size, there will be

**Train & Evaluate:** For each combination of hyperparameters, train the neural network and evaluate its performance on a validation set.

**Select Best Combination**: After evaluating all combinations, select the hyperparameters that gave the best performance on the validation set.

**Test:** Finally, test the performance of the model with the best hyperparameters on a separate test set to evaluate its generalization capability.

**Challenges & Considerations:**

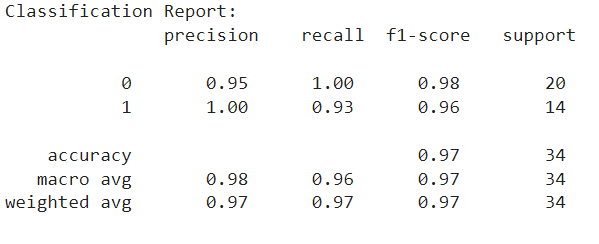
**Computational Cost**: Grid search can be computationally expensive, especially when there are many hyperparameters with a broad range of values. It requires training and evaluating a model for each combination.

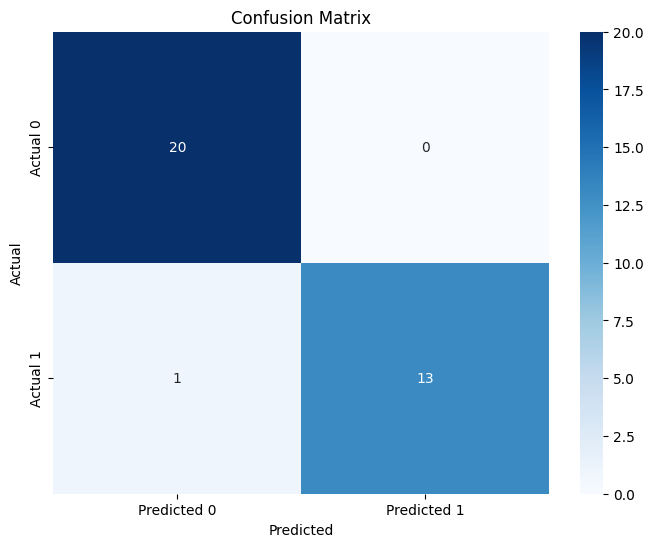
**Curse of Dimensionality**: As the number of hyperparameters grows, the number of combinations grows exponentially, leading to longer search times.

**Local vs. Global Optima:** There's no guarantee that the grid search will find the global optimum. It may find a local optimum depending on the granularity of the grid and the nature of the hyperparameter space.

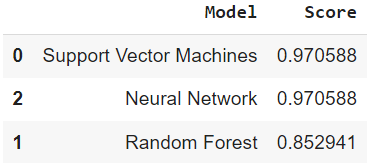
To overcome some of these challenges, researchers often use other methods like random search, Bayesian optimization, or evolutionary algorithms in conjunction with or in place of grid search.

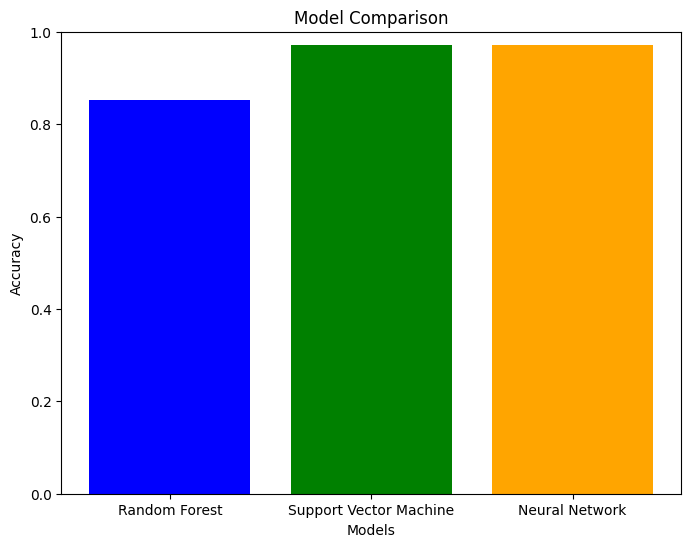
**Results of Neural Networks Classification:**





**Comparison of three models:**

****



In the bar chart shows that the SVM model has the highest accuracy, followed by the neural network model and the random forest model. The SVM model achieves an accuracy of 97.05%, while the neural network model and the random forest model achieve accuracies of 97.05% and 85.29%, respectively.

This result suggests that the SVM model is the best performing model for this particular classification task. However, it is important to note that the difference in performance between the three models is relatively small. Additionally, the performance of a machine learning model can vary depending on the specific dataset and task.

Here are some possible explanations for the difference in performance between the three models:

* The SVM model may be better at learning the underlying patterns in the data.
* The SVM model may be more robust to overfitting.
* The neural network model may be underfitting the data.
* The random forest model may be overfitting the data.

It is also possible that the difference in performance is due to chance.

Overall, the bar chart shows that the SVM model is the best performing model for this particular classification task. However, it is important to note that the difference in performance between the three models is relatively small.