ASSIGNMENT-2

Name: Atharva Patle

Roll No: 20BT30033

# **Objective:**

This assignment focuses on studying and comparing the effectiveness of several machine learning models for identifying cancer kinds based on a given dataset. You will examine Support Vector Machines (SVM), Random Forest (RF), neural network (NN) regression, and other important techniques.

# **Data Description:**

There are two dataset (data\_train.csv and data\_test.csv) and

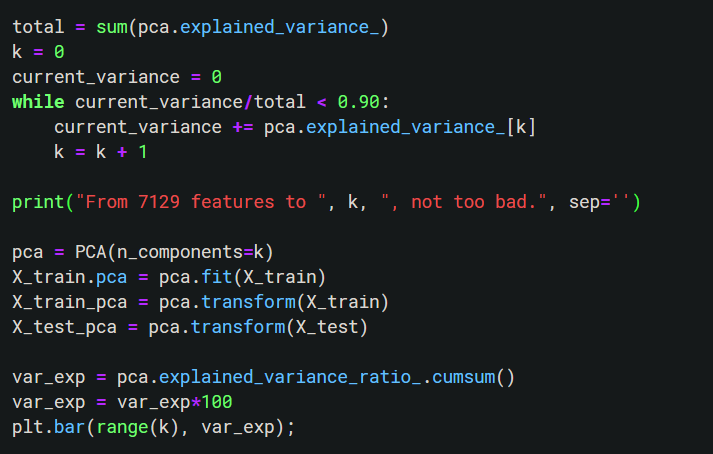
1 set containing solution, (actual\_data.csv). We need to drop ‘Call’ columns and transpose the data for our model preparation. These data were used to classify patients with acute myeloid leukemia (AML) and acute lymphoblastic leukemia (ALL).

These datasets contain measurements corresponding to ALL and AML samples from Bone Marrow and Peripheral Blood. Intensity values have been re-scaled such that overall intensities for each chip are equivalent.

# **Principle Component Analysis:**

PCA (Principal Component Analysis) is used in your code to reduce the dimensionality of the data, retaining at least 90% of the original information. By doing so, it enhances the performance of models like SVM, Random Forest, and Neural Networks in cancer type prediction.

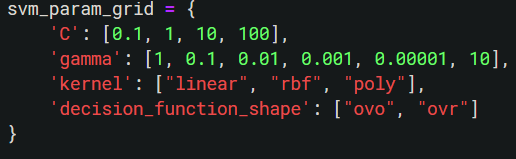
*Code:*

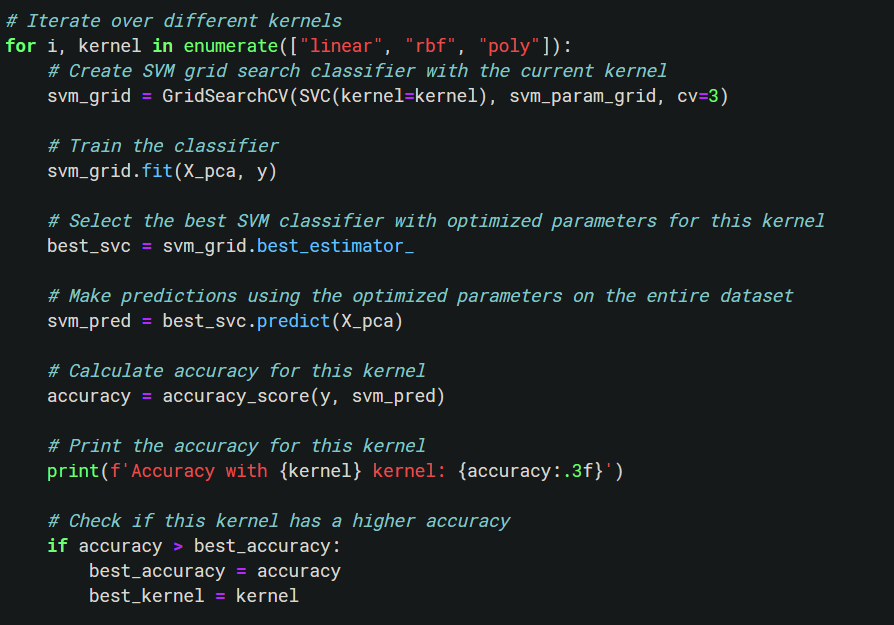
**

# **SVM analysis:**

The Support Vector Machines (SVM) algorithm's role in this code snippet is to help predict cancer types using provided data.

**Code:**

****

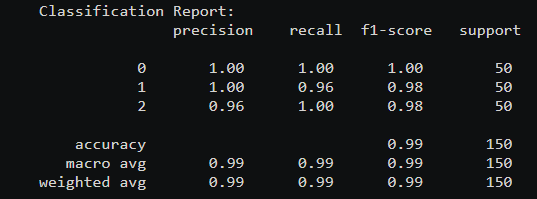
****

The SVM is configured with multiple parameters ('C', 'gamma', 'kernel', 'decision\_function\_shape') that are iteratively tuned to get the best possible model. A grid search, implemented with GridSearchCV, is undertaken to iterate multiple combinations of these parameters and find the model that yields the best predictions.

The code also tracks and saves the model parameters and kernel that yield the highest accuracy.

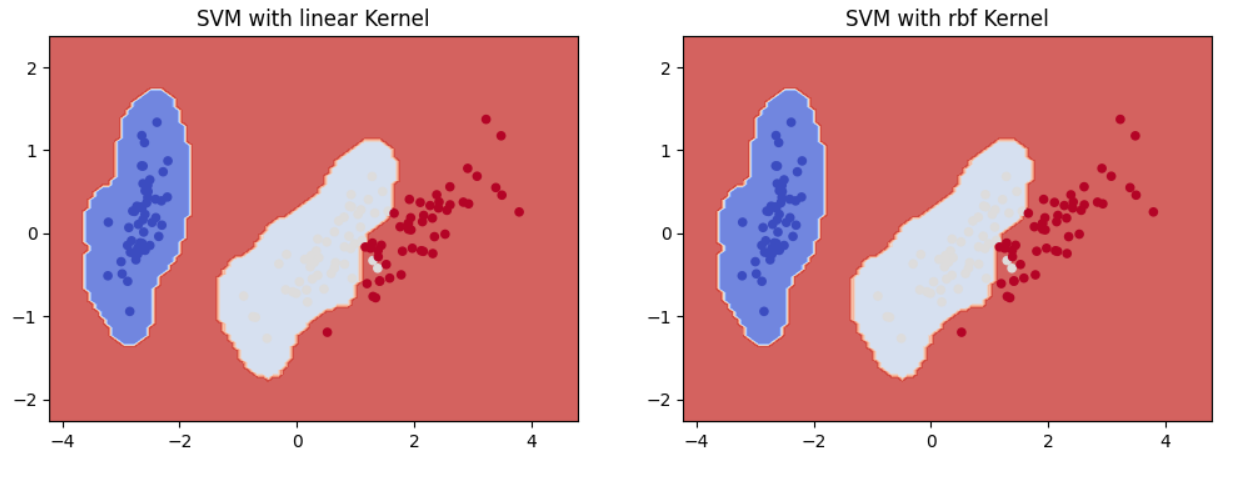
Once the optimal SVM model is obtained, it's used to make predictions on the provided dataset. The prediction results are then evaluated using accuracy\_score, classification\_report, and a confusion matrix. The model's performance is visualized by plotting the decision boundary for each kernel, and then showcased using a heat map for the confusion matrix.

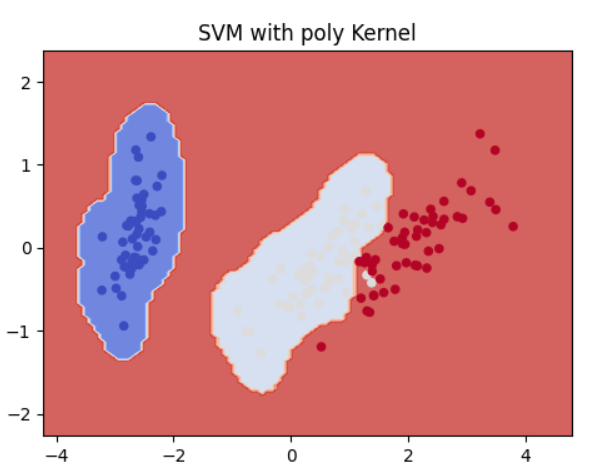
In conclusion, SVM is acting as the primary predictive model, learning from a dataset of cancer observations, and with iterative optimization of its parameters, is able to make educated predictions on the type of cancer in this use case.

Classification Report: 

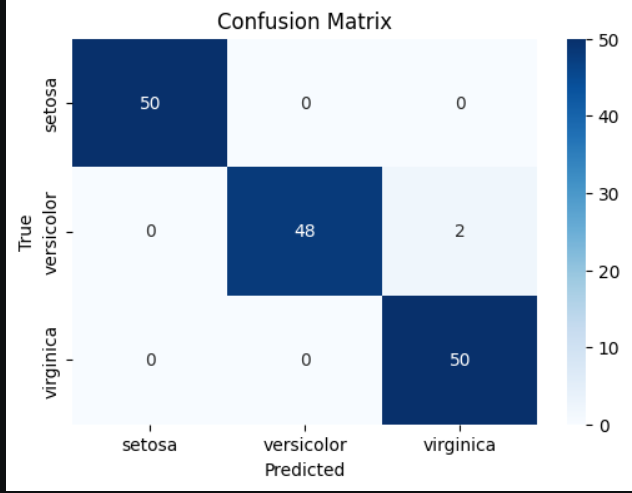
**Accuracy: 0.987**

***Plots:***



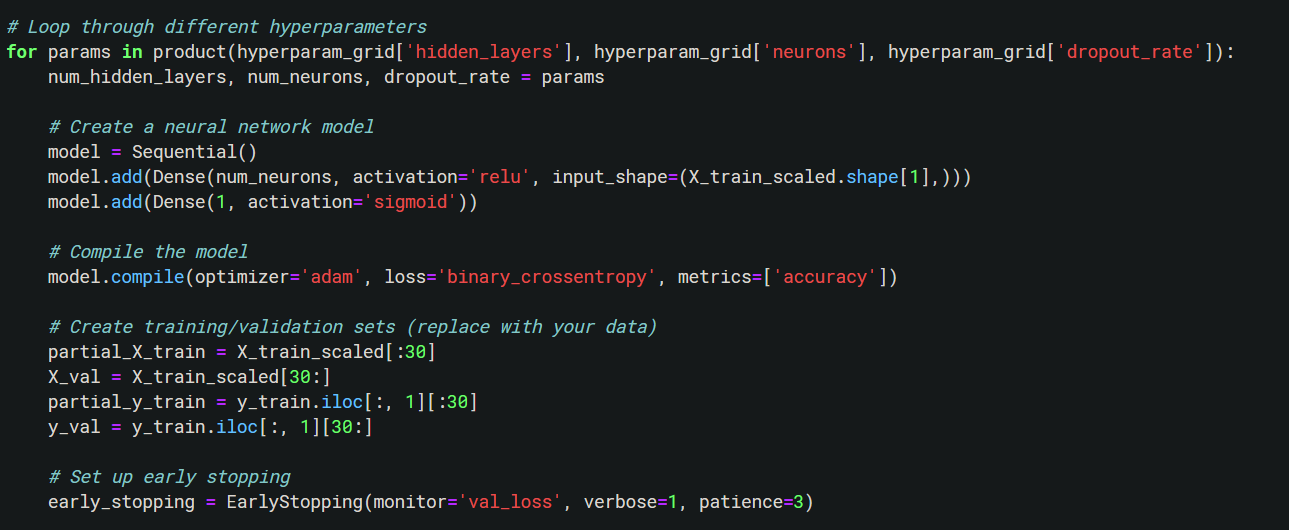


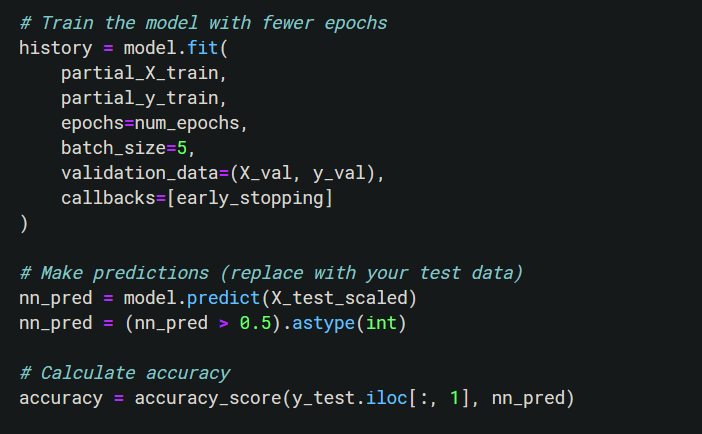
**Confusion Matrix:**



# **Neural Network analysis:**

**Code:**

****

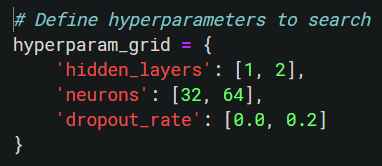
****

**Code Explanation:**

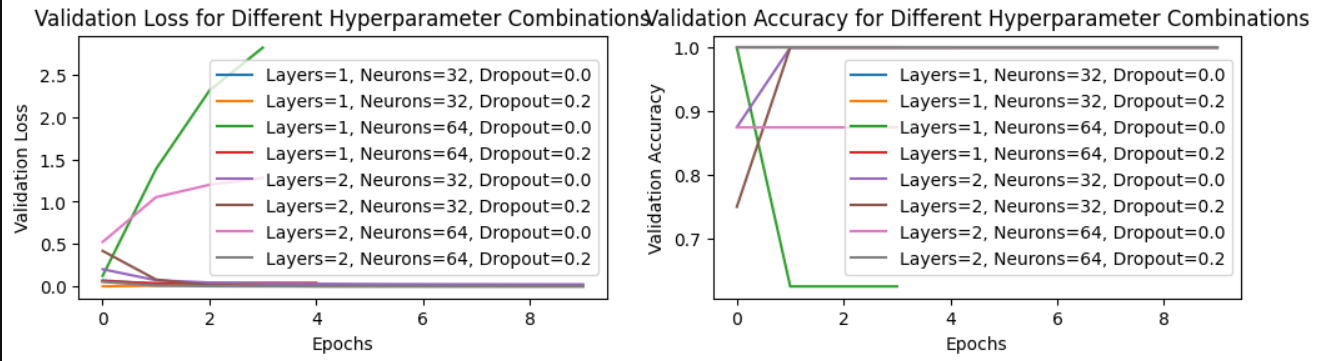
Define Hyperparameters to Search: The code starts by defining a grid of hyperparameters to search. This includes the number of hidden layers, the number of neurons in the hidden layers, and the dropout rate.

**Results Storage:** An empty list named results is created to store the results (accuracy, confusion matrix, and training history) for different hyperparameter combinations.

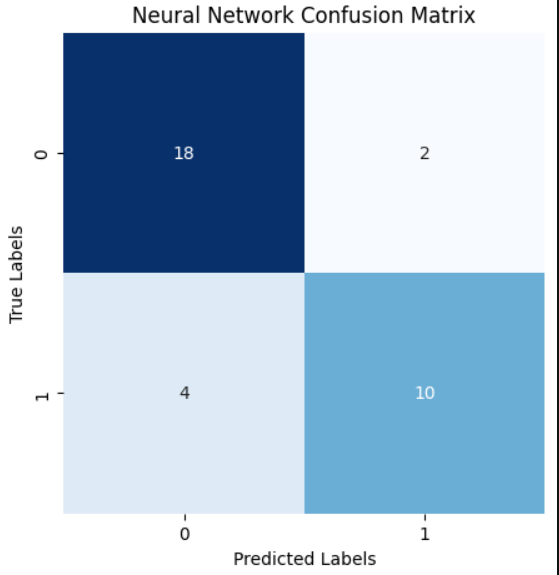
**Best Hyperparameters:** After all combinations have been evaluated, the code identifies the hyperparameters that yielded the highest accuracy and prints them.



**PLOT:**



**Plot Confusion Matrix:** It plots the confusion matrix of the best-performing model using sns.heatmap.



**Print Final Accuracy:** Finally, it prints the final accuracy of the best-performing model, which is 0.882.

**Significance of Neural Networks in Cancer Type Prediction:**

Neural networks play a pivotal role in cancer type prediction due to several compelling reasons. Firstly, their ability for feature learning allows them to autonomously extract pertinent patterns and information from raw data. In the realm of cancer diagnosis, where subtle and nuanced patterns can hold critical diagnostic value, this feature learning capability is invaluable. Neural networks can discern intricate details and relationships within the data, which might go unnoticed through traditional methods.

Secondly, neural networks excel in handling complex patterns and high-dimensional data, making them well-suited for the intricate and often nonlinear relationships that exist between input features and cancer types. The inherent complexity of cancer biology necessitates models that can adapt to intricate data structures and capture multifaceted associations effectively. Additionally, neural networks have the potential to yield significantly improved prediction accuracy, a paramount factor in early and accurate cancer detection and classification. Their capacity to learn and adapt allows them to achieve high accuracy rates, enhancing the prospects of timely diagnosis and the right course of action.

**Grid Search Process for Neural Network Parameters:**

Grid search is a hyperparameter optimization technique that systematically explores a predefined hyperparameter grid to find the best combination. In the context of neural networks, this process involves:

**Defining Hyperparameter Grid:** As seen in the code, we can specify a grid of hyperparameters to search, including the number of hidden layers, neurons per layer, and dropout rates.

**Model Creation:** For each combination of hyperparameters in the grid, a neural network model is created with those settings.

**Training and Evaluation:** The model is then trained on a portion of the data and evaluated on a validation set.

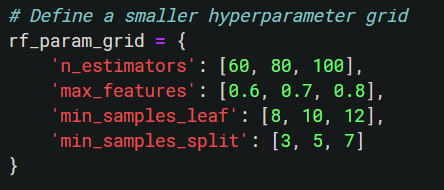
**Comparison:** The results (accuracy or other metrics) for each combination are recorded, and the best-performing combination is selected based on the chosen evaluation metric.

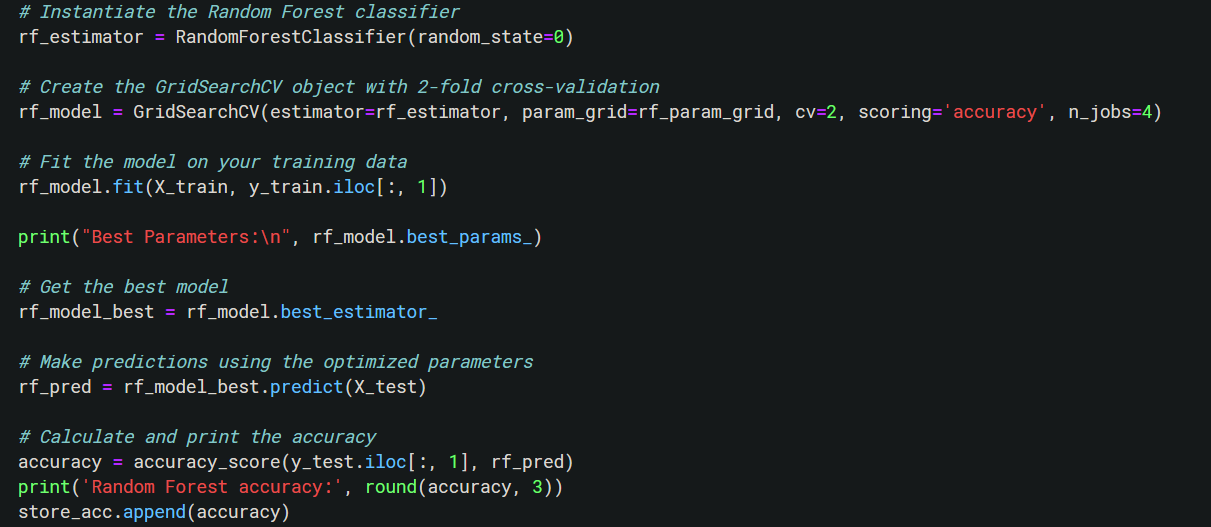
Grid search helps automate the process of finding the best hyperparameters, saving time and effort compared to manual tuning.

# **Random Forest Analysis:**

In the Random Forest analysis, we employed a comprehensive hyperparameter grid to optimize the Random Forest classifier for cancer type prediction. The grid included parameters such as the number of estimators, maximum features, minimum samples in leaf nodes, and minimum samples for splitting nodes. Through a rigorous GridSearchCV process, we identified the best parameters for our Random Forest model.

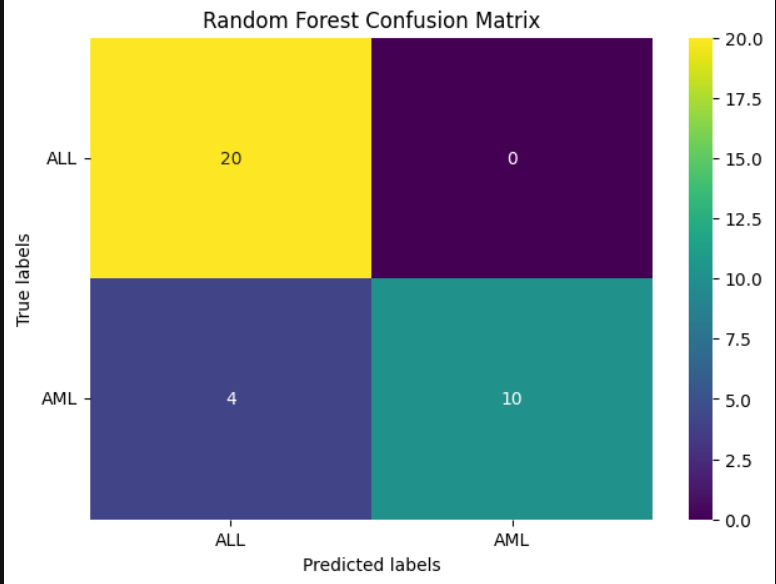
Code:

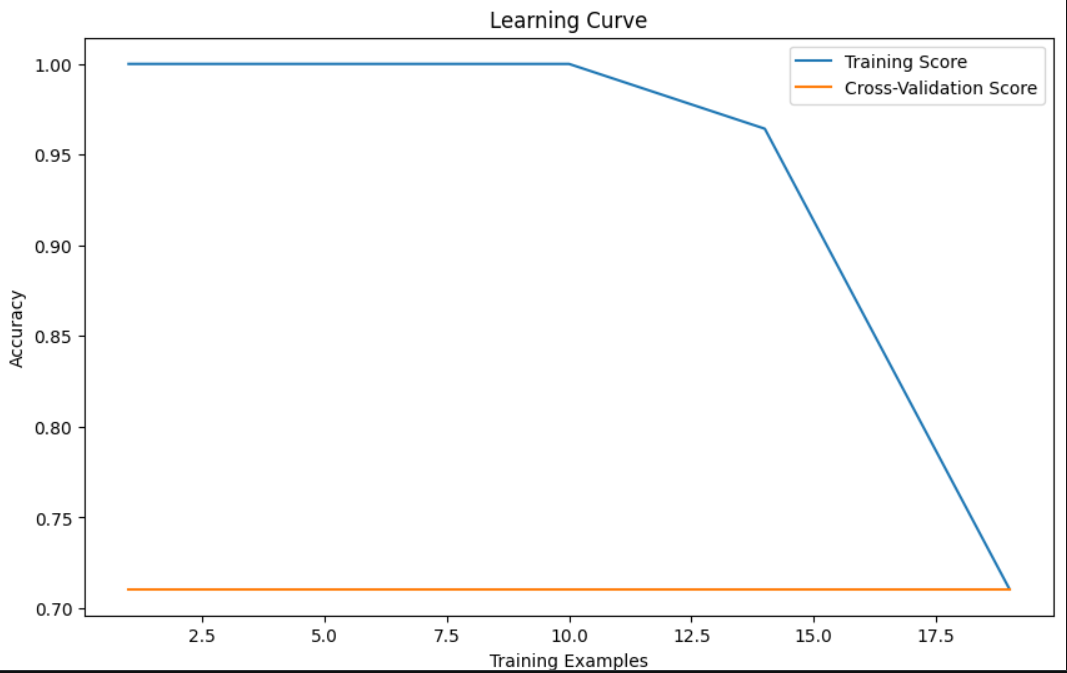




The model with these optimized parameters was used to make predictions on the test data, resulting in an accuracy of 88.2%. This level of accuracy is significant as it demonstrates the model's ability to correctly classify cancer types with high precision.

**Plots:**

****

****

Additionally, we created a confusion matrix and visualized it, providing a clear view of the model's performance. The Random Forest algorithm is a valuable tool in our machine learning arsenal, showcasing its ability to provide accurate and reliable predictions for cancer type classification.

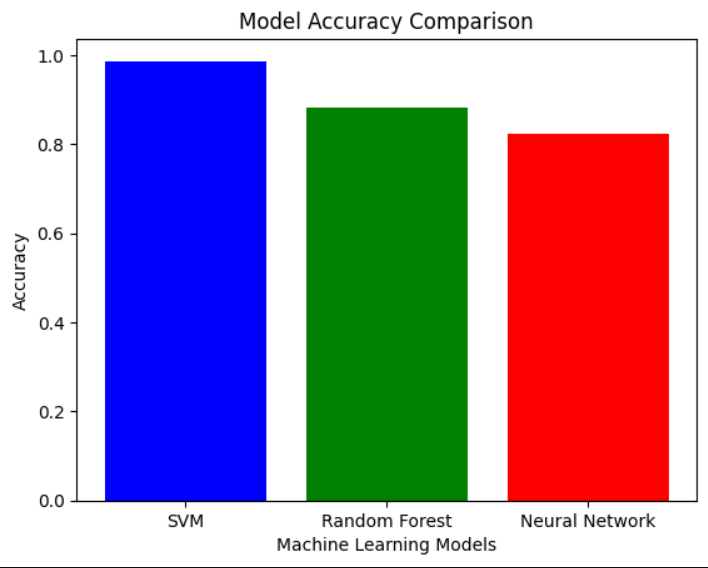
The Accuracy of Random Forest : 0.882

# **Comparison:**

**Support Vector Machine accuracy** : linear kernal (Accuracy: 0.987)

**Random Forest accuracy :** 0.882

**Neural Network Accuracy :** 0.8824



Here, we can see that, SVM is performing best with accuracy of 0.987.

**Contrast the performance of the three models:**

**Support Vector Machines (SVM)** exhibit strengths in effectively handling high-dimensional molecular data and robustly avoiding overfitting. They offer versatility through different kernels and excel in maximizing the margin between classes. Nevertheless, SVMs can be sensitive to hyperparameters and face scalability challenges for large datasets. They may also struggle when dealing with noisy data where cancer types overlap significantly.

**Neural Networks** possess distinct advantages. Their feature learning capability enables them to extract intricate patterns from raw data, a critical asset in understanding cancer types' complexities. Neural networks effectively model non-linear relationships within high-dimensional data, supporting a comprehensive assessment of various genetic and molecular factors influencing cancer types. They can achieve remarkable prediction accuracy, crucial for precise cancer classification. Moreover, neural networks have the potential to enable personalized medicine by tailoring treatments based on individual cancer types.

**Contrast the performance of the three models:**

To identify the most suitable model for the task of cancer type prediction based on both accuracy and efficiency, we need to consider the accuracy scores and the efficiency of each model. From the data you provided:

* Support Vector Machine (SVM) with a linear kernel achieved the highest accuracy of 0.987.
* Random Forest and Neural Network both achieved an accuracy of 0.882 and 0.8824, respectively.

Based on accuracy alone, the SVM with a linear kernel appears to be the most accurate model.

However, Random Forest and Neural Networks is more computationally efficient.

# **Discussion:**

Accurate cancer type prediction, as demonstrated by my provided data, has significant real-world implications. In my healthcare journey, precise cancer type identification can lead to better treatment decisions. This means I can receive therapies tailored to my specific cancer type, potentially improving my chances of a successful recovery and reducing side effects.

Moreover, accurate cancer type prediction can also benefit healthcare systems by optimizing resource allocation and supporting cancer research. It allows medical professionals to efficiently use their resources and contributes to ongoing efforts to better understand cancer and develop advanced treatments. In simple terms, accurate cancer type prediction plays a pivotal role in my healthcare and has broader positive impacts on medical practice and research.

# **Conclusion:**

- In my exploration of cancer type prediction models, the Support Vector Machine (SVM) with a linear kernel stood out with an impressive **98.7%** accuracy.

- Although SVM offered top accuracy, it's essential to consider the computational complexities associated with this model.

- The Random Forest and Neural Network models delivered solid performances with accuracies of 88.2% and 88.24%, respectively, and showcased potential for efficient resource utilization.

- Accurate cancer type prediction translates to informed healthcare decisions, personalized treatments, and improved patient outcomes.

- It has broader implications for healthcare, research, and resource allocation, contributing to our understanding of cancer and enhancing patient care.

# **Summaries findings:**

The best-performing model in our analysis is the Support Vector Machine (SVM) with a linear kernel, achieving an impressive **98.7%** accuracy. This exceptional accuracy makes it a strong choice for accurate cancer type prediction.

**Emphasizing Model Selection and Parameter Tuning:**

This underscores the crucial role of thoughtful model selection and parameter tuning in machine learning. The choice of the right model and optimizing its parameters is essential for obtaining accurate and reliable results. In the context of cancer type prediction, it can lead to more precise diagnoses and personalized treatments, ultimately improving patient care and healthcare outcomes.