M2-Bioinformatics DLAD
Course : extensive biostatistics
Student ID : O21227743
2022-2023

Cancer type prediction using Machine Learning algorithms

1. Introduction

Cancer is one of the most devastating, fatal, dangerous, and unpredictable diseases. To reduce the risk of fatality in this disease, we need some ways to predict the disease, diagnose it faster and precisely, and predict the prognosis accurately.

The incorporation of artificial intelligence (AI), machine learning (ML), and deep learning (DL) algorithms into the healthcare system has already proven to work wonders for patients. Artificial intelligence is a simulation of intelligence that uses data, rules, and information programmed in it to make predictions.

The science of machine learning (ML) uses data to enhance performance in a variety of activities and tasks. It is a technique of artificial intelligence which consists in "teaching" a machine, from data to make predictions.

All of these are required to improve patient's quality of life, increase their survival rates, decrease anxiety and fear to some extent, and make a proper personalized treatment plan for the suffering patient.

In our case, ML models play a key role in taking into consideration effective features which conduct to a precise cancer type. Here, we are interested into the phenotype determination (=cancer type) using gene expression levels of a variety of genes of interest) in the most occurring cancer types.

2. Material and methods

2.1. Dataset

The dataset is composed of two files one containing the expression of 20531 genes of 801 patients with five different tumors which are the labels present in the second file. These five different tumors are breast cancer (BRCA), Colon Adenocarcinoma (COAD), Kidney Renal Clear Cell Carcinoma (KIRC), Lung Adenocarcinoma (LUAD) and Prostate Adenocarcinoma (PRAD). The data was obtained from the UCI Machine Learning Repository (Fiorini, 2016) which contains 622 datasets for the Machine Learning community to practice.The goal here is to implement a model allowing the diagnosis of a cancer based on this dataset.

2.2. Material

In our work we used the Python programming language with the version 3.10 of the interpreter and the following library version ： pandas 1.5.2 , scikit_learn 1.2.0 , matplotlib 3.6.2 .

All the code is available in the gihub repository(Ouertani, 2022).

2.3. Methods

The overall goal of this work was to predict the cancer type based on the gene-expression (features) using labeled data : its then a supervised classification challenge.

Different ML algorithms were implemented, which follow various strategies when it comes to classification of cancer-type.

Data preprocessing

The first step consisted of loading the data then proceeding to cleaning it by removing the columns and rows containing only zeros.

Data splitting

Once our data was ready, the next step consisted of splitting the data into 70% training and 30% testing. This is a typical step to prevent the model later from learning all the data “by heart”.

Algorithm choice

For this work we chose to compare the prediction capabilites of 3 of the most used classification algorithms :

1) Multiclass logistic regression

Multiclass logistic regression is also called multinomial logistic regression and softmax regression. It is used when we want to predict more than 2 classes. It is an extension of logistic regression that involves changing the loss function to cross-entropy loss and predicting the probability distribution to a multinomial probability distribution to natively support multi-class classification problems.

2) Support Vector Machine (SVM)

SVM (Support Vector Machine) is a supervised machine learning algorithm that can be used for both classification and regression. The goal of the SVM algorithm is to use a training set of objects (samples) separated into classes to find a hyperplane in the data space that produces the largest minimum distance (called margin) between the objects (samples) that belong to different classes. So the hyperplane is known as the maximum margin hyperplane. SVM only uses the objects (samples) on the edges of the margin (called support vectors) to separate objects (samples) rather than using the differences in class means. Since the separating hyperplane is supported (defined) by the vectors (data points) nearest the margin, so the algorithm is called SVM.

3) Random Forest Classifier (RFC)

Random forests is one the most flexible and easy to use algorithms. A forest is comprised of trees. It is said that the more trees it has, the more robust a forest is. Random forests creates decision

trees on randomly selected data samples, gets prediction from each tree and selects the best solution by means of voting. It also provides a pretty good indicator of the feature importance.

Model training and optimization

Once we chose the appropriate algorithms for our use we proceeded to fitting/training these models to the training data.

In order to find the best fit for each model, we needed to optimize the hyperparameters of each model by trying different combinations and finding the best one. This is done using the grid search method which we also combined with cross validation to optimize the training accuracy at the same time.

Model testing and evaluation

Once we had our models trained using the optimal hyperparameters we tested them on the testing dataset (remaining 30%). Then we proceeded to evaluating and comparing the models by generating the full classification report including the following metrics : accuracy , precision, recall and f1 score. We also visualized the learning curves of each model using the best parameters found by grid search to inspect the difference between testing and training accuracy.

3. Results and discussion

After cleaning the dataset, only 20264 of the initial 20531 features were left while all the samples remained.

Next the training and optimization of the models using grid search and cross validation led to the results obsereved in Table 1.

We can see that the Multiclass logistic regression model has perfect scores in all the metrics followed by score of 0.99 for the Random Forest Model and finally the Support Vector Machine model which has the lowest scores for accuracy, recall, precision and f1 score with values around 0.92.

The following scores indicate that the Multiclass regression model and Random Forest models are a very good fit for our data and use case. These very high prediction scores could be due to the very high number of features (20264) compared to the number of samples (801) which could be causing an overfitting issue that should be resolved by a reduction in the number of the features taken into account. Another hypothesis is that the raw data used in this project is simply “too perfect” that the algorithm find ease in learning its specificities.

Table 1 : Evaluation metrics for all 3 models and their best parameters using grid search

	Multiclass logistic regression	Support Vector Machine	Random Forest
Accuracy	1.00	0.92	0.99
Precision	1.00	0.93	0.99
Recall	1.00	0.92	0.99
F1 score	1.00	0.92	0.99
Best parameters	'C': 0.001	'bootstrap': True, 'max_depth': 110,	'C': 10, 'gamma': 0.0001,

'max_features': 3, 'kernel': 'rbf'
'min_samples_leaf': 3,
'min_samples_split': 8,
'n_estimators': 500

To verify the overfitting problem we visualized the learning curves for all 3 models which can be seen in Figure 1.

We can see that the training and testing accuracy is identical no matter the number of samples used for the Multiclass logistic regression model which denies an overfitting problem. Whereas the Random Forest model seems to have a slight difference between training and testing accuracy for a low number of samples which deminishes rapidly by increasing the number of samples deniying also in this case the overfitting problem.

The SVM model on the other hand, has a perfect training accuracy score but there is a big difference with the testing accuracy score which could in this case indicate an overfitting problem that needs to be addressed.

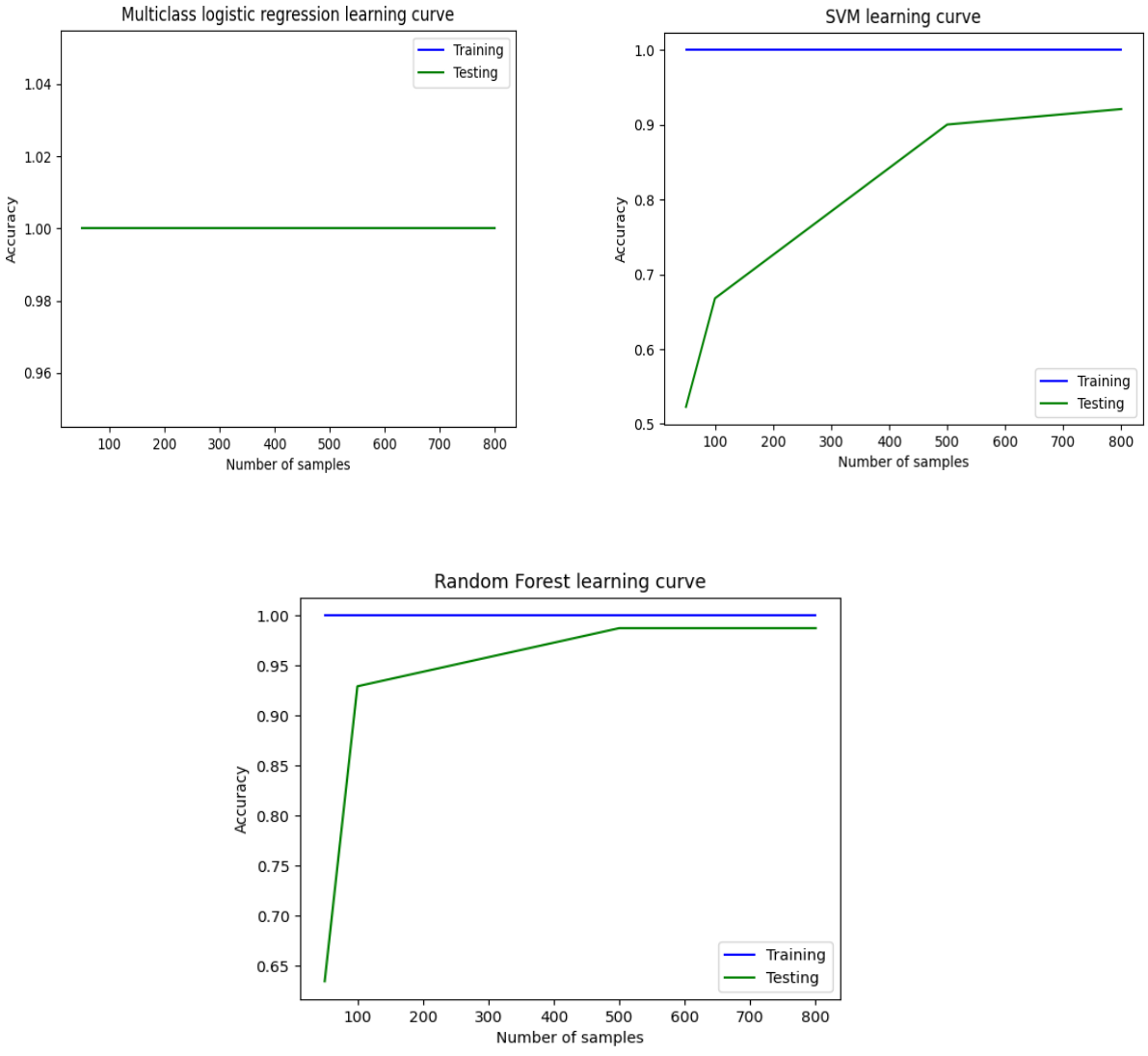

Figure 1: Learning curve graph for all 3 models

4. Conclusion

The Multiclass logistic regression as well as the Random Forest models seem to be a great fit for predicting the cancer types from this dataset with the first one having perfect scores on all types.

This is not always the case as the model to be used depends on the data input as well as the final expected results from using that model which shows the importance of the algorithm choice in the prediction process and also the importance of having good quality data to work on in the first place.

5. References

Fiorini, S., 2016. Dataset link
https://archive.ics.uci.edu/ml/datasets/gene+expression+cancer+RNA-Seq

Ouertani, M., 2022. Github machine learning project
https://github.com/Ouertani95/Machine_learning_project