CRC prediction based on early microbiome predictors in mice.

Date: 26.4.2020
Sharon Komissarov, Yoram Louzoun.

Abstract

Given a microbiome of several mice in six different time intervals and more basic information,
We manage to correctly predict whether a mouse will have a tumor in time interval five, based on its microbiome in time interval zero in more than 90% of the cases.

Introduction

Four independent experiments were performed.
In each experiment, samples were taken from mice on six different time intervals.
Every sample has its related microbiome which consists of thousands of bacteria.
The data also includes the taxonomy of all bacteria hence, it's possible to understand the similarity between bacteria.
By preprocessing the data and using machine learning tools we want:
First, predict in early stages whether a mouse will have CRC in advanced intervals or not (Binary).
next to determine how many tumors it will have (Discrete).
And lastly, to provide a continuous prediction on the number of tumors.

File description

"exported feature table for Yoram" - Contains the microbiome of some samples (only 168 out of 458 samples have a corresponding microbiome) and the identification is made by the OTU_ID value.

"mapping file with data Baniyahs Merge" - A table that maps every microbiome to a mouse and some information about it.
Including Treatment, sample time point, number of tumors(tumor_load), and spleen info.
There are 458 samples in the table.

"taxonomy"- In this file, every bacterium translated to its seven-layered matching taxonomy

Methods

Preprocessing

At first, the taxonomy of each bacterium was reduced to level six.
afterward, the microbiome file was merged (all rows with the same corresponding taxonomy value were replaced with the mean).
This action highly decreased the dimension of the table.
(2062,169) -> (109,168).

In order to gather all the information on each sample, the reduced microbiome is merged with the corresponding mapping table.

For each row (microbiome of a sample) classification is being added (tumor-load at time 5 multi-class or binary)

Zero columns and columns with an absolute correlation of 0.8 or higher were removed (only one per time).

Multiple normalizations and standardization methods were performed however, only log normalization was able to balance the data.

Dimension reduction

The data was reduced to five dimensions using ICA and PCA.
It's possible to implement the dimensionality reduction methods on either all-time points or only on time point zero.

Machine learning

The classification was made by KNN and performance evaluation with cross-validation and the elbow method.

SVR was used to predict continuous tumor values but resulted with poor performance.

Results

In the preprocessing stage, the correlations (Spearman) of bacteria and Immune system parameters with tumor values were inspected.

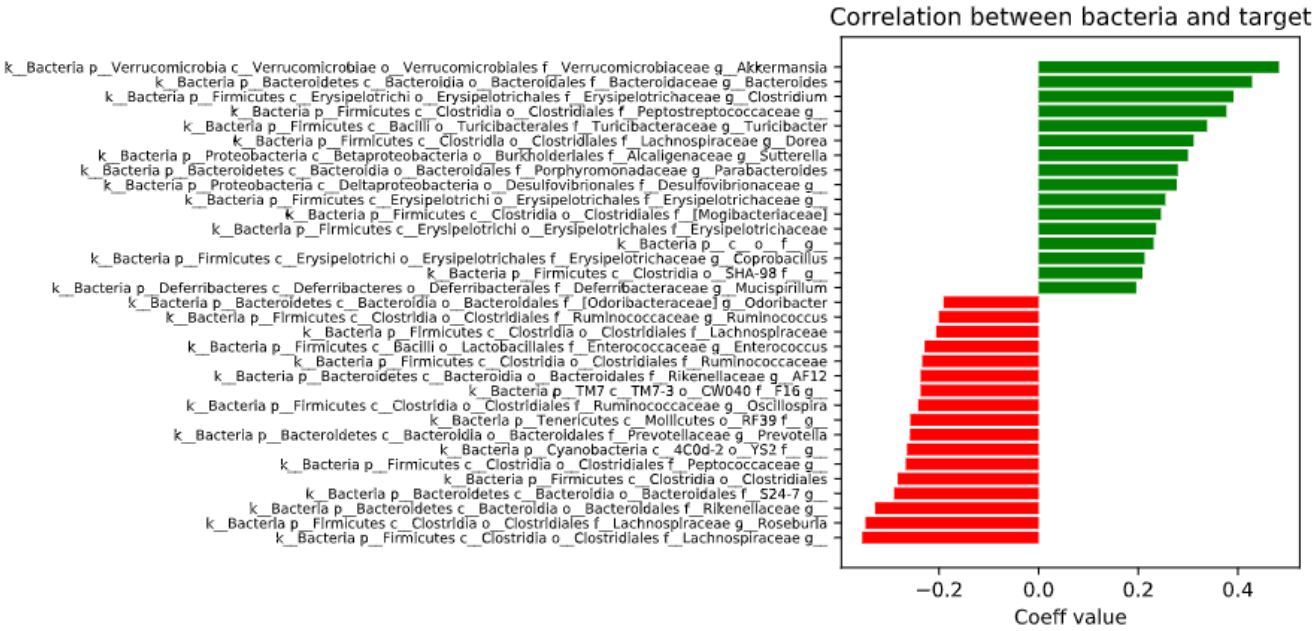

Figure 1 Bacterias with the highest absolute correlation with the target (descending order).

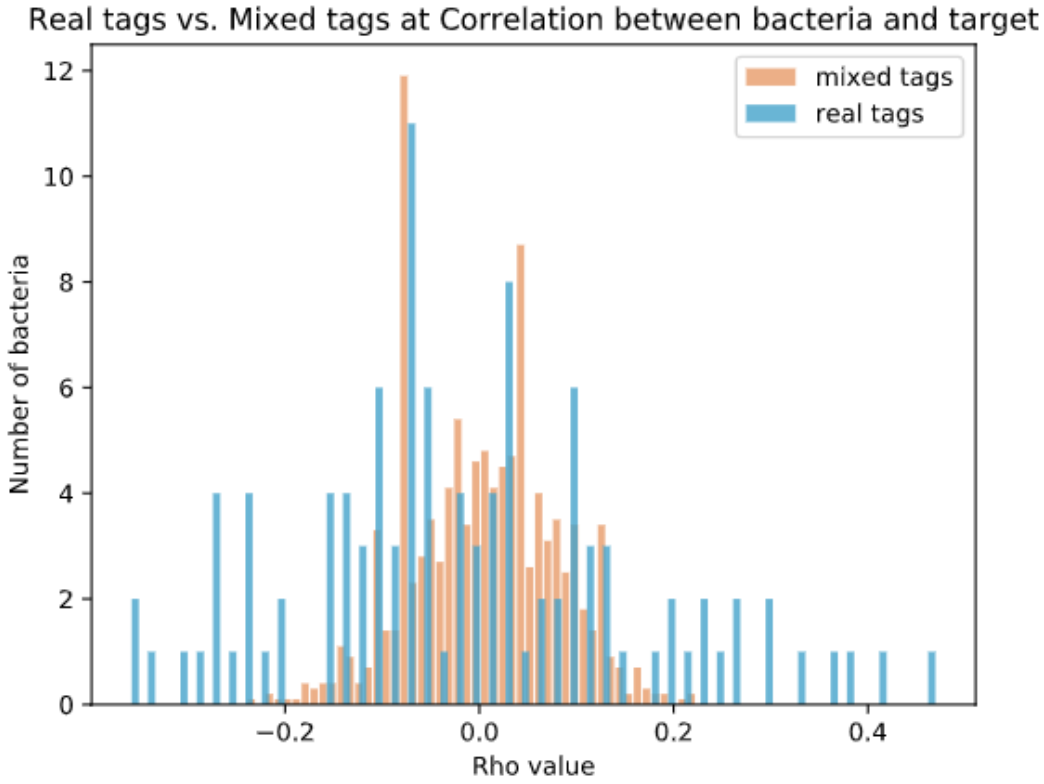

Figure 2 Comparing the correlations when the data is shuffled and not to check their stability.

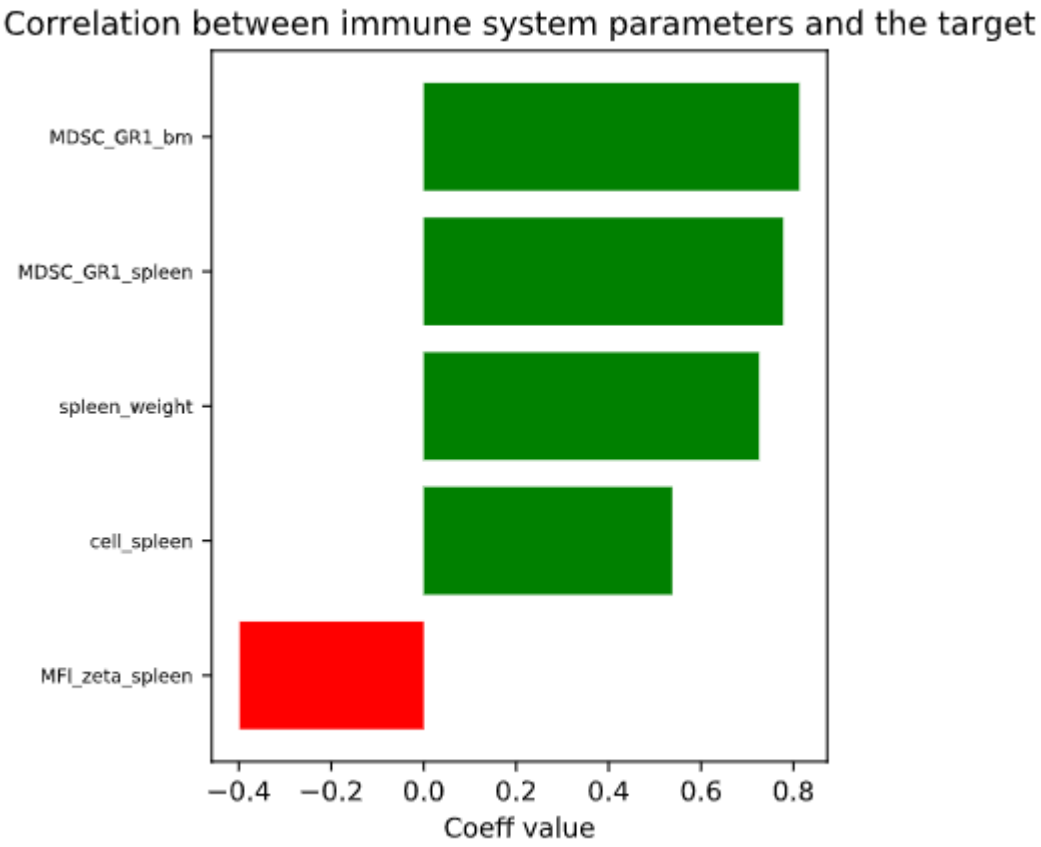

Figure 3 Correlations between immune system features and the target (descending order).

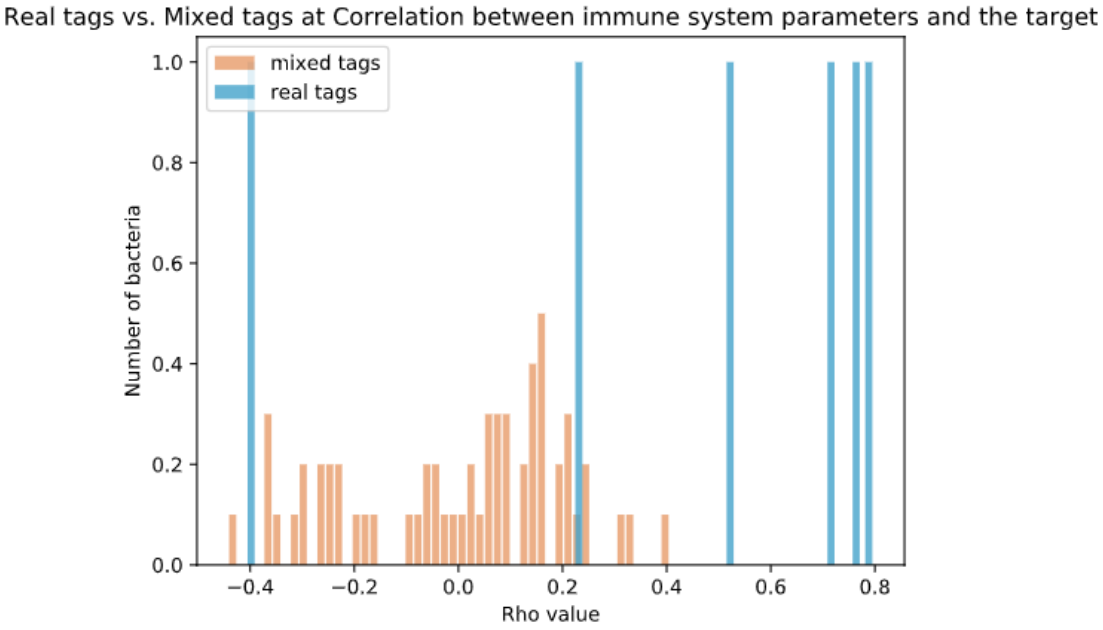

Figure 4 Comparing the correlations when the data is shuffled and not to check their stability.

After the decomposition, we wanted to examine if it's possible to separate between the samples with CRC and the ones without.

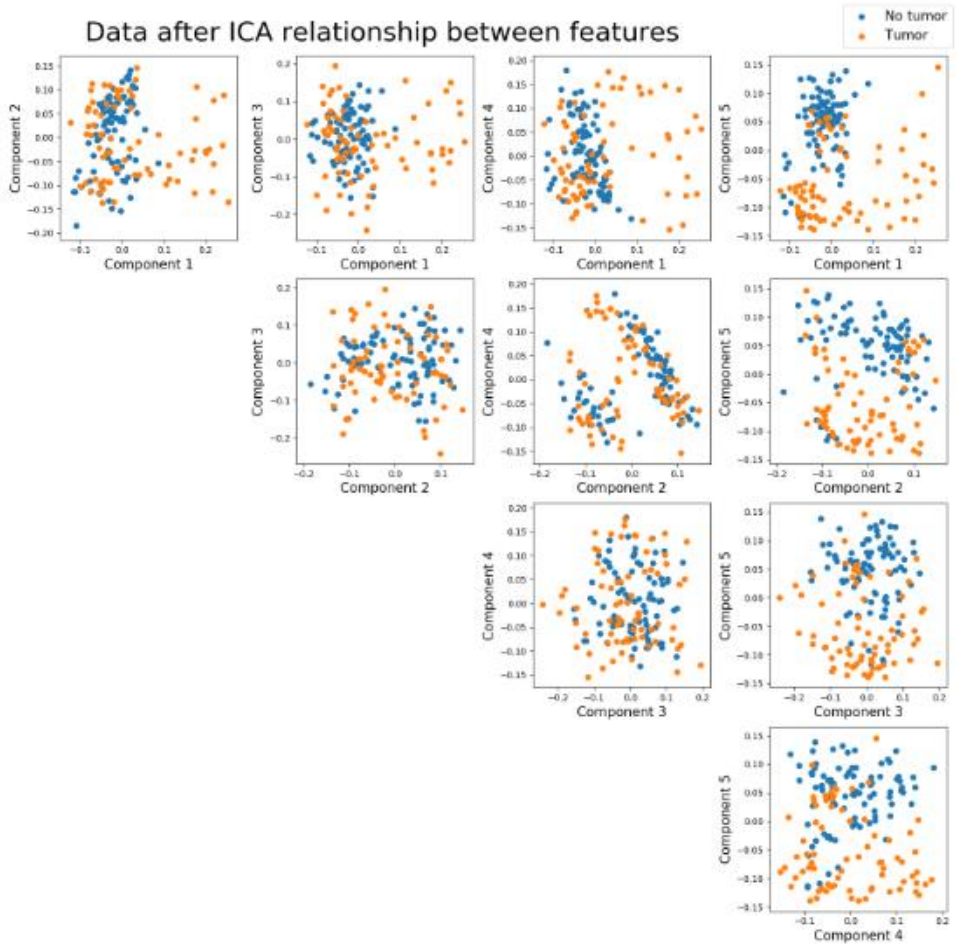

Figure 3: Relationship between every pair of components after the ICA decomposition.

Colored by Binary tumor load.

Figure 4: Relationship between every pair of components after the PCA decomposition.

Colored by Binary tumor load.

It's visible that the groups mentioned above are separable in most pairs of components after both decompositions.

It also can be seen, that all projections are well spread, an indication of successful preprocessing.

In addition, we wanted to analyze the progress of the mean of each component in time. And, to see whether there is a significant difference between the CRC and Normal groups mean for all time-points.

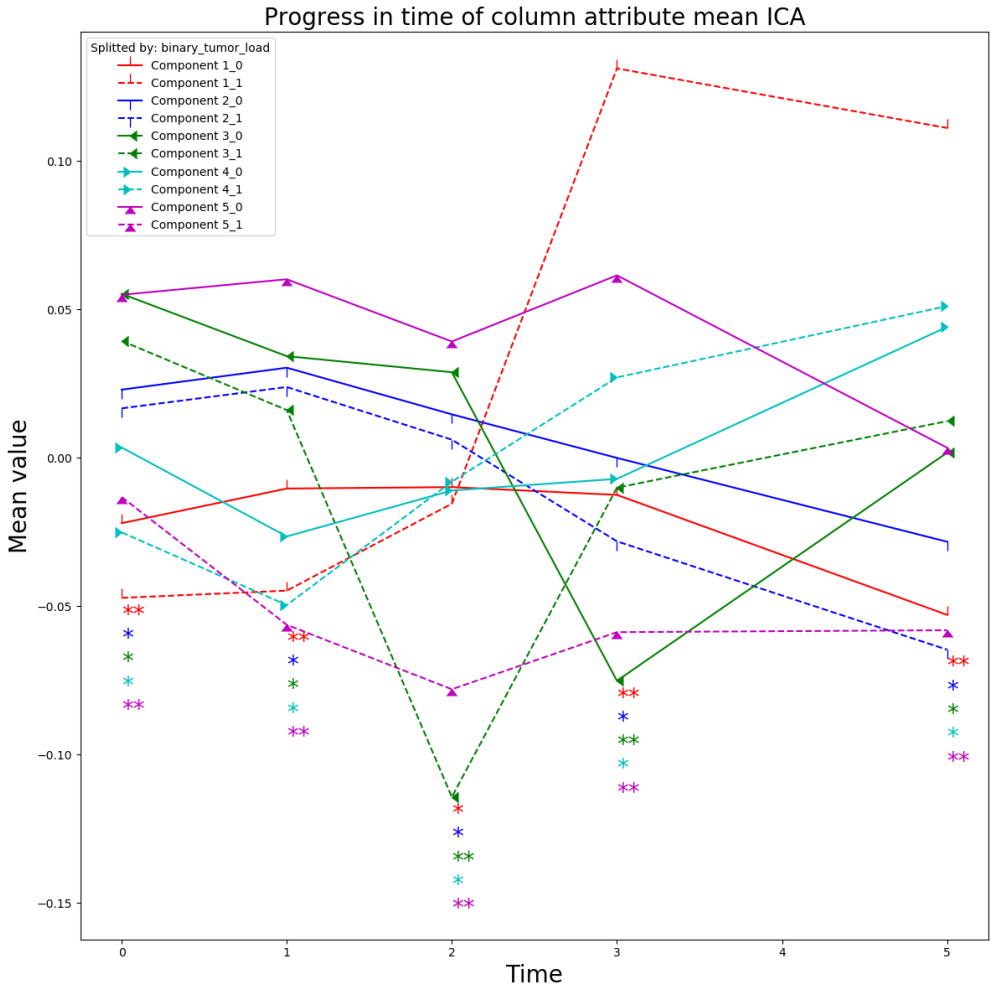

Figure 5 Each component was split to two groups by its corresponding tumor load, and the progress of the means in time of both groups is plotted. T-test was performed between every pair of groups.

*** - P-value<0.05 * else*

As reflected from the figure, there is a major difference between the CRC and Normal projections of component one and five in almost all time-points.

Because of the small amount of data, it wasn't possible to use complicated models (e.g. Tress or neural networks).

Therefore, KNN was trained on 70% of the samples and evaluated by cross-validation and the elbow method on the rest 30%.

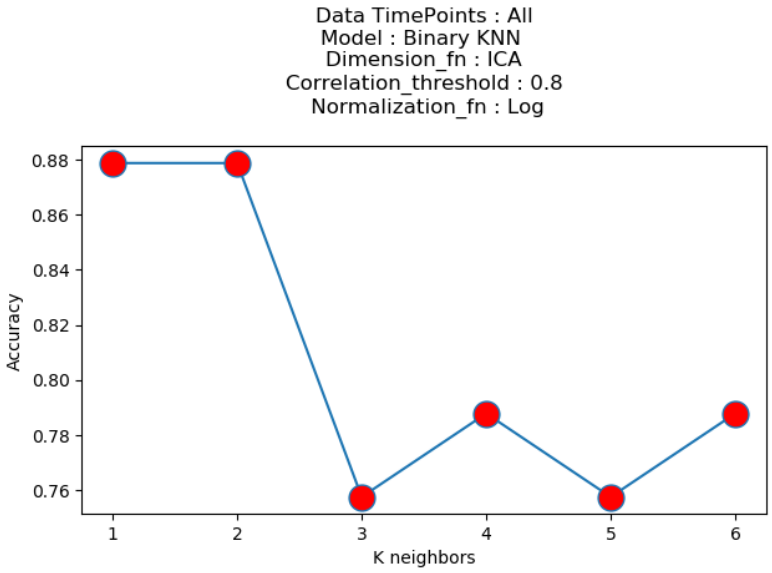

Figure 7: Accuracy performance of KNN after ICA decomposition.

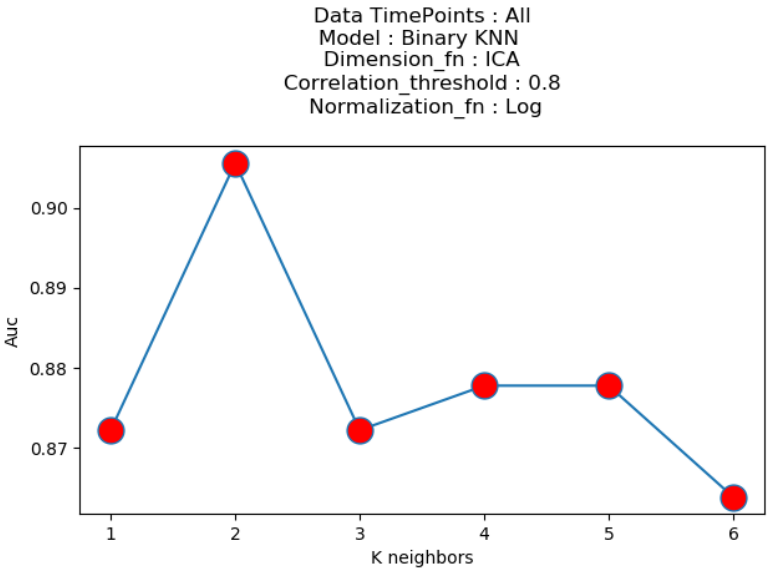

Figure 6: Auc performance of KNN after ICA decomposition.

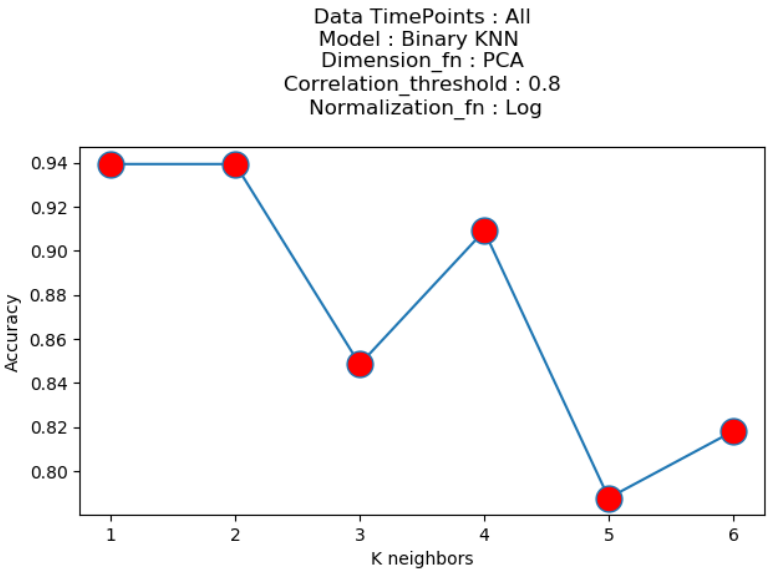

Figure 9: Accuracy performance of KNN after PCA decomposition.

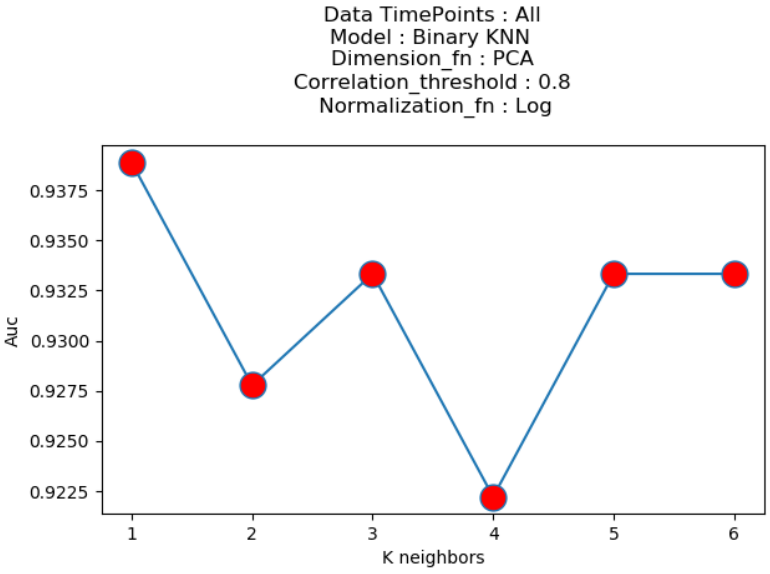

Figure 8: Auc performance of KNN after PCA decomposition.

In total, the KNN model achieved high accuracy in both decompositions.

Also, we can observe that PCA decomposition performed better than ICA in final results for each K value.

Regression models such as SVR, Lasso, and Ridge produced poor results on the test set, no matter the hyper-parameters due to the small amount of data.