IITM-CS5691 : Pattern Recognition and Machine Learning | Release Date: 22 Oct, 2021
New Method Writeup | Due Date: 10 Nov, 2021

Roll No: ME18B181 | Name: Aniruddha R Gandhewar
Roll No: ME18B182 | Name: Vasudev Gupta

1. (2 points) [Name of the Method chosen for each descriptor, and its Paradigm: paradigm could be linear/non-linear models, etc.)]:

Solution:

1. **Descriptor 1** (*co1*) :
   - METHOD: PCA → RFE-SVM → Bagged SVM Classifier
   - PARADIGM: Non-linear
2. **Descriptor 2** (*co2*) :
   - METHOD: PCA → RFE-SVM → Bagged SVM Classifier
   - PARADIGM: Non-linear
3. **Descriptor 3** (*co3*) :
   - METHOD: PCA → RFE-SVM → Bagged SVM Classifier
   - PARADIGM: Non-linear
4. **Descriptor 4** (*co4*) :
   - METHOD: SelectKBest → PCA → RFE-SVM → Bagged SVM Classifier
   - PARADIGM: Non-linear
5. **Descriptor 5** (*co5*) :
   - METHOD: SelectKBest → PCA → RFE-SVM → Bagged SVM Classifier
   - PARADIGM: Non-linear
6. **Descriptor 6** (*co6*) :
   - METHOD: SelectKBest → PCA → RFE-SVM → Bagged SVM Classifier
   - PARADIGM: Non-linear

2. (5 points) [Brief description on the dataset: could show graphs illustrating data distribution or

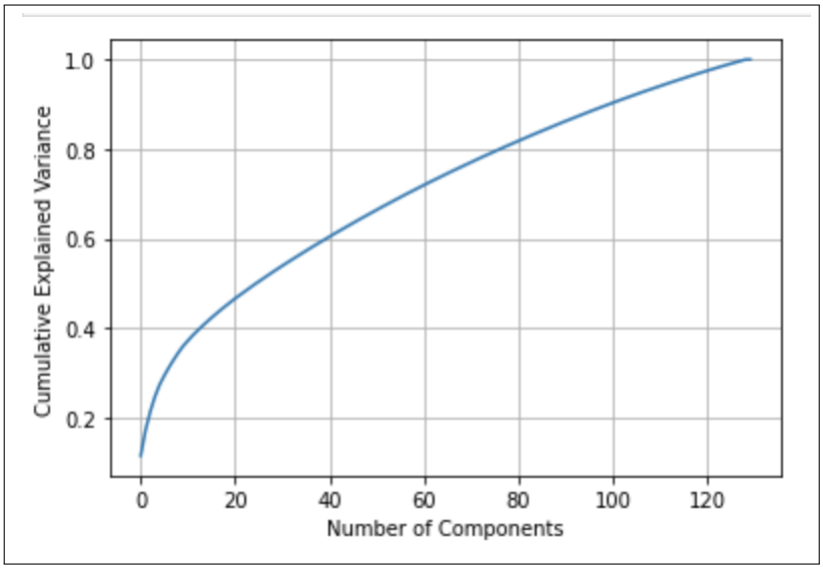
brief any additional analysis/data augmentation/data exploration performed]

Solution:

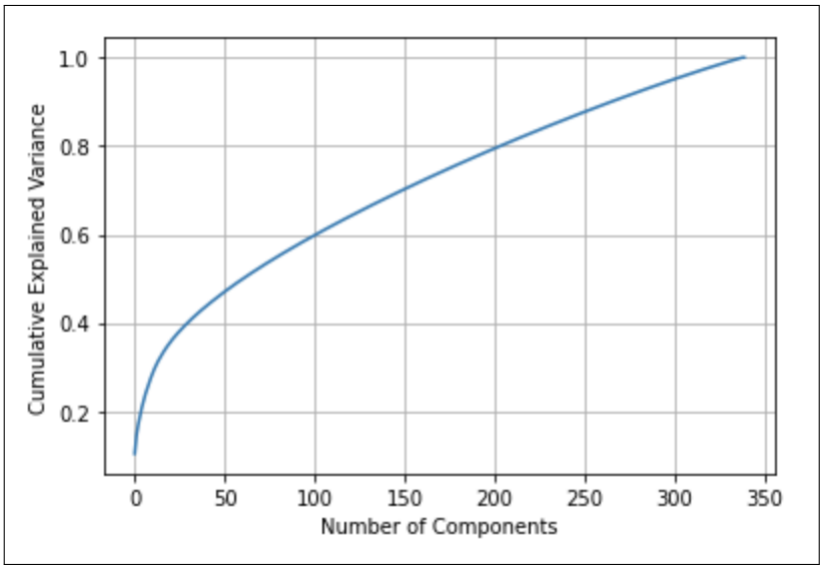
Dataset I consists of 130 samples, and dataset II consists of 340 samples for training the classifiers. Dataset I includes 2 clinical descriptors, and Dataset II includes 4 descriptors.

We applied Principal Component Analysis (PCA) for dimensionality reduction on datasets I & II and visualized the dependence of cumulative explained variance on the number of components.

Variance in Data-set I can be explained by Principal Components as follows:

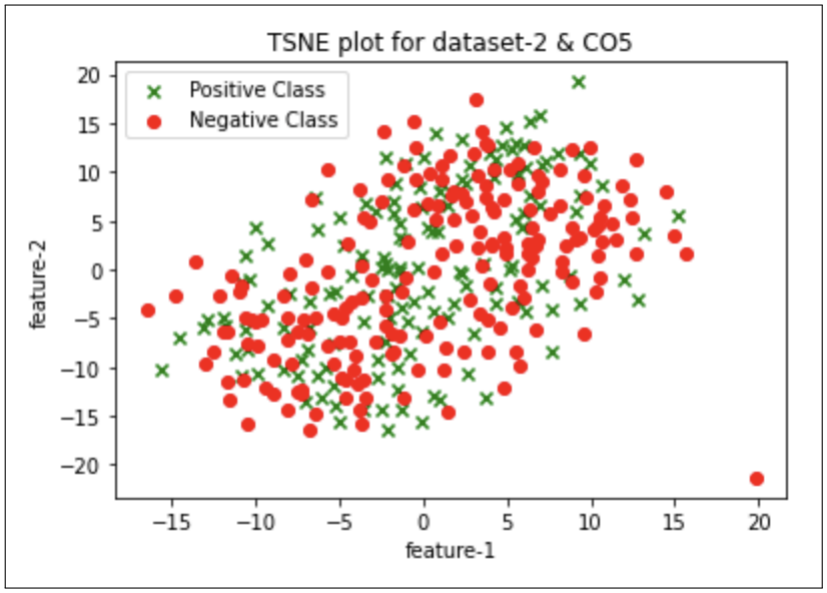


Variance in Data-set II can be explained by Principal Components as follows:

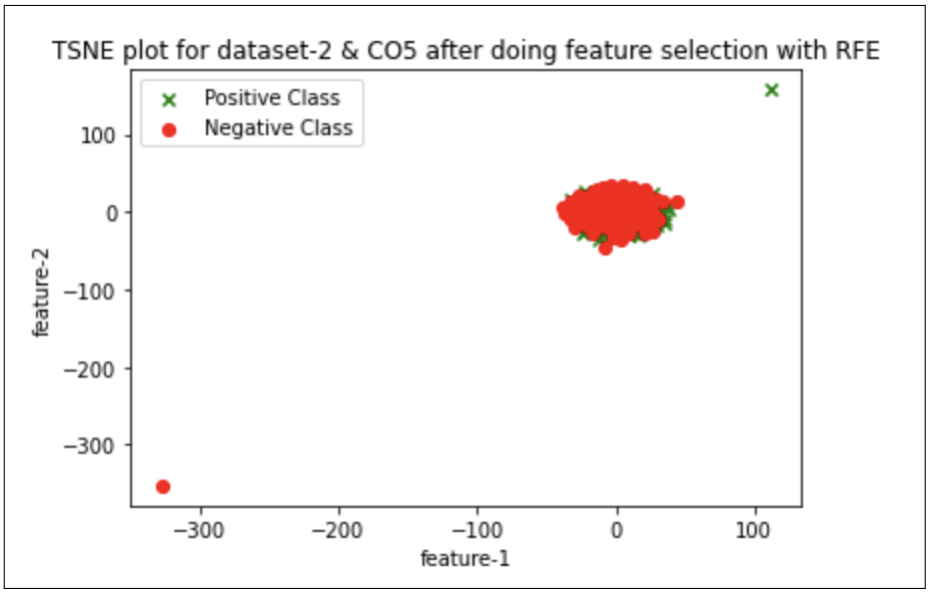


We further applied T-distributed Stochastic Neighbor Embedding (t-SNE) for clustering and visualization for all the predictors. For demonstration purposes in this work, we are showing graphs only for `co5` predictor.

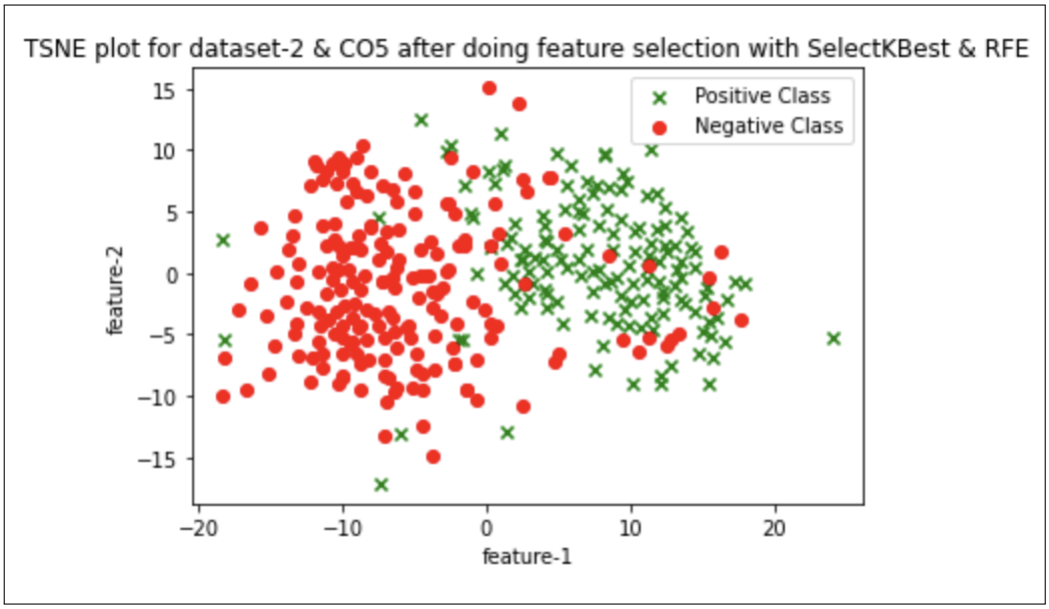
The following plot is obtained after applying the t-SNE dimensionality reduction technique directly on the dataset-2 without any preprocessing and feature selection:



The following plot is obtained after applying the t-SNE dimensionality reduction technique after applying Recursive feature elimination on the dataset-2:



The following plot is obtained after applying the t-SNE dimensionality reduction technique after applying `SelectKBest` feature selection technique and Recursive feature elimination on the dataset-2:



Using the above plots, we concluded that feature selection techniques like `SelectKBest` and Recursive feature elimination together could help to separate the feature space of the data and have the potential to improve the model's performance.

3. (5 points) [Brief introduction/motivation: Describe briefly the reason behind choosing a specific method for predicting each of the descriptors (could show plots or tables summarizing the scores of training different model)]

Solution:

1. METHOD A:
   Used for **Descriptor 1** (*co1*), **Descriptor 2** (*co2*), **Descriptor 3** (*co3*)

   (a) DATA PREPROCESSING:
   Since $\mathbf{p} >> \mathbf{n}$ (where **n** is the number of training samples and **p** is the number of features), dimensionality reduction is the first step that springs to mind. We use **Principal Component Analysis - PCA** to reduce the number of features. As $\mathbf{p} > \mathbf{n}$, we need at-most **n** principal components to represent the data. However, we only keep the principal component features which explain **90%** of the variance in the data.

   (b) FEATURE SELECTION:
   Although we have considerably reduced the number of features, not all of them are useful in predicting the targets. In order to extract the useful features, we perform feature selection. to perform feature selection, we will use a wrapper type feature selection technique. After going through literature on the feature selection in micro-array gene expression data, we narrowed down on Recursive Feature Elimination with SVM as our method. Not only was it shown to effective but also easy to implement and understand. **Recursive feature elimination**, in short **RFE**, is a greedy optimization algorithm which aims to find the best performing feature subset. At each iteration, it trains a classification model (in this case SVM) and eliminates the worst performing feature until all the features are exhausted or a stopping criteria is satisfied.

(c) CLASSIFICATION MODEL:
We tested different classification models on our processed data. We obtained the following results.

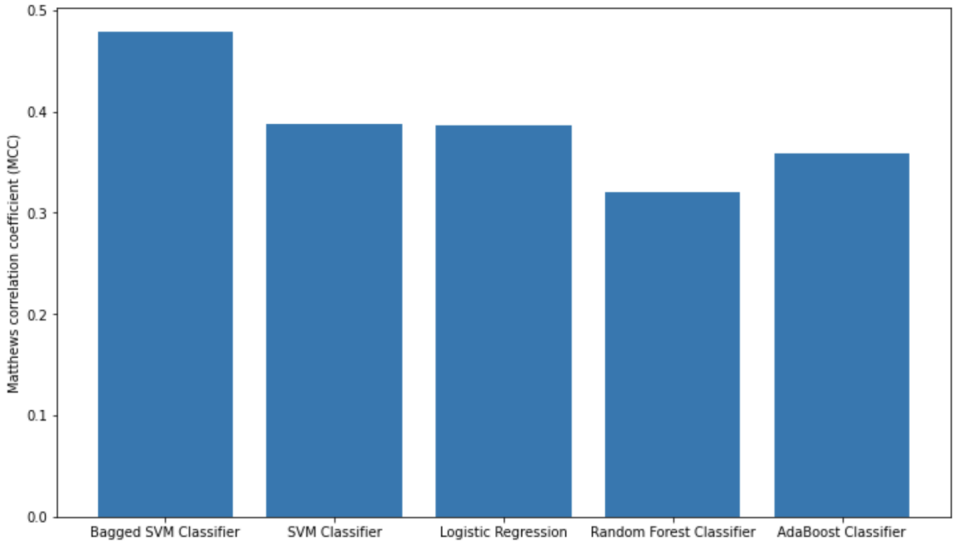


Figure 1: Performance of Different Classification Models on *co1*

Although SVM outperformed all other classification models, it seemed to to overfit the data (giving a training accuracy score of 1). This motivated the use of bagging. Bagging averages the result of multiple independently trained high-capacity learners. SVM, a high-capacity learner is a good candidate for bagging. Bagging SVMs gave us a small increase in our validation score.

2. METHOD B:
Used for **Descriptor 4** (*co4*), **Descriptor 5** (*co5*), **Descriptor 6** (*co6*)
Method A performed poorly on Descriptors 4, 5 and 6. This could be attributed to the dimensionality reduction step where we try to represent 54,675 features by just 340 principal components. This incorporates a lot of noise inherent in the data. Instead, we perform feature selection first to eliminate the majority of the irrelevant features.

(a) FEATURE SELECTION 1: We can not directly use RFE-SVM on the raw data. Being wrapper type feature selection technique, it would not only be computationally expensive but also time consuming. Hence, we use a two-fold feature selection process. In the first step we use a filter type feature selection technique - **SelectKBest** which removes all but the **k** highest scoring features where the ANOVA F-value is used as the scoring criteria. We keep around 3000 features which have significant scores.

(b) DIMENSIONALITY REDUCTION: After removing all irrelevant features, we perform dimensionality reduction using PCA. We only retain the principal component features which explain **90%** of the variance in the data.

(c) FEATURE SELECTION 2: We now perform RFE-SVM to further narrow down on the subset of features to use (size of the subset being a hyper-parameter we tune).

(d) CLASSIFICATION MODEL: We again discover that SVMs out perform all classification models. To make our model more robust we use bagging.

4. (4 points) [As the data comes from a clinical setting, model interpretation is also an important task along with prediction. Describe briefly your preferred choice of methods if you are asked to determine the significant genes behind the regulation of the endpoints or restrict the number of features/genes.]:

Solution:
We would prefer to use wrapper type feature selection techniques to determine significant genes. A wrapper method overcomes the shortcomings of filter and embedded feature selection methods. Although simple and interpret-able, filter methods fail to capture the relationship of two or more features with the target, we might lose important genes in the process. Embedded methods like feature selection by LASSO regularization of Logistic Regressor or Linear SVM not only have low interpret-ability but also turned out to be unstable, giving different number of features every time you ran them.
RFE-SVM is an easy to understand and easy to implement wrapper type feature selection technique. This method has been extensively studied, and several modifications have been proposed to improve its performance. The only shortcoming is the methods is its high time complexity. In the clinical context where the risk of making an incorrect prediction is very high, high time complexity is a small price to pay.

5. (2 points) [Share your thoughts on each of the endpoints: whether easy/difficult to predict]:

Solution:
All ratings are relative to each other.

- *co1*: Moderate
- *co2*: Moderate
- *co3*: Difficult
- *co4*: Difficult
- *co5*: Easy
- *co6*: Very Difficult

6. (2 points) [Add any additional information: like challenges faced or some details that would help us to better understand the strategies that you have utilized for model development]:

Solution:
Challenges faced:

- **Very High Dimensional Data**: The data has many more features than it had samples. Machine learning algorithms are designed assuming the data has many more samples than it has features. Hence, we could not directly applying classification algorithms on the data set. Hence, data preprocessing and feature selection formed a very important part of the overall model.
- **Imbalanced Data**: Due to imbalance in the data, the model is biased towards the majority class. Oversampling methods like `SMOTE` and `ADASYN` did not seem to improve the model's performance. This again can be attributed to the high dimensionality of the data.
- **Noisy Data**: Majority of the features were not useful and added noise to the principal components when using PCA. We had to come up with an alternative approach to efficiently reducing the dimensionality.