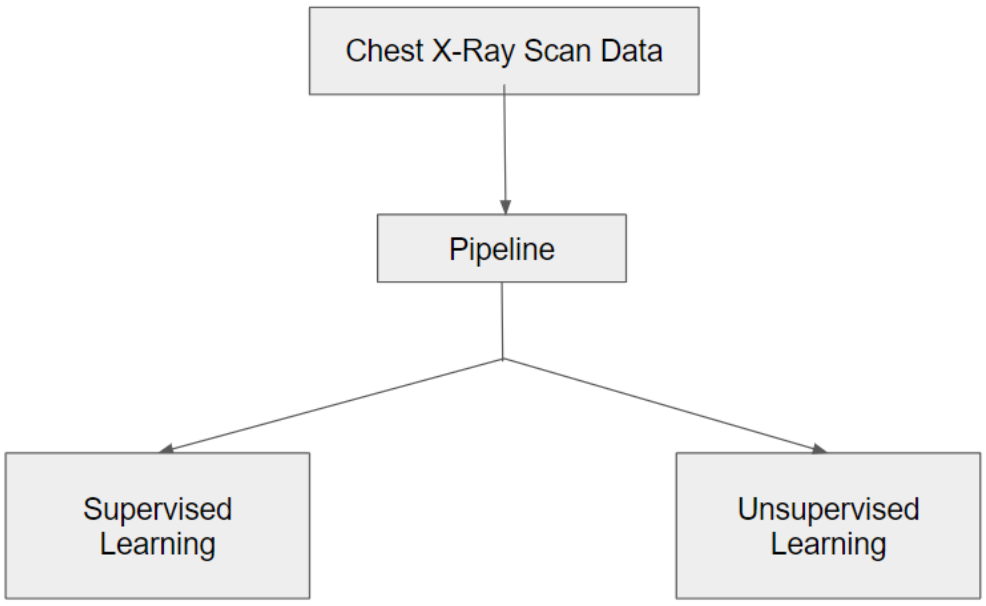
PRML Course Project: COVID Detection Using X-Ray Images

Abstract *– The global pandemic of COVID- 19 affected global health care and lifestyle, and its early detection is crucial for controlling the spread of cases and mortality. The current diagnostic leader is the Reverse Transcription Polymerase chain reaction (RT-PCR), but result times and costs are high, necessitating the development of additional rapid and inexpensive diagnostic instruments. This paper describes our experience with COVID-19 detection using chest X-ray (CXR) images. On CXR images, we attempt to implement unsupervised, supervised, and deep learning methods for binary image classification (COVID or Non-COVID). Data preprocessing pipeline included reduction techniques - PCA and LDA. We analysed models such as Random Forest classifier, Decision Tree classifier, GaussianBayes, and XGBoost in supervised learning. For the Deep Learning approach, we created a custom pipeline that utilises segmentation and CNN models. The best models obtained a COVID-19 detection accuracy of approximately 98%.*

I. Introduction

Classification is a learning technique in which the predicted values (categorical variables) belong to a specific number of classes. It is used for detecting fraudulent emails, recognising images, determining the presence or absence of a disease, and recognising digits. The output variable can belong to either two cases (binary classification) or multiple classes (multi-class classification).

II. Overview of the pipeline



III. Importing the Dataset

The dataset comprises of chest X-ray scans for COVID, non-COVID (including Viral Pneumonia) cases. The data has been derived from Github, SIRM, Github-CXNET and Eurorad.

Retrieval of data from these sources is performed using various parsing techniques (including wget and git cloning of repositories). The image data is then converted into numpy arrays using PIL.Image library and all the images are resized for uniformity in the standard (250*250) shape.

To create the dataset, all the images are combined in an array.

IV. Data Preprocessing and Analysis

The array is then converted to a Pandas dataframe. Class labelling (for supervised purposes) is done on the basis of the category of the dataset the image belongs to - COVID or Non-COVID. Each datapoint in this dataframe has 62500 attributes - the number of pixel values per image.

Considering the complexity of the data and the computational cost involved, the number of attributes in the dataset must be reduced. For this, dimensionality reduction techniques such as PCA and LDA are used.

The advantages of PCA are :

- Reducing the time and storage space required.
- Removing multi-collinearity which improves the interpretation of the parameters of the machine learning model. Considering the complexity of the data and the computational cost involved, the number of principal components is set as 20.

The advantages of ICA are :

- Separating a multivariate signal into independent sources that are not correlated..
- Does not require any prior knowledge of the underlying sources or mixing process.

Differences between PCA and ICA :

- The underlying assumption of PCA is that the dataset is obtained by a linear combination of uncorrelated variables. PCA aims at finding these uncorrelated orthogonal features, called as ‘principal components’, by use of linear methods.
- On the other hand, ICA assumes that the dataset is obtained from linear combination of independent variables. ICA finds out these independent variables by using non-linear methods.

Considering the computational cost involved, relevance of the dataset and the overall impact of the transformation, the number of attributes are reduced to 20 using these reduction techniques.

V. Supervised Learning Techniques

Gaussian Naive Bayes: It is based on Bayes' theorem. It is a probabilistic method that assumes independent Gaussian characteristics. It is "naive" because it thinks features are independent, which is not always true. The algorithm chooses the class with the highest probability based on input features. Gaussian Naive Bayes is fast and easy for high-dimensional data.

Decision Tree Classifier: It does classification and regression using supervised learning. The decision tree starts at the root node and displays a feature with branches representing its possible values. Using feature values, the algorithm chooses a path at each node. The process continues until a leaf node makes the ultimate decision. Interpretability and ability to handle category and numerical data make the decision tree classifier quite useful in these scenarios.

Random Forest Classifier: It is an algorithm for supervised learning that applies ensemble learning for classification. Ensemble learning is a technique that integrates the predictions of multiple machine learning algorithms to produce predictions that are more precise than a single model. Decision Trees would represent these algorithms in the case of Random Forest.

XGBoost: It is a popular machine learning algorithm for classification and regression tasks. Xtreme Gradient boosting creates a strong learner from numerous weak learners. Iteratively adding trees that minimise error optimises the loss function using decision trees as weak learners. Industry uses XGBoost for fraud detection and customer churn prediction due to its speed and accuracy.

VI. Unsupervised Learning Techniques

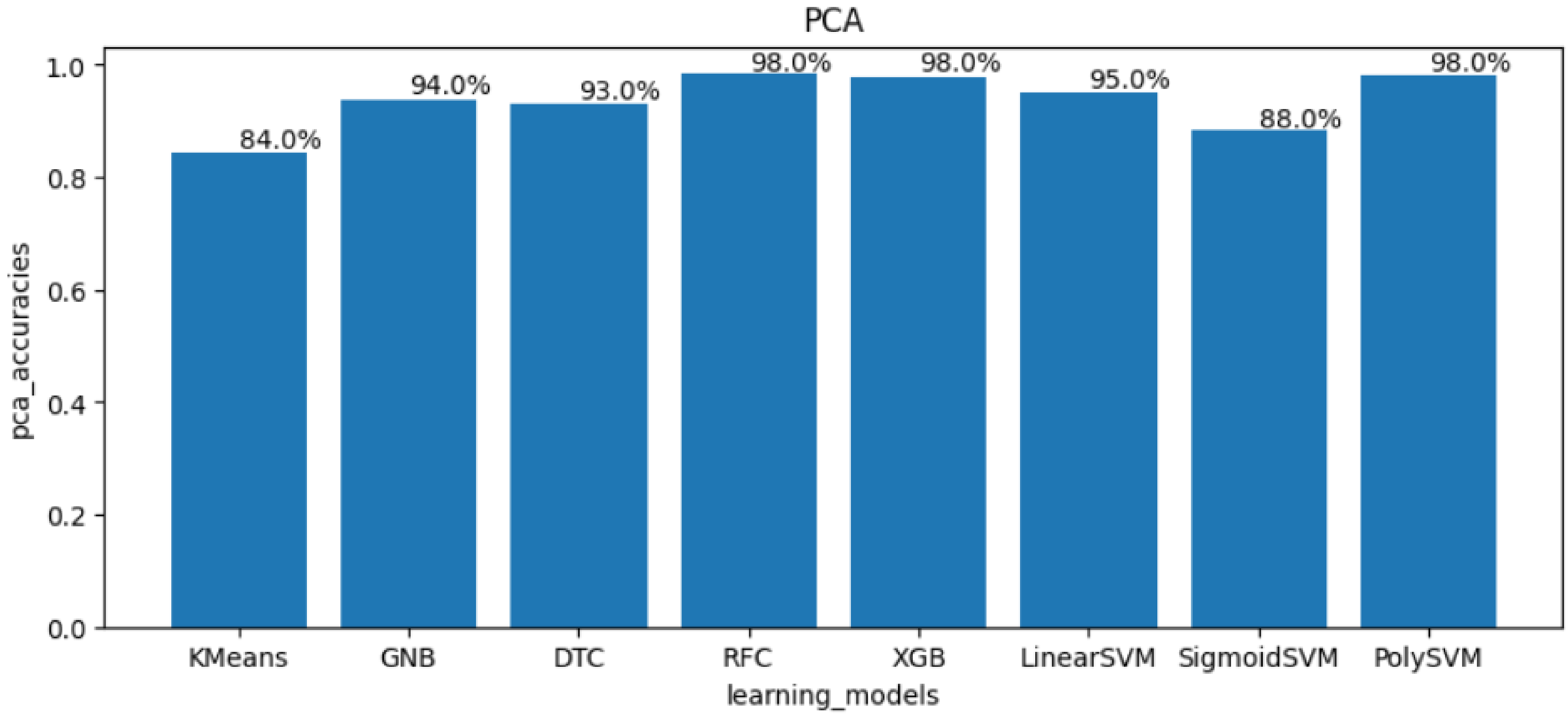
K Means: It organises data points into k clusters based on their similarity or distance. The algorithm initializes k centroids, assigns data points to the closest centroid, updates centroids, and repeats until convergence. It assumes spherical clusters, identical size, and similar densities and converges the clusters to local optimum condition.

VII. Comparative Study of Models

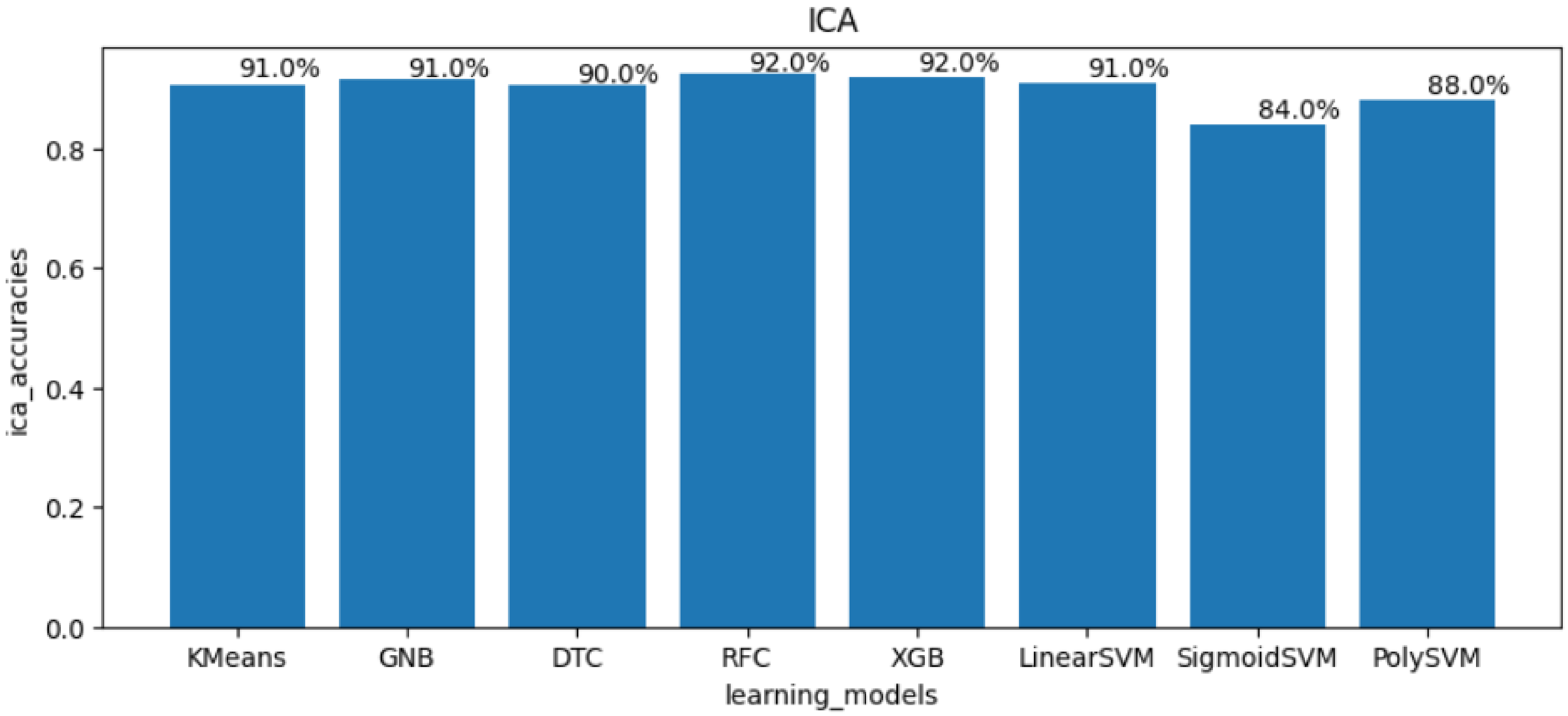
The comparison of these models is done by means of performance metrics such as accuracy score, precision, and recall. As for the clustering models (KMeans), we shall use the accuracy and Silhouette score as a metric.

The analysis of these variations will be done using a bar plot.

For PCA dataset -



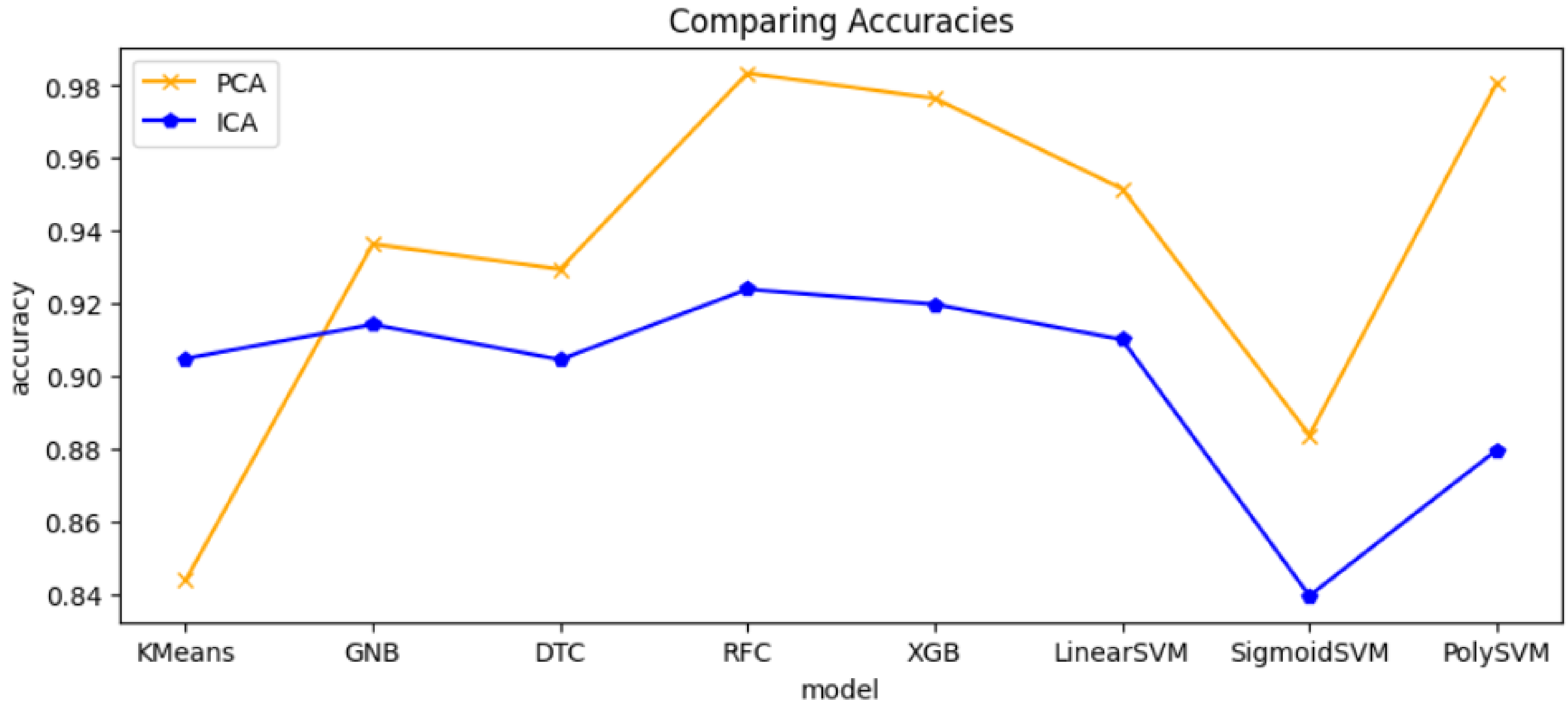
For ICA dataset -



The graph shows how the performance of these models varies. We can see that for almost all the models, we have received practically great results.

For the combined analysis of both the datasets, we can refer to the graphical representation below -

Combined analysis of performance -


Comparing Accuracies
PCA
ICA
accuracy
KMeans
GNB
DTC
RFC
XGB
LinearSVM
SigmoidSVM
PolySVM
model