# ASSIGNMENT 2

**Group Details:**

#### Group - 002

20CS10039 – Nikhil Saraswat

20CS30003 – Amit Kumar

November 1, 2020

# Procedure:

1. We have used the standard K-Means algorithm for clustering in question of this assignment. The data used for it was provided in wine.data file. We have shuffled the data. We have handled missing values with most frequent values.

The attributes are as follows: -

* Alcohol
* Malic acid
* Ash
* Alcalinity of ash
* Magnesium
* Total phenols
* Flavanoids
* Nonflavanoid phenols
* Proanthocyanins
* Color intensity
* Hue
* OD280/OD315 of diluted wines
* Proline

1. **Some important Helper functions:**

**train\_test\_split:** This function splits the data into two sets train and test **load\_csv:** This function reads a .csv file and stores the data in a row-format **cross\_validation\_split:** splits the input data and returns the indices of the validation set for each iteration of cross-validation

**accuracy\_metric:** evaluates the accuracy of predicted values with respect to the actual values. Accuracy is defined as the number of correct predictions divided by total length of predictions

**summarize\_dataset:** it returns the mean, standard deviation, number of entries, threshold value for outlier for each column of the dataset

**calculate\_probability:** it returns the probability density of the data point on a Gaussian with given input mean and standard deviation handle\_missing\_data: this function imputes the missing value(s) of a feature with the most\_frequent value of the feature

**encoded\_dataset**: this function encodes the categorical columns in the dataset with LabelEncoder from scikit-learn

**sequential\_backward\_selection:** this function uses sequential backward selection method to remove features

**remove\_outliers:** this function removes those samples from dataset which contain the number of outlier features more than the parameter 'outlier'

prepare: this functions preprocesses the dataset and prepares for splitting and training purposes. It imputes the missing data, encodes the categorical values, converts each entry into float and normalizes each feature.

**separate\_by\_class:** this function is used to split the input dataset as per the classes defined by the target class. The target class will be the last column of the dataset.

**calculate\_class\_probabilities:** Helper function to calculate the probabilities of each class given the current data value. This function helps in calculating the a posteriori for each class.

**predict:** make predictions for the given vector of data (row) and return the class label that should be assigned to this data element

## Principal Components Analysis (PCA)

* 1. **Principal component:** component along the direction **w** such that its variance is maximum among all possible projections. Additionally, the subsequent principal components are perpendicular to the prior principal components.
  2. **X** is the random variable whose instance is **x** with mean **m** and covariance matrix **Σ**. Then the first component is evaluated by the following optimization problem (where *l* is Lagrangian coefficient):

*w*1 =

*argmaxw*{*wT* Σ*w*

− *l*.(*wT w*

− 1)}

Σ.*w*1 = *l*.*w*1

* 1. The first component is evaluated by the following optimization problem (where

*l*1 and *l*2

are Lagrangian coefficients ):

*w*2 =

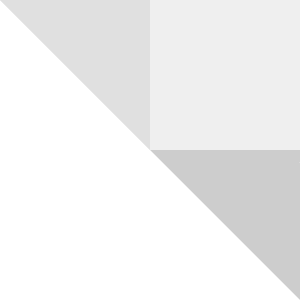
*argmaxw*{*wT* Σ*w*

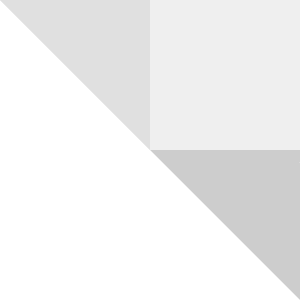
− *l*1.(*wT w*

− 1)

− *l*2.(*w*2*T w*1 −

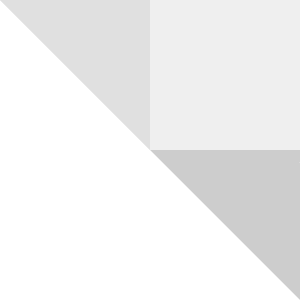
0)}

Σ.*w*2 = *l*1.*w*2



# Results

### Handling missing values

The missing values are handled by replacing the **most\_frequent value** for categorical attributes and **mean value** for the continuous attributes.

### Application of PCA

PCA is applied to the dataset and the required graphs were plotted. We observed that for preserving 95% of total variance, 10 principal components were captured. The required plot is attached below, which justifies our statement:

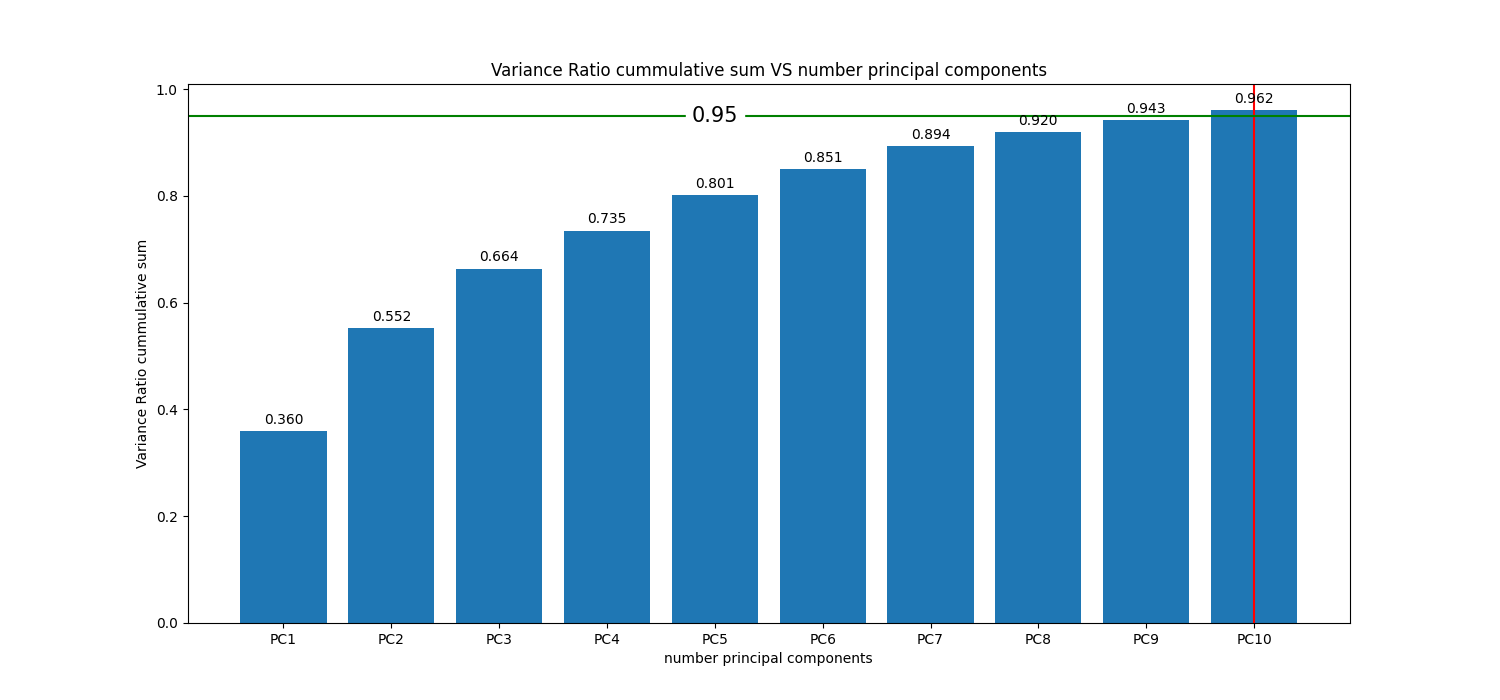


FIG 1: Variance ratio cumulative sum vs the number of principal components

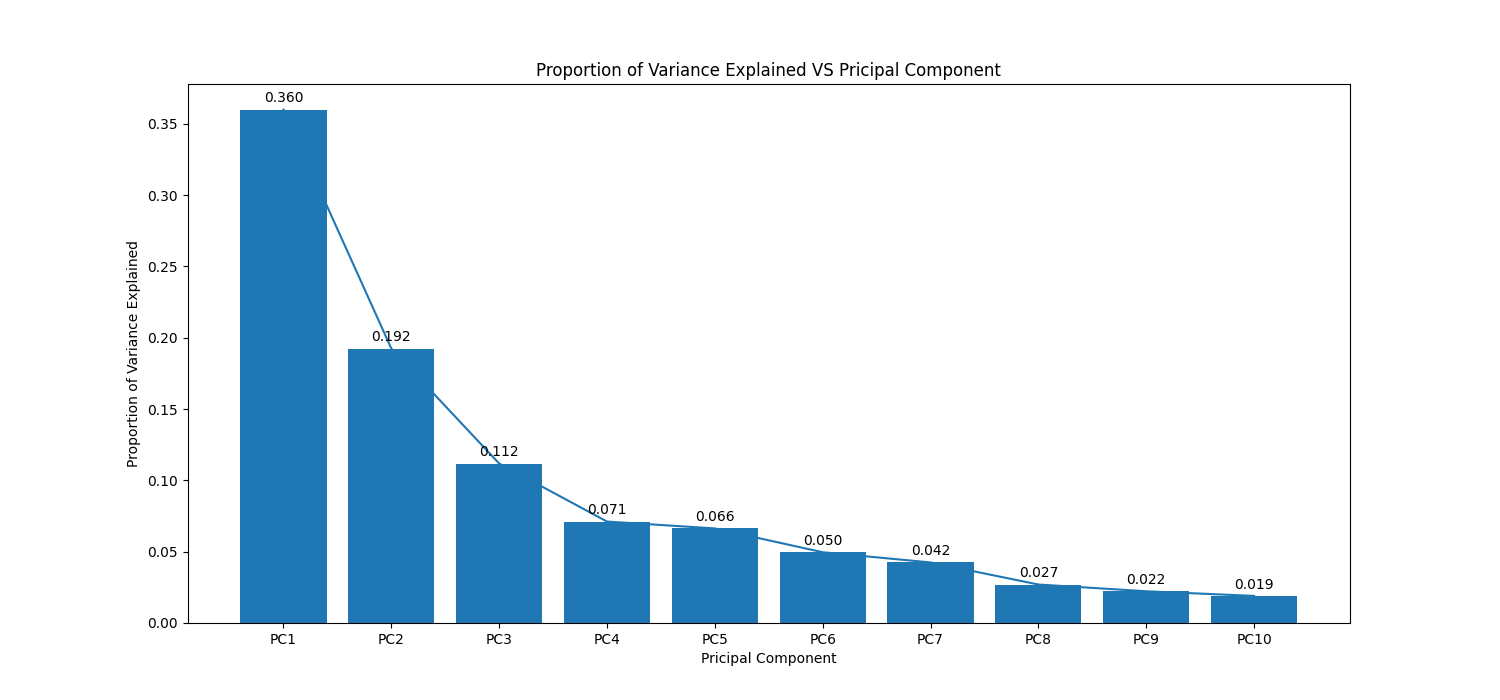


FIG 2: Proportion of variance explained vs Principal Component

