Forest Cover Type Project

Module : CS 6405 Data Mining

Haoyu Wang (116220323)

Abeer Almabdi(116221065)

Introduction

This data set is the Colorado vegetation type data, the only one care about the real data of forest. Each record contains many indicators describe each piece of land in **Roosevel National forest, Colorado,USA**.

Dataset Description

- A total of **581012** records. Each record has **55** column, one of the column is the type of soil, other **54** column is the input features.

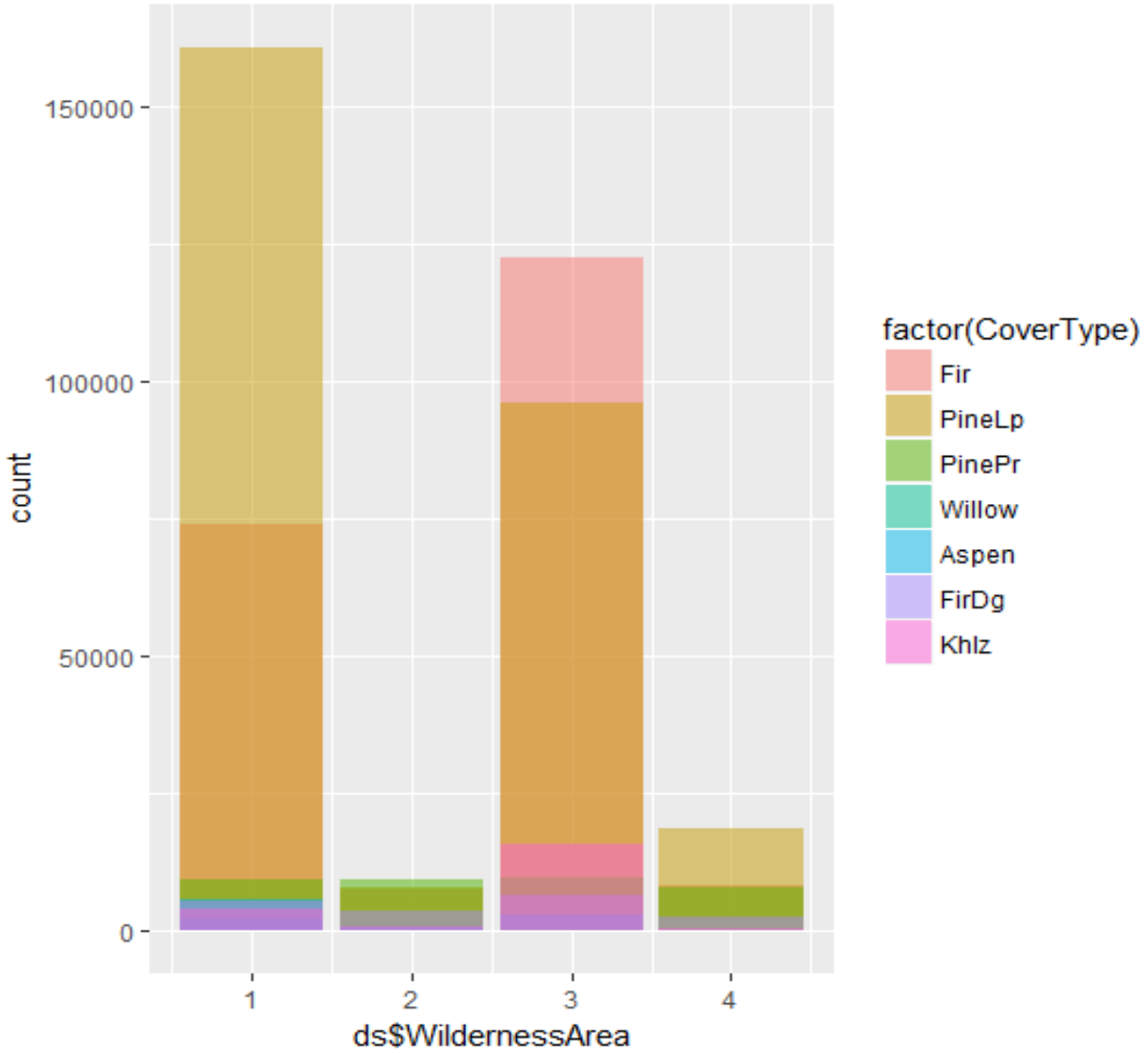
Soil type (x40) | Wilderness area (x4) | Elevation | Aspect | Slope | Horizontal Distance To Hydrology | Vertical Distance To Hydrology | Horizontal Distance To Roadways | Hillshade 9am | Hillshade Noon | Hillshade 3pm | Horizontal Distance To Fire Points

Cover Types

Spruce/Fir
Lodge Pole Pine
Ponderosa Pine
Cottonwood
Willow
Aspen
Douglas fir
Krummholz

We can see there are 7 different types of trees distributing in the following four Areas.

Dataset Description



Cleaning Data

We will apply those 10 numeric variables in next section.

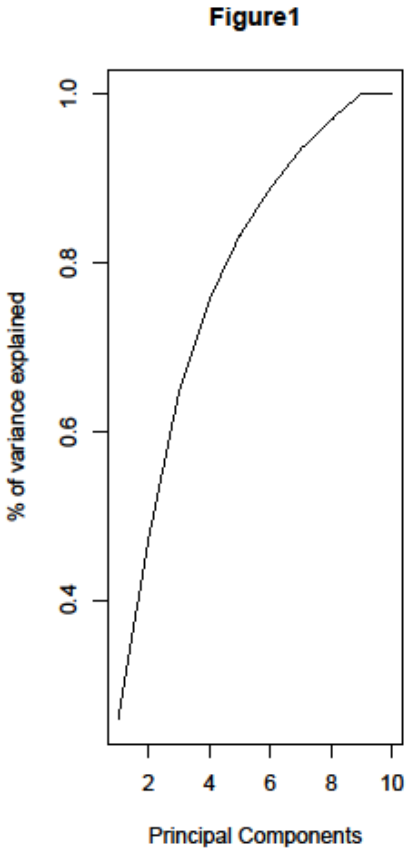
remove any empty or NAs row in this data-set

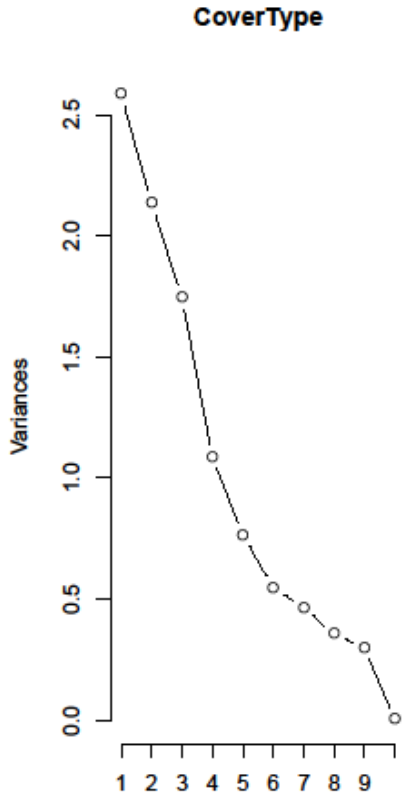
Soiltype (x40) →Soiltype (x1)
Wilderness area (x4) →Wilderness area (x1)

Rename CoverType and remove original columns

Finally we get 10 numeric variables and 4 factor variables

Dimensionality Reduction
PCA



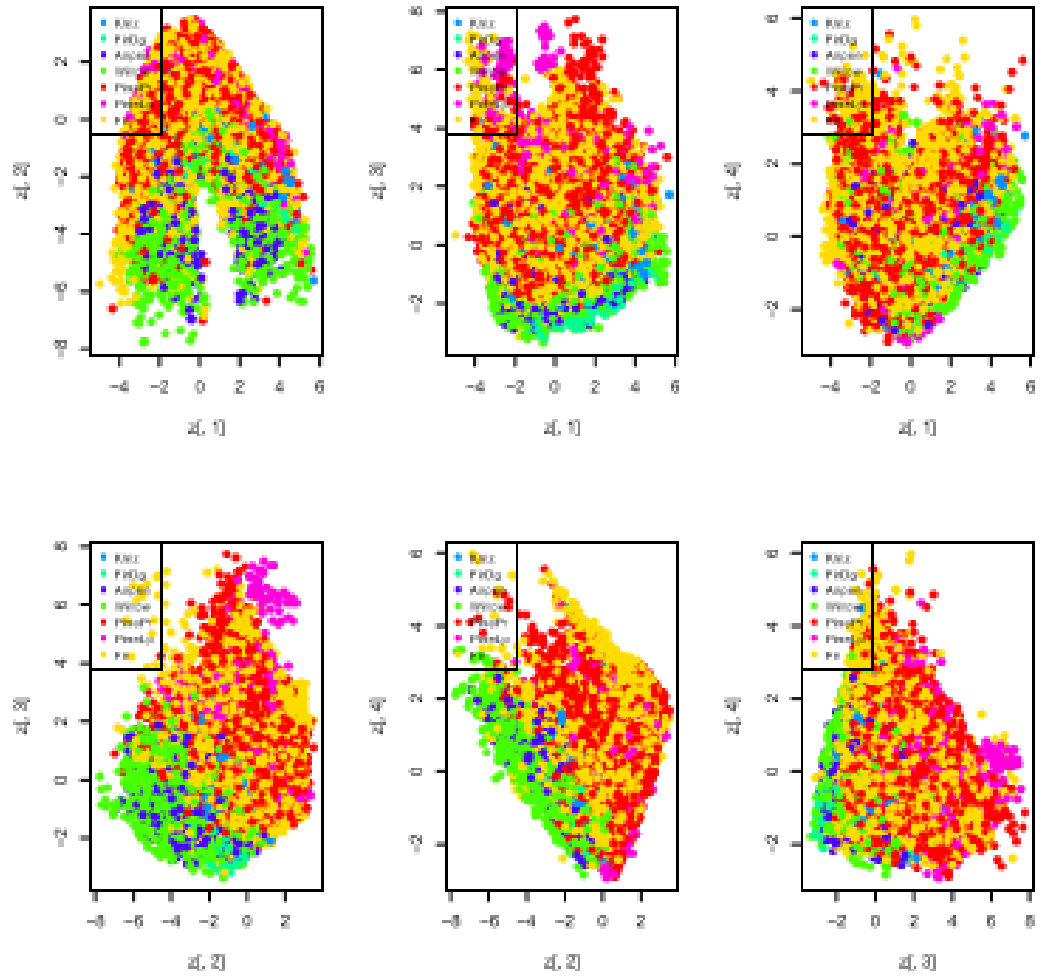


cumulative probability of principle components

explained variances by each principal components

Dimensionality Reduction
PCA

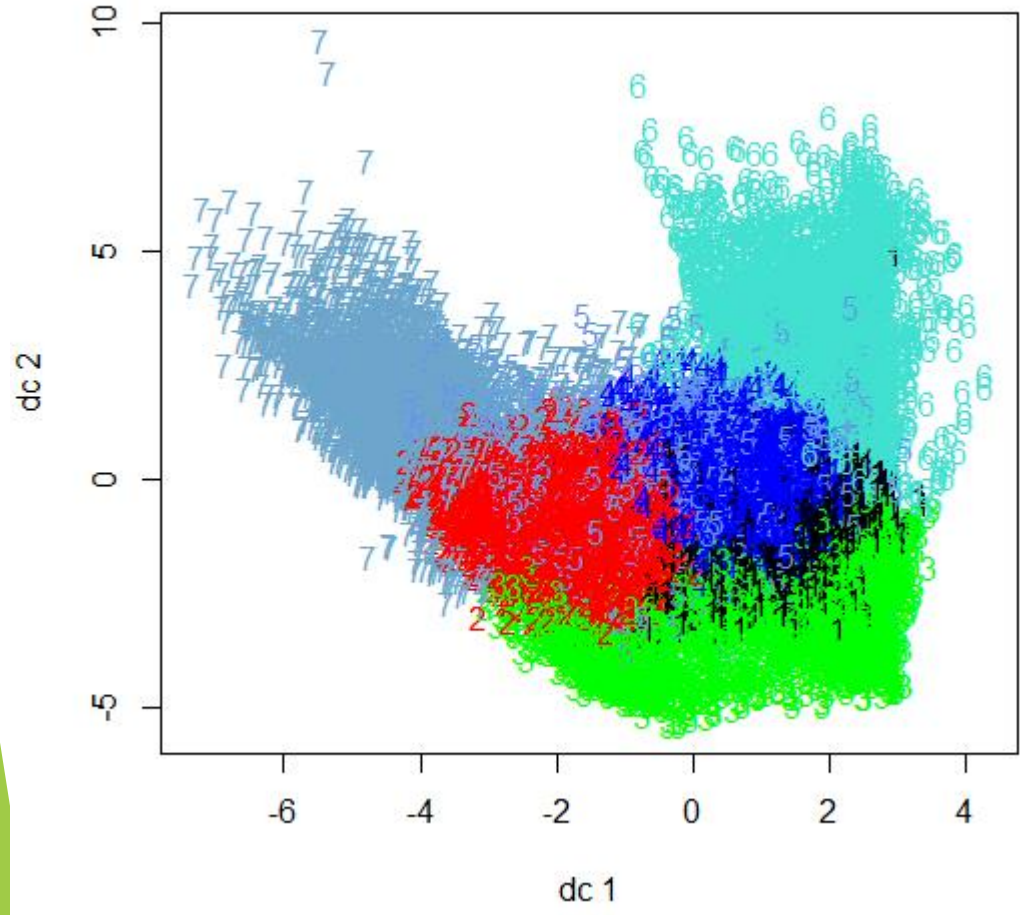
Pairwise Scatterplot of principle components 1-4



- the first Five principle components explained over 80% variances.
- the data would be essentially 10-D data, spread out in 5-D space.

In the next section, we will choose first 5 components and do success rate comparison between within without PCA and different sample size of dataset.and see what will happen.

Clustering
K-Means



Sample size 50000

Dataset	Success rate
Normal	0.18588
After PCA	0.11598

Sample size 100000

Dataset	Success rate
Normal	0.10339
After PCA	0.09929

If Sample size increasing
Accuracy rate decreasing

K-means clustering focuses on partitioning n observations into k clusters
We can see here Sample size incresed, the correct rate decreased and its not normal.

when we processing data after PCA, rate will lower then previous
We can see that when we apply PCA to data will decrease variables and it will affect the clustering result.

Sample size increased to 100000, we can see correct rate of LDA only have small increment from 0.68676 to 0.68115.And around 2% increment in QDA.

After PCA: We used first five components, and LDA performed not well, decreasing around 8%, and QDA even increase slightly.

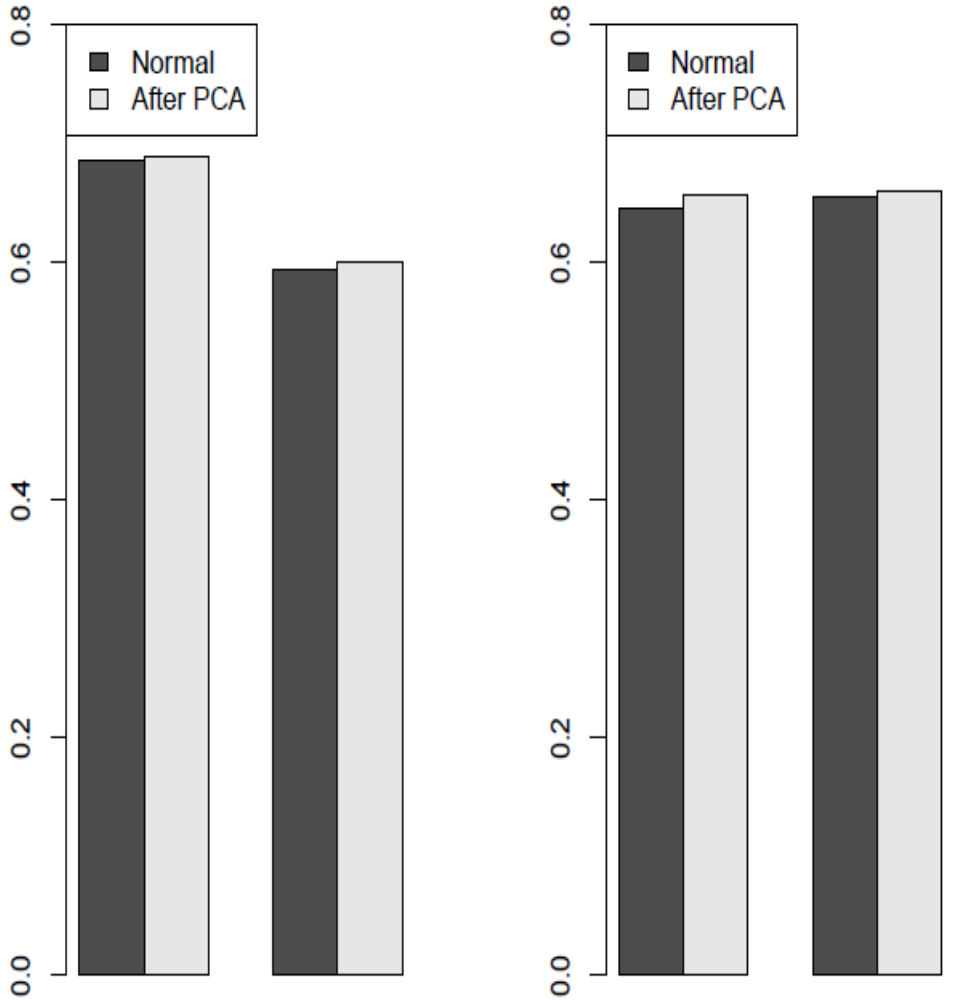
When sample size increase to 100000, correct rate of LDA decrease from 0.68115 to 0.6 and QDA keep stable around 0.6596.

Classification

- **LDA & QDA**

Sample size 50000	Success Rate	Processing Time
LDA	0.68676	0.4142950
QDA	0.6455987	0.3662629
LDA(after PCA)	**0.5968**	0.4403100
QDA(after PCA)	**0.650653**	0.3772540

Sample size 100000	Success Rate	Processing Time
LDA	0.68115	0.9176369
QDA	0.6565	0.9266598
LDA(after PCA)	**0.6**	0.7845550
QDA(after PCA)	**0.6596**	0.7015018



Choose training-set Manually which means we choose first 35000 and 750000 rows data without shuffling

KNN

We can see when sampling size increase to 100000, success rate are increase around 4%. But when dataset is processed by PCA, success rate have largely decrease around 15%.

Manually selection **VS.** Randomly selection

We choose 70% data from dataset randomly become Training data rather then manually selected.

Sample size 50000	Success Rate	Processing Time
KNN Manually	0.8453333	9.9020591
KNN Randomly	0.8438	9.4957359
KNN Manually(after PCA)	**0.6956**	**6.9109299**
KNN Randomly(after PCA)	**0.6902**	**7.4102521**

Sample size 100000	Success Rate	Processing Time
KNN Manually	0.88757	38.8315940
KNN Randomly	0.887	37.8809490
KNN Manually(after PCA)	**0.7212**	**30.6508050**
KNN Randomly(after PCA)	**0.72193**	**31.6665480**

Success rate is lower after PCA

Processing time on data set which reduce dimension is shorter after PCA

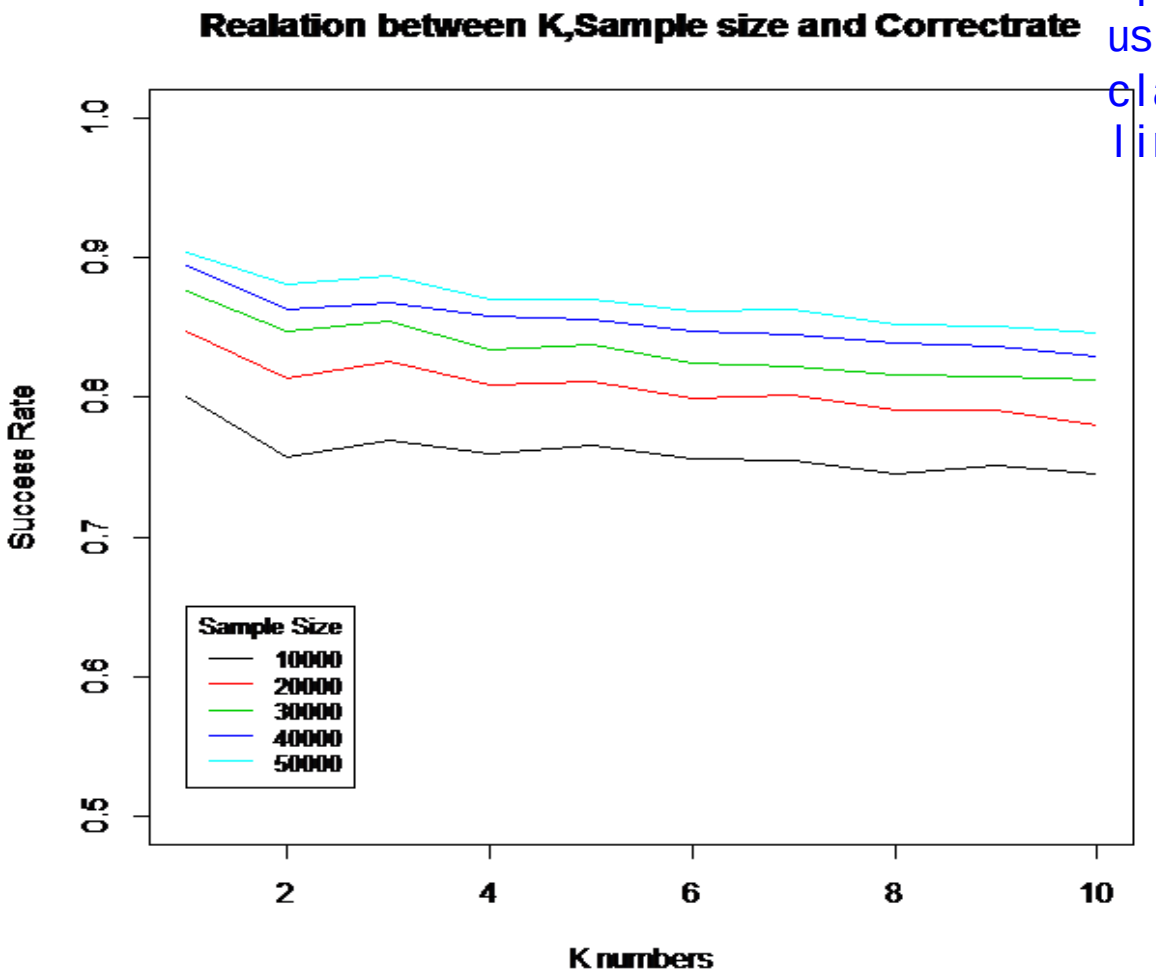
KNN

we can see sample size have strong relationship with success rate in KNN

- KNN Cross-Validation

the success rate will be decreasing with increasing of K.

if we want high success rate, one possible way is using huge data set. From this plot, we conclude classification correct rate is negative linearrelationship with K numbers.

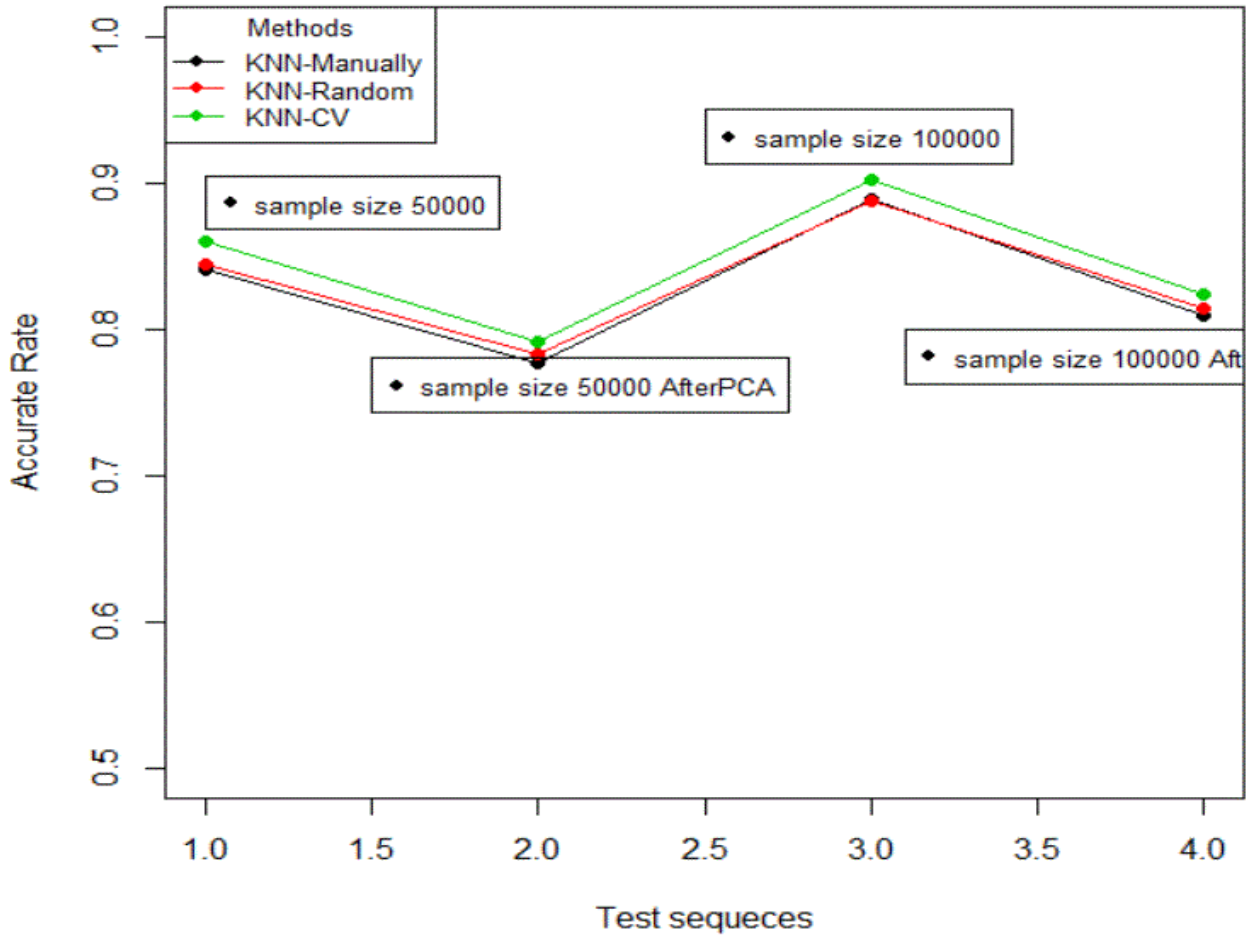


We choose the best performance model (KNN-CV) use **5** different sample size and 1:10 k numbers to test the relationship between **K** ,**success rate** and **sample size**.

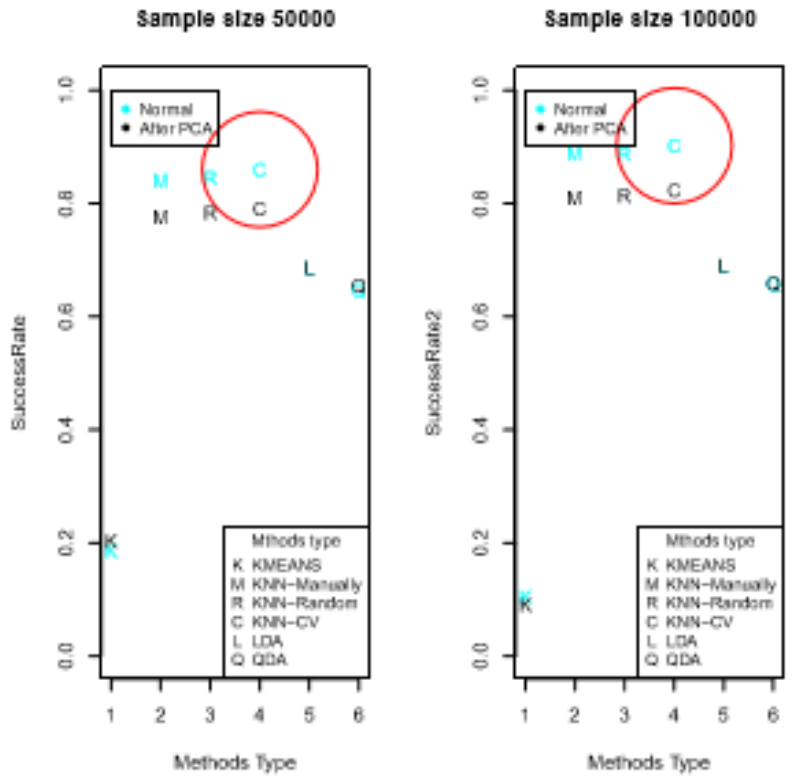
KNN

- Compering between KNN methods

We can see three different KNN have similar performance, so we suggest manual choosing,random choosing, Cross validation have same success rate when sample size is enough big.



Result & Discussion

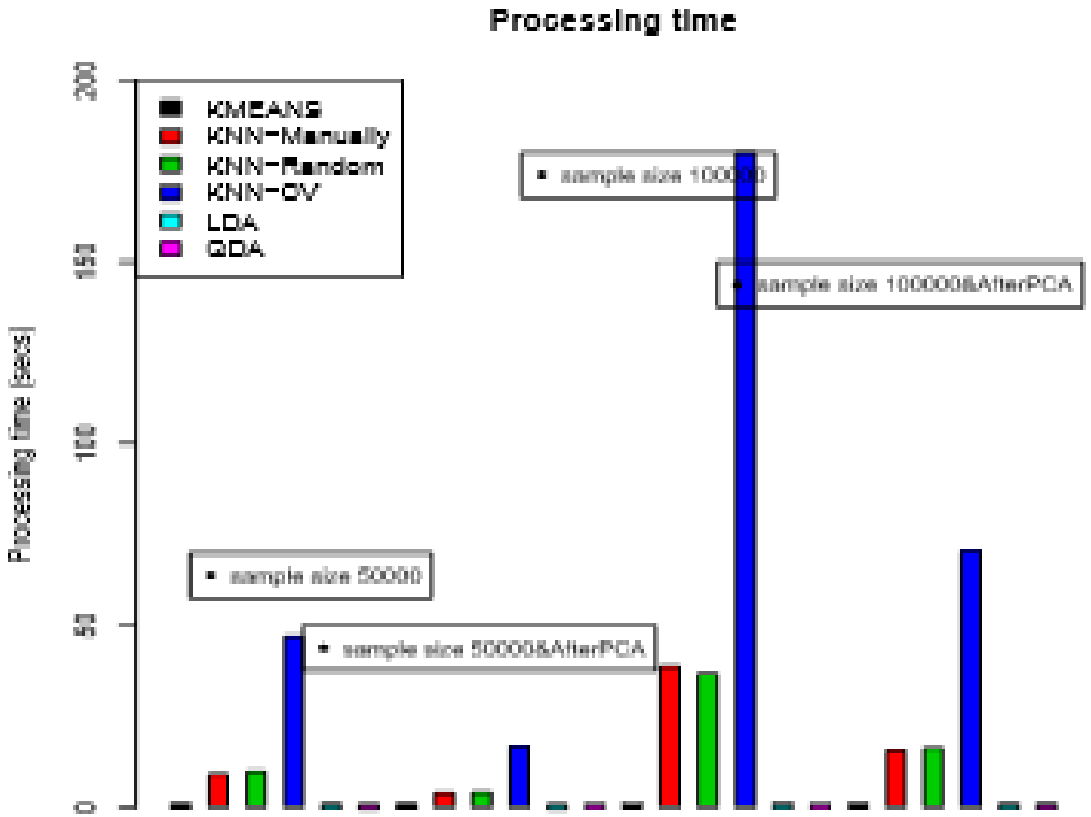


KNN-CV classifier and KNN-Random algorithms performed quite well with 90% and 88%. While clustering algorithm K-means is the lowest around 10%

The best methods is
KNN Cross-Validation

Result& Discussion

We can see from here, Cross validation use longest time to process data. And PCA is benefit for reduce processing time.



- If Sample Size increasing, processing time increasing
- majority of algorithms which process data-set **after PCA** will be faster than the normal data-set

Conclusion

- In this project we can see KNN Cross validation is the best model.
- PCA only have small effect on LDA and nearly no impact on QDA in this experiment.
- PCA not only can largely decrease processing time but also reduce success rate.

Thank you for your attention

Any Questions