Classifying Cancer

Computational Research Project Using Machine Learning Methods to Correlate Genes with Lung & Prostate Cancer

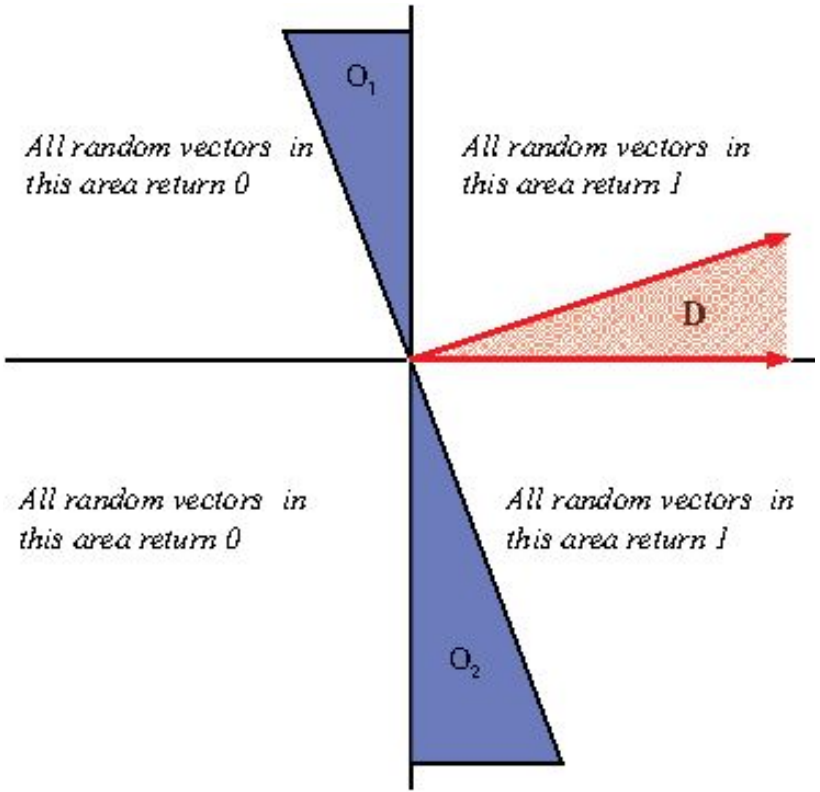
Ria Doshi

Data

Leukemia	Lung Cancer	Prostate Cancer
Acute myeloid leukemia (AML) (25 samples)	Adenocarcinoma (AD): 139 samples Normal lung (NL): 17 samples Small Cell Lung Cancer: 6 samples	Normal tissue: 50 samples
Acute Lymphoblastic Leukemia (47 samples)	Squamous cell carcinoma (SQ): 21 samples Pulmonary carcinoid (COID): 20 samples	Prostate tumor: 52 samples

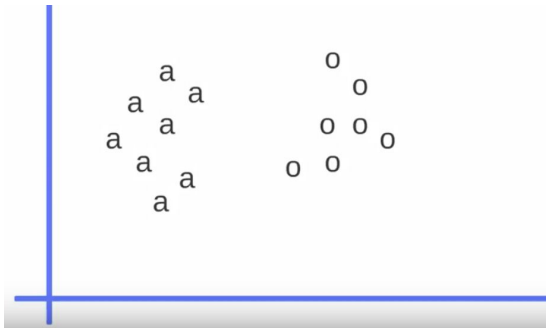
Methods

Random Hyperplane Projection

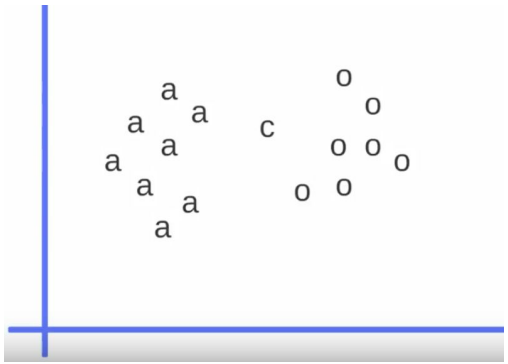


K-nearest Neighbors

1) Training Data is plotted.

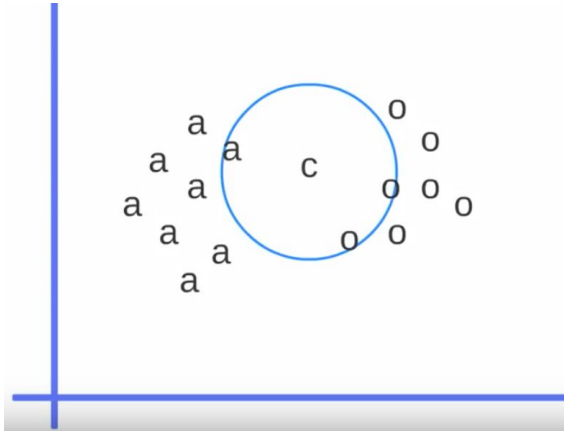


2) Choose a new input value, let's say "c" and plot it onto the graph, ignore labels.

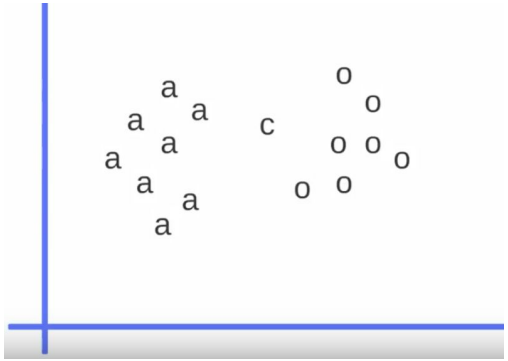


K-nearest Neighbors

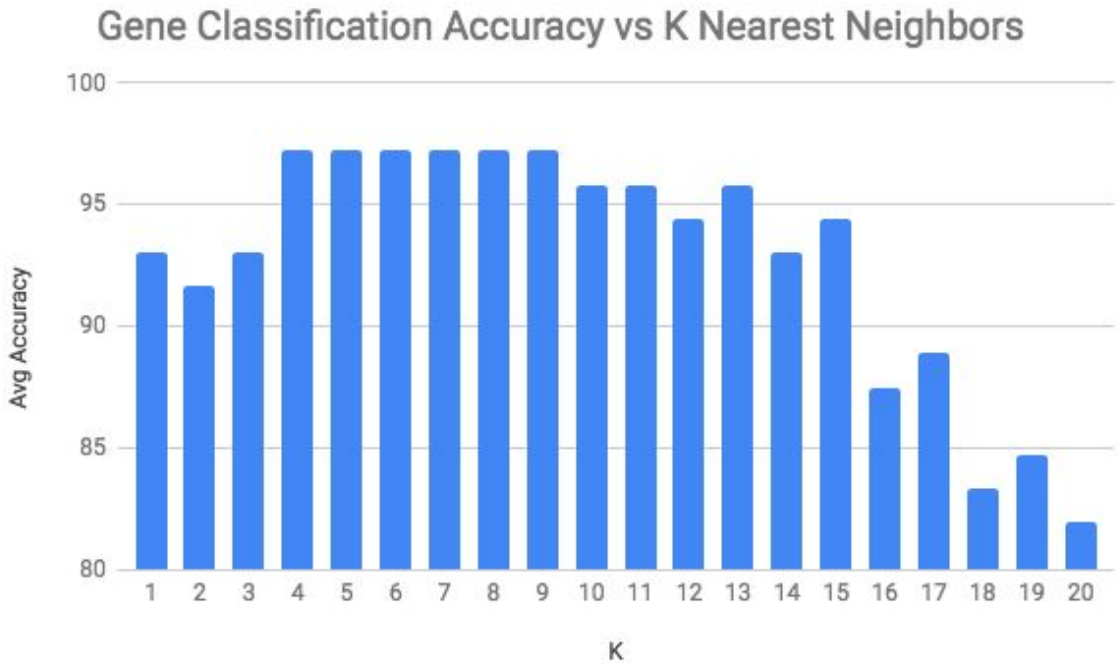
3) Set an integer value “k” and find the k number of closest points to the new input value.



4) Look at the “k” number of nearest points to “c” and choose whichever classification is in majority.



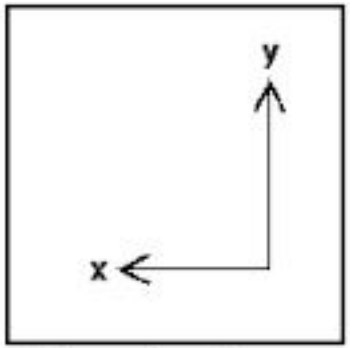
K-nearest Neighbors



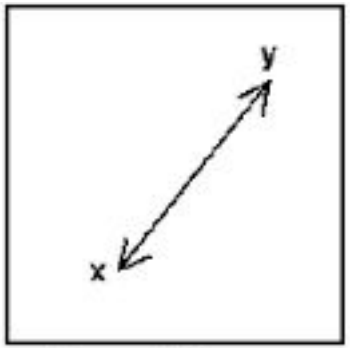
-When we set the value of “k” from between 4 to 9, we got an **accuracy score of 96%**

K-nearest Neighbors

- Method for regression and classification
- Finding distances between points
- Implemented with K = 5 or 7
- 96% accuracy

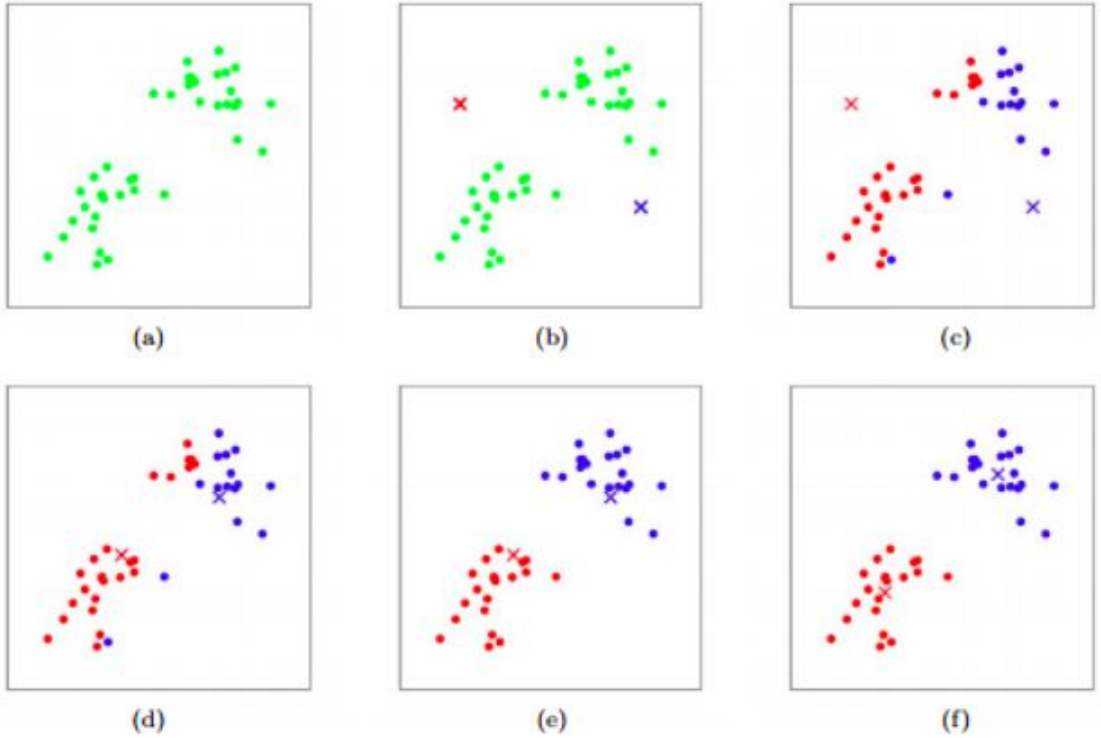


Manhattan

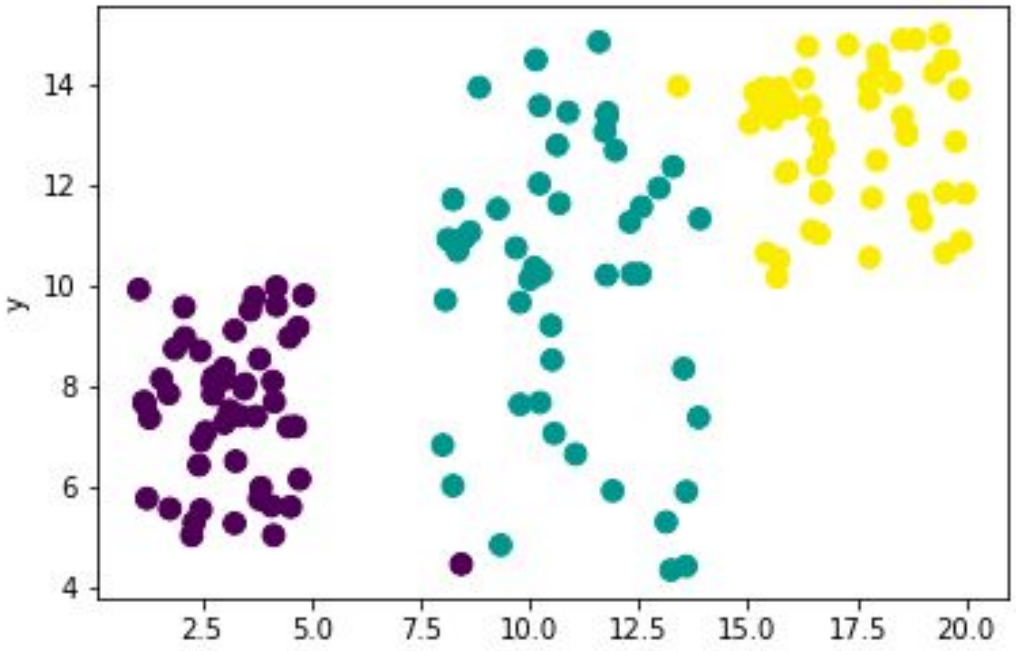


Euclidean

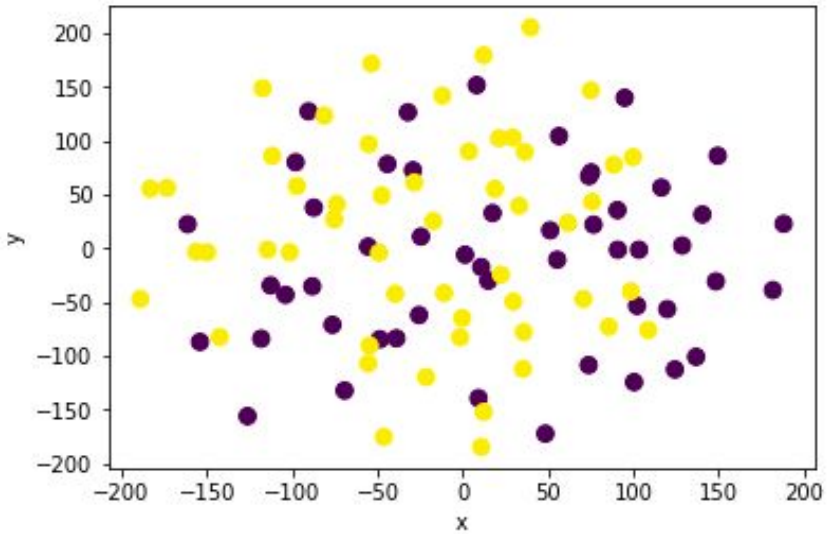
K means



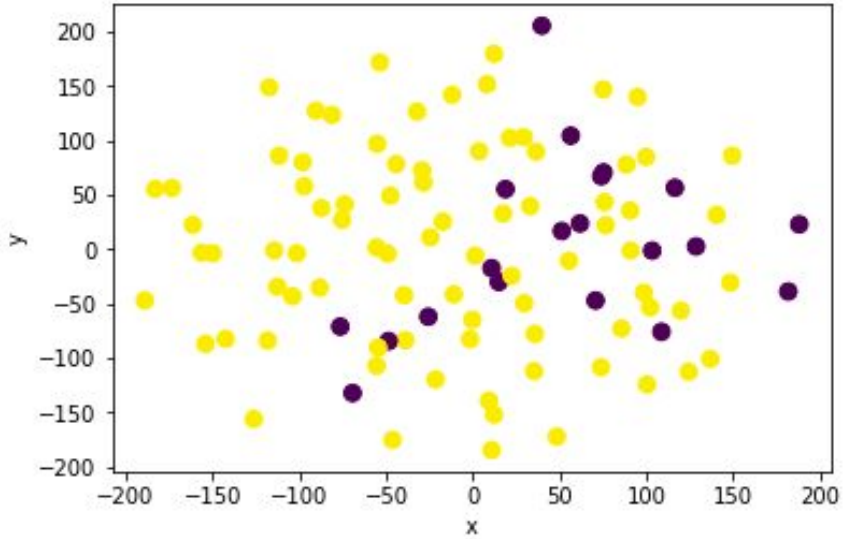
K means



K means

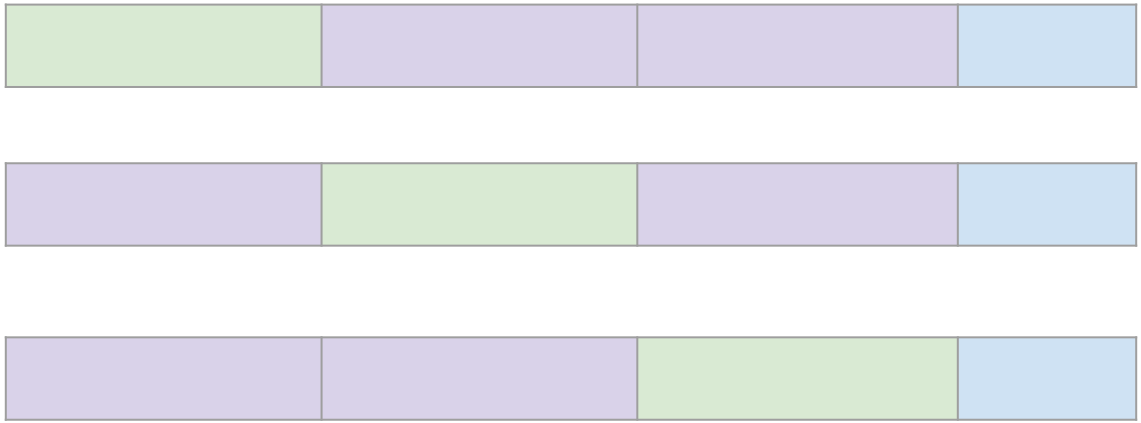


Actual Labels



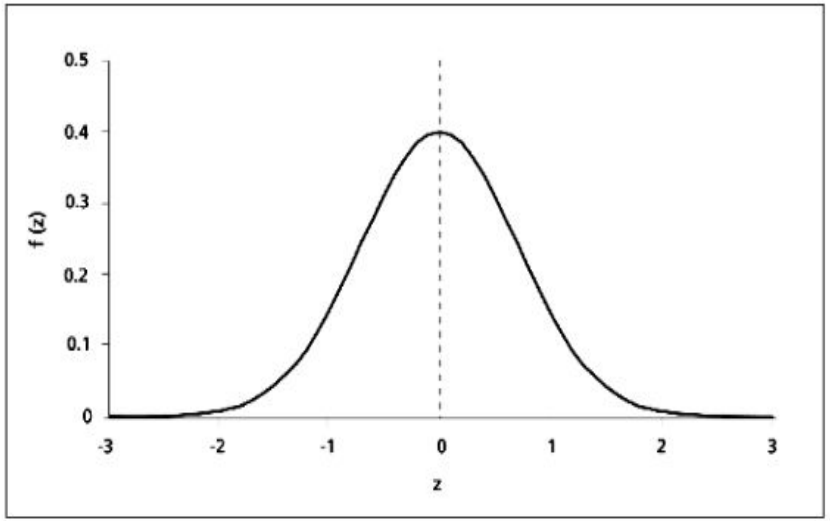
K-Means

Cross Validation



	Validation
	Test
	Train

Noise & Normalization



$$\text{2D: } |\mathbf{v}| = \sqrt{x^2 + y^2}$$

$$\text{3D: } |\mathbf{v}| = \sqrt{x^2 + y^2 + z^2}$$

Statistics

$$t = \frac{\overline{X} - \mu}{\frac{s}{\sqrt{N}}}$$

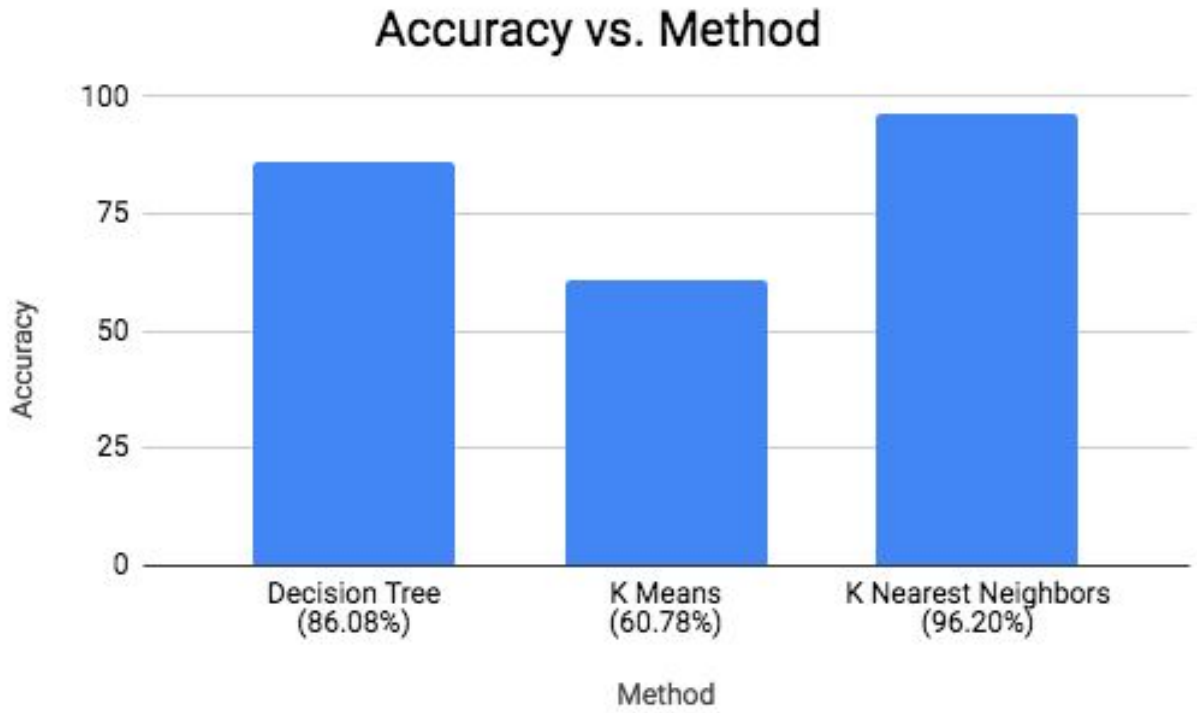
ONE SAMPLE T-TEST

TWO SAMPLE T-TEST

$$t = \frac{\overline{X}_1 - \overline{X}_2}{\sqrt{\frac{s_1^2}{N_1} + \frac{s_2^2}{N_2}}}$$

Results

Accuracies


Accuracy vs. Method
100
75
50
25
0
Accuracy
Decision Tree
(86.08%)
K Means
(60.78%)
K Nearest Neighbors
(96.20%)
Method

Accuracies

1317

Informative Genes

4-9

Neighbors

3

Folds

Classifying Cancer

Computational Research Project Using Machine Learning Methods to Correlate Genes with Lung & Prostate Cancer

Ria Doshi