

Pattern Recognition Classification Project

BY

HARIKANTH GHANTA

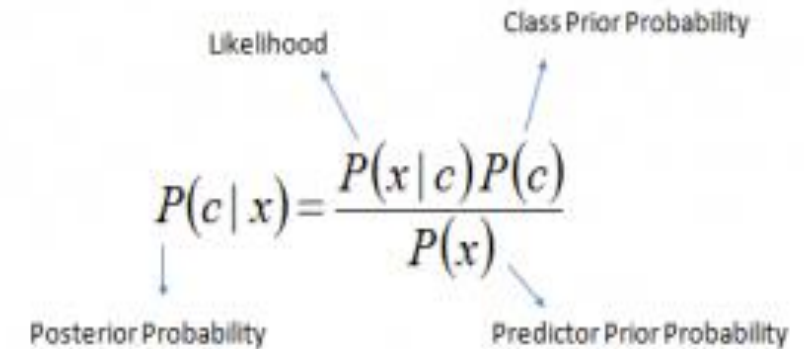
CHARAN LELLABOYENA

Contents

1. Naïve Bayes
2. Decision Tree
3. Datasets
4. Results

Naive – Bayes Classification

- It is a classification technique based on [Bayes' Theorem](#) with an assumption of independence among features.
- In simple terms, a Naive Bayes classifier assumes that the presence of a particular feature in a class is unrelated to the presence of any other feature.
- Naive Bayes model is easy to build and particularly useful for very large data sets. Along with simplicity.

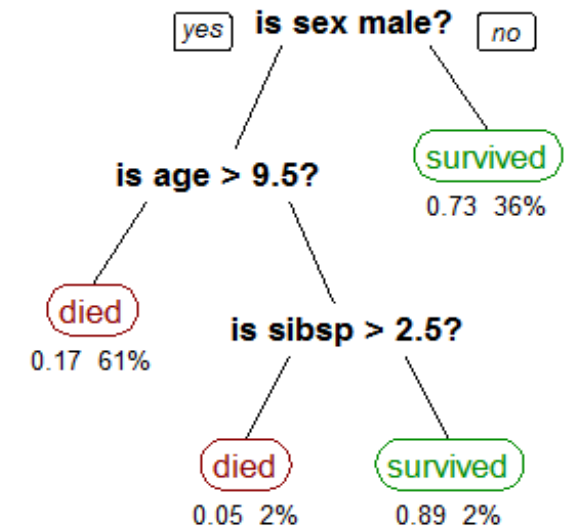


A diagram showing the Bayes' Theorem formula $P(c|x) = \frac{P(x|c)P(c)}{P(x)}$ with four labels and arrows: 'Likelihood' points to $P(x|c)$, 'Class Prior Probability' points to $P(c)$, 'Posterior Probability' points to $P(c|x)$, and 'Predictor Prior Probability' points to $P(x)$.

$$P(c | X) = P(x_1 | c) \times P(x_2 | c) \times \dots \times P(x_n | c) \times P(c)$$

Decision Tree

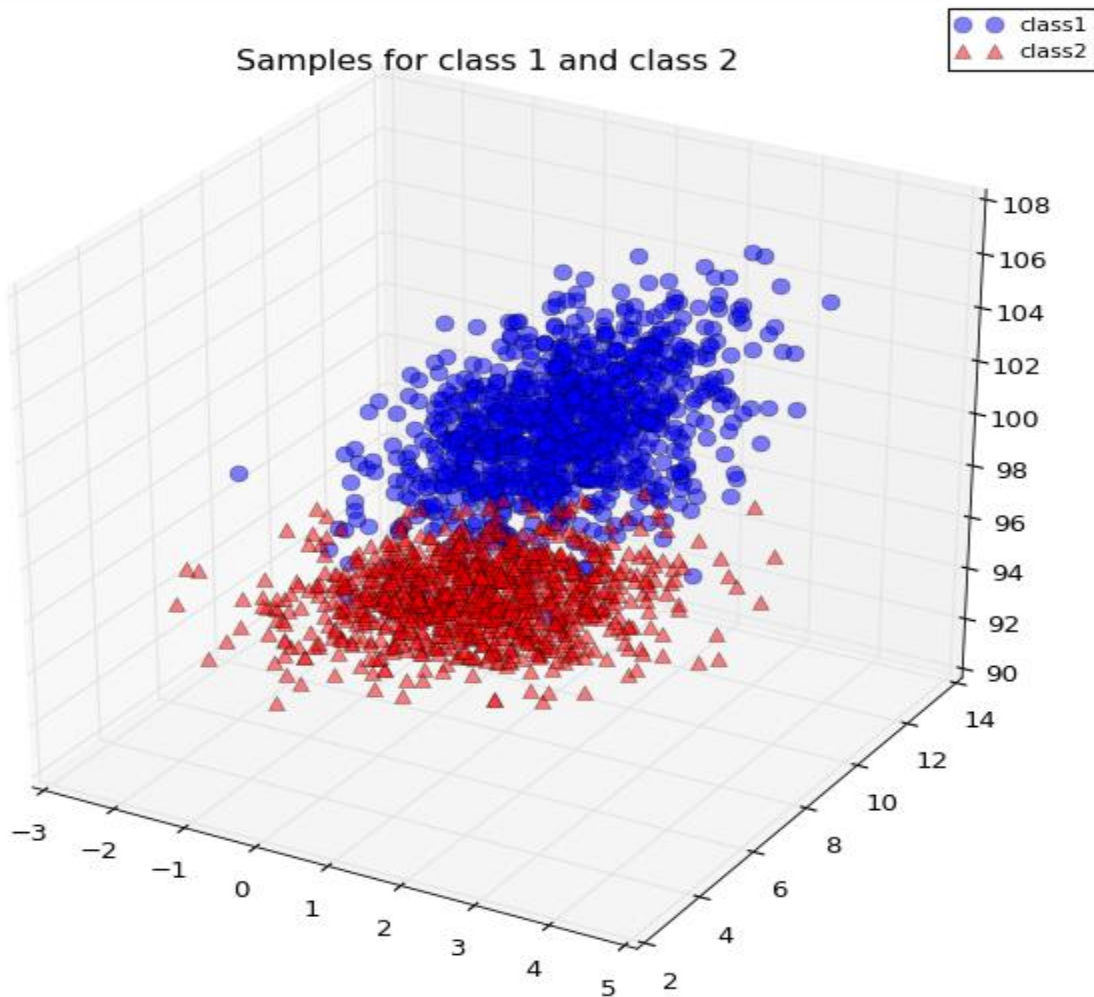
- Decision tree builds classification or regression models in the form of a tree structure. It breaks down a dataset into smaller and smaller subsets.
- The final result is a tree with **decision nodes** and **leaf nodes**.
- We used Gini index method in building the decision tree.
- Gini index is a measure of how often a randomly chosen element from the set would be incorrectly labeled if it was randomly labeled according to the distribution of labels in the subset.
- We split the dataset at a particular value where we get minimum gini index.



Datasets

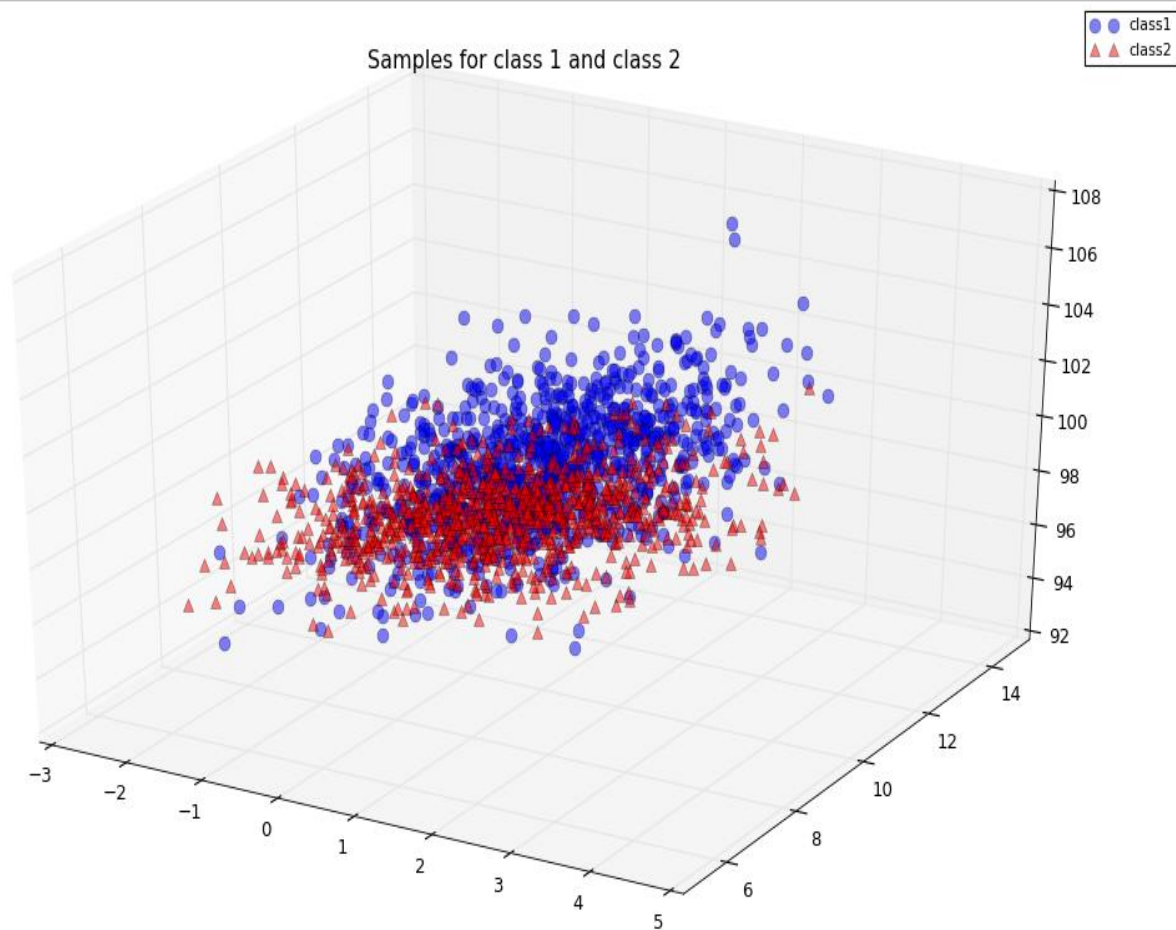
1. Generated dataset with few overlaps
2. Generated dataset with more overlaps
3. Census dataset
 1. Predict whether income exceeds \$50K/yr based on census data.
 2. Mixed Data Set with both categorical and continuous data
 3. Features: 14
4. Sonar(Mine vs Rocks) dataset
 1. The task is to train a network to discriminate between sonar signals bounced off a metal cylinder and those bounced off a roughly cylindrical rock.
 2. Continuous Data
 3. Features: 60

1. Generated dataset(low convergence)



	Accuracy(min)	Accuracy(max)
Naïve-Bayes	93.345	95.15
Decision tree	92.6	94.3

2. Generated dataset(high convergence)

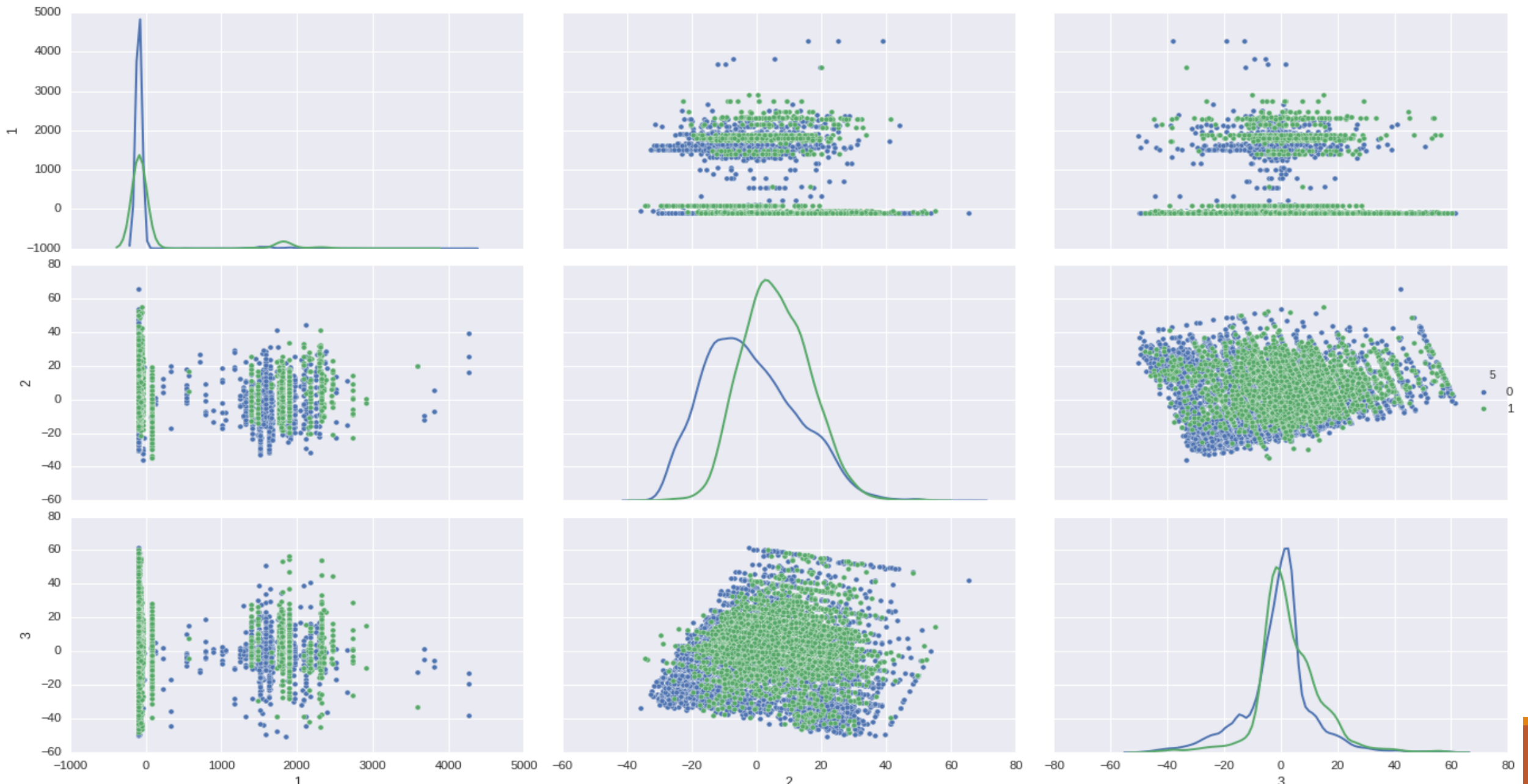


	Accuracy(min)	Accuracy(max)
Naïve-Bayes	75.92	79.24
Decision tree	59.85	60.05

Principle Component Analysis

- Principle Component Analysis: A statistical technique used to examine the interrelations among a set of variables in order to identify the underlying structure of those variables. Also called factor analysis.
- Regression determines a line of best fit to a data set, factor analysis determines several orthogonal lines of best fit to the data set.
- These two properties can be regarded as weaknesses as well as strengths.
 - Since the technique is non-parametric, no prior knowledge can be incorporated.
 - PCA data reduction often incurs a loss of information.

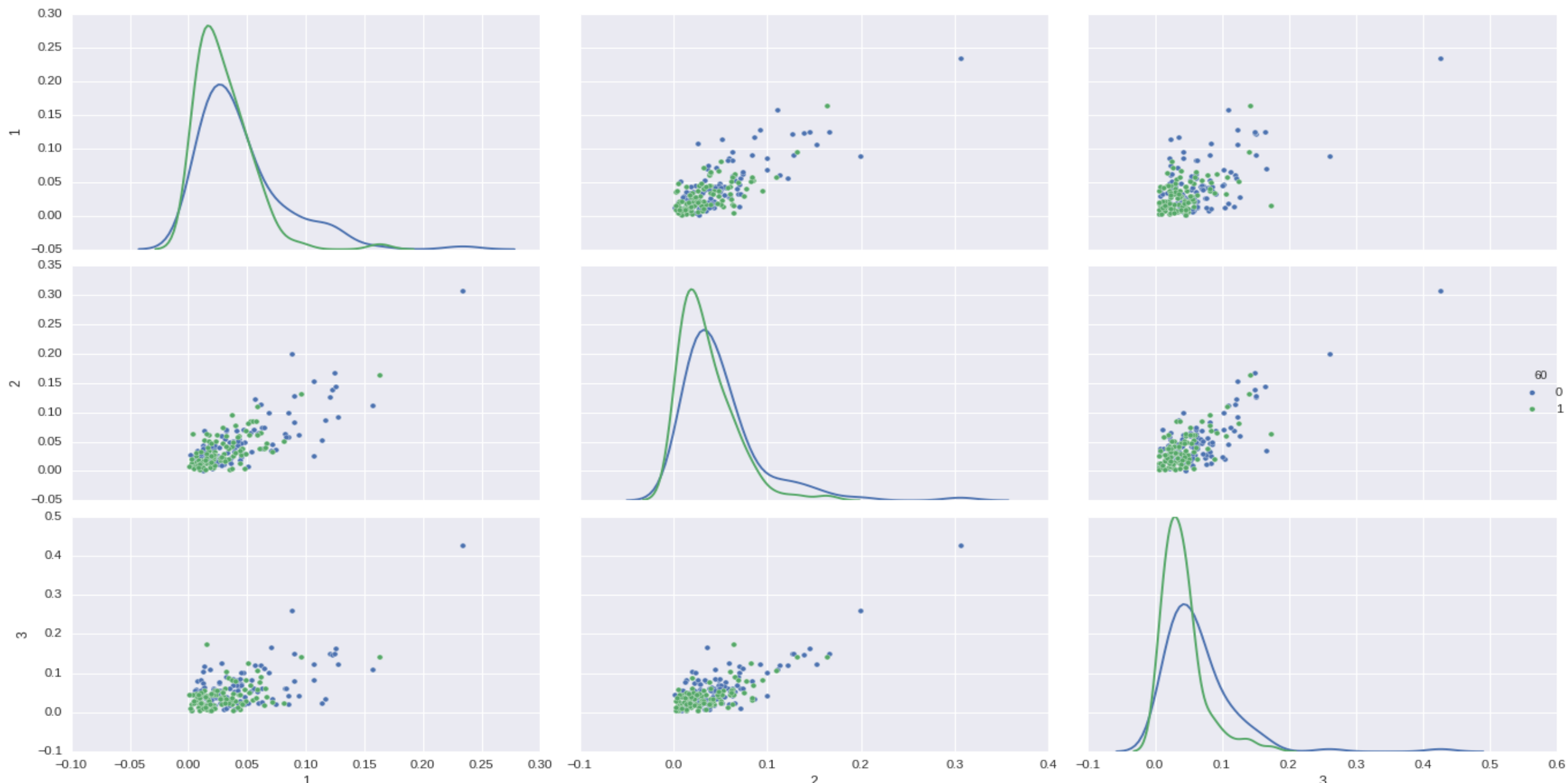
3. Census dataset



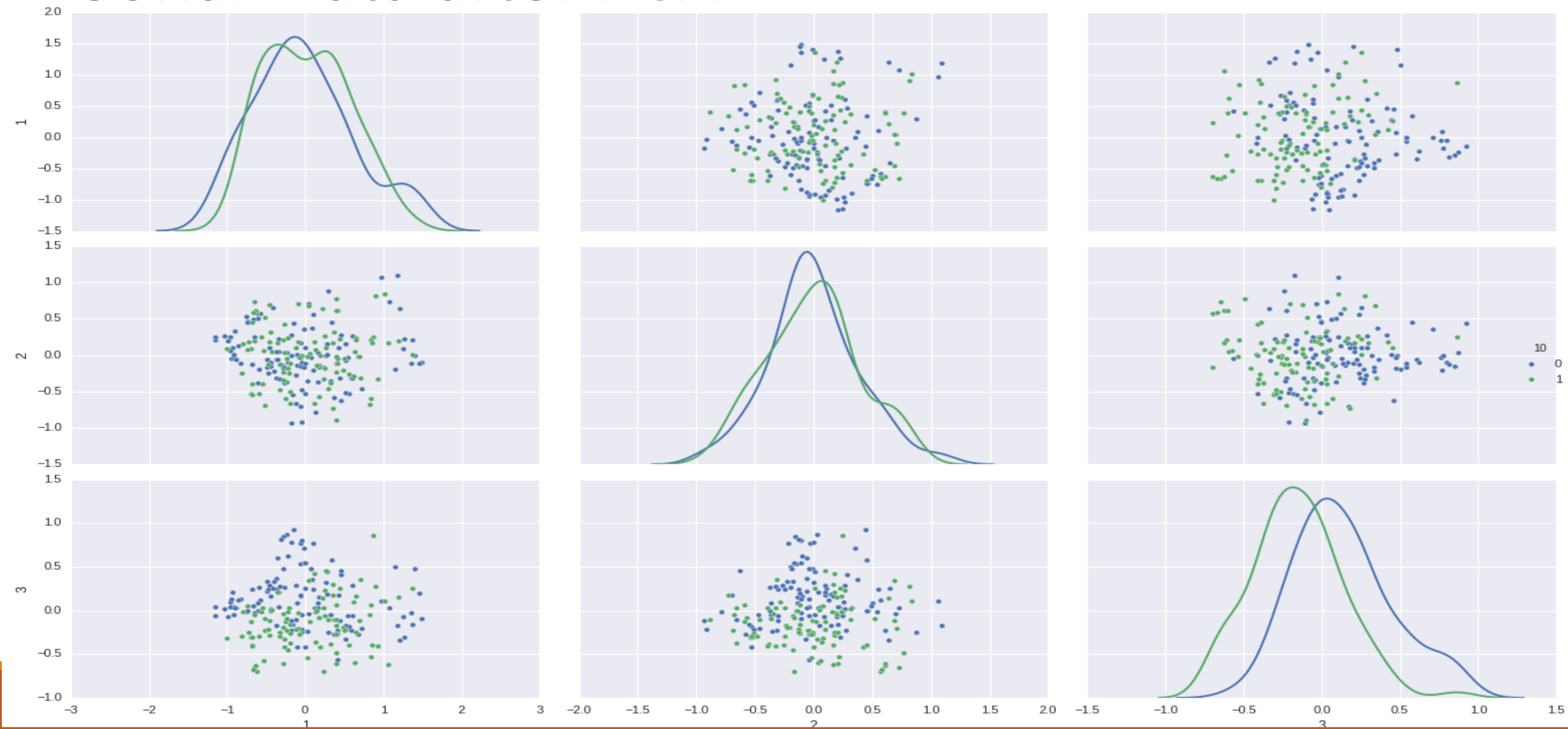
Results For CensusData

Classifier	Before PCA	PCA(5 components)
Naïve-Bayes	81.69	80.07
Decision tree	77.984	80.19

4. Sonar dataset



Sonar Data after PCA



Results for Sonar Data

Classifier	Before PCA(60 features)	PCA(10 components)	PCA(25 components)	PCA(40 components)
Naïve-Bayes	76.811	81.159	84.057	79.71
Decision tree	69.230	65.38	68.26	63.461

Thank You

