­

**Ramez hassouna – 20170639**

**Submitted to:**

**Eng. Bushra O. Alqarout**

Data Science Course’s Project

First Semester of Year 2020/2021

-

.ty

**12/2020**

job-classification

# Table of Contents

[Table of Contents 2](#_Toc60926552)

[Phase 1: Data Selection 3](#_Toc60926553)

[Introduction 3](#_Toc60926554)

[Problem Statement 3](#_Toc60926555)

[Dataset Description 3](#_Toc60926556)

[phase 2: Data Preparation and EDA 5](#_Toc60926557)

[Data Reading and Exploring 5](#_Toc60926558)

[Data Transformation 7](#_Toc60926559)

[EDA 8](#_Toc60926560)

[phase 3: Modeling and Evaluation 21](#_Toc60926561)

[Data Splitting 21](#_Toc60926562)

[Models and Evaluation 21](#_Toc60926563)

[Comparison 26](#_Toc60926564)

# Phase 1: Data Selection

## Introduction

The intent is to use machine learning classification algorithms to predict PG from Educational level through to Financial budget information.

Typically, job classification in HR is time consuming and cumbersome as a manual activity. The intent is to show how machine learning and People Analytics can be brought to bear on this task.

## Problem Statement

This is a dataset containing some fictional job class specs information. Typically job class specs have information which characterize the job class- its features, and a label- in this case a pay grade - something to predict that the features are related to.

**Aim**

* The aim was to identify patterns between variables in the dataset through Principle Component Analysis (PCA)
* Apply and evaluate different classification models and see how each one performs based on the 'Accuracy' and 'Kappa' metrics

I expected the Random Forest model to perform the best, so just felt like exploring different parameter tuning methods (random Search through caret)

## Dataset Description

The data is a static snapshot. The contents are

* ID column - a sequential number
* Job Family ID
* Job Family Description
* Job Class ID
* Job Class Description
* Paygrade- numeric
* Education Level
* Experience
* Organizational Impact
* Problem Solving
* Supervision
* Contact Level
* Financial Budget
* PG- Alpha label for Paygrades

That's where the source data was taken from <https://www.kaggle.com/HRAnalyticRepository/job-classification-dataset>

How the data was collected

Job data is collected from more than one service provider to collect data, as in the following example : <https://datarade.ai/data-categories/job-postings-data>

# phase 2: Data Preparation and EDA

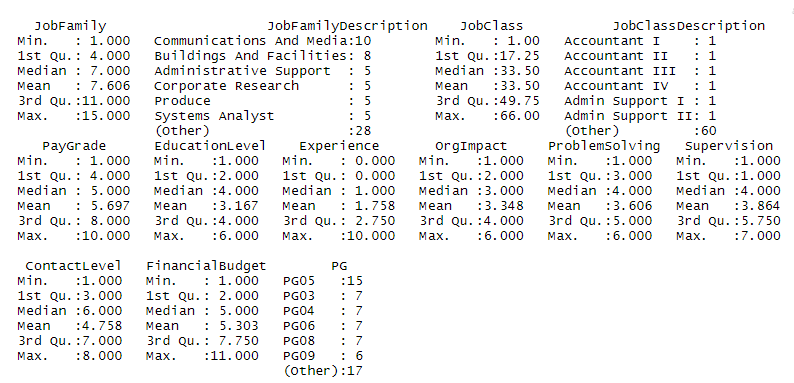
## Data Reading and Exploring

The file has been read in csv format, and the file contains the following characteristics,

In terms of the number of columns and rows it contains 13 col and 66 rows, and the column names were as follows:

* JobFamily
* JobFamilyDescription
* JobClass
* JobClassDescription
* PayGrade
* EducationLevel
* Experience
* OrgImpact
* ProblemSolving
* Supervision
* ContactLevel
* FinancialBudget
* PG

The following summary contains the file properties



Where the following commands were used to implement the above:

* To read file:



* To show data from file:



* To show count column:



* To show count rows:



* To show name column:



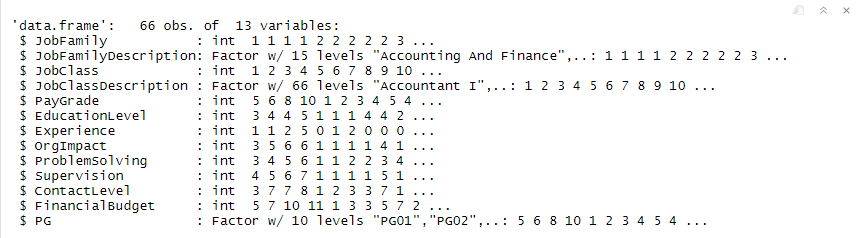
* To show summary:



* To show structure:



His structure was as follows



## Data Transformation

Some transformations were made in the file as the PG column was converted as it contained the values PG1 -PG10, where they were converted to 1-10 by removing the PG from the text.

A further conversion was made in the dataset, where the zero values were converted into an empty value to check the usability of the columns in the analysis and whether or not their removal will be part of the conversion process or not, but no column came out for any removable column because the empty values are few and the values have been returned Tare back to its original state.

Some columns have been deleted because they will not affect the analysis process and the columns that have been deleted are:

* ID
* JobFamilyDescription
* JobClassDescription

Where the following commands were used to implement the above:

* To delete column:





* To convert 0 to null:



* To convert null to 0:



* To transform PG:



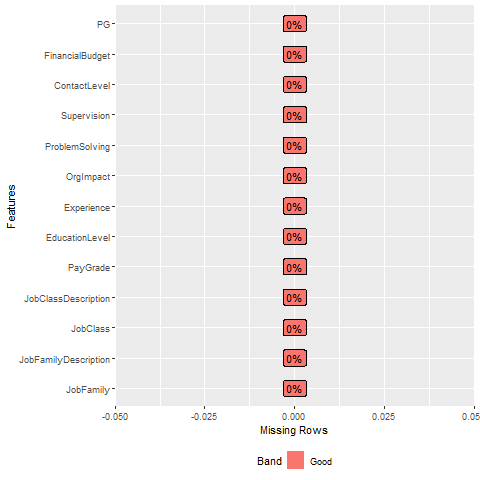


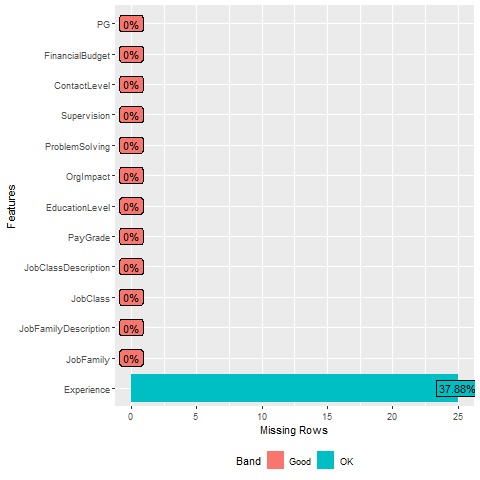
## EDA

The following libraries were used to implement the project

* corrplot
* ggplot2
* DataExplorer
* e1071
* caret
* dplyr
* lattice
* GGally
* PerformanceAnalytics

Blank values have been checked by function 

The results were the blank values before converting the zero values to empty values as follows, as it was observed that there are no blank values in dataset

After the conversion process, the practical results were as follows, as we noticed after the conversion process the zero values into empty values, which led to the emergence of a column containing the empty values, which is the column of experience, where the percentage of empty values did not exceed 37.88% of the total value and accordingly the column was reserved and not deleted

Next, we measured the relationship of the columns to each other and the strength of the values of the columns that correlate with the rest.

His result was as follows, where the relationship between the family and the function and its classification was a strong relationship with each other, but it was weak with the rest of the characteristics and was a correlation PayGrade Powerful in descending order from largest to smallest with next :

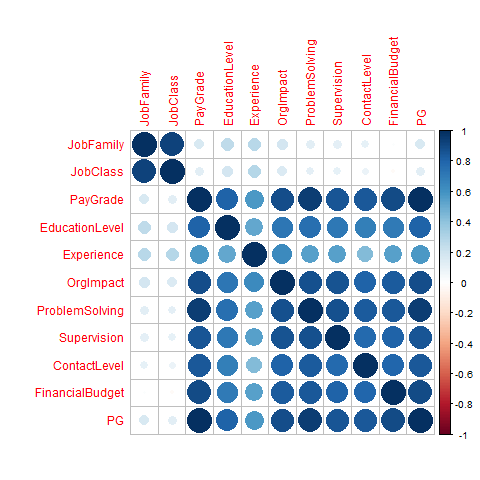
* PG
* ProblemSolving
* FinancialBudget
* OrgImpact
* Supervision
* ContactLevel
* EducationLevel
* Experience

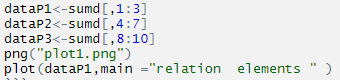
PG

Powerful in descending order from largest to smallest with next

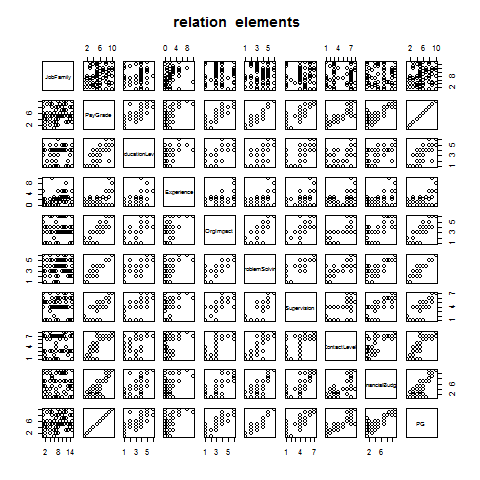
* PayGrade
* ProblemSolving
* FinancialBudget
* OrgImpact
* Supervision
* ContactLevel
* EducationLevel
* Experience

And since the relationship with the rest of the characteristics is as shown in the following analytical drawings



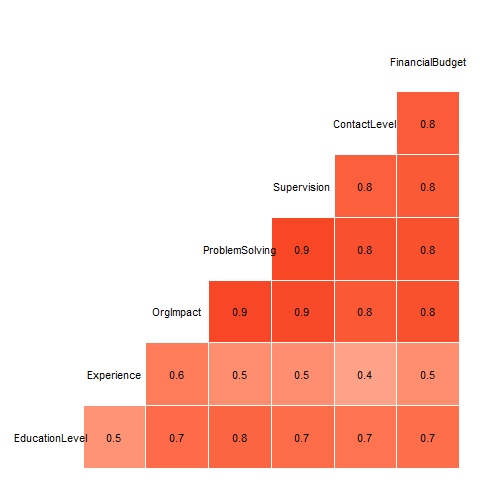


|  |  |
| --- | --- |
|  |  |
|  | |

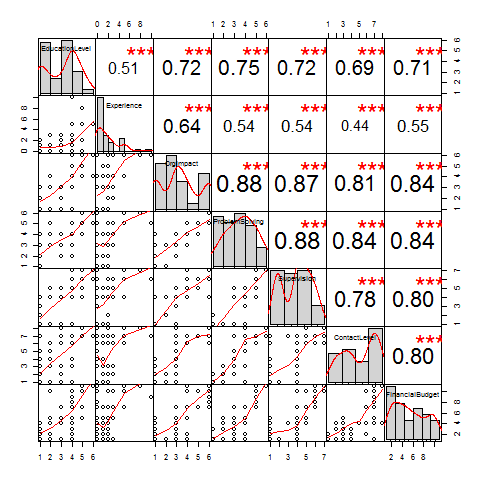


The following command was used to execute the following diagram

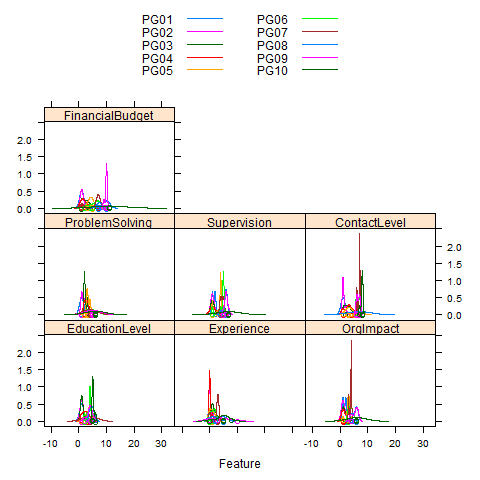




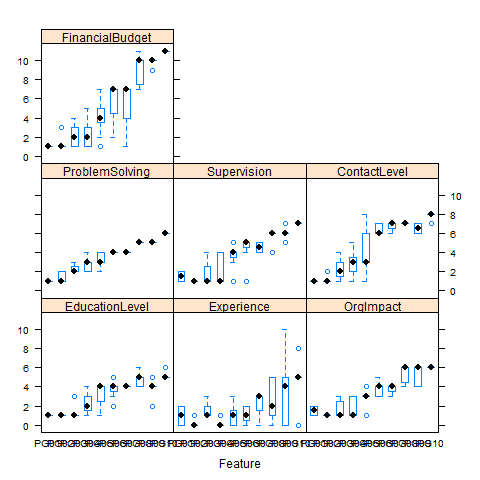
The following command was used to execute the following diagram











The following command was used to execute the following analysis graphics, where the variable X was changed with each property

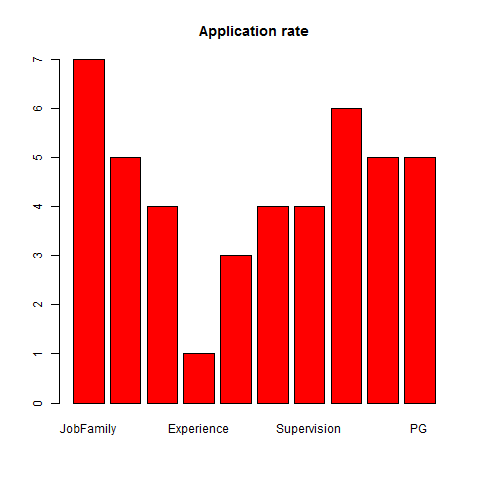


|  |  |
| --- | --- |
|  |  |
|  |  |
|  |  |

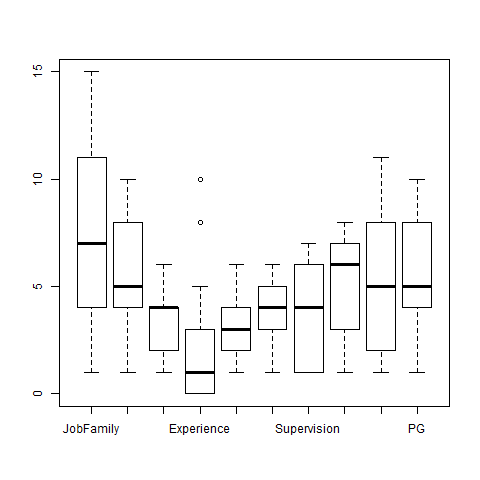
ح

Where what is illustrated in this drawing according to the values most used in each property in dataset



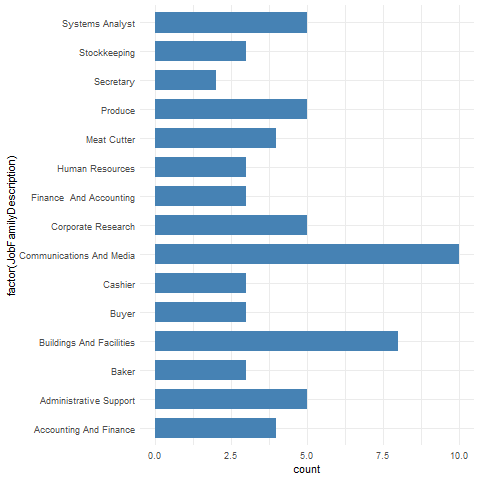


In this diagram, what has been explained are the values contained in each property in dataset where about the concentration of all values 



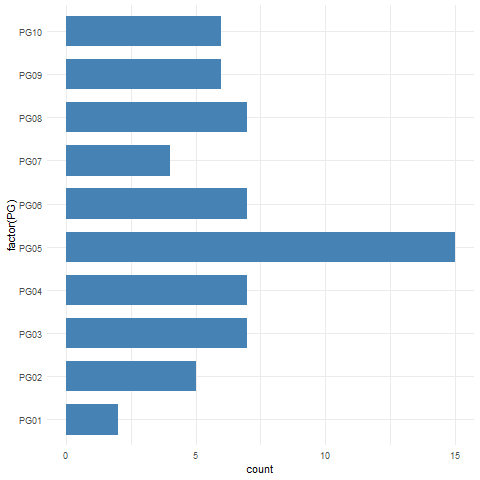
In this illustration, what was explained in it is the number of repetition or use of values in the jobfamilyDescription property





In this illustration, what was explained in it is the number of repetition or use of values in the PG property





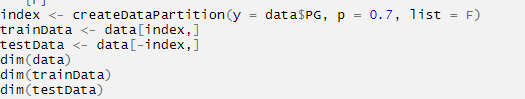
# phase 3: Modeling and Evaluation

## Data Splitting

* Split the samples:

The data was divided into two:

* test samples: Where the test sample is at 30% and is separated from the training samples and used to verify the success and training of the network in a proportional way to verify the performance of the algorithm
* data training samples: Where the training sample is 70% separated from the test samples and used to train the algorithm used



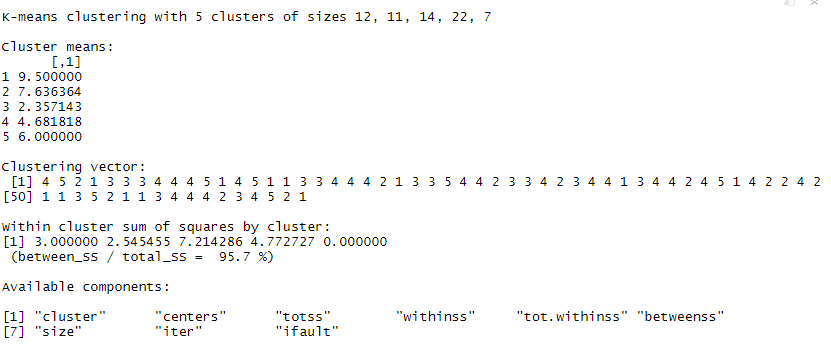


## Models and Evaluation

**The k-nearest neighbors (KNN)**: algorithm is a simple, supervised machine learning algorithm that can be used to solve both classification and regression problems. It's easy to implement and understand, but has a major drawback of becoming significantly slows as the size of that data in use grows.



We made five groups in the \_ to find the closest five identical groups out of ten groups and this was the following result



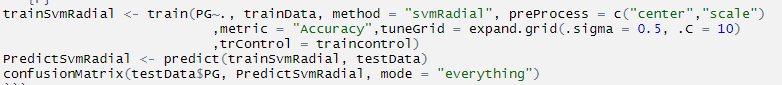
**Cross-validation** is a statistical method used to estimate the skill of machine learning models.

It is commonly used in applied machine learning to compare and select a model for a given predictive modeling problem because it is easy to understand, easy to implement, and results in skill estimates that generally have a lower bias than other methods.

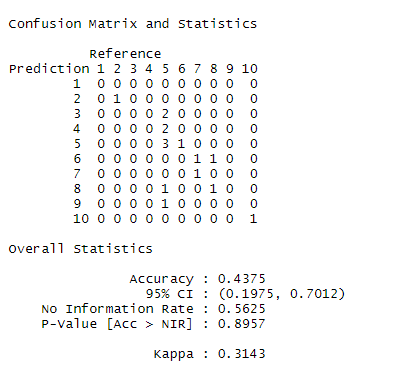
We have used a controller in the training process of algorithms and the controller used in the training process a method kfold because of that dataset is small in size and a command was given to divide them into 10 groups to complete the training process



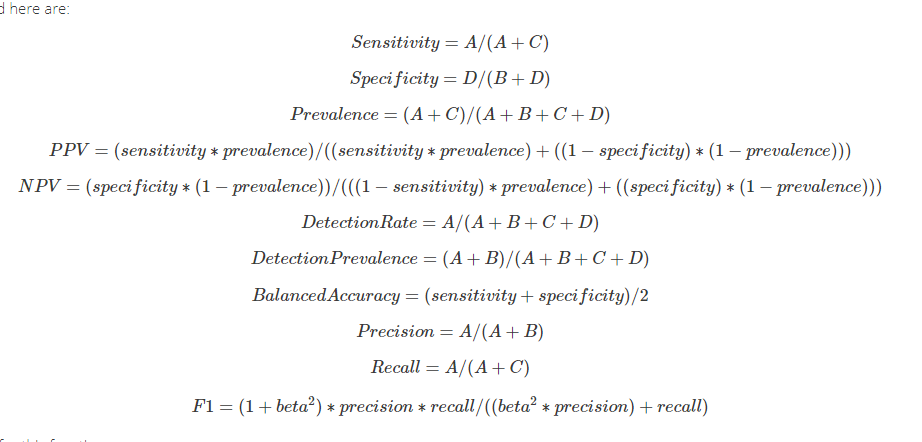
**support machine vectors**: fall under the monitored machine learning algorithms for analyzing the data in order to classify it statistically and make the necessary regression analysis. The input to the process is a piece of data for the algorithm's flexibility and coded for its classification into one of the two factions; Belonging to type X or not



Where the algorithm was created by the training function that the training process can be based on it through the function entering it which is a function svmRadial and the main result of the search is defined and it is accuracy and the controller in the training process is passed in order to train the function and pass the values for the function output is:



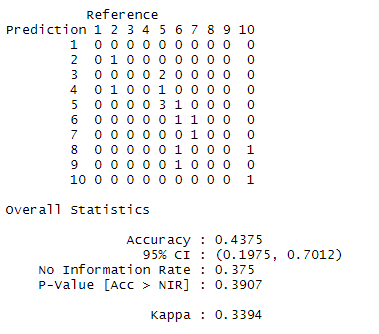
* accuracy = (true positive + true negatve) / all (100 times this is the same as percentCorrect)
* sensitivity = true pasitive rate = true positive / all positive (sensitivity is also called recall)
* specificity = true negative rate = true negative / all negative
* precision = positive predictive velue = true positive rate.



**Naive Bayes**: It is a classification technique based on Bayes’ Theorem with an assumption of independence among predictors. 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.



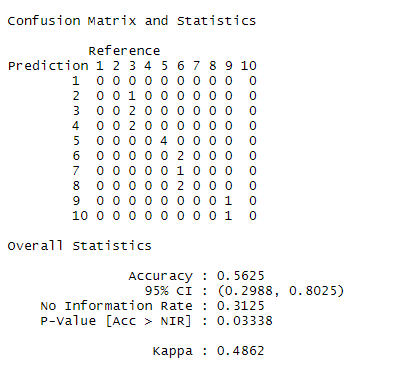
Where the algorithm was created by the training function that the training process can be based on it through the function entering it which is a function Naive Bayes and the main result of the search is defined and it is accuracy and the controller in the training process is passed in order to train the function and pass the values for the function output is:



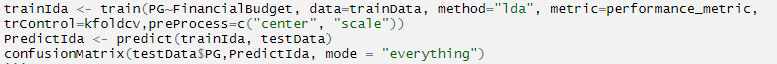
**Classification and Regression Tree (CART)**: is a predictive algorithm used in machine learning. It explains how a target variable's values can be predicted based on other values. It is a decision tree where each fork is a split in a predictor variable and each node at the end has a prediction for the target variable.



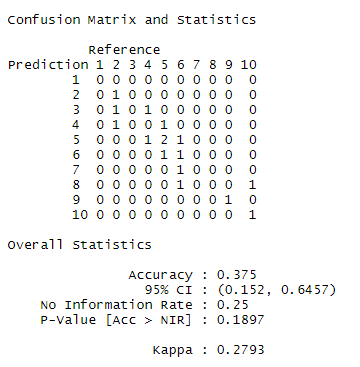
Where the algorithm was created by the training function that the training process can be based on it through the function entering it which is a function Classification and Regression Tree and the main result of the search is defined and it is accuracy and the controller in the training process is passed in order to train the function and pass the values for the function output is:



**Linear discriminant analysis (LDA)**: discriminant function analysis is a generalization of Fisher's linear discriminant, a method used in statistics and other fields, to find a linear combination of features that characterizes or separates two or more classes of objects or events



Where the algorithm was created by the training function that the training process can be based on it through the function entering it which is a function Linear discriminant analysis and the main result of the search is defined and it is accuracy and the controller in the training process is passed in order to train the function and pass the values for the function output is:



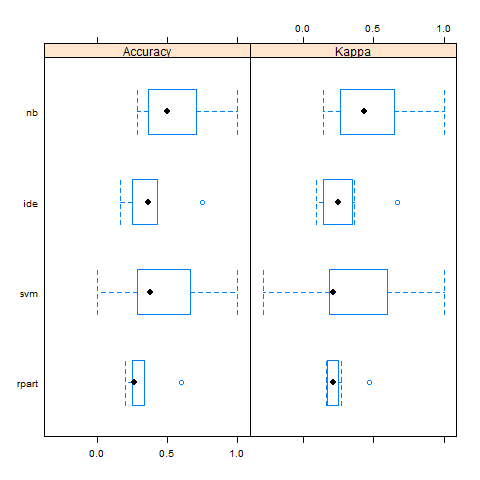
## Comparison

I invoke the training algorithms and review their comparison results, whichever results are better in performance, by this code:

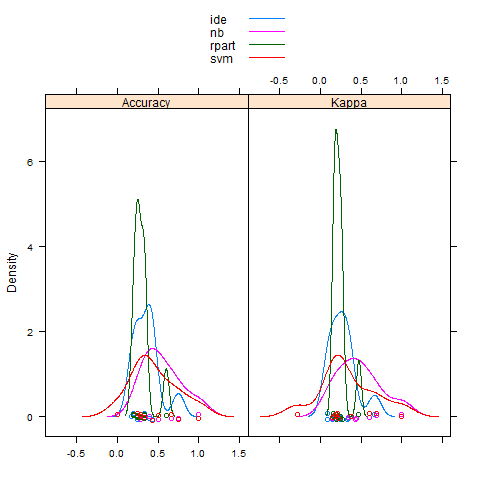


Boxplot algorithms and this code and its results









Accordingly, we decide that the algorithm **Classification and Regression Tree** she is achieved much better results than the rest of the algorithms