

**DEPARTMENT OF COMPUTER SCIENCE ENGINEERING**

Course code – 15CS4171

**MACHINE LEARNING**

III B. Tech – 6th Semester

Academic Year 2018-2019

Project Based Lab Report

ON

**BANK MARKETING**

Using R Studio

Presented by

**Section – 23 (Batch-8)**

**Name ID NO**

D. Likhitha -160030309

A.Naveen Kumar -160030041

R.Bala narasimha -160031160

**Under the Guidance of**

Dr. K. Swarna

\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_

**KONERU LAKSHMAIAH EDUCATION FOUNDATION**

**DEPARTMENT OF COMPUTER SCIENCE ENGINEERING**

**(DST-FIST Sponsored Department)**



**CERTIFICATE**

This is to certify that the course based project entitled **“BANKK MARKETING”** is a bonafide work done by **D.Likhitha(160030309), A.NaveenKumar (160030041), R.Bala Narasimha(160031160)** in partial fulfilment of the requirement for the award of degree in **BACHELOR OF TECHNOLOGY** in **COMPUTER SCIENCE ENGINEERING** during the academic year **2018-2019.**

**Faculty In Charge Head of the Department**

**DR. K. SWARNA DR. HARI KIRAN VEGE**

**DEPARTMENT OF COMPUTER SCIENCE ENGINEERING**

**(DST-FIST Sponsored Department)**



**DECLARATION**

We hereby declare that this project-based lab report entitled **“BANK MARKETING”** has been prepared by us in partial fulfilment of the requirement for the award of degree “**BACHELOR OF TECHNOLOGY in COMPUTER SCIENCE ENGINEERING**” during the academic year 2018-2019.

We also declare that this project-based lab report is of our own effort and it has not been submitted to any other university for the award of any degree.

**Date:**

**Place: Vaddeswaram**

|  |  |
| --- | --- |
| **Name** | **Student ID** |
| D.Likhitha | 160030309 |
| A.Naveen Kumar | 160030041 |
| R.Bala Narasimha | 160031160 |

**ACKNOWLEDGMENTS**

My sincere thanks to **DR. K. SWARNA** in the Lab for his outstanding support throughout the project for the successful completion of the work

We express our gratitude to **DR. HARI KIRAN VEGE,** Head of the Department for Computer Science and Engineering for providing us with adequate facilities, ways and means by which we can complete this project-based lab work.

We would like to place on record the deep sense of gratitude to the honourable Vice Chancellor, K L University for providing the necessary facilities to carry the concluded project-based lab work.

Last but not the least, we thank all Teaching and Non-Teaching Staff of our department and especially my classmates and my friends for their support in the completion of our project-based lab work.

|  |  |
| --- | --- |
| **Name** | **Student ID** |
| D.Likhitha | 160030309 |
| A.Naveen Kumar | 160030041 |
| R.Bala Narasimha | 160031160 |

**INDEX**

|  |  |  |
| --- | --- | --- |
| **S. NO** | **TITLE** | **PAGE NO** |
| 1 | Abstract | 6 |
| 2 | Introduction | 7 |
| 3 | Description   1. Introduction to Bayes theorem 2. Naïve Bayes classifier 3. Representation used by Naïve Bayes classifier models | 8, 9 |
| 4 | Source Code | 10 |
| 5 | Output |  |
| 6 | Requirements |  |
| 7 | Result |  |
| 8 | Conclusion |  |

**ABSTRACT:**

This document utilizes Data Classification to examine a dataset related with direct marketing campaigns (phone calls) of a Portuguese banking institution. The objective of the classification is to predict if the client will subscribe to a Term Deposit. “Data Classification” is the use of Machine Learning techniques to organize datasets into related sub-populations, not previous specified in the dataset. This can uncover hidden characteristics within data, and identify hidden categories that new data belongs within.

This document mainly utilizes the Data Science technique of “Data Classification” to examine a dataset related with direct marketing campaigns (telemarketing phone calls) of a Portuguese banking institution. The objective of the classification is to predict if the client will subscribe to a Term Deposit. “Data Classification” is the use of Machine Learning techniques to organize datasets into related sub-populations, not previous specified in the dataset. This can uncover hidden characteristics within data, and identify hidden categories that new data belongs within.

The Data Science techniques used within this research effort are Exploratory Data Analysis, Data Classification via K-means Clustering, Data Correlation Testing, Predictive Analytics, and Machine Learning with Cross-Validation.

**INTRODUCTION:**

With the rapid development of the techniques in computer science, artificial neural network which in inspired by the bio-neural network is introduced to solve certain problems in real life. With years of development of the neural network technology, the neural network can be implemented to be capable of completing certain complicated tasks. For example, Apple Inc. started to use deep neural network for face detection and face ID. The new product iPhone X released by Apple in 2017 uses the deep neural network along with the deep camera system to recognize certain faces to unlock the phone, which is much more convenient to the finger prints in the previous generation iPhone. Neural network can also be used in many areas, which is a great value to the society, with the neural network, individuals can lead a convenient life.

The dataset called Bank Marketing Data Set’ is chosen from UCI machine learning repository. It has 60 features and 208 instances and is used for classification task. This dataset is chosen because the data set contains only numeric data and the target values only contains 0 and 1, which meets the basic requirements of the Naïve Bayes classification.

**DESCRIPTION:**

In **machine learning** and statistics, **classification** is a supervised **learning** approach in which the computer program learns from the data input given to it and then uses this **learning** to classify new observation. ... Linear **Classifiers**: Logistic Regression, Naive Bayes **Classifier**.

## Introduction to Bayes’ Theorem

In machine learning we are often interested in selecting the best hypothesis (h) given data (d).

In a classification problem, our hypothesis (h) may be the class to assign for a new data instance (d).

One of the easiest ways of selecting the most probable hypothesis given the data that we have that we can use as our prior knowledge about the problem. Bayes’ Theorem provides a way that we can calculate the probability of a hypothesis given our prior knowledge.

Bayes’ Theorem is stated as:

P(h|d) = (P(d|h) \* P(h)) / P(d)

Where

* **P(h|d)** is the probability of hypothesis h given the data d. This is called the posterior probability.
* **P(d|h)** is the probability of data d given that the hypothesis h was true.
* **P(h)** is the probability of hypothesis h being true (regardless of the data). This is called the prior probability of h.
* **P(d)** is the probability of the data (regardless of the hypothesis).

You can see that we are interested in calculating the posterior probability of P(h|d) from the prior probability p(h) with P(D) and P(d|h).

**Naive Bayes Classifier**

Naive Bayes is a classification algorithm for binary (two-class) and multi-class classification problems. The technique is easiest to understand when described using binary or categorical input values.

It is called *naive Bayes* or *idiot Bayes* because the calculation of the probabilities for each hypothesis are simplified to make their calculation tractable. Rather than attempting to calculate the values of each attribute value P(d1, d2, d3|h), they are assumed to be conditionally independent given the target value and calculated as P(d1|h) \* P(d2|H) and so on.

This is a very strong assumption that is most unlikely in real data, i.e. that the attributes do not interact. Nevertheless, the approach performs surprisingly well on data where this assumption does not hold.

### Representation Used By Naive Bayes Models

The representation for naive Bayes is probabilities.

A list of probabilities are stored to file for a learned naive Bayes model. This includes:

* **Class Probabilities**: The probabilities of each class in the training dataset.
* **Conditional Probabilities**: The conditional probabilities of each input value given each class value.

**SOURCE CODE:**

getwd

setwd("C:\\Users\\Lenovo\\Desktop\\skilling")

mydata=read.csv("bankk.csv")

str(mydata)

dim(mydata)

tindex = sort(sample(nrow(mydata), nrow(mydata)\*.7))

mtraining=mydata[tindex,]

mtesting=mydata[-tindex,]

library(e1071)

NB=naiveBayes(y~.,data=mtraining)

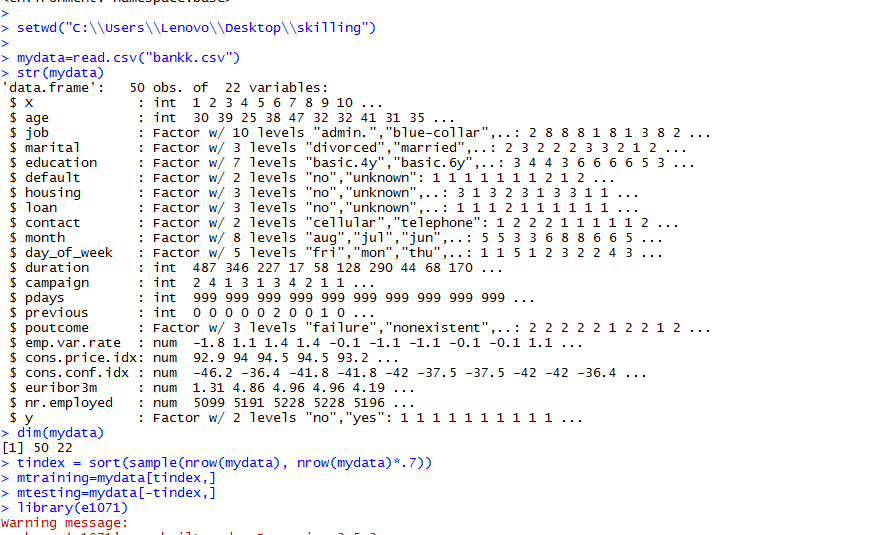
print(NB)

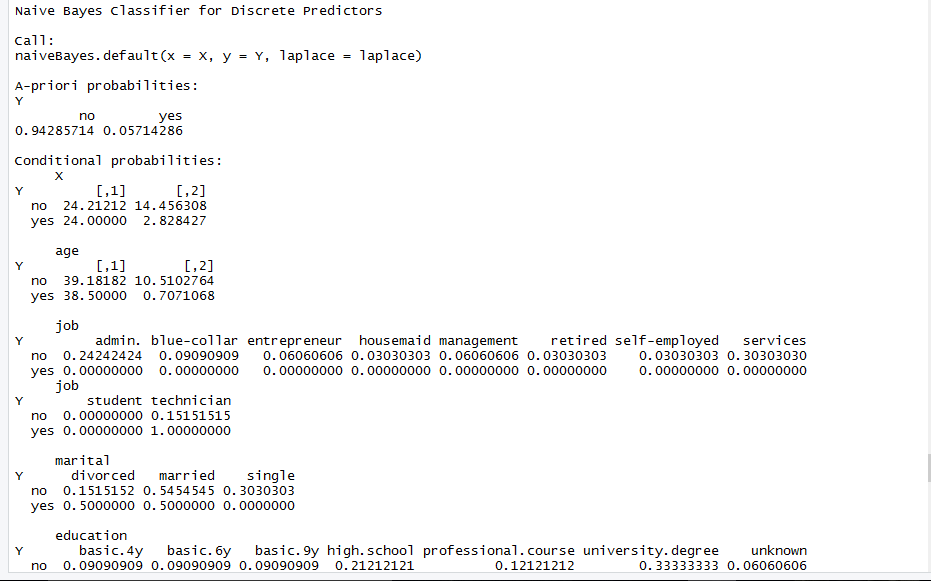
predNB1=predict(NB,mtesting,type=c("class"))

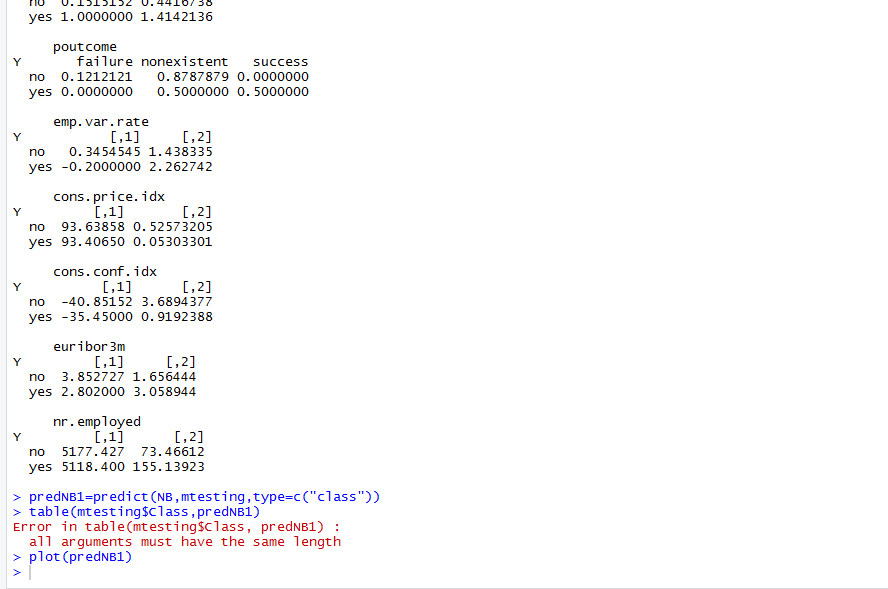
table(mtesting$Class,predNB1)

plot(predNB1)

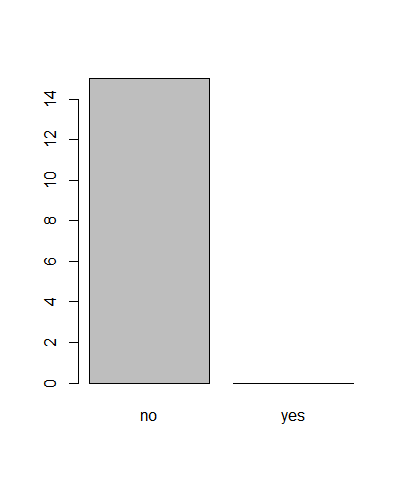
**OUTPUT:**

****

****

****

**RPLOT GRAPH:**

****

**REQUIREMENTS:**

* Windows (7, 8, 8.1, 10)
* R language
* R studio

**RESULT:**

The problem with comparing these results with the results from the original paper is that they did not used a proper train, validation and test splitting. Even with a simple classifier such as NBC, we can achieve better results. The neural networks show some sort of indifferentiable behaviour but do perform reasonably well on 3 of the 4 datasets.

**CONCLUSION:**

We could say that though the validation accuracies (and therefore in a way the test results) that the models here generalize more than the models in the original publication because we used a much bigger test set as well as a proper set for cross-validation. If we would overfit a bit more to the training set and perhaps iterate over all features and leave them out iteratively, we may end up with better results. However, we would end up with a model that is somewhat disconnected from what data we are receiving (in case we would deploy such a model).