SN Computer Science (2020) 1:96
https://doi.org/10.1007/s42979-020-0102-0

ORIGINAL RESEARCH

Forecasting and Avoiding Student Dropout Using the K-Nearest Neighbor Approach

Mahesh Mardolkar[1] · N. Kumaran[1]

Published online: 24 March 2020

Abstract

All India School Education Survey of 2018–2019 has put a total of 13,06,992 schools and enrollment figure as 22,67,19,283 number of students in these schools in India, and the problem is continuing studies at school or college for many students is difficult due to reasons such as financial problem, domestic problems at home, being unable to cope up with studies, language or medium of education and many other problems like this. The dropout ratio in 2018–2019 for secondary level classes IX and X is at 17% as compared to 4% in primary classes I to V, and to minimize the dropout ratio, the technique of prediction is proposed; in the field of machine learning and data mining, the most widely used prediction method is K-nearest neighbors. This method is more versatile and simple and can handle different types of data in the process of prediction. The prediction of students is classified into dropout or no-dropout category and hence enables the teacher to counsel the students who are at risk of dropping out.

Keywords K-nearest neighbors · Dropout · RStudio · ggplot2 · R Shiny

Introduction

In the human resource development, education plays an important role and also a remedial role in balancing the socioeconomic fabric of the country; since the most valuable resource is the citizens of India, for better quality of life they need nurture and care in the form of basic education, and on September 26, 1985, the Ministry of Human Resource Development (MHRD) was created. Currently, the MHRD works through two departments Department of School Education and Literacy and Department of Higher Education. The development of school education and literacy in the country is taken care of by Department of School Education and Literacy, and the development of one of the largest higher education systems of the world is taken care of by Department of Higher Education. The Ministry of Human Resource Development has initiated Web-based All India Survey on Higher Education (AISHE) since 2010. All the institutions in the country are covered under the survey; on different parameters, data are being collected, such as student enrollment, teachers, examination result, programs, infrastructure, institution density, gross enrollment ratio, pupil–teacher ratio and gender parity index; per student expenditure is calculated through AISHE; also, the number of universities established stands at 789 with over 37,204 colleges and another 11,443 stand-alone institutions in India [1]; and the Eighth All India School Education Survey (AISES) of 2018–2019 conducted by the National Council of Educational Research and Training (NCERT) has put a total of 13, 06,992 schools and enrollment figure as 22,67,19,283 number of students in the schools [2]; the problem is continuing studies at school or college for many students is difficult due to reasons such as financial problem, domestic problems at home, being unable to cope up with studies, language or medium of education and many other problems like this, and under such circumstances, the students drop out further studies. The dropout ratio in 2018–2019 for secondary level classes IX to X is at 17% as compared to 4% in primary classes I to V, 4% in upper primary classes VI to

This article is part of the topical collection "Advances in Computational Intelligence, Paradigms and Applications" guest edited by Young Lee and S. Meenakshi Sundaram.

✉ Mahesh Mardolkar
principal@bharateshbca.com

N. Kumaran
kumaran81@gmail.com

[1] Department of Computer Science and Engineering, Annamalai University, Annamalai Nagar, Tamil Nadu, India

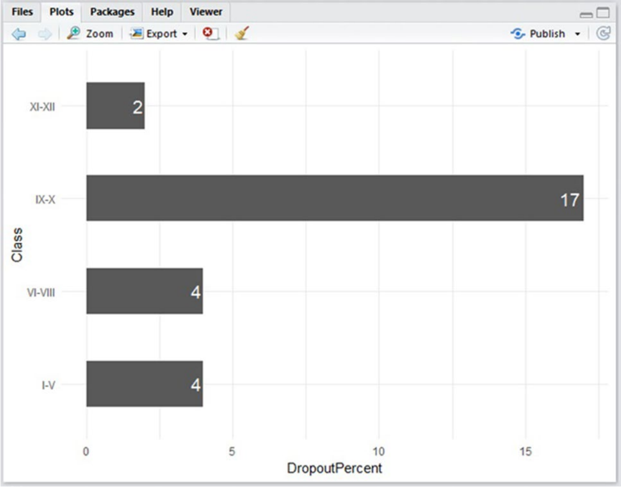


Fig. 1 Class-wise dropout ratio

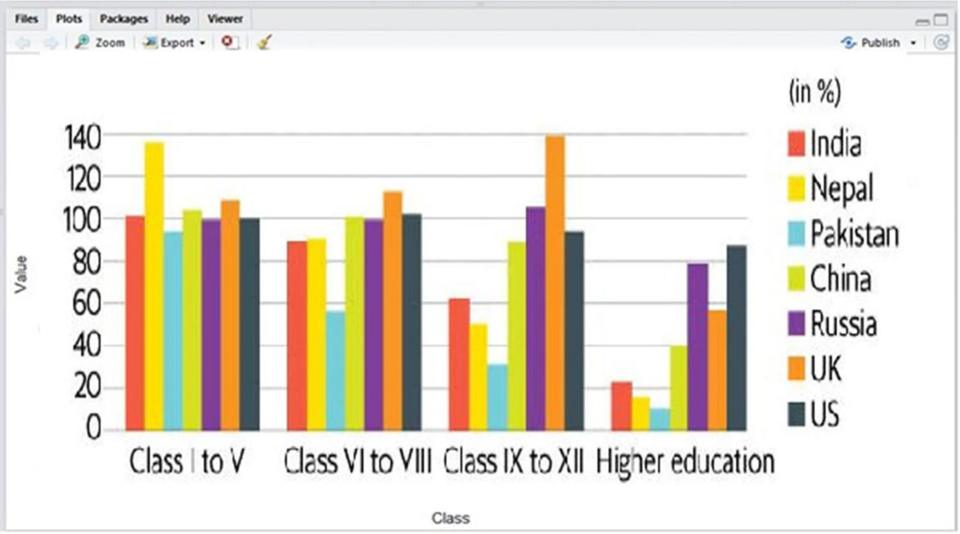
VIII and 2% in upper secondary classes XI to XII (Fig. 1). The transition rate from class X to class XI is at 69% since many students are held back or have dropped out of school, and until the completion of elementary education, a child cannot be expelled or detained under the RTE Act.

Literature Survey

A gross enrollment ratio (GER) is used in education sector and is a statistical measure, it is mainly used to determine the number of students enrolled at different levels such as elementary, middle and high school, and it is the ratio of numbers of students in the country to those who are enrolled. The enrollment more than 100% means that students are younger or older than the specified age group. In comparison with developed countries, India's enrollment rate in primary education classes I–V seems to be better and fall behind after class VI. India's enrollment rate stands at 23% against 87%, 57% and 39% in the USA, the UK and China, respectively, in higher education [3] (Fig. 2).

Domestic activities, marriage and lack of interest are the key reasons for female students to drop out; on the other hand, lack of interest, economic activities and financial constraints are the reasons for male students to drop out (Fig. 3). The majority of dropouts are seen in classes IX and X in higher primary education, and with data mining technique, the student activity of these classes is analyzed, which helps the teacher and the counselors quickly find and focus their attention to the students who are on the verge of dropping out, and based on their activity, the goal is to predict dropout or no dropout, so at an early stage a tool can alert teachers, giving teachers an idea of the students who are at the risk of dropping out. In this paper, the need of assistance for the underperforming students is discovered which helps the teacher to identify the students, and we extensively evaluate and propose K-nearest neighbor (KNN) method to predict students' performance at an early stage of study, we develop and experimentally evaluate settings and variants of KNN method, and we report experimental results of IX class students of English medium school. The KNN technique [4] applied here generalizes well; also this would enable the teachers to take immediate action to improve their students' welfare and academic performance, which would prevent students dropping out, and also, in case of regression and

Fig. 2 Country-wise comparison of enrollment



classification, the KNN technique of assigning weights to the contribution of neighbors can be used, so that nearer neighbor has more average than distant ones. The common scheme for weighting is 1/d weight of each neighbor, where d is the distance to the neighbor.

A student drops out of school is a process, this process does not happen overnight, and the dropping process starts much before a child is admitted to school. The predication is poor academic achievement at elementary schooling, there are different factors which also put the student at risk of dropping out of school, and the risk factors are different among different categories of students, which lead student to drop out of school. The following are the list of risk factors: *Academic performance is poor, Parent engagement is lacking, Economic need, Supportive adult is lacking, No individual attention* and *Student engagement is low.* Also, there appear different major reasons for a student's dropping out of school.

Parent engagement is very important for a child to be successful in school, both financial and emotional are key factors for a child to stay in school, and parents' aspiration for their child's education is also required for doing well in school. Parent engagement in early stages of child's education helps in obtaining a successful result. Parents should also feel child's education is important.

Academic performance is an important factor which has a higher influence for a child to stay in school, the main focus in lower classes is to make child learn to read and also help students who are struggling to read, and the reading proficiency will help the child to be more prepared and successful in the future. The success in middle classes is the key indicator of whether a child will continue or drop out of school; sometimes, the relationship with their teacher is not strong, and also handling multiple subjects at a time makes it difficult to get required attention which student needs.

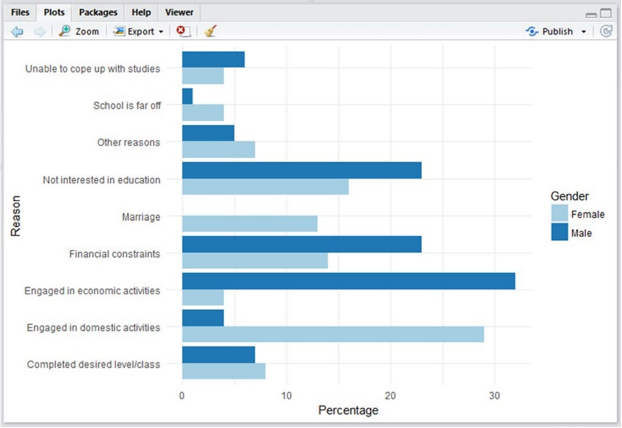


Fig. 3 Other reasons for dropout

Family economic status also contributes to school dropouts, and it was found that students with lower socioeconomic status are at higher risk of dropping out of school than students with higher socioeconomic status, also students working part time have a certain risk of dropping out, and there are various factors that affect the student's behavior, due to which over a time they lack interest toward studies. Delinquency—students get fascinated by fancy things and try to acquire them which move toward delinquency, they also take up theft, buying and selling, and they are helping in unethical crimes and many more such things where they get trapped in. Lack of interest—students in classes refuse to show any interest in subject being taught to them, they give lack of attention in whatever is taught to them, students cannot be forced and pressurized for learning, which does not last long, and they also feel whatever is taught to them is not important and hence prefer to drop out. Working in family shops—many students have an option of working in family business, they feel family business is more important than spending six to seven hours in school, and their decision is supported by their family members. Supporting family—Due to economic reasons, students give up their education, their inability to pay fees is one of the reasons, and they also follow the notion that the number of helping hands is always better. Constant failure—failing repeatedly in class test or semester examination causes lack of self-confidence in students and repeated failure results in giving up school education. Support for ailing family member—students take leave to take care of their ailing parents or grandparents who are alone at home and long leaves divert them from going back to school. Serve bullying—students face extreme humiliation at school which they do not discuss with anybody, they are also feared of being judged and laughed and they start making excuses from going to school. Too much of academic pressure—the recent generation is not quipped in handling stress situations, they resort to alcoholism or substance abuse to overcome this problem of excessive anxiety, and this inability to deal with academic pressure makes them drop out of school. Grabbing hold of other opportunities—students find many opportunities in advertising and modeling, and also participation in auditions and competition more often put them away from school. Unable to fit in—some students adapt changes within and outside school pretty easily; on the other hand, there are students who do not cope up with any changes, making friends and establishing good rapport with teachers are difficult for them, they feel like an outsider and they give up. When discussed with dropouts and interviewed of what school can do to improve chances of students' stay in school, we found the percentage of suitable means to improve the chances of students staying in school (Table 1), the student engagement in school (Table 2) and the reasons for dropout (Table 3).

Table 1 Chances of students staying in school

Improve chances of students staying in school	Percent (%)
Need help in problems learning	55
Want smaller classes and one-to-one attention of teachers	75
More tutoring, extra classes with teachers	70
There should be real-world and experimental learning so that the student should see opportunity further	81
Want better teachers	81

Table 2 Students engagement in school

Students engagement in school	Percent (%)
Making students to attend classes	68
Maintain classroom discipline	62
Help students from violence	57

Table 3 Reason for dropout

Reason for dropout	Percent (%)
Failing in test and examinations	35
Too much of freedom	38
No friendly environment	42
Low attendance	43
Classes were not interesting	47

Prediction Problem

The problem that we want to solve is the following: Can we predict with reasonable accuracy the performance of the students and which students will drop out? The various features of the students are defined and categorized into two as student welfare feature (Table 4) and student performance feature (Table 5) [5]. Student welfare feature defines parent involvement, medium of instructions, qualified parent, earning members in family, annual income of parents, time spent with friends and playing, working in family shop, doing part-time job, liberty given and financial assistance. Student performance feature defines previous examination performance, everyday reading and writing activity, academic pressure, need for extra classes and re-examination performance.

Euclidean Distance

Euclidean distance is probably the most well-known distance used for numerical data; for isolated and compact clusters [6, 7] when deployed with datasets Euclidean distance performs well; and in clustering, Euclidean distance is very common, with few drawbacks such as the distance may be smaller if two data vectors have no common attributes as compared to a data vector having common attributes [8] and a large-scale feature would dominate others and the solution to this problem is normalization [9]. Euclidean measure is defined as:

Table 4 Student welfare feature

Student welfare feature	
Parent involvement	Do parents know the performance of their wards?
Medium of instructions	Is the medium of instruction same as mother tongue?
Qualified parent	Mentioning the qualification of parents
Earning members in family	The number of members earning in the family
Annual income of parents	Annual income of earning members in the family
Time spent in playing	Number of hours spent in playing everyday
Working in family shop	Engaging in family business or work in shops
Doing part-time job	Part-time job for earning purpose
Liberty given	Liberty at home and school
Financial assistance	Financial assistance from government or any other source to pay fee

Table 5 Student performance feature

Student performance feature	
Previous examination performance	Percentage- and grade-wise category of students
Everyday reading and writing activity	Number of hours spent in reading and writing everyday
Academic pressure	Checks whether the student has any academic pressure and does not understand the subjects
Need for extra classes	Engagement of students in tuitions and extra classes apart from regular classes
Re-examination performance	Checks whether re-examination is to be given and the performance of re-examination

$$d(x,y)=\sqrt{\sum_{i=1}^{n}\left(x_i-y_i\right)^2}. \quad (1)$$

Proposed Work

In educational institutions, several studies have been carried out to analyze student performance and without any assumption about prior probabilities of training data, and the most simplistic method is K-nearest neighbor (KNN) algorithm; in real-world classification, the performance of this method has shown to be satisfactory, in student performance prediction and machine learning, KNN is found to be more accurate and competitive with more complex method such as support patterns, kernel method and support vector machine, when genetic learning of feature is considered, its performance is very close and found to be better than competing methods, in case of noisy and incomplete data where prediction needs to be made KNN seems to be more suitable and robustness, and simplicity of KNN makes it more suitable for classification in many situations. The proposed work is implemented on the RStudio to generate graphical presentation of the dropouts, and the complete concept is discussed in the following sections.

RStudio

It is an integrated development environment (IDE) which is free and is open-source software developed for R programming. R is a programming language mainly used for statistical computing and for graphics. J J Allair the creator of ColdFusion programming language also founded RStudio, and two editions of RStudio are available for user: RStudio desktop where user can run local desktop applications and RStudio Server, which uses a Web browser to access RStudio a remote linux server. Distributions of RStudio are available for Linux, Windows and Mac OS, and commercial and open-source editions of RStudio are available for the users. RStudio is created using C++ programming language, and RStudio also used Q+ framework for its graphical user interface. The first public beta version was announced in February 2011, in November 2016 version 1.0 was released, and in October 9, 2017, version 1.1 was released.

Packages

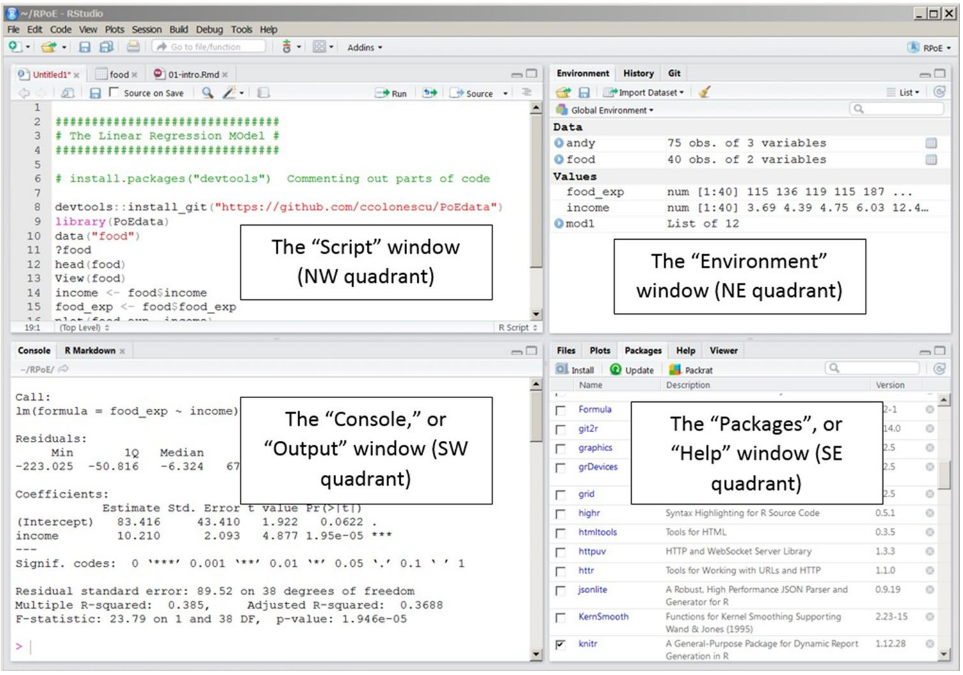
Through user-created packages, the different capabilities of R are extended. These packages are developed in R and other languages such as JAVA, C, C++ and Fortran. Packages allow specialized statistical techniques to be incorporated easily such as graphical devices (ggplot2), Import/Export capabilities and reporting tools (knitr, Sweave). R installation includes a core set of packages and additional packages of more than 7801 available at Comprehensive R Archive Network(CRAN), Bioconductor, Omegahat, GitHub and other repositories. ggplot2 package is used for data visualization. Hadlay Wickham created ggplot package in 2005. Leland Wilkinson's grammar of graphics is implemented in ggplot. The basic graphics in R can be replaced by ggplot2 which contains print display and number of defaults for Web. ggplot2 is one of the most popular packages of R license under GNU GPL. If you are using R much, you will likely need to read in data at some points. While R can read excel.xls and.xlsx files, these file types often cause problems. Comma-separated files (.csv) are much easier to work with. It is best to save these files as csv before reading them in R. If you need to read in csv with R, the best way to do it is with the command read.csv. Here is an example of how to read CSV into R:

```
#Read CSV into R
MyData < - read.csv(file="c:/Dropout.csv" , header = TRUE, sep =",")
```

The above reads the file Dropout.csv into a data frame that it creates called MyData. header = TRUE specifies that these data includes a header row and sep="," specifies that the data are separated by commas (Fig. 4).

Updates

ggplot2 version 0.9 was released in March 2, 2012; after longtime of maintenance mode operations, ggplot2.0.0 was released in December 21, 2015. Comparison with other packages and base graphics—with the high level of abstraction ggplot2, allows user to alter, add or remove components in a plot. The cost of ggplot2 is lower than that of lattice graphics, more complex plotting can be done in ggplot2, both multivariate and univariate and also categorical and numerical data can be used to create graph in ggplot2, and color, size, symbol and transparency can be used for grouping. ggplot2 is a well-known visualization package in R, it is easy and simple, to draw graphs there is a huge repository of packages in R available for download from various locations, and there are different ways to produce graphical output in R programming. The package which is called ggplot2 is more versatile and powerful than plot(). plot() comes with basic version of R. The required packages need to be installed in R, and to install ggplot2 package, the command install.packages("ggplot2") is executed, every package has a unique name to be mentioned or identified correctly before installation, and once the package is installed, it requires activation which is done with the command



The “Script” window (NW quadrant)

```
#################################
# The Linear Regression MOdel #
#################################

# install.packages("devtools")  Commenting out parts of code

devtools::install_git("https://github.com/ccolonescu/PoEdata")
library(PoEdata)
data("food")
?food
head(food)
View(food)
income <- food$income
food_exp <- food$food_exp
```

The “Environment” window (NE quadrant)

```
Data
 andy            75 obs. of 3 variables
 food            40 obs. of 2 variables
Values
 food_exp        num [1:40] 115 136 119 115 187 ...
 income          num [1:40] 3.69 4.39 4.75 6.03 12.4…
 mod1            List of 12
```

The “Console,” or “Output” window (SW quadrant)

```
Call:
lm(formula = food_exp ~ income)

Residuals:
     Min       1Q   Median
-223.025  -50.816   -6.324   67

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)   83.416     43.410   1.922   0.0622 .
income        10.210      2.093   4.877 1.95e-05 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 89.52 on 38 degrees of freedom
Multiple R-squared:  0.385,     Adjusted R-squared:  0.3688
F-statistic: 23.79 on 1 and 38 DF,  p-value: 1.946e-05

> |
```

The “Packages”, or “Help” window (SE quadrant)

	Name	Description	Version
□	Formula		2-1
□	git2r		14.0
□	graphics		2.5
□	grDevices		2.5
□	grid		2.5
□	highr	Syntax Highlighting for R Source Code	0.5.1
□	htmltools	Tools for HTML	0.3.5
□	httpuv	HTTP and WebSocket Server Library	1.3.3
□	httr	Tools for Working with URLs and HTTP	1.1.0
□	jsonlite	A Robust, High Performance JSON Parser and Generator for R	0.9.19
□	KernSmooth	Functions for Kernel Smoothing Supporting Wand & Jones (1995)	2.23-15
☑	knitr	A General-Purpose Package for Dynamic Report Generation in R	1.12.28

Fig. 4 RStudio desktop interface

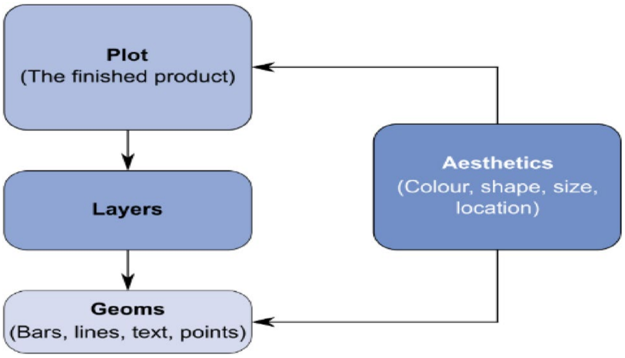
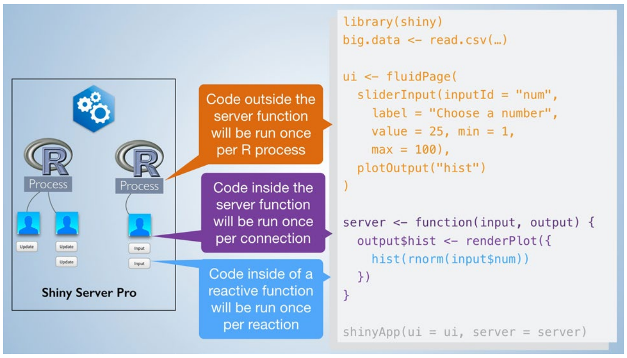


Fig. 5 ggplot2 in R programming


Process
Process
Update
Update
Update
Input
Input
Shiny Server Pro
Code outside the server function will be run once per R process
Code inside the server function will be run once per connection
Code inside of a reactive function will be run once per reaction

```
library(shiny)
big.data <- read.csv(…)

ui <- fluidPage(
  sliderInput(inputId = "num",
    label = "Choose a number",
    value = 25, min = 1,
    max = 100),
  plotOutput("hist")
)

server <- function(input, output) {
  output$hist <- renderPlot({
    hist(rnorm(input$num))
  })
}

shinyApp(ui = ui, server = server)
```

Fig. 6 R Shiny scripting architecture

library(ggplot2). The other way to activate the packages is to select the packages in graph panel, the list of packages which are downloaded appears with a checkbox for selection, the user can select the appropriate checkboxes to install the packages, there are series of layers to be defined in ggplot2 which makes up the graph like geoms which defines the visual elements called as geometrical objects such as points and bars and so on, and also, the location and appearance of these geoms mentioned as color, size and so on are controlled by properties called as esthetics; variables that are to be plotted are referred as aes() (Fig. 5).

R Shiny

Interactive Web apps can be built easily in R Shiny Package, Shiny app is mainly used in building and deploying Web apps, it also supports reactive programming, which is more correct, robust and efficient, and it also supports interactive plots and interactive documents having shiny components. R Shiny helps user to prepare analyzed reports into interactive Web application without knowledge of HTML, CSS or JavaScript (Fig. 6).

K-Nearest Neighbor (KNN) Method

In the field of machine learning and data mining, the most widely used prediction method is K-nearest neighbor (KNN), and the method is more versatile and simple and can handle different types of data in the process of prediction. It was initially introduced by E. Fix and J. L. Hodges in the report for US Air Force School of Aviation Medicine in 1967 which was unpublished, and the main properties and original idea were formalized. KNN is a lazy or instance-based method; since the original training data do not require building a model to represent underlying distribution and statistics, it directly

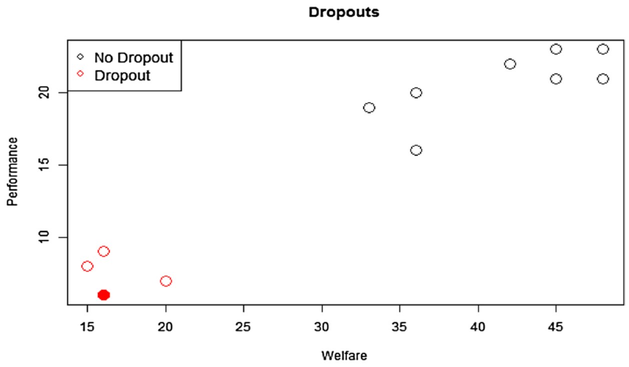


Fig. 7 Graphical representation of dropouts. $D=\text{Sqrt}[(16–33)^2+(6–19)^2]=21.40 \gg \text{Dropout}=N$

works on training data and their actual instances. The formal definition of KNN was described by Covert and Hart [4]. The method has twice upper error bound as compared to Bayes' error probability. The K most similar instances in the training set are found based on the performance score of student welfare and student performance feature, the baseline KNN algorithm makes a prediction once the performance of a student is known, we calculate similarity with simple Euclidean distance between the student welfare feature and corresponding student performance feature in the training set, also for every student welfare and performance feature it is evaluated on 1 to 5 scale, 1 represents poor response and 5 represents excellent response, and then, we decide two classes for prediction: The first is dropout and the other no dropout; the k instance with shortest distance is found to predict the class. Figure 7 shows 0 means no dropout and 3 dropout for the vote with three neighbors ($k=4$); for the best result, the required majority threshold can be tuned [10].

Results

By first inspecting the data, we can choose optimal value of K, higher value of K provides better results and also large K value is more precise as it reduces the overall noise, and when a student case is considered for prediction, the welfare score and performance score are calculated and the Euclidian distance with distance of other students in the dataset is computed; by selecting optimal value of K, the result is compared with neighbors to categorize the case into dropout or no dropout and consider the following data representation concerned with dropout. Welfare and Performance are two numerical variables (predictors), and dropout is the target. Euclidean distance [6] can be used on training set to classify an unknown case of Welfare score = 16 and Performance score = 6 nearest neighbor in the training set if K = 4; out of three close neighbors, three have Dropout = Y; and the prediction for unknown case is again Dropout = Y. The graphical output on RStudio with the ggplot2 packages seems to be more interactive and readable, the graphical output makes the process of prediction easier, and ggplot2 is a powerful package to generate graph in R [11] (Table 6).

Table 6 Sample class IX students' dataset

Student ID	Welfare	Performance	Dropout	Distance
A012	16	6		0.00
A004	15	8	Y	2.24
A007	16	9	Y	3.00
A008	20	7	Y	2.24
A005	33	19	N	3.00
A002	36	16	N	4.12
A011	36	20	N	24.41
A006	42	22	N	30.53
A009	45	21	N	32.65
A003	45	23	N	33.62
A001	48	21	N	35.34
A010	48	23	N	36.24

Conclusion

The prediction of students based on various related features who are studying in IX class was studied in this paper. The observation provides teachers a better way to identify the non-performing students; with this prediction, the teacher focuses on features of a student that need attention the most. KNN method is applied to predict students and classify them into dropout or no-dropout category, and we also check their welfare and performance score and identify the reason for dropout. RStudio is found to be more helpful in graphical analysis of data mentioned in this paper. The results of the research work done here will assist teachers to counsel student who are at risk, and early step will benefit the student to overcome the problems and continue their studies.

References

1. All India Survey on Higher Education Homepage. http://aishe.nic.in/aishe/home. Accessed 22 Feb 2019.
2. National Council of Educational Research and Training Homepage. http://www.ncert.nic.in/index.html. Accessed 4 Mar 2019.
3. Mardolkar M, Kumaran N. School dropout analysis with R programming charts. Int J Res (IJR). 2018;05(04):1042–7. ISSN: 2348-6848.
4. Covert T, Hart P. Nearest neighbor pattern classification. IEEE Trans Inf Theory. 1967;13(1):21–7.
5. Kotsiantis S, Pierrakeas C, Pintelas P. Preventing student dropout in distance learning systems using machine learning techniques. In: Proceedings of seventh international conference on knowledge-based intelligent information and engineering systems, lecture notes in artificial intelligence, vol 2774. Springer; 2003. p. 267–274.
6. Mao J, Jain AK. A self-organizing network for hyperellipsoidal clustering (HEC). IEEE Trans Neural Netw. 1996;7:16–29.
7. Jain AK, Murty MN, Flynn PJ. Data clustering: a review. ACM Comput Surv. 1999;31:264–323.

8. Legendre P, Legendre L. Numerical ecology. Amsterdam: Elsevier; 2012.
9. Wang H, Wang H, Wang W, Wang W, Yang H, Yang H et al. Clustering by pattern similarity in large data sets. In: 2002 ACM SIGMOD international conference on management of data. New York, New York, USA: ACM Press; 2002. p. 394.
10. Hstie T, Tibshirani R, Friedman J. The elements of statistical learning. 1st ed. Berlin: Springer; 2001.
11. Mardolkar M, Kumaran N. Universal comparison of school education in RStudio. In: International conference on research trends in engineering, applied science and management, vol 08. Issue-XII, December 2018 ISSN: 2249-7455.

Publisher's Note Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.