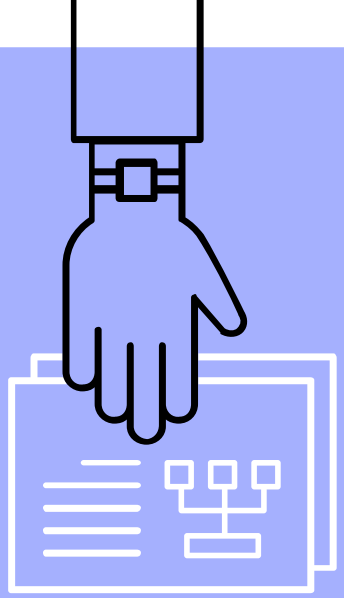
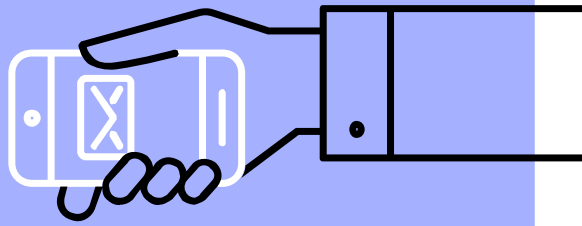
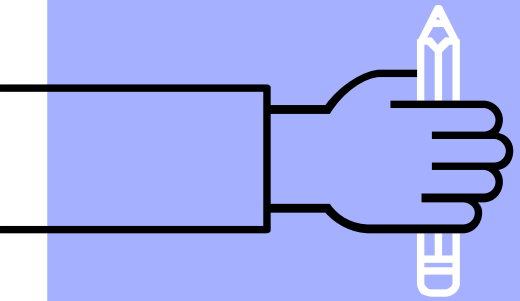
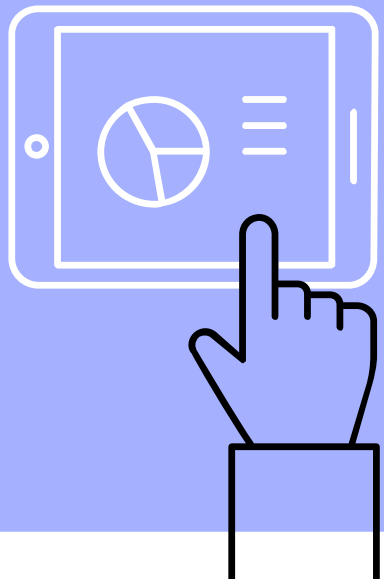


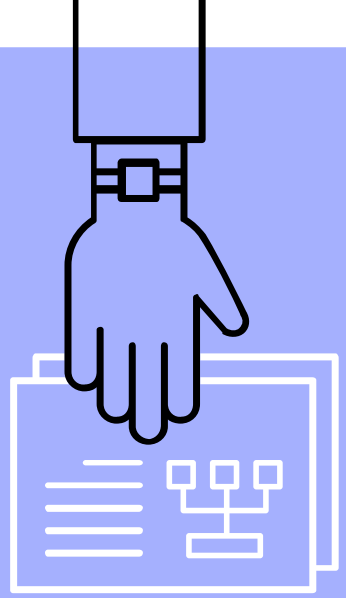


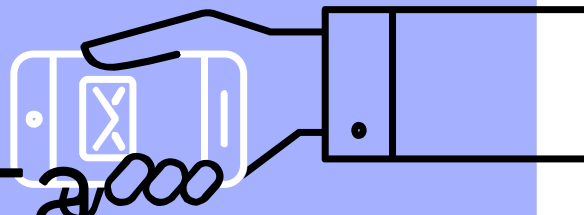
DATA SCIENCE Olympics

By Smera Gora and ShwinyG on GitHub

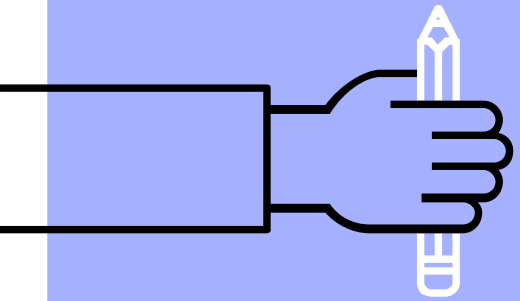


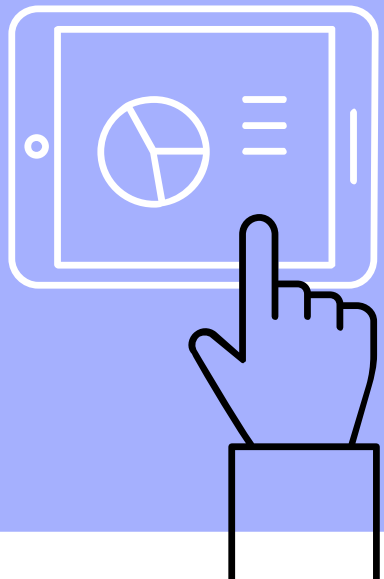




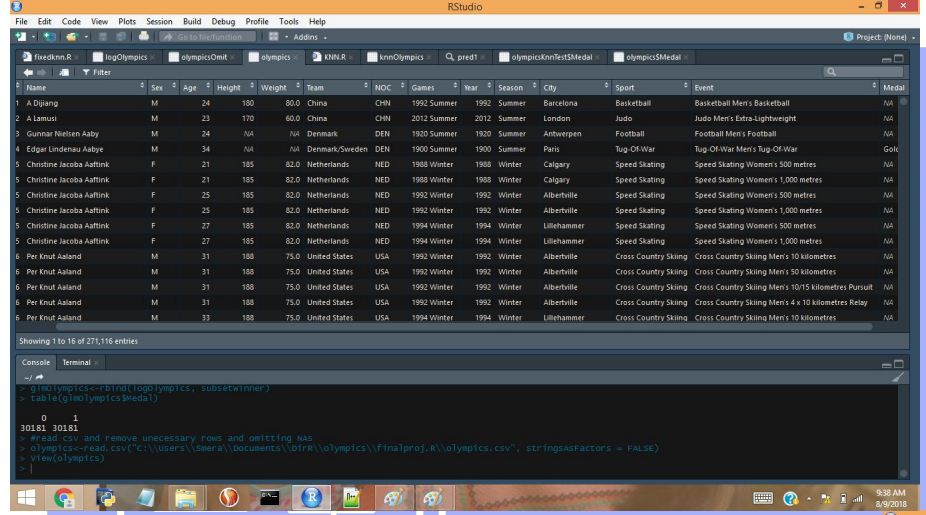


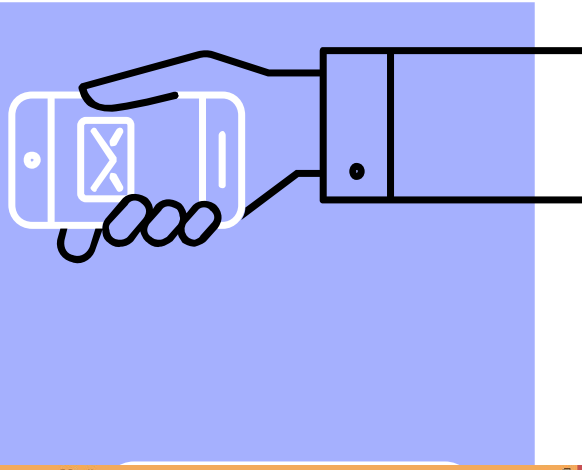
We got our data on Kaggle 270,000 x 15

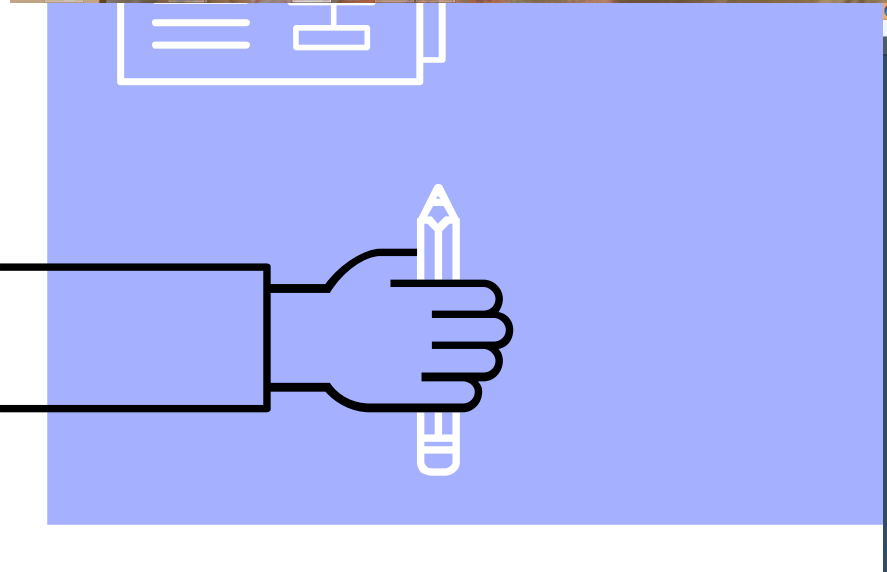


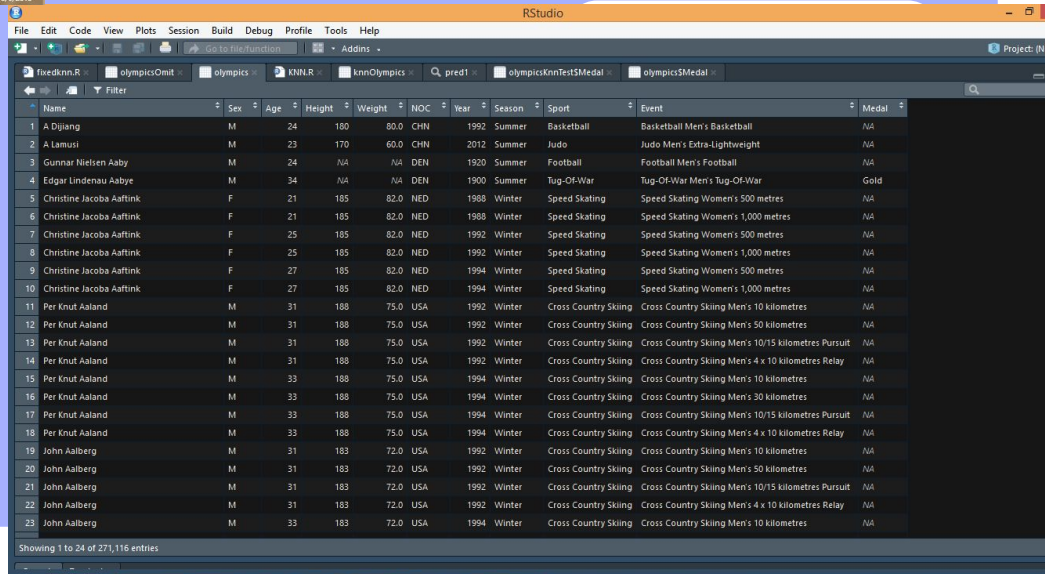


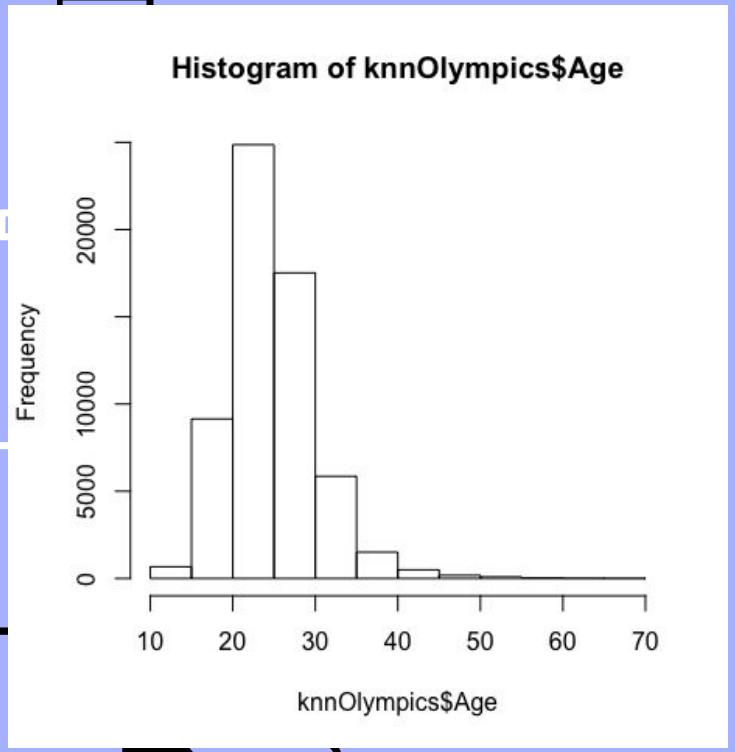
Variety of different data with categorical and numerical predictors

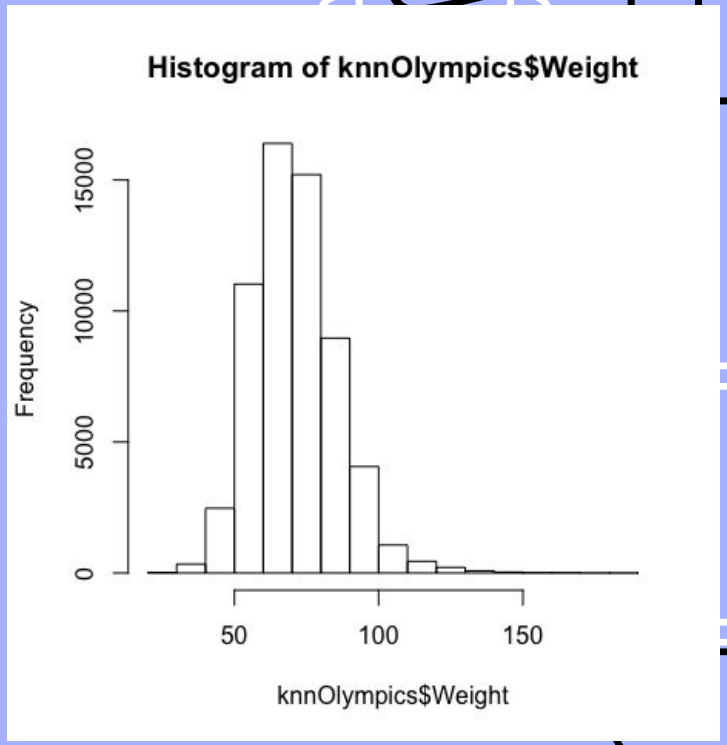

RStudio
File Edit Code View Plots Session Build Debug Profile Tools Help
Name Sex Age Height Weight Team NOC Games Year Season City Sport Event Medal
1 A Dijiang M 24 180 80.0 China CHN 1992 Summer 1992 Summer Barcelona Basketball Basketball Men's Basketball NA
2 A Lamusi M 23 170 60.0 China CHN 2012 Summer 2012 Summer London Judo Judo Men's Extra-Lightweight NA
3 Gunnar Nielsen Aaby M 24 NA NA Denmark DEN 1920 Summer 1920 Summer Antwerpen Football Football Men's Football NA
4 Edgar Lindenau Aabye M 34 NA NA Denmark/Sweden DEN 1900 Summer 1900 Summer Paris Tug-Of-War Tug-Of-War Men's Tug-Of-War Gold
5 Christine Jacoba Aaftink F 21 185 82.0 Netherlands NED 1988 Winter 1988 Winter Calgary Speed Skating Speed Skating Women's 500 metres NA
5 Christine Jacoba Aaftink F 21 185 82.0 Netherlands NED 1988 Winter 1988 Winter Calgary Speed Skating Speed Skating Women's 1,000 metres NA
5 Christine Jacoba Aaftink F 25 185 82.0 Netherlands NED 1992 Winter 1992 Winter Albertville Speed Skating Speed Skating Women's 500 metres NA
5 Christine Jacoba Aaftink F 25 185 82.0 Netherlands NED 1992 Winter 1992 Winter Albertville Speed Skating Speed Skating Women's 1,000 metres NA
5 Christine Jacoba Aaftink F 27 185 82.0 Netherlands NED 1994 Winter 1994 Winter Lillehammer Speed Skating Speed Skating Women's 500 metres NA
5 Christine Jacoba Aaftink F 27 185 82.0 Netherlands NED 1994 Winter 1994 Winter Lillehammer Speed Skating Speed Skating Women's 1,000 metres NA
6 Per Knut Aaland M 31 188 75.0 United States USA 1992 Winter 1992 Winter Albertville Cross Country Skiing Cross Country Skiing Men's 10 kilometres NA
6 Per Knut Aaland M 31 188 75.0 United States USA 1992 Winter 1992 Winter Albertville Cross Country Skiing Cross Country Skiing Men's 50 kilometres NA
6 Per Knut Aaland M 31 188 75.0 United States USA 1992 Winter 1992 Winter Albertville Cross Country Skiing Cross Country Skiing Men's 10/15 kilometres Pursuit NA
6 Per Knut Aaland M 31 188 75.0 United States USA 1992 Winter 1992 Winter Albertville Cross Country Skiing Cross Country Skiing Men's 4 x 10 kilometres Relay NA
6 Per Knut Aaland M 33 188 75.0 United States USA 1994 Winter 1994 Winter Lillehammer Cross Country Skiing Cross Country Skiing Men's 10 kilometres NA
Showing 1 to 16 of 271,116 entries
> glmOlympics<-rbind(logOlympics, subsetWinner)
> table(glmOlympics$Medal)
0 1
30181 30181
> #read csv and remove unecessary rows and omitting NAs
> olympics<-read.csv("C:\\Users\\Smera\\Documents\\DirR\\olympics\\finalproj.R\\olympics.csv", stringsAsFactors = FALSE)
> view(olympics)






RStudio
File Edit Code View Plots Session Build Debug Profile Tools Help
Name Sex Age Height Weight NOC Year Season Sport Event Medal
1 A Dijiang M 24 180 80.0 CHN 1992 Summer Basketball Basketball Men's Basketball NA
2 A Lamusi M 23 170 60.0 CHN 2012 Summer Judo Judo Men's Extra-Lightweight NA
3 Gunnar Nielsen Aaby M 24 NA NA DEN 1920 Summer Football Football Men's Football NA
4 Edgar Lindenau Aabye M 34 NA NA DEN 1900 Summer Tug-Of-War Tug-Of-War Men's Tug-Of-War Gold
5 Christine Jacoba Aaftink F 21 185 82.0 NED 1988 Winter Speed Skating Speed Skating Women's 500 metres NA
6 Christine Jacoba Aaftink F 21 185 82.0 NED 1988 Winter Speed Skating Speed Skating Women's 1,000 metres NA
7 Christine Jacoba Aaftink F 25 185 82.0 NED 1992 Winter Speed Skating Speed Skating Women's 500 metres NA
8 Christine Jacoba Aaftink F 25 185 82.0 NED 1992 Winter Speed Skating Speed Skating Women's 1,000 metres NA
9 Christine Jacoba Aaftink F 27 185 82.0 NED 1994 Winter Speed Skating Speed Skating Women's 500 metres NA
10 Christine Jacoba Aaftink F 27 185 82.0 NED 1994 Winter Speed Skating Speed Skating Women's 1,000 metres NA
11 Per Knut Aaland M 31 188 75.0 USA 1992 Winter Cross Country Skiing Cross Country Skiing Men's 10 kilometres NA
12 Per Knut Aaland M 31 188 75.0 USA 1992 Winter Cross Country Skiing Cross Country Skiing Men's 50 kilometres NA
13 Per Knut Aaland M 31 188 75.0 USA 1992 Winter Cross Country Skiing Cross Country Skiing Men's 10/15 kilometres Pursuit NA
14 Per Knut Aaland M 31 188 75.0 USA 1992 Winter Cross Country Skiing Cross Country Skiing Men's 4 x 10 kilometres Relay NA
15 Per Knut Aaland M 33 188 75.0 USA 1994 Winter Cross Country Skiing Cross Country Skiing Men's 10 kilometres NA
16 Per Knut Aaland M 33 188 75.0 USA 1994 Winter Cross Country Skiing Cross Country Skiing Men's 30 kilometres NA
17 Per Knut Aaland M 33 188 75.0 USA 1994 Winter Cross Country Skiing Cross Country Skiing Men's 10/15 kilometres Pursuit NA
18 Per Knut Aaland M 33 188 75.0 USA 1994 Winter Cross Country Skiing Cross Country Skiing Men's 4 x 10 kilometres Relay NA
19 John Aalberg M 31 183 72.0 USA 1992 Winter Cross Country Skiing Cross Country Skiing Men's 10 kilometres NA
20 John Aalberg M 31 183 72.0 USA 1992 Winter Cross Country Skiing Cross Country Skiing Men's 50 kilometres NA
21 John Aalberg M 31 183 72.0 USA 1992 Winter Cross Country Skiing Cross Country Skiing Men's 10/15 kilometres Pursuit NA
22 John Aalberg M 31 183 72.0 USA 1992 Winter Cross Country Skiing Cross Country Skiing Men's 4 x 10 kilometres Relay NA
23 John Aalberg M 33 183 72.0 USA 1994 Winter Cross Country Skiing Cross Country Skiing Men's 10 kilometres NA
Showing 1 to 24 of 271,116 entries


Histogram of knnOlympics$Age
Frequency
0
5000
10000
20000
10
20
30
40
50
60
70
knnOlympics$Age


Histogram of knnOlympics$Weight
Frequency
0
5000
10000
15000
50
100
150
knnOlympics$Weight

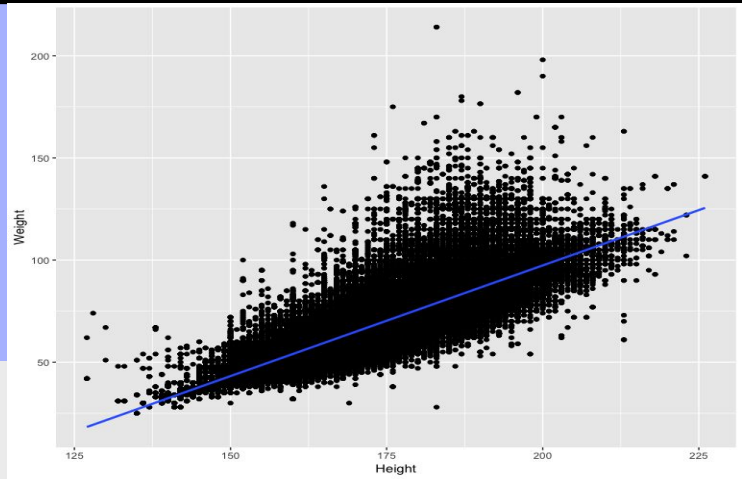
PIE CHART

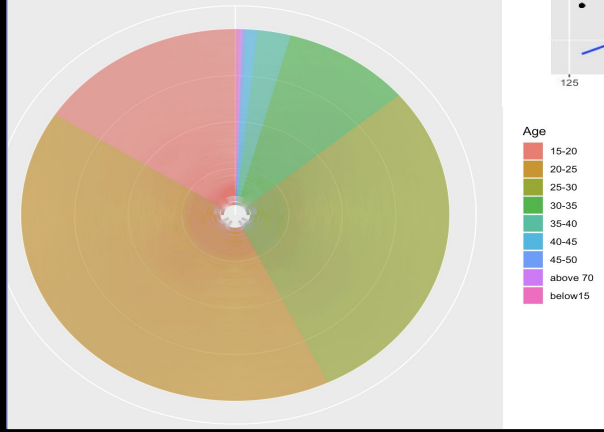
- For KNN and Naive Bayes we had to make all our predictors categorical by splitting them up into groups

SCATTER PLOT

- We were able to reduce the dimensions of our df by noticing that height and weight were

Height to Weight
Correlation=0.796





Size of Age Groups in increments of 5

The three different models

GLM:

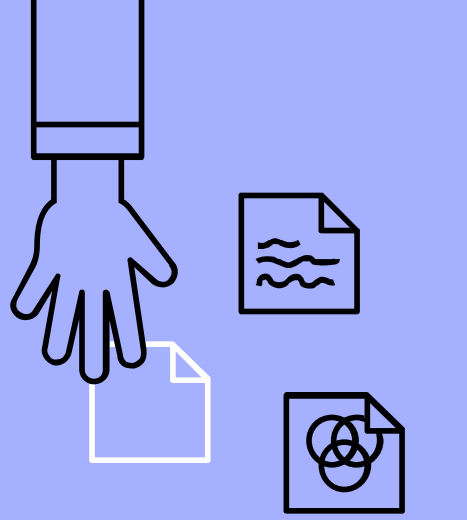
- Fits Generalized Linear Models.
- Puts coefficients to each predictor
  - A negative coefficient: less likely to win
  - A positive coefficient: more likely to win

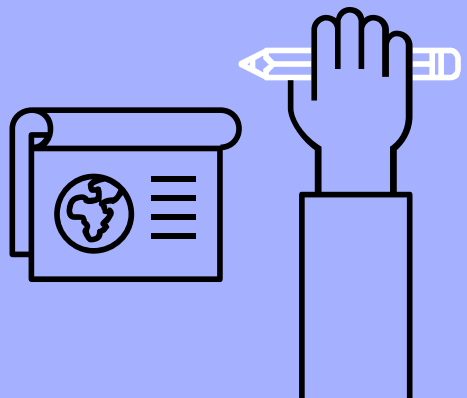
Naive Bayes:

- Computes probabilities based on the given categorical predictor variables.
- Bayes Theorem

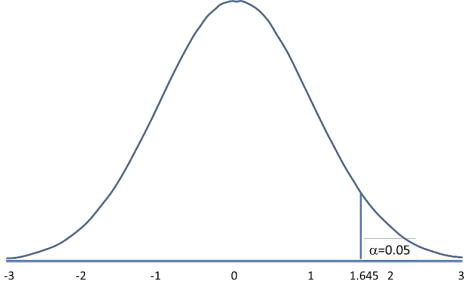
KNN:

- K Nearest Neighbors
- K value is determined with seeing the clusters with the lowest error
- Naive classifier: belong to the same class
- Euclidean Dist: takes the distance between the points



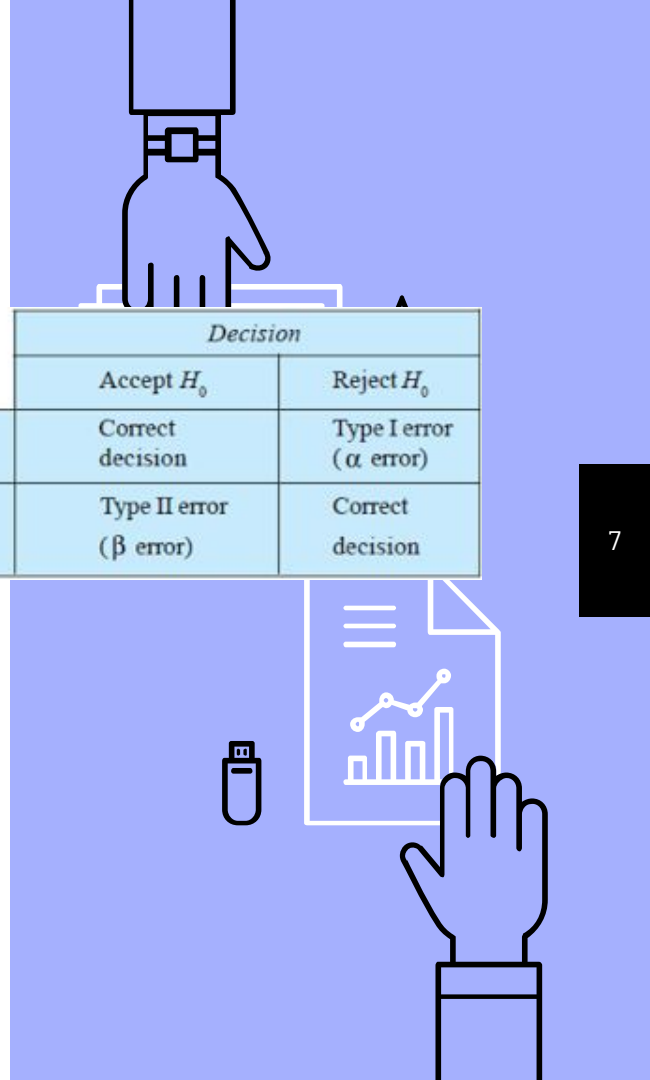


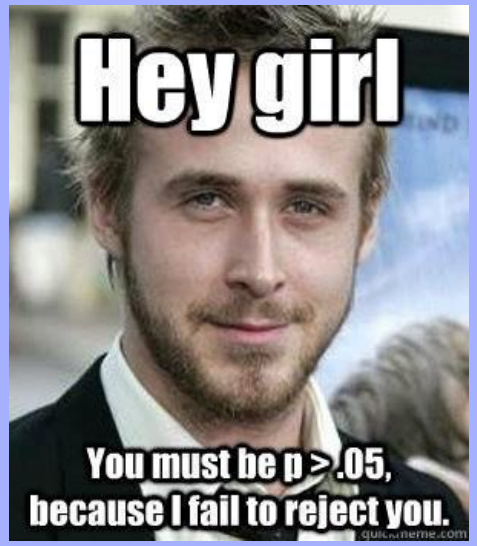
Vocabulary



- P-Value: if P is low reject Ho
- receiver operating characteristic curve, ROC: sensitivity vs 1- specificity
- Area Under Curve, AUC: Area Under the ROC. The range is from .5 to 1 and the closer the area is to one the better
- True Positive: predicting the condition positive correctly
- True Negative: predicting condition in negative correctly
- False Positive: Predicting condition is positive incorrectly
- False Negative: Predicting the condition is negative incorrectly
- Train: data set we used to run the model
- Test: data set tested the model on
- Accuracy: correct identification
- Sensitivity : correctly identify true positive
- Specificity: correctly identify true negative
- F1: Measures rate of performance using recall and precision

	Decision	
	Accept H_0	Reject H_0
H_0 (true)	Correct decision	Type I error (α error)
H_0 (false)	Type II error (β error)	Correct decision





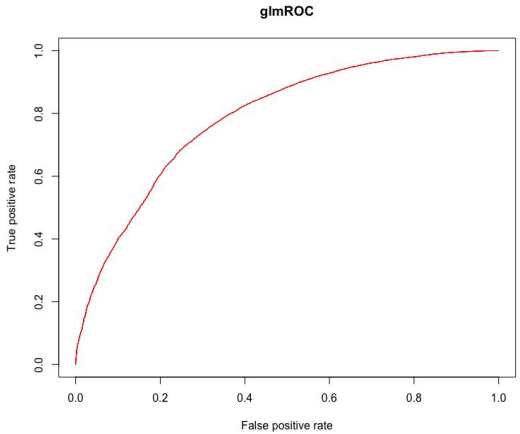
1. Generalized Linear Model

GLM: Generalized Linear Model

	False	True
0	6245	2915
1	2162	6785

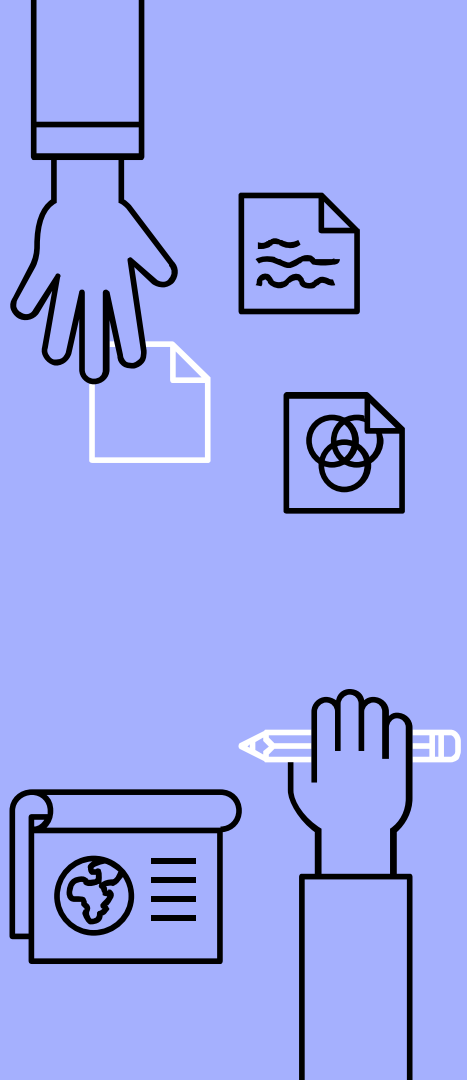
Accuracy: 0.7196
Sensitivity: 0.6995
Specificity: 0.6818
F1: 0.6905

AUC: 0.7201



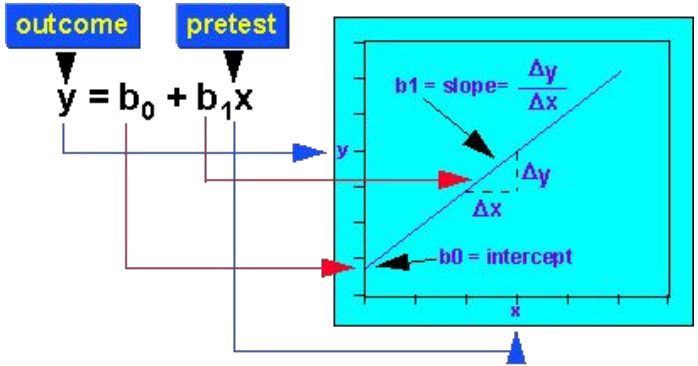
Predictor examples::

- Height
- Gender
- Age
- Women's Archery team
- Women's 4 x 100
- Women's 4 x 400
- Event Archery Men's team
- Men's 1600m relay
- Men's 3 mile
- Men's 3000 miles
- Men's 4 x 100
- Men's 4x400
- Men's CC
- Men's High Jump
- Etc

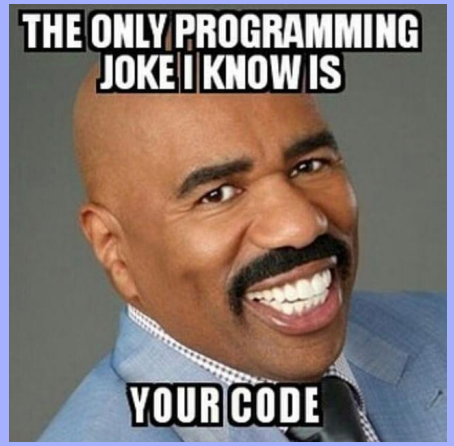


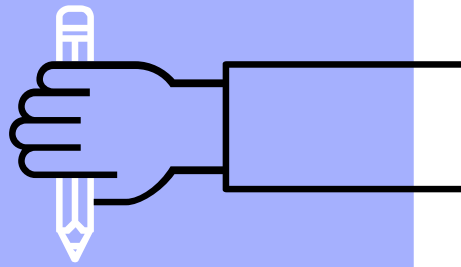
GLM (problem) cont.

- Made our response binary
- Got rid of NAs
- Ran glm
- It wouldn't work because we had to many levels in some of our obs variables
- Once we did some dimension reduction it worked and we got rid of the risk of overfitting
- We were getting really low sensitivity: it was about 8%
- The frequency of wins was extremely low compared to the loss counts.
- Another dataframe with equal frequencies of each

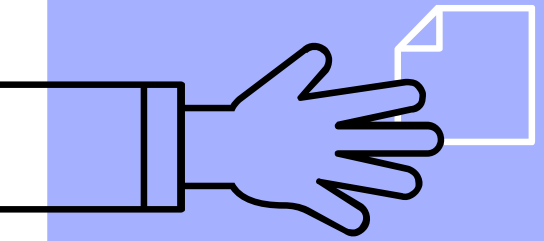



LOOK AT OLD CODE FROM A YEAR AGO TO REFRESH MEMORY
DIDN'T WRITE A SINGLE COMMENT


THE ONLY PROGRAMMING JOKE I KNOW IS
YOUR CODE



Naive Bayes



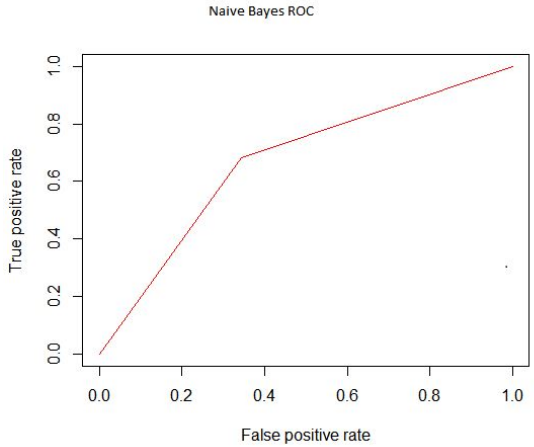

WROTE R CODE NOT USING GOOGLE
WORKED ON THE FIRST TRY

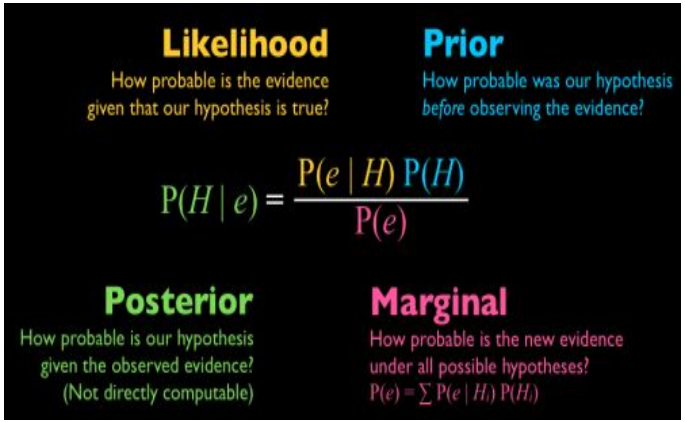

I ARE PROGRAMMER
I MAKE COMPUTER BEEP BOOP BEEP BEEP BOOP

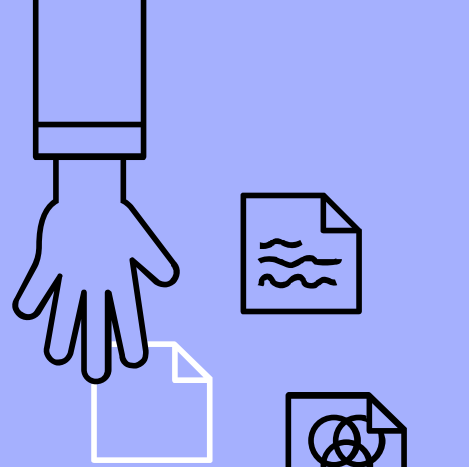
Naive Bayes

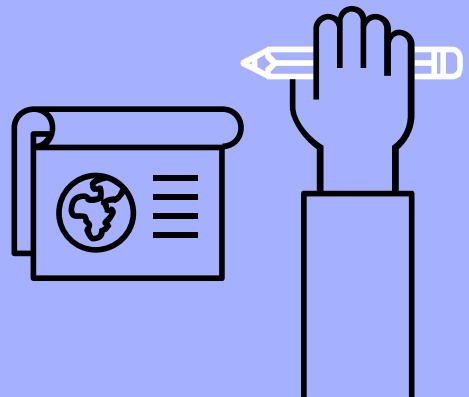
	False	True
0	5981	2837
1	3054	6234

Accuracy: 0.6746
Sensitivity: 0.6872451
Specificity: 0.6782717
F1: 0.6827289
AUC: 0.6692



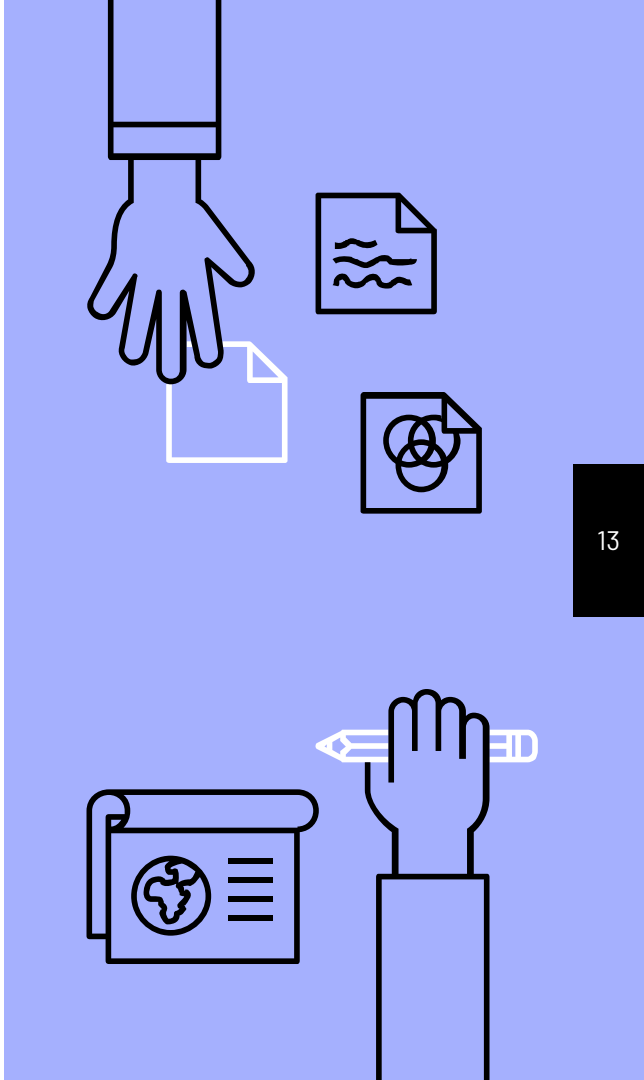






Naive Bayes

- Reduced levels by putting the numerical values in categorical interval and everything else was in factors
- Installed the package e1071
- We ran the test with the categorical variables




IS MY R CODE WORKING
OR THE WINE?

KNN


RUN, R CODE
RUN
memegenerator.net

KNN

Accuracy:0.9883464

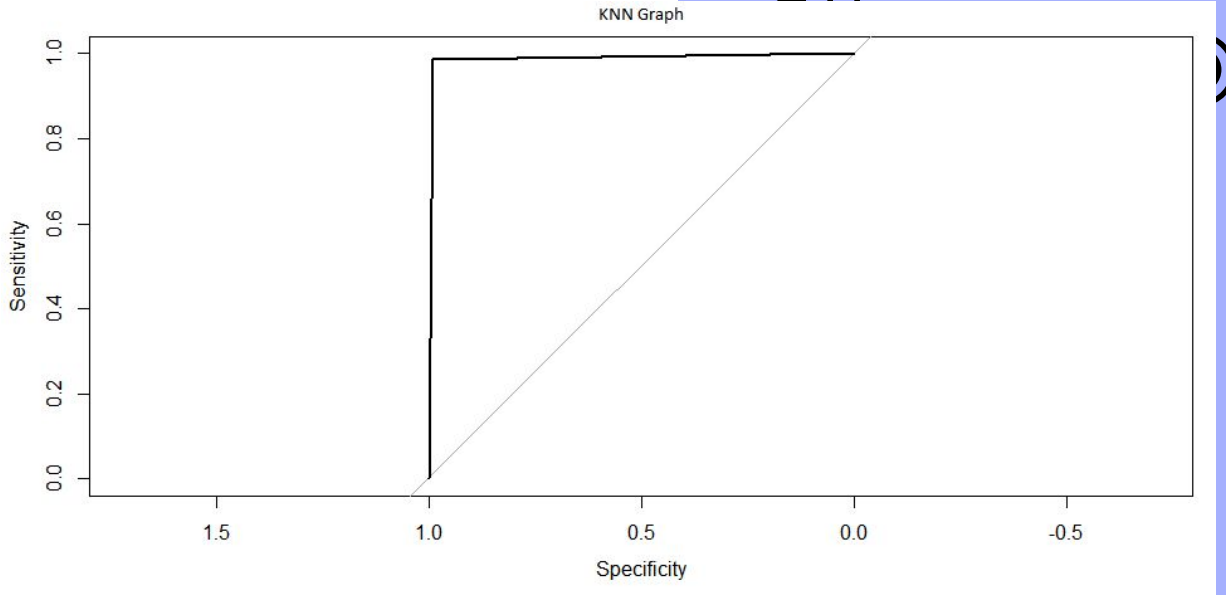
Sensitivity: 0.9870273

Specificity: 0.9869397

F1: 0.9869835

AUC: 0.9884

	False	True
0	8917	118
1	93	8978

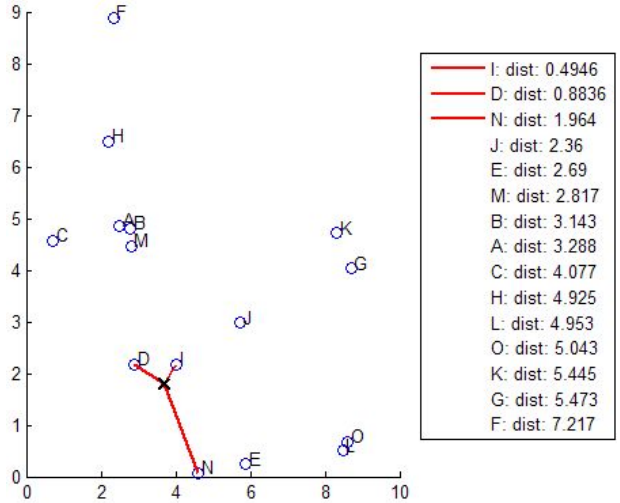


KNN

- Just like the other models we started by reducing dimensions
- Then we found out what would be the best k value
- Then we ran our model

Here is a picture we found

To better e explain the model

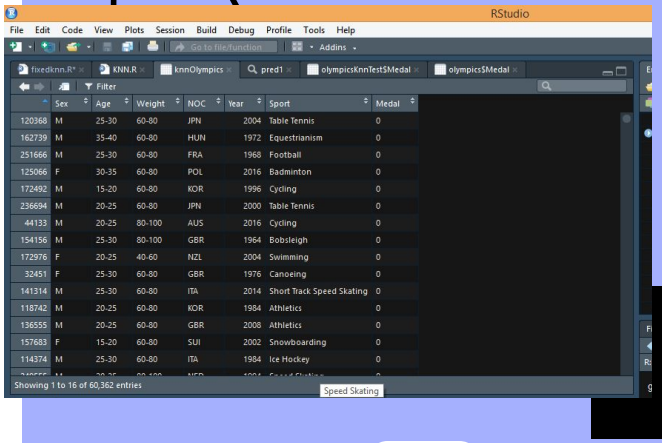


Best Model??

KNN was the best model for our data by far. We used the variables: Sex, Weight, Age, NOC, Year, and Sport. We further broke down the Weight and Age variables by assigning them to intervals.

Next best was our GLM with the predictors: Height, Event, Sex, Age, NOC. The sensitivity and specificity were pretty average around 70%

Lastly, the Naive Bayes fell in last place because of it's low accuracy. We used the predictors Sex, Age, Weight, Height, NOC Year, and Sport.



THANKS!

Any questions?

You can find us at:

@smeragora on GitHub

And

@ShwinyG on GitHub