NBA advance to the playoffs prediction

Final report – Big Data project

שמות המגישים:

שלומי אוחנה (305606899)

צבי קוגן (311674212)

רקע קל – מהו כדורסל ?

משחק הכדורסל הומצא בשנת 1891 בארצות הברית על ידי ד"ר ג'יימס נייסמית', מאמן ספורט מקולג' ימק"א שבספרינגפילד, מסצ'וסטס. מטרתו של נייסמית' הייתה לשמור על כושרם של שחקני קבוצות הקולג' בתקופה שבין סיום עונת הפוטבול לבין פתיחת עונת הבייסבול.

נייסמית' לקח שני חישוקים והרכיבם על הקירות בשני צדי המגרש ומטרת המשחק הייתה לקלוע לתוכם. הוא חילק את שחקניו לשתי קבוצות בגודל של תשעה שחקנים ובהמשך הפך זאת למשחק של חמישה שחקנים בכל קבוצה. עפ"י חוקיו של נייסמית', לשחקנים אסור היה להתקדם עם הכדור שלא באמצעות מסירות מהאחד לשני. עם השנים תפס הספורט תאוצה ולמרות שחלק מן הכללים של המאמן השתנו, גודלה של הקבוצה נשאר קבוע עד ליום זה (וככל הנראה כבר לא ישתנה).

משחק הכדורסל הראשון נערך בינואר 1892 באולם ימק"א שבספרינגפילד והסתיים בתוצאה 0-1 (מדהים לראות כמה שזה השתנה). 4 שנים לאחר מכן בשנת 1896 נפגשו מכללות איווה ושיקגו למשחק המכללות הראשון בארצות הברית, שנערך באיווה. באותה השנה בדיוק התקיים משחק הכדורסל המקצועני הראשון במסגרת הליגה ה"לאומית", שנערך בטרנטון, ניו ג'רזי.

במלחמת העולם הראשונה משחק הכדורסל הוצג לראשונה באירופה ע"י חיילים אמריקאים ומאז נקבע מספר השחקנים בכל קבוצה להיות חמישה. בשנת 1936 התווסף הכדורסל לענפים האולימפיים. בשנת 1949 התאחדו שתי ליגות הכדורסל של ארצות הברית – ליגת ה-BAA וליגת ה-NBL ושינו את השם ל-NBA, שקיימת עד היום ונחשבת לליגת המובילה והטובה בעולם.

במהלך השנים הוכנסו שינויים רבים אל תוך הכדורסל על מנת להגביר את קצב המשחק. בעונת 1954/55 נעשה לראשונה שימוש בשעון זריקות המחייב את הקבוצה התוקפת לזרוק אל הסל תוך 24 שניות. שינוי זה היה אחד מבין רבים אשר התווספו במהלך השנים כגון הגדרת "אזור הצבע", שבו לשחקן אסור לשהות יותר משלוש שניות ברצף. שינויים אלו הובילו למהפכה בעולם הכדורסל שהכירו עד אז ודוגמה לכך נוכל לראות בעובדה שתקופה אשר קדמה לשעון הזריקות, מספר הנקודות אותו קלעה קבוצה עמד בממוצע על 79.5, אולם כבר בעונה הראשונה אליה הוכנס השינוי עמד הממוצע על 93.1 נקודות למשחק.

ד"ר נייסמית' עם הסל הראשון של משחק הכדורסל

משחקי העונה הסדירה בליגת ה-NBA

בליגת ה-NBA יש סה"כ 30 קבוצות המחולקות לשני מחוזות של שלושה בתים כאשר בכל בית ישנן 5 קבוצות. כל קבוצה בליגה משחקת 82 משחקים במהלך העונה, המחולקים שווה בשווה בין משחקי בית לחוץ. כל קבוצה משחקת 4 פעמים נגד קבוצה מהבית שלה, 4 פעמים נגד שש קבוצות מאותו אזור, מבתים אחרים (נקבע ע"י הגרלה), 3 פעמים נגד שאר הקבוצות מאותו האזור ופעמיים נגד כל קבוצה מהאזור השני.

בסיומה של העונה הסדירה, מכל מחוז מעפילות למשחקי הפלייאוף שמונת הקבוצות בעלות המאזן הטוב ביותר. 8 הקבוצות בכל אזור מתמודדות על אליפות האזור בשלושה סיבובים: רבע גמר, חצי גמר וגמר.

סטפן קארי ולברון ג'יימס, שני השחקנים המובילים בליגה. נפגשו 3 שנים אחרונות בסדרת הגמר (צפוי גם בשנה הנוכחית).

מהי שאלת המחקר ?

לאחר שהצגנו מהי שיטת הפלייאוף נוכל להציג את שאלת המחקר שלנו המתבססת על כך. שאלת המחקר שלנו הינה האם ניתן לנבא את הקבוצות אשר יעלו לשלבי הפלייאוף על סמך סטטיסטיקות למשחק של כל אחת מן הקבוצות.

אנו נשתמש במידע שברשותנו על מנת לנתח את תוצאות העבר ובכך ננסה להשתמש ולנבא האם עפ"י אותם נתונים נעפיל לפלייאוף או שלא.

תיאור המידע בו השתמשנו :

בעבודה זו נעזרנו בבסיס הנתונים אשר נמצא באתר basketball reference (https://www.basketball-reference.com) ומכיל מידע רב אודות ליגת ה-NBA.

השתמשנו בטבלאות סטטיסטיות עונתיות של כל קבוצה בליגה מעונת 1982/83 ועד עונת 2016/17. מהעונות הראשונות ועד האחרונות בהן השתמשנו התרחשו שינויים רבים במספר הקבוצות המשחקות בליגה שעלה מ-23 בעונת 1982/83 ל-30 קבוצות בעונת 2016/17.

לקחנו עונות אשר תואמות לאותם נתונים אותם לקחו בשנים האחרונות, לדוגמה, שנים שכללו קליעות מהשלוש אשר נכנס לחוקי ליגת ה-NBA לראשונה בשנת 1979. הטבלאות מכילות את הסטטיסטיקה של כל קבוצה למשחק לפי כל עונה.

ההחלטה על בסיס הנתונים בו נשתמש להמשך העבודה לא הייתה פשוטה, שכן היינו צריכים להיעזר בנתונים שעליהם אכן נוכל להתבסס על מנת לענות על שאלת המחקר שלנו. בסיס הנתונים שלנו כולל עונות רבות אשר אוחדו לכדי טבלה אחת ובה כלל המידע. שמות רבים של קבוצות שונו, על מנת ליצור תיאום בין השמות החדשים של קבוצות שונות לשמות הישנים (שונו מהשמות הישנים אל החדשים).

הטבלאות הסטטיסטיות כוללות בתוכן את העמודות הבאות:

שם העמודה	תיאור
Rk	דירוג הקבוצה עפ"י מספר הנקודות הממוצע בעונה
Team	שם הקבוצה
G	מספר משחקים
MP	מספר דקות אשר שוחקו ע"י כל שחקני הקבוצה למשחק
FG	סלי שדה למשחק
FGA	מספר הניסיונות לקליעה מהשדה
FG%	אחוזי הקליעה לזריקות מהשדה
P3	מספר הקליעות מה-3 שצלחו
PA3	מספר הניסיונות לקליעה מ-3
P%3	אחוזי ההצלחה לקליעה מה-3
P2	מספר הקליעות מה-2 שצלחו
PA2	מספר הניסיונות לקליעה מ-2
P%2	אחוזי ההצלחה לקליעה מה-2
FT	מספר הקליעות מהעונשין שצלחו
FTA	מספר הניסיונות לקליעה מה-3
%FT	אחוזי ההצלחה לקליעה מן העונשין
ORB	ריבאונדים בהתקפה
DRB	ריבאונדים בהגנה
TRB	סה"כ ריבאונדים
AST	אסיסטים
STL	חטיפות
BLK	בלוקים/חסימות
TOV	איבודי כדור
PF	עבירות אישיות
PTS	נקודות
Playoff	עלייה לפלייאוף (1 אם עלה, 0 אם לא עלה)

מה היו שלבי הביצוע בהכנת הפרויקט ?

השלב המקדים של העבודה בוצע בטיוטות רבות בהן חשבנו והצענו על איזה בסיסי מידע אנו צריכים להתבסס על מנת שנוכל להפיק את התוצאות/מסקנות אליהן אנו מעוניינים להגיע, כיצד עלינו לגשת אל החומר וכיצד ניתן להשתמש בו בצורה נכונה כך שנוציא מן המידע שלנו את הטוב ביותר.

לאחר שתכננו בקפידה את הדברים אותם אנו מעוניינים לעשות התחלנו בשלבי הכנת המידע איתו נוכל לעבוד בעבודה, שלב שכלל בתוכו את טעינת טבלת בסיס הנתונים, טבלה חדשה ובה את המידע הרלוונטי מתוך טבלת הנתונים המקוריים אותם טענו אל העבודה, הפיכתה לנומרית על מנת להשתמש בה בצורה מסוימת וכו'.

השלב השני של העבודה כולל בתוכו את הצגת הנתונים הרלוונטיים לשאלת המחקר שלנו אשר עוסקת בניבוי עלייה של קבוצות לפלייאוף ה-NBA הקבוצות על סמך סטטיסטיקות של העונה הסדירה. הצגנו גרפים שונים העוסקים בסטטיסטיקות שונות המציגות בפנינו את הנתונים של שנים רבות של קבוצות אשר עלו לפלייאוף וקבוצות, שאינן עלו לשלבי הפלייאוף, ומכך ניסינו להסיק מסקנות לגבי המשך העבודה.

השלב השלישי של העבודה מכיל בתוכו את פירוק המידע וחלוקתו ל-training set ו-testing set. זהו שלב חשוב, שעליו יבנה השלב הבא של העבודה, בו אנו מציגים ומשתמשים בניבויים שונים על מנת למצוא את הניבוי הטוב ביותר, על כן חשוב מאוד שאנו ניצור חלוקה נכונה בין ה-training set וה-testing set.

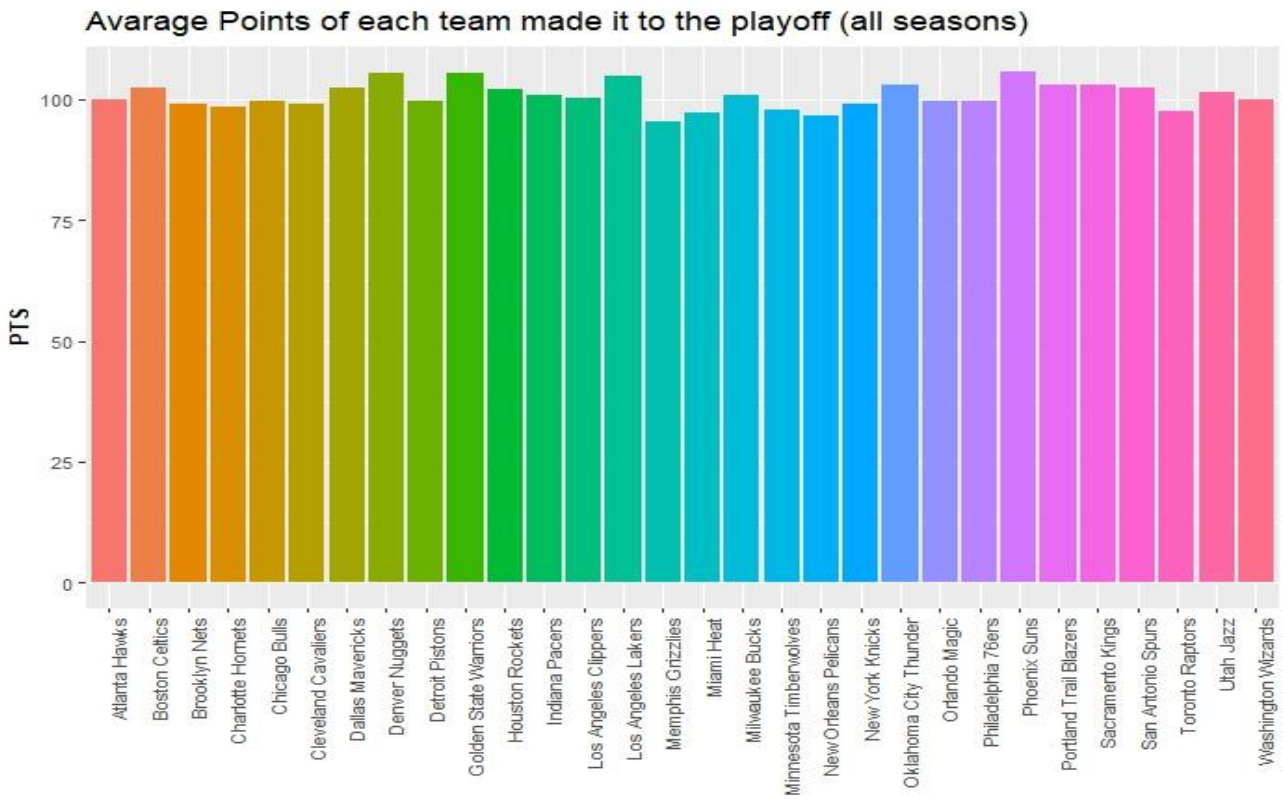
השלב הרביעי של העבודה הינו בניית המודלים של ניבוי וחיזוי לשאלת המחקר. חשוב ששלב זה יעשה בצורה טובה על מנת שנוכל להסיק מסקנות נכונות וזהו בעצם ליבו של העבודה, מפני שחלק זה נותן לנו את התוצאות הסופיות וזהו בעצם החלק שעונה על שאלת המחקר. בשלב זה השתמשנו במספר אלגוריתמי ניבוי בכדי שנוכל להבין איזה מודל מתאים לנו יותר לעבודה ולשאלת המחקר הספציפית, וכמו שנראה בהמשך הבנו כי לא כל מודל עוזר לנו בצורה טובה כמו המודלים האחרים, שכן התוצאות לפעמים אף שונות בצורה משמעותית.

השלב החמישי והאחרון של העבודה הוא בעצם שלב הסקת המסקנות. זהו שלב בו התבוננו על התוצאות אליהן הגענו עם סיום העבודה וניסינו לנתח בצורה הטובה ביותר ולהסיק את המסקנות הנכונות בכל הנוגע לשאלה עליה אנו מעוניינים לענות. לטעמנו, זהו השלב הקשה ביותר בפרויקט, שכן זהו השלב הקריטי ובעל החלק הקשה יותר בעבודה בעצם – ניתוח התוצאות והסקת המסקנות בצורה נכונה ולא פזיזה.

קצת סטטיסטיקות ...

(קובץ Visualization.R)

מספר הנקודות הממוצע של כל קבוצה כאשר עלתה לפלייאוף במהלך כל השנים שנבדקו.

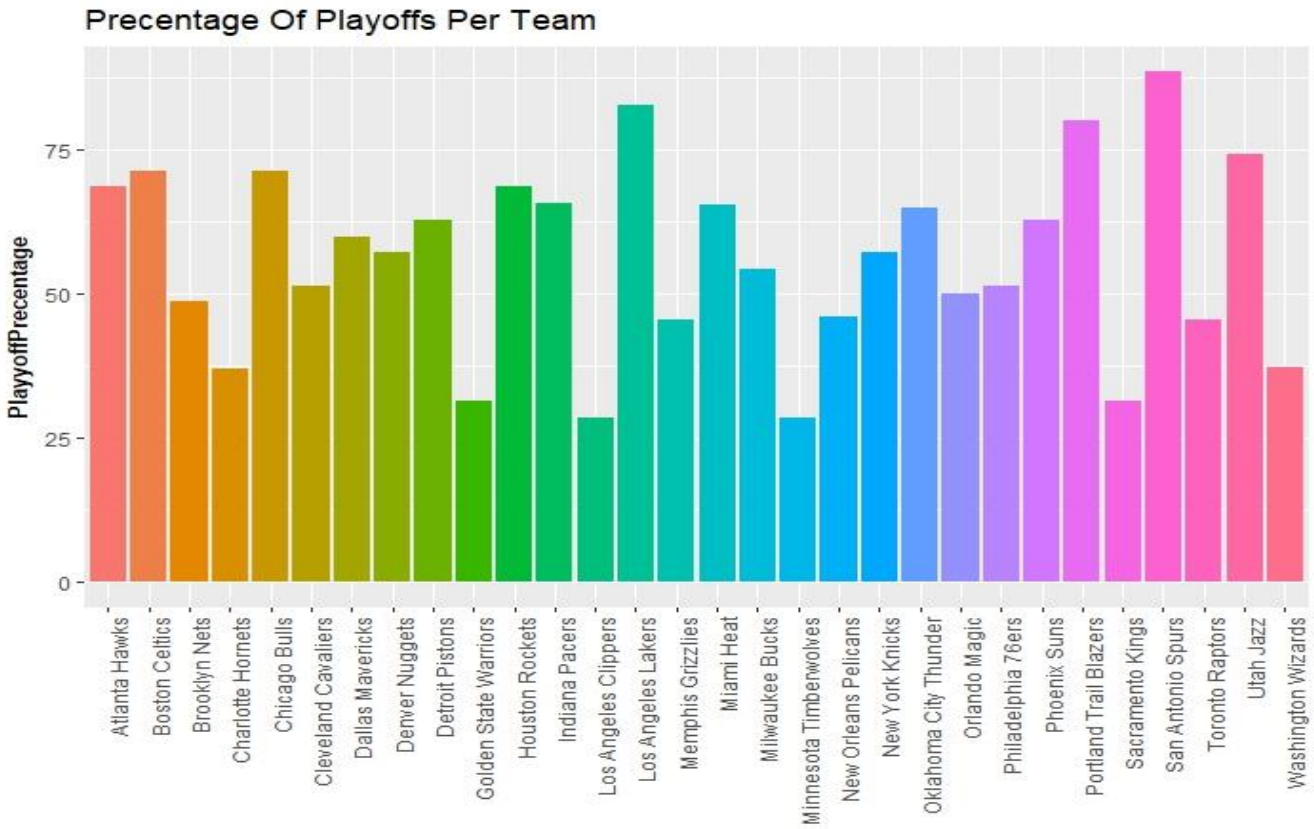



```
seasons <- read.csv("NBA data base 82-83_16-17.csv")


#Average points of each team from all seasons made it to the playoffs round
teamMeanPoints <- seasons[seasons$Playoff == 1,] %>% group_by(Team) %>% summarise(PTS = mean(PTS))
teamMeanPoints$PTS <- round(teamMeanPoints$PTS,digits = 2)
avgPtsPlot <- ggplot(teamMeanPoints, aes(x=Team, y=PTS)) +
  geom_bar(aes(fill=Team),stat = "identity") +
  scale_x_discrete()+ ggtitle("Average Points of each team made it to the playoff (all seasons)")  +
  theme(plot.title = element_text(hjust = 0)) +
  theme(axis.title.x = element_blank(),legend.text=element_text(size=8),legend.position="none") +
  theme(axis.text=element_text(size=8),

        axis.title=element_text(size=10,face="bold"))+
  theme(axis.text.x = element_text(angle = 90, hjust = 1))
avgPtsPlot
```

אחוז ההצלחה בעלייה לשלבי הפלייאוף ה-NBA לכל קבוצה.




```
# Precentage of playoff appearances per team
numberOfPlayoffPerTeam <- seasons %>% group_by(Team) %>% summarise(Playoff = sum(Playoff),
                                                                   numberOfSeasons = sum(Rk-Rk+1))
numberOfPlayoffPerTeam$PlayyoffPrecentage <- (numberOfPlayoffPerTeam$Playoff/numberOfPlayoffPerTeam$numberOfSeasons)*100
numberOfPlayoffPerTeam.Plot <- ggplot(numberOfPlayoffPerTeam,aes(Team,PlayyoffPrecentage)) +
  geom_bar(aes(fill=Team),stat = "identity") +
  ggtitle("Precentage Of Playoffs Per Team")  +
  theme(plot.title = element_text(hjust = 0),legend.text=element_text(size=8)) +
  theme(axis.title.x = element_blank()) +
  theme(axis.text=element_text(size=9),
        axis.title=element_text(size=10,face="bold"))+
  theme(axis.text.x = element_text(angle = 90, hjust = 1),legend.position="none")
numberOfPlayoffPerTeam.Plot
```

אז מה בעצם עשינו ?

העבודה החלה כמובן בטעינת קובץ בסיס הנתונים שלנו.

לאחר טעינת הקובץ, הפעלנו את פונקציית הכנת המידע, בה הפכנו את שמות הקבוצות למספרים (כאשר כל מספר קבוע לקבוצה מסוימת) ולאחר מכן הורדנו את עמודות ה- rank אשר הופיעו בטבלת הנתונים שלנו, שכן אין בה צורך והדירוג מבוסס עפ"י מספר הנקודות הממוצע של הקבוצה בעונה ספציפית.

```
all_Data <- read.csv('nba_seasons_82-83_17.csv')

# Vector of the NBA team names (helps us later)
teamNames = c('Atlanta Hawks', 'Boston Celtics', 'Brooklyn Nets', 'Charlotte Hornets', 'Chicago Bulls',
              'Cleveland Cavaliers', 'Dallas Mavericks', 'Denver Nuggets', 'Detroit Pistons', 'Golden State Warriors',
              'Houston Rockets', 'Indiana Pacers', 'Los Angeles Clippers', 'Los Angeles Lakers', 'Memphis Grizzlies',
              'Miami Heat', 'Milwaukee Bucks', 'Minnesota Timberwolves', 'New Orleans Pelicans', 'New York Knicks',
              'Oklahoma City Thunder', 'Orlando Magic', 'Philadelphia 76ers', 'Phoenix Suns', 'Portland Trail Blazers',
              'Sacramento Kings', 'San Antonio Spurs', 'Toronto Raptors', 'Utah Jazz', 'Washington Wizards')

# Prepare the data.table for the future use
data_preparation <- function(data_table){
  data_table$Team = as.numeric(factor(data_table$Team,levels = teamNames,labels = 1:30))
  data_table <- subset(data_table, select = -c(1))
  return(data_table)
}

season.games <- data_preparation(all_Data)
```

לאחר שהכנו את הקובץ כפי שאנו רוצים הגיע השלב המקדים לשלב בו נפעיל את אלגוריתמי החיזוי – שלב חלוקת הנתונים ל- training set ול- testing set.

את הטבלה חילקנו בעזרת פונקציות split ו- subset. בנוסף לכך, בחלק מהאלגוריתמים השתמשנו בנתונים לאחר שהפעלנו עליהם את פונקציית scale, כאשר את הפונקציה בעצם הפעלנו על המשתנים הבלתי תלויים שלנו, כאשר עמודת ה- playoff היא בעצם המשתנה התלוי שלנו.

```
# Splitting of the training test and the testing set
set.seed(123)
split <- sample.split(season.games$Playoff,SplitRatio = 0.75)
training_set <- subset(season.games,split == TRUE)
testing_set <- subset(season.games,split == FALSE)
training_set_scaled <- training_set
testing_set_scaled <- testing_set
training_set_scaled[,-25] <- scale(training_set[,-25])
testing_set_scaled[,-25] <- scale(testing_set[,-25])
training_set$Playoff <- factor(training_set$Playoff)
testing_set$Playoff <- factor(testing_set$Playoff)
```

Decision Tree

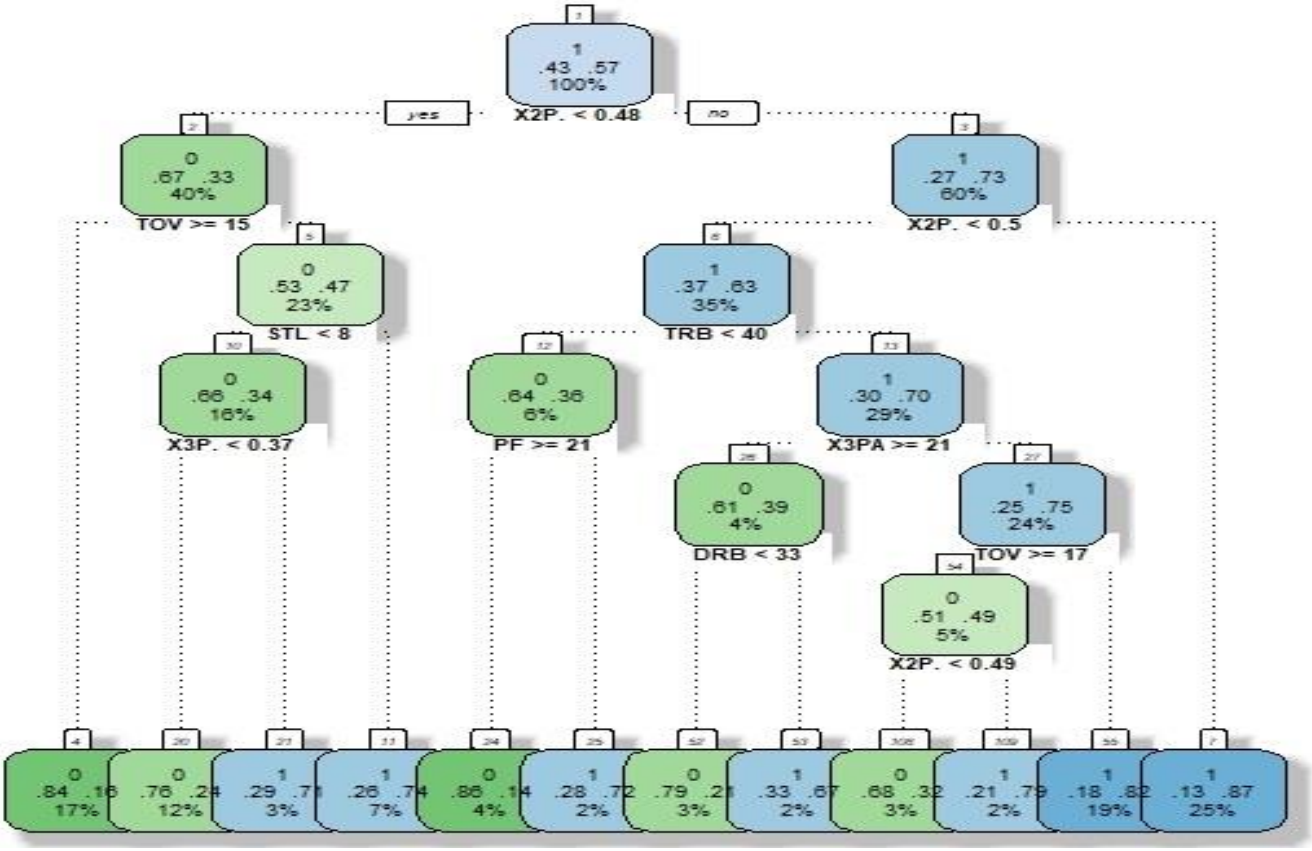
כעת הגענו אל החלק המעניין של העבודה, שזהו כמובן שלב הפעלת אלגוריתמי החיזוי ובעצם לראות ולחוש את המושג שנקרא Machine Learning.

אלגוריתם החיזוי הראשון בו החלטנו להשתמש הוא אלגוריתם Decision tree. כפי שאנו רואים, לקחנו את ה- training set שברשותנו ואימנו אותו על מנת למצוא את החיזוי האופטימלי ביותר עבורנו.

```
#Predict using decision tree model
decisionTree = rpart(Playoff~., method="class", data=training_set, na.action = na.pass)
plot(decisionTree)
text(decisionTree)
```

בשורות 2-3 נוכל לראות את עץ ההחלטה אותו יצרנו בעזרת האלגוריתם. עץ ההחלטה הוא גרף המייצג את הבחירות ואת תוצאותיהן בצורה של עץ. העץ בחר את חמשת המשתנים הבלתי תלויים, אשר לדעתו הם החשובים והמשמעותיים ביותר להצלחתו של הניבוי. עץ החלטה הוא עץ בינארי מלא המורכב מצמתי החלטה שבכל אחד מהם נבדק תנאי מסוים על מאפיין מסוים של התצפיות ועלים המכילים את הערך החזוי עבור התצפית המתאימה למסלול שמוביל אליהם בעץ.

לצורך דוגמה נוכל לקחת את המסלול השמאלי ביותר עפ"י העץ, כ- 17% אחוז מקבוצת האימון שלנו הם קבוצות אשר קולעות זריקות ל-2 באחוז נמוך מ-48 ועם מספר איבודי הכדור של 15 ומעלה והינן קבוצות שלא עלו לפלייאוף. האלגוריתם צדק ב- 84% מן המקרים הללו.



לאחר אימון ה- training set עשינו prediction על ה- testing set. בשורה הבאה בעצם יש לנו בתורך משתנה את ה- confusion matrix ובתוכו יש לנו טבלה המכילה את הניבויים שלנו עפ"י המודל ואת מספר הפעמים שפגענו או לא עם הניבוי. מתוך טבלה זו מסיקים את אחוזי ההצלחה של האלגוריתם, שאצלנו נמצא בשורה האחרונה ומניב לנו 75.82% אחוזי הצלחה בחיזוי העולות לשלבי הפלייאוף.

```
predictDesicionTree = predict(object = decisionTree, newdata = testing_set, type = "class")
treeConfusionMatrix =  confusionMatrix(data = predictDesicionTree,reference = testing_set$Playoff)
decisionTreeAcuuracy = treeConfusionMatrix$overall['Accuracy']
```

Random Forest

אלגוריתם זה הוא שיפור של האלגוריתם הקודם בו השתמשנו וזאת בשל העובדה שהוא משתמש במספר רב של עצי החלטה על מנת לחזות עפ"י הנתונים העומדים לרשותו.

בשורה הראשונה הפעלנו את האלגוריתם על ה- training set השני שלנו, שלו עשינו scale. לאחר מכן השתמשנו בפונקציית החיזוי על testing set (גם הוא scaled) ובשורה שלאחריו העברנו כל תוצאה הגדולה מ- 0.5 להיות מיוצגת ע"י 1, כאשר נסמן ב-0 אחרת.

על מנת למצוא את אחוזי הדיוק של האלגוריתם עשינו ממוצע לכל מה בתואם בין binaryRandomForest לבין ה- testing set עליו ערכנו את החיזוי, מה שהניב לנו 78.28% הצלחה בכל הנוגע לקביעת זהות העולות לפלייאוף.

```
# Predict using random forest model
randomForest = randomForest(as.factor(training_set_scaled$Playoff) ~., data = training_set_scaled)
predictRandomForest = predict(randomForest, testing_set_scaled[,-25], type="prob")[,2]
binaryRandomForest = ifelse(predictRandomForest > 0.5,1,0)
randomForestAccuracy = mean(binaryRandomForest == testing_set_scaled$Playoff)
randomForest.confusionMatrix = randomForest$confusion
```

Linear Regression

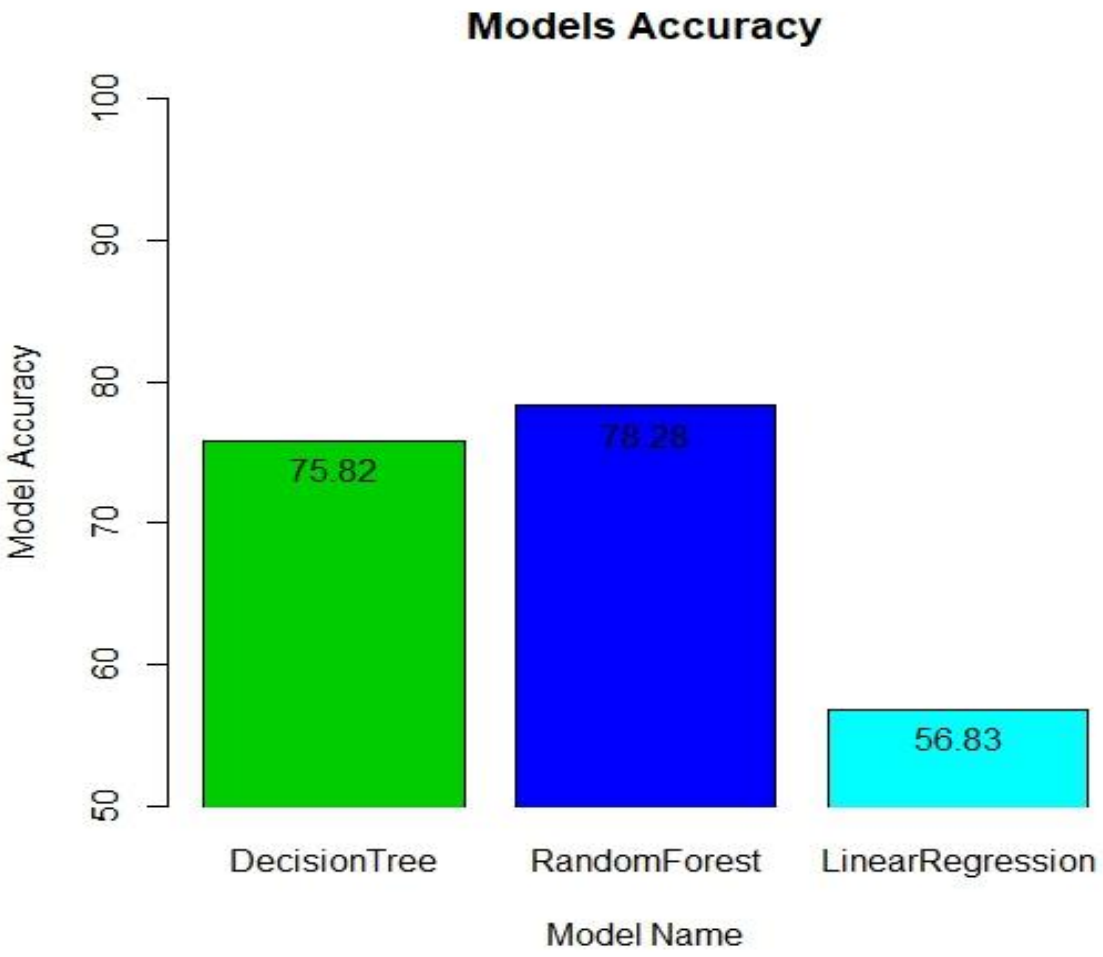
האלגוריתם האחרון בו השתמשנו הוא הרגרסיה הלינארית.

אל המשתנה relation הכנסנו את הנתונים לאחר הפעלת האלגוריתם על ה- training set ולאחר מכן הפעלנו על משתנה זה את החיזוי. כמו באלגוריתם Random forest גם כאן הפכנו את התוצאה שבידינו לבינארית, כאשר כאן כל תוצאה אשר קטנה מ-0 הפכנו ל-0 וכל תוצאה אחרת שגדולה מ-0 הפכנו ל-1.

בחלק הבא השווינו את התוצאות בין התוצאה הבינארית שיצרנו לבין עמודת הפלייאוף ב- testing set עליו הפעלנו את החיזוי וכך הצלחנו בשורה האחרונה להסיק מהם תוצאות הניבוי של האלגוריתם כאשר חילקנו את מספר ההצלחות בסך הכולל של הניבויים, מה שנתן לנו אחוז הצלחה של 56.83%, שנמוך משמעותית משני האלגוריתמים הקודמים בהם השתמשנו לחיזויים.

```
# Predict using linear regression model
relation <- lm(Playoff ~., data = training_set_scaled);
predictRegression <-  predict(relation, data = testing_set_scaled);
#Binary prediction
predictRegression[predictRegression<0] <- 0;
predictRegression[predictRegression>0] <- 1;
correctPredictions <- predictRegression == testing_set_scaled[,25];
predictionRate = sum(correctPredictions)/length(predictRegression)
```

תוצאות הפרויקט :



כפי שניתן לראות וכפי שכבר הבנו בהסברים המקדימים, התוצאות מראות כי המודל אשר ינבא לנו בצורה הטובה ביותר את הקבוצות אשר צפויות לעלות לשלבי הפלייאוף הוא מודל ה- Random_forest אשר חוזה את זהות העולות בהצלחה של 78.28% ובכך נותן לנו חיזוי ברמה טובה (אולם לא הייתי שם את כל הכסף על זה). לא רחוק ממנו נמצא מודל ה- Decision_tree אשר נותן לנו הצלחה צפויה של 75.82% בכל הנוגע לזהות העולות לשלבי הפלייאוף.

למרות הניבויים הטובים עליהם דיברנו עד כה ישנם מודלים אשר לא מתאימים לניבוי מסוג זה, זאת נוכל לראות עפ"י אחוזי הניבוי אותם הניב לנו אלגוריתם הניבוי Linear_regression אשר הצלחתו עומדת על 56.83% בלבד (!), דבר עליו לא ניתן להסתמך כלל וכלל. אנו יכולים לדעת כי אלגוריתם זה אינו מתאים על סמך העובדה כי ישנם אלגוריתמים אחרים המנבאים ומראים לנו תוצאות שונות לגמרי (לטובה) ממה שנתן לנו אלגוריתם ה- Linear_regression.

כעת, לאחר שבחנו את תוצאות שלושת האלגוריתמים בהם השתמשנו על מנת לחזות את זהות הקבוצות העולות לפלייאוף נוכל להתמקד באלגוריתם Random_forest אשר הניב לנו את התוצאה הטובה ביותר, נרצה להסיק מהם המשתנים אשר תרמו הכי הרבה להצלחה בניבוי מתוך מספרם הרב של הסטטיסטיקות בהם נעזר האלגוריתם החיזוי.

נקודת מבט על האלגוריתם המנצח (!)

כיצד חושבו אחוזי הדיוק ?!

אם כן, האלגוריתם אשר נתמקד בו הינו אלגוריתם Random_forest אשר ניבא בצורה הטובה ביותר ועם האחוזים הגבוהים ביותר את הקבוצות אשר יעלו לפלייאוף.

נתבונן על ה- Confusion_matrix :

	0	1	class.error
0	226	88	0.2802548
1	67	351	0.1602871

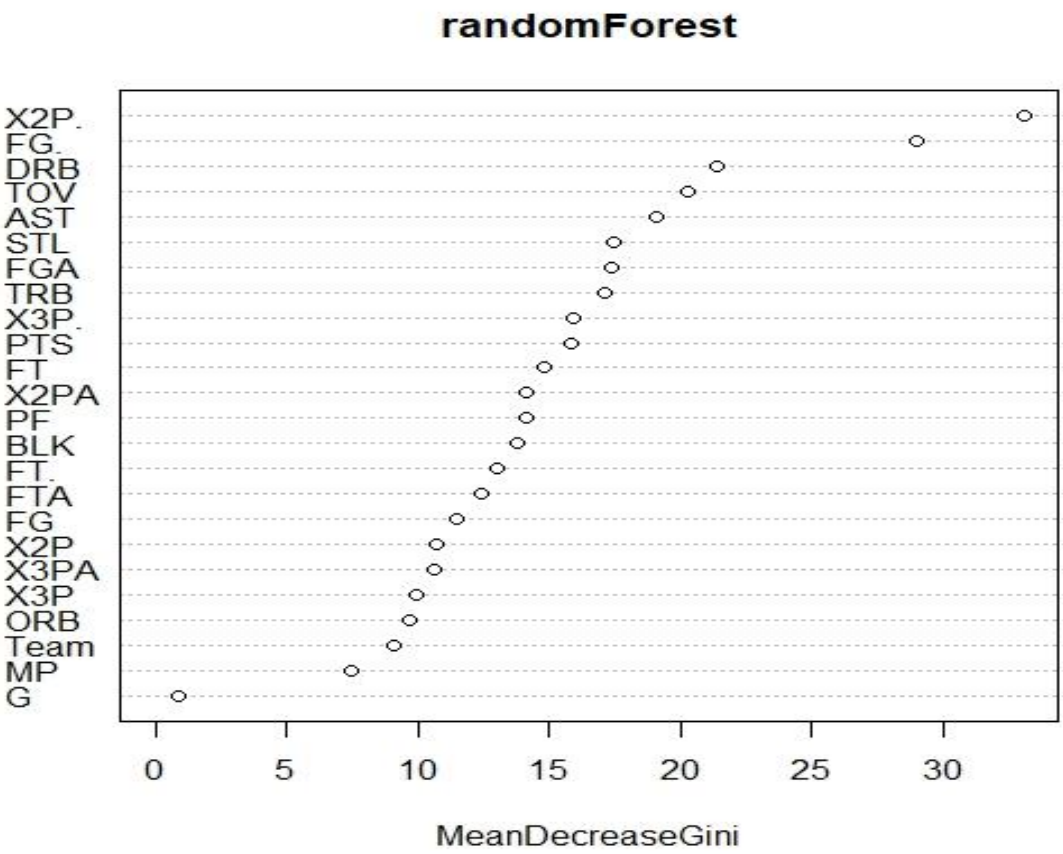
בטבלה זו אנו רואים כי ניבאנו 226 פעמים כי קבוצה לא תעלה לפלייאוף והיא אכן לא עלתה, לעומת 88 ניבויים כי קבוצה תעלה לפלייאוף, אולם היא כשלה לעלות. (שורה ראשונה)

בנוסף, כאשר ניבאנו 67 פעמים כי קבוצה לא תעלה לפלייאוף בסופו של דבר היא אכן עלתה וכאשר ניבאנו 351 פעמים שקבוצה תעלה לפלייאוף היא אכן עלתה. (שורה שנייה)

מכאן שאם נחשב את מספר הפעמים אותם ניבאנו מסך כל הניבויים אשר קרו או שלא קרו נגיע לאחוז דיוק של 78.28% כמו שאנו כבר יודעים.

מהם הגורמים המשפיעים ביותר ?

כעת, לאחר שאנו מבינים כיצד חושב אחוז הדיוק של האלגוריתם אנו רוצים לדעת מה היו הגורמים המרכזיים להצלחה של אלגוריתם זה (חוץ מהשוני בין האלגוריתמים עצמם), כלומר, מה היו הפיצ'רים, שהשפיעו בצורה הגדולה ביותר על הצלחת הרצת האלגוריתם.




```
varImpPlot(randomForest)
```

בגרף שלפנינו (שורה 68) נוכל לראות בבירור את חשיבותם של כל המשתנים אשר ניתנו לאלגוריתם ובעזרתם הוא נתן את החיזוי שלו. אפשר לראות כי הדבר החשוב ביותר לחיזוי נכון עפ"י Random forest הינו אחוזי הקליעה מ-2, במקום השני זה אחוזי הקליעה מן השדה ופער רב אחריהם אלו הם הריבאונדים בהגנה כמו גם איבודי כדור ואסיסטים.

כל מה שהוזכר למעלה הם חמשת הקריטריונים החשובים ביותר לחיזוי נכון והם המבדילים בין עלייה לשלבי הפלייאוף לבין ישיבה מול המסך בשלב מוקדם יותר של העונה.