**פרוייקט למידת מכונה באמצעות R**

**מגיש: רואי גיגי**

**תעודת זהות: 211359187**

**מרצה: דוקטור רוני וייצמן**

**סוגי למידת מכונה:**

**למידה מונחית** (Supervised Learning) – מודל לומד ממדגם הכולל קלטים ותשובות.

**למידה לא מונחית** (Unsupervised Learning) – המודל מנסה למצוא מבנים פנימיים בנתונים ללא תשובות ידועות מראש.

**למידה חיזוקית** (Reinforcement Learning) – סוכן מקבל החלטות על פי תגמול חיובי או שלילי.

**עץ החלטה** (decision tree) הינו הבסיס ליער רנדומלי (random forest) . עץ החלטה הינו אלגוריתם יחסית פשוט ואינטואיטיבי בלמידה מונחית (כאמור, למידה ממדגם שבו קלטים ותשובות). האלגוריתם מבצע סדרת שאלות בינאריות על המשתנים (features) במטרה לסווג דוגמה לקבוצה או לחזות ערך. בכל "צומת" נשאלת שאלה (למשל במקרה שלי, בו משתנה המטרה שאני רוצה לחזות הוא איכות האלכוהול על סמך המשתנים המספריים: "האם כמות האלכוהול > 10.5?"), אם התשובה היא כן – הדוגמה תעבור לענף אחד; ואם לא – לענף אחר, כמו עץ שמסתעף עם ענפים, כשכל ענף הוא עוד שלב ושאלה חדשה. בסוף התהליך, נבנית "מפת החלטות" שממשיכה להתפצל עד שמגיעים לעלים (השלב הסופי, בסוף כמות השאלות שנקבעו, כמות הענפים) בהם מתקבלת התחזית הסופית (למשל במקרה שלי: "יין מצוין", "יין טוב", או "יין גרוע"). האלגוריתם בוחר את השאלות (הפיצולים) באופן שממקסם את ההפרדה בין הקבוצות – לרוב לפי קריטריון מתמטי כמו מידע הדדי (Information Gain) או אימפוריטה של ג'יני .(Gini Impurity) עם זאת, עצים בודדים נוטים להתאים את עצמם יותר מדי לנתוני האימון (Overfitting) ולתת ביצועים פחות טובים על נתונים חדשים.

האלגוריתם שבו השתמשתי,  **Random** **Forest**הוא שיפור של עץ החלטה בודד, ומשלב בתוכו עצים רבים, אשר מצמצמים את הטעויות על ידי כך שהם מייצר גיוון ומדייקים את ההחלטות. אלגוריתם זה נחשב לאחד האלגוריתמים החזקים והפופולריים ביותר בתחום למידת המכונה המונחית. האלגוריתם הוצג לראשונה על ידי ליאו בריימן (Leo Breiman) בשנת 2001. כאמור, הרעיון המרכזי הוא לא להסתמך על עץ יחיד, אלא לבנות אוסף של עצים – כל אחד מהם לומד דגימה שונה מהנתונים ובוחר תת-קבוצה אקראית של משתנים בכל פיצול. כך, נוצרת מעין "הצבעה דמוקרטית" של כל העצים. בנוסף יורדת ההשפעה מנתוני קצה, שעשויים להשפיע על עץ יחיד.

**עקרונות העבודה של : Random Forest**

**- Bootstrap Aggregating** בור כל עץ ביער, נבחרת דגימה אקראית מתוך הדאטה (עם החזרה). כלומר, כל עץ רואה סט שונה.

**- Random Feature Selection**בכל צומת (שאלה), נשקלת רק תת-קבוצה אקראית של משתנים (features), ולא כולם.

**ריבוי עצים -** האלגוריתם יוצר מאות עצים ולעיתים אף אלפים. התחזית הסופית מתקבלת ע"י ממוצע (בבעיות רגרסיה) או הצבעת רוב (בבעיות סיווג).

**הספרייה randomForest :**

הספרייה randomForest של R מבצעת את התהליך הנזכר בצורה פשוטה ונוחה. הספרייה מאפשרת לבנות מודלים של יער רנדומלי הן לבעיות סיווג (classification) והן לבעיות רגרסיה ( (regression.

**פרמטרים חשובים בספרייה** (אפשר לשחק איתם לשיפור המודל):

ntree - מספר העצים ביער (ברירת מחדל: 500).

mtry- מספר המשתנים שייבחרו באופן אקראי בכל פיצול (ברירת מחדל בבעיות קלסיפיקציה – שורש ממספר המשתנים המלא).

nodesize - מספר מינימלי של דגימות בכל עלה.

- maxnodesמספר מרבי של עלים בעץ.

- importance = TRUEמאפשר לקבל את תרומת כל משתנה לתחזיות המודל.

בנוסף, הספרייה גם כוללת פונקציות נלוות להערכת ביצועי המודל, ניתוח חשיבות משתנים, וגרפים.

**יתרונות השיטה:**

* דיוק גבוה יחסית למודלים פשוטים.
* פחות רגישות לרעש או נתונים חריגים.
* יכולת לעבוד עם הרבה משתנים ולגלות משתנים חשובים.

**חסרונות השיטה:**

* קשה לפרש את התחזיות (מודל "קופסה שחורה").
* דורש יותר זמן חישוב וזיכרון ממודל בודד.
* לא מתאים כאשר נדרש הסבר פשוט וברור למקבלי החלטות.

**מקורות:**

Breiman, L. (2001). Random Forests. Machine Learning, 45(1), 5-32 (1)

Liaw, A., & Wiener, M. (2002). Classification and Regression by randomForest. R News (2)

Hastie, T., Tibshirani, R., & Friedman, J. (2009). The Elements of Statistical Learning (2nd ed.) (3)

**תיאור ה־Dataset**

**\*\*\*בעבודה מצורפים גם הקובץ של העבודה המלאה, וגם קבצים המתארים כל חלק בנפרד (ונושאים את שם החלק בכותרת). בקבצים אשר מתארים כל חלק בנפרד, ישנם עוד ניתוחים וגרפים שאינם נכנסו לקובץ העבודה המלאה מפני שהיו פחות רלוונטים.\*\*\***

ה־ Dataset שבחרתי הוא ה‑Wine Quality Data Set (כולל יינות אדומים ולבנים) מתוך אתר . Kaggle מאגר נתוניום זה כולל כ‑4,898 דגימות יין לבן ו‑1,599 דגימות יין אדום, שנסרקו באמצעות בדיקות פיזיקו‑כימיות של יינות. Portugal Vinho Verde

בכל דוגמה מתועדת כמות של תכונות כמו pH, חומציות, סוכר, גופרית חופשית וסך הגופרית, צפיפות, ריכוז אלכוהול ועוד - סך הכל 12 משתנים + סוג היין (אדום או לבן) בנוסף מיוצגת גם איכות יין מדורגת (Quality) בטווח של 0–10, אשר מקורם מטעימות מומחים.

המטרה המקורית של ה־Dataset היא לבחון איך תכונות פיזיקו‑כימיות משפיעות על איכות היין, וכיצד ניתן לחזות את איכות היין בעזרת מודלים סטטיסטיים או של למידת מכונה (מאגר זה הוא מאגר נוח ומוכר לתרגול למידת מכונה).

**שלבי ניתוח הנתונים**

**א. הכנת הנתונים לניתוח**

השלב הראשון היה טעינת הקובץ *WineQuality-WhiteAndRedWine.csv*, עליו תואר בתחילת העמוד. ביצעתי עליו מספר פעולות הכנה וטרנספורמציות:

* **המרה לפקטורים**: כמה עמודות הוסבו לפקטורים (כגון type, excellent\_good\_badכדי שיהיה ניתן להשתמש באלגוריתם הלמידה, כיאה ללמידה מונחית).

- **הכנת משתנים חדשים**:

- alcohol\_level אשר גם הוא הומר לפקטור של רמות אלכוהול (Low/Medium/High) .

- excellent\_good\_bad קטגוריזציה חדשה לציוני איכות יין (excellent/good/bad) אשר תשמש לאלגוריתם למידת המכונה איתו עבדתי.

יצירת מספר מדדים כמו next\_level\_score שנועדו לנסות ולמדוד איכויות סמויות ביין - על בסיס שילוב של מספר משתנים קיימים.

* **בדיקות איזון** : בוצעה בדיקה שחלוקת הדוגמאות בין הקטגוריות (excellent\_good\_bad) סבירה, ואין הטיות קיצוניות.
* **בדיקות התפלגות**: בדיקת התפלגות נורמלית בכמה משתנים מרכזיים (pH, sulphates, alcohol)l וכן את קיומם של חריגים.

**EDA+Visualizations:**

בשלב ה־EDA (ניתוח נתונים ראשוני), המטרה היא להבין את מבנה הדאטה, לבדוק חריגות, ולאתר תובנות ראשוניות שיכולות להנחות את תהליך בניית המודל. השלב הזה עוזר להבין את איכות הדאטה.  
הבדיקות שבוצעו בשלב זה כוללות:

**סיכום סטטיסטי של המשתנים** (summary(wine\_data)) – להצגת ממוצעים, חציון, מינימום, מקסימום וכו'.

* **בדיקת מבנה הנתונים** (str(wine\_data)) - כדי להבין אילו סוגים יש לכל עמודה (פקטור, מספרי וכו)
* **חישוב שונות לכל משתנה** var(wine\_data)) -לזיהוי משתנים עם שונות גבוהה או נמוכה מדי.
* **בדיקת ערכים חסרים** - (colSums(is.na(wine\_data))) לוודא שאין NA בדאטה.
* **בחינת עמודות נבחרות** - (select(type, alcohol\_level, quality)) לבדוק קטגוריות מרכזיות.
* **ספירת מופעים לפי איכות היין** (count(quality)) - לראות את ההתפלגות של ערך המטרה.
* **היסטוגרמות וגרפים** - לניתוח התפלגות של משתנים כמותיים (כמו alcohol, pH, density, citric.acid, quality) ) ולזיהוי חריגים.
* **תרשימי עמודות** - להשוואה בין קטגוריות כמו alcohol\_level ו־excellent\_good\_bad ולהבנת המבנה של הפילוח.

**Heatmap - -** ניתוח של קשר בין alcohol\_level לבין איכות היין.

בנוסף, בגרפים בדקתי את הקורלציה, בניסיון למצוא קורצליה גבוהה בין המשתנה או המדד שנבנה לאיכות היין.

**שימוש באלגוריתם למידת מכונה: Random Forest**

המודל שנבחר לניתוח הנתונים הוא **Random Forest** עליו פורט בהרחבה. ההגדרה בוצעה כך:

* **בחירת משתנים רלוונטיים** - הוחלט שהמשתנים שיוכנסו למודל יהיו כל משתני מאגר הנתונים למעט סוג היין (אדום או לב) וכן עמודת הפקטור שמחלקת את היין לשלוש דרגות איכות.
* **פיצול הנתונים**: בוצע פיצול ל־train/test ביחס של 80/20. המשמעות היא ש80% מהדאטה המקורי ישומש לאימון המודל והשאר לנתוני מבחן.
* **הגדרת המודל**: מודל עם ntree = 500. כמו כן בוצעה הפיכת הנתונים הנתונים לפקטורים לצורך שימוש.

**שיפור ביצועים**: במהלך העבודה ניסיתי לשנות ולהוסיף פרמטרים על מנת לשפר את אחוז הדיוק. לשם כך,

העלתי את אחוז הנתונים שיילכו לאימון המודל (train) מ0.8 ל0.9. בנוסף, נתתי משקל גבוהה יותר ליינות שסווגו כמצויינים, מכיוון שמספרם היה מועט ביחס לשאר הציונים classwt = c("bad" = 1, "good" = 1 "excellent" = 3.2). כמו כן, הורדתי את מספר העצים מ500 ל300 לשיפור יעילות המודל וזמן הריצה (ראיתי לפי plot(rf\_model) שהורדה זו בכמות העצים אינה משפיעה על השגיאה).

יש לציין שניסיתי גם לבדוק האם הורדת משתנים שנכנסים או לחילופין להוסיף מדדים חדשים למודל יכול להשפיע, והתוצאה הכי טובה הייתה כאשר כל המשתנים נמצאים במודל.

**בדיקת המודל :** לבסוף, הכנסתי יין חדש עם מדדים מומצאים כפלט, והמודל סיווג את היין לפי אחוזים (מה הסיכויים שהיין מצויין, טוב או רע) .

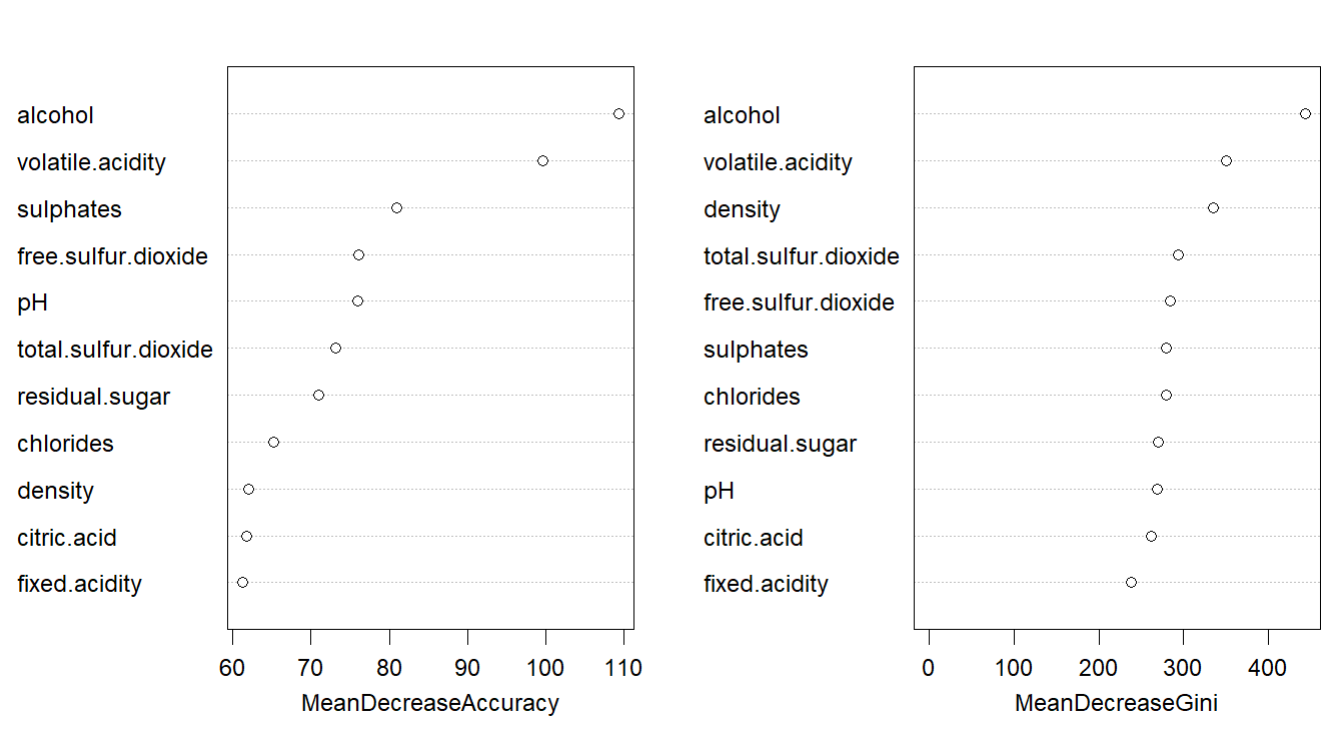
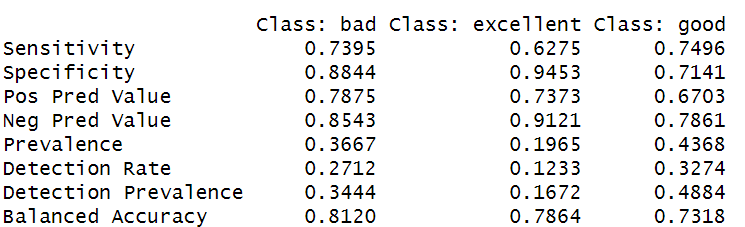
**תוצאות ותובנות מעניינות מתוך ניתוח המודל:**

**תובנות כלליות מתוך הגרפים והבדיקות:**

* **רמות אלכוהול גבוהות** תואמות לרוב ל-**איכות יין גבוהה יותר**, כפי שצוין גם בגרף הליניארי, במפת החום, ובגריד.
* אין יינות שהם בעלי איכות מקסימלית (9), והם עם רמת אלכוהול נמוכה.
* **היחס בין אלכוהול לצפיפות** (alcohol\_density\_ratio) נמצא בקורלציה חיובית לאיכות.
* **volatile acidity** (חומציות נדיפה) ככל שהיא גבוהה יותר - איכות היין נוטה לרדת.
* נמצא מדד בשם **next\_level\_score** אשר מורכב מחלק מהפרמטרים ולו קורלציה גבוהה מאוד עם איכות היין (כ-0.515).
* דווקא היינות המועטים שקיבלו את ציון האיכות המקסימלי, 9, הם לא בעלי ציון ה next\_level\_score הגבוה ביותר.

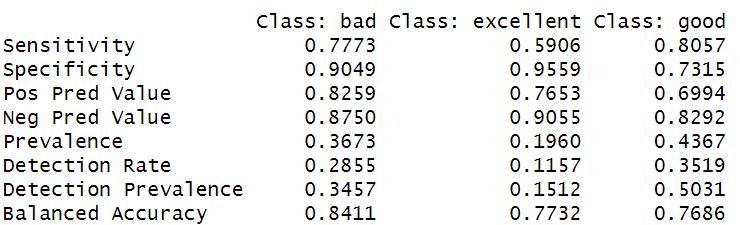
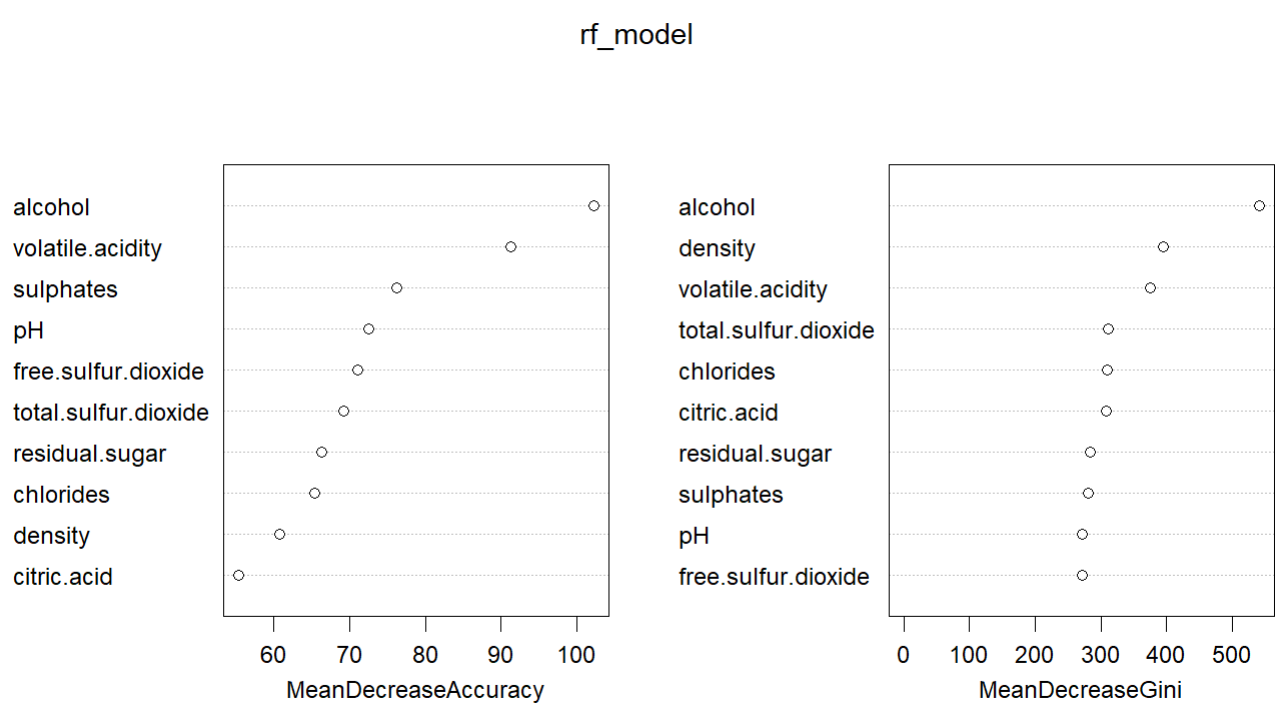
**תובנות על התפלגות משתנים:**

* לא נמצאו ערכים חסרים (NA) – הדאטה היה **נקי**.
* איכות היין התפלגה בקירוב **נורמלית**.
* במדדים כמו pH, sulphates, fixed.acidity , ההתפלגות נראית בקירוב **נורמלית**.
* מדדים כמו total.sulfur.dioxide, chlorides, citric.acid, volatile.acidity זוהו ככאלו שכנראה מתפלגים בצורה לא אחידה או כוללים ערכים קיצוניים (חריגים).
* תהליך הקטגוריזציה של רמות האלכוהול לשלוש קטגוריות (Low, Medium, High) סייע לחשיפת קשרים ברורים יותר מול איכות היין.
* מטריצת הבלבול לפני השיפור וניתוח חשיבות המשתנים:

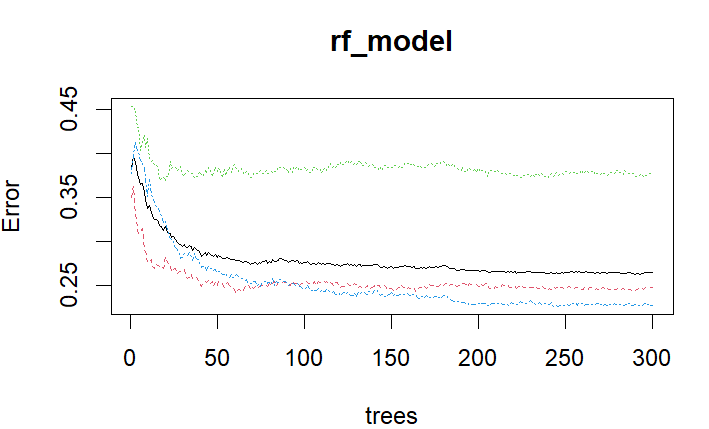


**אמינות המודל: 72.1%**

* מטריצת הבלבול אחרי השיפור וניתוח חשיבות המשתנים:

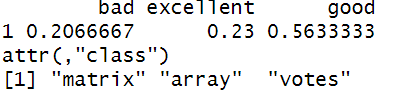
 

**אמינות המודל: 75.3%**



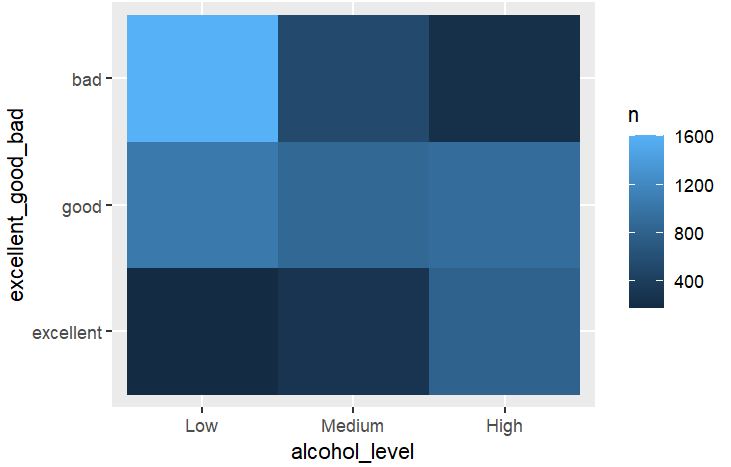
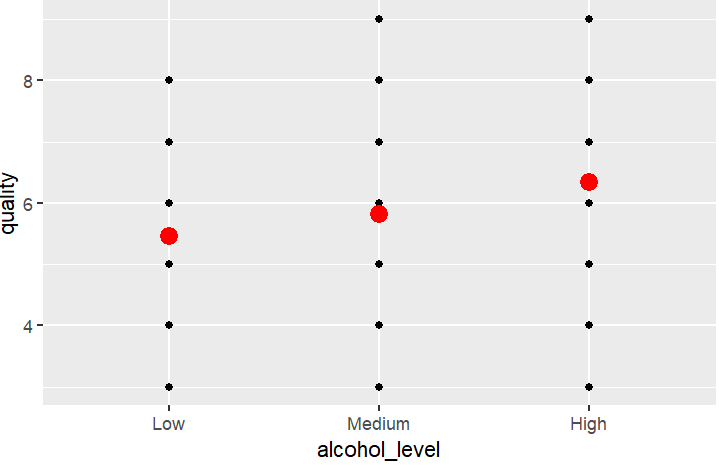
גרף המתאר את השפעתם של כמות העצים על הטעות באחוזים, מחולק לפי ציוני היין. השחור הוא איחוד. כלומר, אין סיבה להשתמש בכמות יותר גדולה של עצים מכיוון שתהיה לכך אותה השפעה אולם זמן הריצה יהיה איטי יותר.

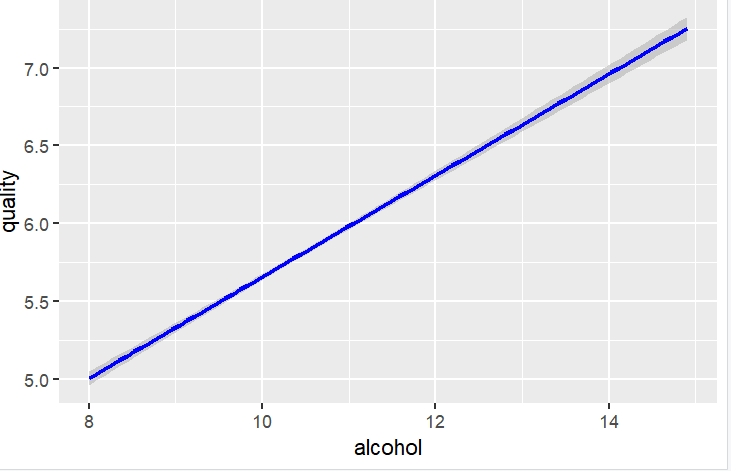
הקלט לאחר השיפור, מציג את הסיכוי באחוזים שהיין שהוכנס כפלט משתייך לכל אחת מהקטגוריות:



לאחר ביצוע שיפורים במודל, שכללו שינוי פרמטרים ומתן משקל גבוה יותר לקטגוריית "excellent", נצפה שיפור כללי במדדי הדיוק של המודל. המודל מצליח לזהות טוב יותר יינות "רע" ו"טוב", כאשר גם התחזיות עבור "excellent" מדויקות יותר, למרות ירידה מסוימת ביכולת לזהות אותם בפועל. השיפור במדד ה־ Balanced Accuracy בכל הקטגוריות מצביע על כך שהמודל הפך מאוזן ומדויק יותר באופן כללי.

דיאגרמות נבחרות שמקשרת בין ציוני היין לרמת האלכוהול, ניתן לראות שישנה קורלציה גבוהה.



**סיכום:**

במהלך העבודה נחשפתי לעבודה עם מודל Random Forest ויישומו על דאטה אמיתי בתחום היין. למדתי כיצד להכין נתונים ללמידת מכונה, לבצע ניתוחים סטטיסטיים מקדימים, ולהתאים פרמטרים כדי לשפר את ביצועי המודל. העבודה תרמה להבנה מעשית על תהליך שלם של , Data Science מהבנת הנתונים ועד הפקת תובנות.

\*\*\*לסיום, עבודה זו לוקה גם במספר נקודות חולשה:

* לא נעשה שימוש מספק במדדים החדשים שהוגדרו כגון next\_level\_score ועוד, לצורך שיפור מודל למידת המכונה ובגרפים.
* לא בוצעה התייחסות מספקת לערכים חריגים וטיפול בהם. עם זאת, יינות אלו לא התאפיינו בערכי איכות חריגים, מה שהעלה שאלה באשר לטיפול המתאים. (כלומר, הם לא היו בעלי קשר ישיר למשתנה המטרה שלי במודל).
* לא נעשה שימוש בשיטות אוטומטיות לבחירת משתנים כגון feature importance או backward selection אלא רק על בסיס קורלציה ידנית וניסוי וטעיה.