**לימוד מכונה 364-1-1811**

**תרגיל מספר 2**

**הנחיות הגשה:** **דו"ח התרגיל השני וקוד ה-R שכתבתם יוגשו לתיבת ההגשה במודל עד לתאריך ה-23.6.19 בשעה 23:55.**

משקל התרגיל: 25% מהציון הסופי של הקורס.

מטרת התרגיל: בתרגיל נשתמש בנתונים אשר סקרנו והכנו בתרגיל הראשון לצורך אימון ובחינה של מערכות לומדות והשוואה ביניהן. התרגיל יעסוק ויציג שלושה מודלים של מערכות לומדות אשר הוצגו בכיתה ובתרגול, אליהם יש להתייחס כמפורט למטה.

כל ההמלצות שניתנו בתרגיל מס' 1 לגבי השימוש בתוכנה וביצוע העבודה והדיווח, ניתנות גם כאן.

דגשים לדו"ח: אורך הדו"ח **לא יעלה על 11 עמודים** (לא כולל הקודים שנכתבו ועמודים נלווים כמו שער, תוכן עניינים ונספחים), בגודל כתב 12, פונט Arial, רווח שורה וחצי. הדו"ח יוגש כקובץ word. חריגה ממספר עמודים זה תגרור הורדת נק'. יש לשמור על תמציתיות ולהתמקד בתובנות המרכזיות שלכם בכל סעיף.

שאלותבנוגע לתרגיל, יש לפרסם **בפורום הייעודי במודל.**

**Neural Networks (package “nnet”) (26 נק')**

שימו לב – בחבילת nnet, אם אתם נתקלים בבעיה עם הרצה על משתנים קטגוריאליים, עליכם להמיר את הפיצ'רים הקטגוריאליים למשתני דמה (מומלץ לפי קידוד 1/-1)

1. הריצו (אמנו ובחנו) את הרשת בערכי ברירת המחדל של R.

דונו בתוצאות רשת זו:

* + הסבירו את משמעות הקונפיגורציה שנלמדה ע"י המודל (מספר נוירונים בשכבת הכניסה, מספר שכבות חבויות ומספר נוירונים חבויים בכל שכבה).

הרצנו רשת נוירונים בעזרת הקוד:

nn1 <- nnet(factor(y)~.,data = train.data, size = 1)

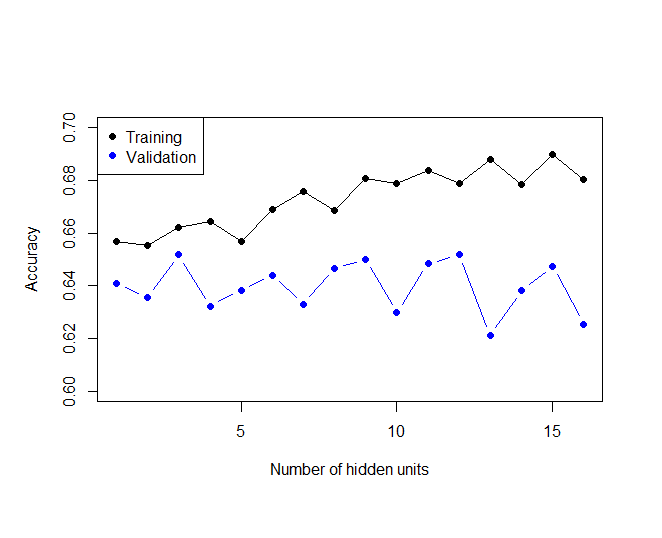
היינו חייבים לספק את ההיפר-פרמטר size שכן לא היה לו ברירת מחדל. לכן מדובר ברשת עם יחידה אחת בשכבה החבויה. על פי הגדרה, הפונציה nnet מחבילת nnet מייצרת רשת עם שכבה חבויה אחת.

לאחר הרצת הלימוד של הרשת ניתן לראות את האיטרציות של הלמידה, הערך וכמות המשקולות הסופית. במקרה של הרשת שלנו ישנם 22 משקולות לרשת.

* + הציגו את אחוזי הדיוק בשלב האימון והבחינה.

דיוק בשלב האימון הינו 64.2% ובבחינה 63.3%

1. בצעו תהליך של "כיוונון פרמטרים" עבור מספר הנוירונים, תוך שימוש בסט אימון, אימות.
   * הציגו גרף של "מספר נוירונים "כפונקציה של "אחוז הדיוק".
   * דונו בהבדלים באחוזי הדיוק (אם קיימים). האם יש בתוצאות הגיון?



לצורך הבדיקה הרצנו אימון ממספר היחידות הנמוך ביותר בשכבה החבויה (1) עד מס' המשתנים (16). ראשית, ניתן לראות כי יש לא מעט רעש, דבר שהינו הגיוני לאור העובדה שכל רשת אומנה על מספר איטרציות בברירת המחדל (100) בלי שום גורם רגולרליזציה (נרמול) כגון decay. עם זאת, עדיין ניתן לראות מגמות הגיוניות בתוצאה – באימון ישנה מגמת עלייה תמידית, שכן ככל שמוסיפים יחידות בשכבה החבויה לומדים את האימון טוב יותר. מצד שני ניתן לראות מגמת עלייה מאוד מהירה שמגיעה לשיא ב3 יחידות ולאחריה מגמת ירידה איטית שנובעת מאוברפיטינג. מעבר לרעש, לדעתנו מדובר בדוגמה קלאסית של הצגת אוברפיטינג ככל שהמודל נהיה "מסובך" יותר.

1. לאחר מציאת מספר הנוירונים המיטבי, הריצו 2 רשתות נוספות בעלות קונפיגורציות שונות הנבדלות זו מזו בתנאי עצירה של הרשת ובאלמנטים אחרים.
2. היעזרו ב-help של התכנה (או גוגל) כדי לבחור קונפיגורציה ולהבין את משמעותה.
   * עבור כל רשת: מה הייתה המוטיבציה בבחירת קונפיגורציה זו? מה היתרונות והחסרונות של הרשת הנ"ל?

על מנת לבחור את הרשתות התנסינו בשינוי מספר האיטרציות ובהיפר פרמטר "decay" למען רגולרליזציה.

הרשת הראשונה אומנה על ידי

nn1 <- nnet(factor(y)~.,data = train.data, size = 3, maxit = 300, decay = 0.01)

והשנייה על ידי

nn2 <- nnet(factor(y)~.,data = train.data, size = 3, maxit = 400, decay = 0.001)

המוטיבציה הייתה לבחון אופציה עם פחות איטרציות ושינויים מעט יותר חדים (אם כי הרבה פחות חדים מברירת המחדל) לעומת רשת המתאמנת עם יותר איטרציות והשינויים למשקלים יותר איטיים.

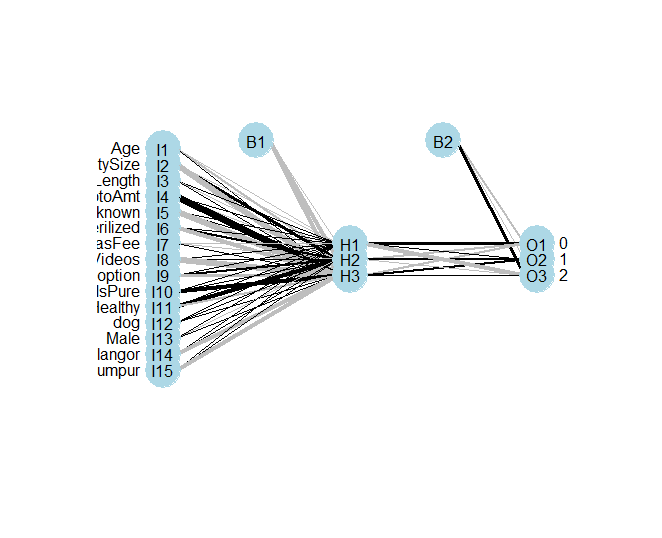
היתרון של הגדלת מס' האיטרציות הינה להתכנס באופן יותר שקול – להציג חלק מתצפיות יותר פעמים על מנת להבין את הדאטה יותר טוב. לעומת זאת הדבר עלול להוביל ל"שינון" של האימון, ולכן היה לנו חשוב להוסיף את הdecay אשר מעודד שינויים קטנים יותר כך שבתקווה נלמד את הדפוס ולא את נתוני האימון. מכיוון שהרשת הראשונה על פחות איטרציות, כלומר תראה פחות את הנתונים גם קנסנו אותה עם פחות decay, לעומת בשנייה שעלולה יותר ליפול לאוברפיטיניג. על מנת להגיע למידה האופטימאלית של decay החזקנו את מס' האיטרציות בכל אחת כקבוע ומצאנו את הדיוק הגבוה ביותר על הוולידציה.

* + תארו את הקונפיגורציה שנלמדה ע"י המודל והסבירו משמעותה (מספר נוירונים בשכבת הכניסה, מספר שכבות חבויות ומספר נוירונים חבויים בכל שכבה).

בשתי המודלים ישנם מס' זהה של נוירונים, שכן ההיפר פרמטר size הינו זהה ובנוסף בשתיהן שכבה חבויה אחת – שכך החבילה nnet עובדת.

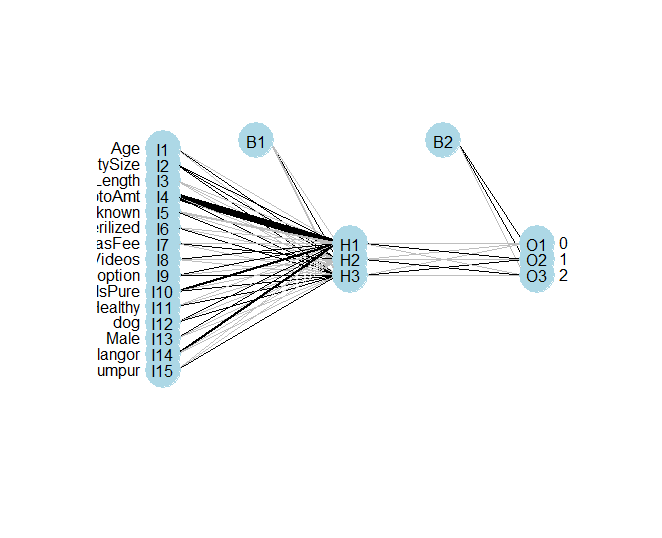
עם זאת, מעניין לראות את הקשרים השונים והמשקולות שנוצרו.

הרשת הראשונה:



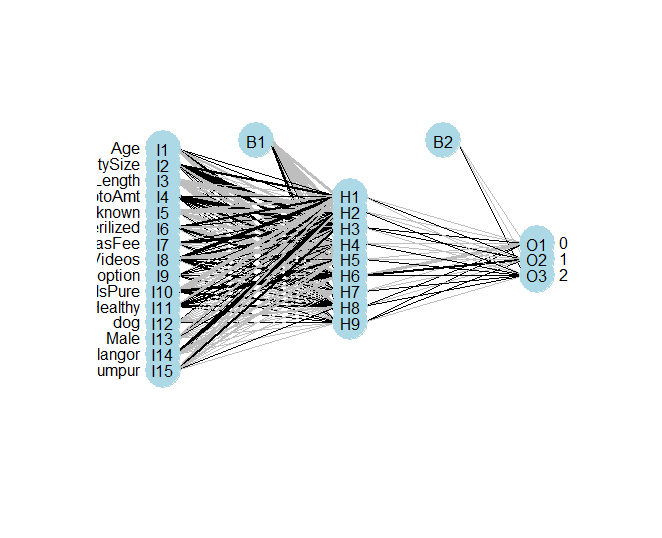
כאשר הקווים מייצגים את המשקולות – ככל שהמשקולת משמעותית יותר באופן יחסי כך היא עבה יותר, אפורה זאת משקולת שלילית ושחורה הינה חיובית.

הרשת השנייה:



ניתן לראות לדוגמה ברשת השנייה כי המשקולת של כמות התמונות בפרופיל בעלת משקל מאוד גדול וחיובי לנוירון החבוי H1 (שם שהחבילה שעושה את הגרף נותנת אוטומטית).

מתוך סקרנות, הרצנו עד אופטימיזציה גם רשת עם 9 נוירונים חבויים, שכן זו גם הייתה נקודת שיא מסוימת בתהליך הוולידציה הראשוני. הרשת השלישית נראית כך:



* + הציגו את אחוזי הדיוק בשלב האימון והבחינה.

עבור הרשת הראשונה: 66.6% על האימון ו66.6% על סט הבחינה

עבור הרשת השנייה: 66.3% על האימון ו67.6% על סט הבחינה

עבור הרשת השלישית: 68.9% על האימון ו64.7% על סט הבחינה

* + השוו בין תוצאות המודלים. איזה מודל היה הטוב ביותר ואיך הגעתם למסקנה הזו? מהם הקריטריונים בהם השתמשתם לבחירת המודל? אם הייתם צריכים לבחור פרמטר נוסף לכיוונון, במה הייתם בוחרים?

ראשית, ניתן לראות כי כפי שחשדנו בהתחלה, העלאת מספר הנוירונים החבויים ל9 הובילה לאוברפיטינג. הרשת השנייה, על אף שהציגה ביצועים מעט נמוכים יותר על סט האימון, הצליחה יותר בשלב הבחינה. מכיוון שברמת הסיבוכיות שני הרשתות הראשונות זהות, נבר את הרשת השנייה שצלחה את הבחינה.

אם היינו צריכים לבחור פרמטר נוסף, היינו רוצים לכוונן את "משקולות מקרה" (case weights) שכן יש לנו מחלקה אחת שאינה מיוצגת מספיק (0 – יאומץ באותו יום) וייתכן כי זהו מקרה שחשוב מאוד לזהות, כך שהיינו רוצים לתת לו משקל גבוה יותר ולראות אם ניתן לשפר את יכולת הניבוי של המודל למקרים אלו.

**Decision Trees (package “rpart”) (23 נק')**

1. בנו את עץ ההחלטה באמצעות סט האימון והאימות יחדיו.
   * מה הסיבה לכך שאנו משתמשים בסט האימות לטובת האימון? (רמז: הדבר קשור לאופן שבו הפקודה rpart עובדת).

למיטב הבנתנו, הסיבה לכך נעוצה בכך שRpart מייצר בעצמו אומדים יעילים לcross validation. ניתן לראותם בprintcp בעמודות של השגיאות (xerror). לכן, כאשר מריצים את אימון המודל על יותר נתונים ניתן להגיע למסקנות יותר מדויקות. לאחר ההרצה הראשונית והכתימה של העץ, ניתן לערוך תהליך של ולידציה על מנת לבחון את בעצמך אם יוצאות תוצאות דומות, אך האלגוריתם עושה זאת עבורך.

1. הריצו את הפקודה printcp().
   * הסבירו אילו מאפיינים מוצגים בפלט?

השורה הראשונה במודל מראה את שורת הקוד שאימנה את המודל, לאחר מכן ניתן לראות את המשתנים אשר מרכיבים את העץ הסופי לפי שם, הטעות של המודל בשורש (כלומר בעזרת הפיצול הראשון בלבד). לאחר מכן ניתן לראות את גודל המדגם ולבסוף את מדד הסיבוכיות (CP) עבור כל עץ לאחר קטימה (prune) של חלק מהענפים. בנוסף ניתן לראות את השגיאות לכל רמת קטימה

* + מהם ערכי הפיצול האפשריים שמופיעים בטבלה? האם ישנם ערכים חסרים בטבלה (כלומר, האם הערכים בטבלה עולים באופן סדיר: 0,1,2,3,4,5 וכו. או שישנן קפיצות, לדוגמא: 0,1,4,5,9)?

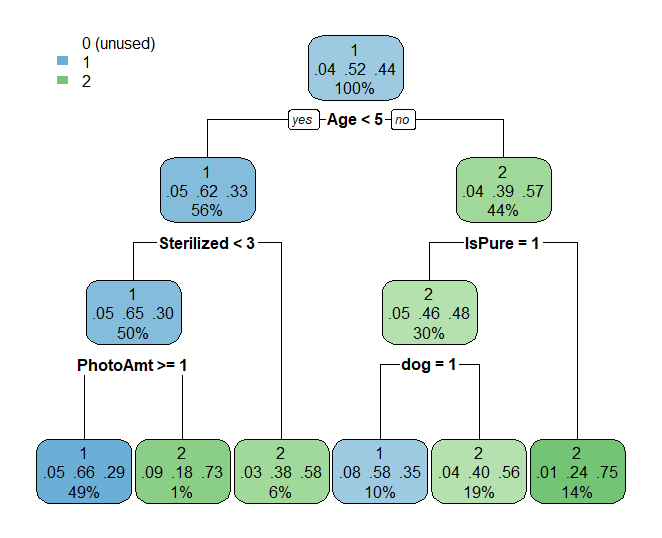
אם ישנן קפיצות – מה הסיבה לכך?.

במקרה של המודל שלנו ישנה קפיצה אחת (על מס' 2). עושה רושם שהסיבה לכך היא ששכבה 2 היא מש

* + מצאו את פרמטר הסיבוכיות (cp) המיטבי. הסבירו כיצד הוא נבחר וקטמו באמצעותו את העץ. מהם אחוזי הדיוק המתקבלים בעזרת העץ הן על סט הבחינה והן על סט האימון? מה ניתן להסיק מאחוזי הדיוק המתקבלים על שני הסטים? אם קיים הבדל משמעותי – מה הסיבה לכך? האם יש לנקוט בפעולות נוספות?

ערך הCP המיטבי (הקטן ביותר) הינו 0.01 וכולל את העץ המלא שנוצר בעזרת Rpart. הפרמטר נבחר באמצעות התרומה היחסית של הפיצול תוך קבלת קנס על העלאת סיבוכיות (כמות פיצולים). על כן, לא קטמנו את העץ. אחוזי הדיוק על סט האימון הינו 64% ועל סט הבחינה 63%. הסיבה להבדל היא כמובן אוברפיטינג, אך אחוז אחד (בקירוב) אינו אוברפיטינג משמעותי, בעיקר לאור העובדה שהדיוק נבדק פעם אחת אל מול סט בלתי תלוי ולא במספר בדיקות. לכן, בעיננו אין צורך בפעולות נוספות. כמובן, ניתן לחזור אחורה לשלב הכנת הדאטה ולנסות לשפר באופן כללי את יכולת הניבוי של המודל, כמו גם לייצר דגימת יתר של המחלקה בעלת הסיכויים האפריוריים הנמוכים ביחס לשאר (0), או רעיונות אחרים שעשויים לשפר את יכולת המודל באופן כללי, ללא קשר לאוברפיטינג.

* + הציגו גרף של העץ שהתקבל (אם גדול מידי – קטמו עוד עד שיהיה ניתן להצגה ברורה).



חשוב לציין כי חלק מהמשתנים פה (לדוגמה גיל) עברו דיסקרטיזציה, לכן משמעותם שונה מהמשמעות המקורית.

* + בחרו רשומה לדוגמא מסט הבחינה, העבירו אותה באופן ידני (על פי המופיע בגרף העץ) דרך עץ ההחלטה הקטום ודווחו מה הסיווג שהתקבל. האם הסיווג תואם את המציאות?

לצורך הבדיקה בחרנו את הרשומה הראשונה בבסיס הנתונים לפני הפיצול לאימון ובחינה – והסיווג תאם את המציאות. הגיל היה 4, מסורס היה 2 והיו תמונות, ואכן הy היה 1.

* + ע"י התבוננות במבנה העץ, אילו תובנות הוא מספק על הבעיה ועל החשיבות השונה של מאפייני הבעיה?

ישנן כמה תובנות "מובנות מאליהן", אך גם מפתיעות. לדוגמה, מאוד הגיוני עבורנו כי הגיל משחק תפקיד קריטי, ולכן מהווה את השורש – שכן אנשים מעדיפים לאמץ חיות צעירות מאשר מבוגרות (כך זה היה גם מהניסיון שלנו בתור מאמצים). בנוסף, הגיוני לנו כי ישנה חשיבות למס' התמונות בפרופיל החיה. אם אין תמונות בכלל הדבר עשוי להיות מאוד מחשיד ולהרתיע מאימוץ.

בניגוד לכך התוצאה של סירוס מעניינת – עושה רושם כי לא מפריע למאמצים שהחיה אינה מאומצת (2), אך מרתיע אם לא מצוין כלל אם מסורסת או לא (3). כלומר חוסר הידיעה הוא שמוביל לחוסר אימוץ, ולא החוסר בסירוס. בנוסף מעניין שכלבים גזעיים מאומצים למרות שהם מבוגרים, כנראה כי ישנו פוטנציאל הרבעה (למרות שזהו ניחוש). בניסוח שונה, באופן מעניין גזעיות "מצילה" רק כלבים, ולא חתולים.

* + השתמשו בפונקציית variable.importance, הסבירו את משמעות בפלט של פונקצייה זו. האם התוצאות מתיישבות עם המסקנות מהסעיף הקודם?

פלט:

caret::varImp(FirstTree)

Overall

Age 213.952707

dog 31.950636

FurLength 53.853992

Gender 7.157753

IsPure 117.979758

MaturitySize 8.904560

PhotoAmt 115.923881

Quantity 58.611549

SingleAdoption 58.611549

Sterilized 150.266982

Vaccinated 19.205042

כפי שציינו בסעיף הקודם, ניתן לראות כי גיל הינו המשתנה החשוב ביותר, דבר שיושב עם ההיגיון שלנו. לאחר מכן, מסורס/מעוקר, גזעיות וכמות תמונות. כל השאר משמעותית נמוכים יותר. הדבר היחיד שהפתיע אותנו הוא המדד הגבוה של גזעיות, אשר קיבלה מקום בעץ רק כאשר כבר יש נטייה לתת ערך 2 ורק עבור כלבים (אשר קיבל ערך נמוך בחשיבות). הערכים הגבוהים של כמות תמונות וסירוס\עיקור הגיוניים גם לפי הסעיף הקודם.

1. מתוך הטבלה שהתקבלה, קטמו את עץ על פי 3 ערכי cp נוספים:
   * 1. ערך ה-cp שהוביל לשגיאה הגבוהה ביותר על סט הולידציה.
     2. ערך cp=0.
     3. ערך cp נוסף לבחירתכם (הסבירו את הבחירה).
   * עבור כל ערך cp, דווחו מהם אחוזי הדיוק הן על סט האימון והן על סט הבחינה. הסבירו את השוני ביניהם עבור כל ערך cp בנפרד.

ראשית, ערך הCP הנוסף שבחרתי הינו 0.023477. הסיבה לכך היא שעד לערך זה כל פיצול תרם משמעותית להורדת השגיאה (בכ4 נקודות כל פעם) ולאחר ערך זה גודל הירידה יורד ל2 עבור כל פיצול. כלומר, בחרתי את העץ שהכי שיצר את השיפור הגדול ביותר עד לפיצולים נוספים שעזרו לעשות "Fine Tuning" – אך עשוי להיות שייצרו אוברפיטינג על הבחינה.

עבור המודל עם השגיאה הגדולה ביותר (רק שורש) מקבלים דיוק של 60% על סט הבחינה ודיוק של 59% על סט הבחינה.

עבור ערך 0 מקבלים את המודל המקורי אשר נותן 64% על סט האימון ועל סט הבחינה 63%.

עבור המודל הנוסף, מקבלים 63.5% על סט הבחינה ו62.5% על סט הבחינה.

ראשית, ניתן לראות כי אין הבדל גדול מאוד בדיוק בין המודלים, דבר המלמד כי המשתנה גיל הינו המשמעותי ביותר בהפרדה באופן מאוד קיצוני. שאר המשתנים עוזרים, אך באופן מינורי ביחד למשתנה גיל. הגיוני (ומשמח) לראות כי ישנה הלימה בין הCP לבין הולידציה על סט הבחינה. כלומר, המדד שRpart נתן לבחירת המודל נמצא בהלימה עם יכולת הניבוי גם על סט הבחינה הבלתי תלוי.

1. השוו בקצרה בין תוצאות ארבעת המודלים. איזה מודל היה הטוב ביותר ואיך הגעתם למסקנה הזו? מהם הקריטריונים לבחירת המודל? כיצד קריטריונים אלה מתקשרים לבעיה העסקית אותה אתם מנתחים?

הדבר תלוי במידה רבה במטרות הארגון – האם המטרה היא להשיג דיוק כללי גבוה ככל הניתן על סט הבחינה, או שקריטי לשמור על סיבוכיות נמוכה? במקרה הנוכחי, נניח כי המטרה היא דיוק גבוה ככל הניתן, דבר שמשאיר את העץ המקורי, בעל 63% דיוק על סט הבחינה אשר הושג על ידי קטימה בCP 0 הינו המודל הטוב ביותר. הקריטריון מתקשר לבעיה העסקית שכן לדעתנו אחת ההשלכות המשמעותיות של יישום המודל תהיה שיווק ונתינת "עדיפות" לפרופילים ספציפיים שהמודל ינבא שיאומצו בקרוב. כלומר, יש ערך בדיוק טוב בכלל הקטגוריות (בניגוד אולי לניבוי חולי שמגיע עם סיכון מוגבר לניבוי שגוי של קטגוריה אחת לעומת אחרת).

**K-Means (package “stats” – default package) (17 נק')**

1. הוציאו את משתנה המטרה מתוך ה- inputs. מהי המטרה של ביצוע פעולה זו?
2. הריצו את המודל עם ערכי ברירת המחדל ("המאשכל הבסיסי"), פרט להגדרת מספר האשכולות (specified number of clusters), אותו תגדירו כך שיתאים לבעיה עמה אתם מתמודדים ועבורה ידוע לכם מספר המחלקות.
3. דונו בתוצאות. מה טיב ההתאמה בין האשכולות למחלקות? כיצד שויכו תצפיות לאשכולות? לוו את ההסברים בגרפים מתאימים.
4. במנותק ממשתנה המטרה של סיפור המקרה, בו יש מספר ידוע של קבוצות, אמנו כעת מחדש שמונה מודלים של אשכול עם ערכי K משתנים (כלומר, עם מספר שונה של אשכולות בכל פעם). מהו ה-K המיטבי? הסבירו את אופן בחירתכם.

האם הערך מתקשר לסיפור המקרה? אם לא, מה יכולה להיות הסיבה לכך?

**השוואה בין מודלים (9 נק')**

1. הציעו דרך שתאפשר השוואת ביצועי המודל הלא-מונחה (K-means) לאלה של המודלים המונחים (NN ו-DT). בעזרת שיטה זו, השוו את ביצועי שלושת המודלים (k-means, המודל הנבחר של NN והמודל הנבחר של DT) והצדיקו פערים ביניהם.

הסבירו **בפירוט** את השיטה בה בחרתם להשוות את המודל הלא-מונחה למודלים המונחים.

מהן מסקנותיכם? על איזה מהמודלים תמליצו? מדוע?

1. ערכו השוואה בין המודלים מבחינת ביצועים, מורכבות, יתרונות וחסרונות.

**המודל הנבחר (9 נק')**

1. הציגו מהו המודל שנבחר לטובת ההגשה לתחרות ופרטו מהי הקונפיגורציה שנבחרה עבור המודל (cp/מס' נוירונים/ מספר איטרציות).
   * הסבירו מדוע בחרתם את מאפייני הקונפיגורציה.
2. בצעו ניתוח של תוצאות במודל באמצעות מטריצת מבוכה (confusion matrix).
   * האם התוצאה תקינה או שישנה הטייה?
   * האם התוצאה מתיישבת עם כך שמשתנה המטרה אינו מאוזן?

**הגשת חיזויים סופיים (9 נק' + 2 נק' בונוס)**

1. בצעו חיזוי באמצעות המודל שבחרתם, על קובץ הנתונים "Test.csv" הנמצא במודל. את החיזויים שיתקבלו העלו למודל כקובץ אקסל, על פי הפורמט שמופיע בקובץ הדוגמא "TestY-Example" הנמצא במודל (הערכים בקובץ זה אקראיים). בדקו שהקובץ שאתם מגישים הוא בעל מבנה זהה ובעל אותו מספר רשומות. (חלוקת הנק': 4 נק' – על הגשת הקובץ באופן תקני. 5 נק' - בהתאם למיקומכם ביחס לשאר הצוותים בקורס. 2 נק' בונוס - לקבוצה אשר תשיג את הביצועים הגבוהים ביותר).

**איכות הדו"ח ורמת שימוש בתכנת R (7 נק')** – נק' אלו יינתנו בהתאם להערכה כללית בנוגע לאיכות הדו"ח ובהתאם לרמת השימוש בתכנת ה-R (האם נעשה שימוש מקיף/חלקי/כלל לא).

**בונוס כללי (5 נק')**

בונוס של עד 5 נקודות לחלק ב' יינתן לקבוצות שיעשו שימוש בבצורה מקיפה ואיכותית בפונקציות/הרחבות של המודלים הקיימים שנדרשו בעבודה ו/או במודלים שלא נדרשים בעבודה.

בהצלחה !