מדעי הנתונים ובינה עסקית, סמסטר ב' תשע"ח

תרגיל בית 2 – Decision Trees in R

קראו בעיון את כל ההוראות לפני ביצוע העבודה

<u>הוראות כלליות</u>:

א. אי עמידה בכל אחת מההוראות יגרור הורדת ציון או פסילת העבודה.

ב. הגשת העבודה ב**<u>זוגות</u>** בלבד .

ג. שפת תכנות – R, סביבת פיתוח – מומלץ להשתמש ב-R studio בגרסתו העדכנית ביותר (גרסת 3.4.3 R וגרסת 1.1.383 R studio ומעלה).

ד. יש להגיש את העבודה **לתיקיית ההגשה הרלוונטית באתר הקורס (Moodle).**
אחריותכם האישית לבדוק לפני הגשה כי כל הקבצים נפתחים כראוי.
<u>רק אחד</u> מבני הזוג יגיש את המטלה!

ה. יש להגיש קובץ zip - שם הקובץ יהיה מורכב משני מספרי תעודות הזהות של המגישים באופן הבא: ID1_ID2.zip
הקובץ יכיל את הקבצים הבאים:

- קובץ הסקריפט המלא, **<u>ללא</u>** קבצי הנתונים (קובץ ID1_ID2.R).
- קובץ PDF המכיל את דוח המטלה, הניתוח והפלטים הנדרשים. יש לציין בפינה השמאלית בכל עמוד את ת"ז ושמות הסטודנטים.

ו. בנוסף, זוהי עבודה תכנותית ולפיכך יהיה משקל לכך בבדיקה. כלומר: יש לדאוג לקוד מסודר, הערות בקוד, לשמות משתנים בעלי משמעות וכדומה. יש לחלק את הקוד לפונקציות (במידת האפשר ולפי הצורך).

ז. תאריך ההגשה: 23:55 29.04.2018

הוראות התרגיל:

בתרגיל זה תבנו עץ החלטה כמודל חיזוי באמצעות הספרייה rpart בשפת R.
כשלב מקדים לבניית מודל החיזוי תבצעו ניקוי נתונים והכנתם לבניית המודל. לאחר מכן תבנו את עץ ההחלטה ותבדקו כיצד משפיע שינוי פרמטרים שונים במודל על ביצועיו. לבסוף, תציגו את עץ ההחלטה המתקבל בצורה ויזואלית.

תיאור הקבצים שלרשותכם:

1. **Description** – קובץ המתאר את המשתנים השונים בקובץ הנתונים, מטרתם וסוגם (נומרי או קטגוריאלי). יש לשים לב מהו משתנה המטרה (class).
2. **קובץ הנתונים (בשם German_Credit) -** קובץ המכיל את הרשומות במבנה הבא: כל שורה מציגה רשימה של ערכי משתנה קלט מסוים עבור כל הרשומות בקובץ.
   - לדוגמא:
     - עבור קובץ עם 3 רשומות ושלושה משתנה קלט: `ClaimID, RearEnd, Fraud`:

ClaimID	RearEnd	Fraud
1	TRUE	TRUE
2	FALSE	FALSE
3	TRUE	TRUE

הקובץ מוצג כך:

ClaimID: 1,2,3

RearEnd: TRUE, FALSE, TRUE

Fraud: TRUE, FALSE, TRUE.

תיאור המשימות שעליכם לממש במסגרת התרגיל:

1 **הכנת הנתונים:**

1.1 מודל החיזוי ילמד על בסיס סט אימון (training set) וייבדק על בסיס סט הבדיקה (test set). מכיוון שנתון רק קובץ אחד, יש לחלק את הקובץ הנתון לשני קבצים ע"י חלוקה רנדומלית של 20% מהקובץ עבור סט הבדיקה והיתר עבור סט האימון.

1.2 יש לטעון את הנתונים בצורה הנכונה למבנה נתונים מסוג data.frame.

- שימו לב לטיפוס של כל עמודה ב-data.frame.
- שימו לב לערכים החסרים שיש בקובץ הנתונים. הקפידו שערכים אלו יהיו **NA** ולא string ריק, כדי שתוכלו להשלימם בשלב הבא.

1.3 יש להשלים ערכים חסרים בקובץ:

1.3.1 עבור **ערכים נומריים**: ערך הממוצע של כל ערכי המשתנה על פני כל הרשומות.

1.3.2 עבור **ערכים קטגוריאליים**: הערך השכיח ביותר (Mode).

1.3.3 ניתן להניח כי **אין** ערכים חסרים בתכונת ה-class.

1.4 יש לבצע דיסקרטיזציה למשתנים הנומריים הבאים:

- Average_Credit_Balance
- Over_draft
- Cc_age

את מספר ה-bins יש לקבוע בצורה הגיונית לפי שיקולכם. לשם כך, בדקו מה משמעות המשתנה וחלקו את הטווח ל-2 עד 5 אינטרוולים. ניתן לחלק על פי Equal-frequency discretization או על פי Equal-width discretization.

📄 ציינו בדוח איזה סוג דיסקרטיזציה ביצעתם ולכמה bins.

2 בניית מודל ההחלטה:

2.1 השתמשו בספרייה `rpart` של R כדי לבנות עץ החלטה מתאים לנתונים (יש להתקין את הספרייה, במידה ואינה מותקנת, בעזרת הפקודה install.packages ולאחר מכן לטעון אותה באמצעות הפקודה library). השתמשו בפקודה `rpart(…)` בשביל לבנות את עץ ההחלטה.

- הגדירו את משתנה המטרה ואת משתני הקלט בעץ ההחלטה.
- הגדירו את מדד הפיצול בעץ. כברירת מחדל, הספרייה משתמשת במדד gini split, כדי לבחור את משתנה הפיצול הבא בקודקוד. נסו להגדיר פעם את **קריטריון הפיצול** gini ופעם את קרטריון ה- information gain. השוו את ביצועי המודל בין שני הקריטריונים.
- הגדירו את פרמטר ה- minsplit המציין מה **מספר הרשומות המינמלי** בקודקוד על מנת שניתן יהיה ניתן לפצל אותו. נסו להגדיר שני ערכים שונים לערך זה (הערך גדול שווה ל-2) ולהשוות את ביצועי המודל בין שני הערכים.
- נסו לשלוט **בסיבוכיות העץ** (שילוב של גודל העץ וטיב הסיווג של משתנה המטרה) על מנת למנוע גידול של עצים עמוקים/מסובכים מדי (שעלולים לגרום לתופעת overfitting).

2.2. הציגו את עץ ההחלטה הנלמד בצורה ויזואלית באמצעות הספרייה `RColorBrewer` והספרייה `rattle` (יש להתקינה ולאחר מכן לטעון לפרויקט). טענו את הפונקצייה

`library(rpart.plot)`. השתמשו בפקודה `fancyRpartPlot(…)` על מנת להציג את העץ עבור כל אחת מהתצורות הבאות:

- קריטריון הפיצול gini או information gain.
- פרמטר minsplit.

🗎 שימרו את תצלומי העצים בדוח.

3. הערכת ביצועי המודל:

בדקו את ביצועי המודל על סט הבדיקה שיצרתם מקובץ הנתונים. השתמשו בפקודה `predict(…)` ובדקו מהו הדיוק (accuracy) של המודל עבור:

- קריטריון הפיצול gini או information gain.
- פרמטר minsplit.

- שימו לב שהפקודה predict מציגה את ה-class החזוי לכל רשומה בסט הבדיקה, יחד עם ההסתברות לקבל כל אחד מערכי ה-class. כדי לחשב את דיוק המודל, יש לבצע חישוב פשוט על פלט הפקודה `predict`.

🗎 הציגו את התוצאות בטבלה פשוטה וברורה בדוח, והסבירו את ההבדלים הקיימים בתוצאות (במידה וקיימים).

הערות מיוחדות:

1. יש לשמור את הסקריפט כקובץ R (סיומת R.). יש לציין בקוד הערות קצרות המציינות את חלקי הקוד והפעולות השונות.
2. את הניתוח של הפלט והשוואת ערכי הפרמטרים השונים יש לציין בדוח בצורה ברורה ולצרף את צילומי העצים שהצגתם.
3. ציינו בדוח אם ביצעתם פעולות נוספות, מעבר למצוין בהנחיות.
4. **אין לשתף קטעי קוד ואין להעתיק פתרונות!**
5. שאלות בנוגע לתרגיל יש לשאול **אך ורק** בפורום השאלות הרלוונטי המופיע ב-moodle (ולא במייל - שאלות במייל לא יענו).

בהצלחה!