תרגיל בית 2 - שפת R

הוראות הגשה

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

הגשת העבודה ב**זוגות** בלבד. רק אחד מבני הזוג יגיש את המטלה!

ב. שפת תכנות – R, סביבת פיתוח Rstudio , הבדיקה תבוצע בגרסה העדכנית ביותר.

ג. יש להגיש את העבודה **לתיקיית ההגשה הרלוונטית באתר הקורס (Moodle).**

אחריותכם האישית לבדוק לפני הגשה כי כל הקבצים נפתחים כראוי.

ד. 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, **ללא** 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.

ה. יש לשמור את הסקריפט כקובץ R (סיומת R.). יש לציין בקוד הערות קצרות המציינות את חלקי הקוד והפעולות השונות.

ו. את הניתוח של הפלט והשוואת ערכי הפרמטרים השונים יש לציין בדוח בצורה ברורה ולצרף את צילומי העצים שהצגתם.

ז. ציינו בדוח אם ביצעתם פעולות נוספות, מעבר למצוין בהנחיות.

ח. אין לשתף קטעי קוד ואין להעתיק פתרונות! במהלך בדיקת העבודות ייעשה שימוש בתוכנה הבודקת העתקות.

ט. 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).

י. שאלות בנוגע לתרגיל יש לשאול **אך ורק** בפורום השאלות הרלוונטי המופיע ב-moodle (ולא במייל - שאלות במייל לא יענו).

I. תאור הבעיה

עקב משבר הקורונה, משרד הכלכלה של ממשלת ארה"ב נותן הלוואות לעצמאים ובעלי עסקים קטנים שנפגעו כלכלית.

תהליך הבקשה לקבלת הלוואות ממשרד הכלכלה היה מאז ומתמיד ידני ולאחר מכן נערכת בדיקה האם בעל העסק זכאי להלוואה על סמך הנתונים שמילא.

על מנת לייעל את התהליך עקב מספר גדול של פונים, מעוניין משרד הכלכלה לפתח אתר להגשת הבקשות אונליין אשר ייתן תשובה תשובה בזמן קצר לבעל העסק לאחר שמילא את פרטיו האישיים ואת פרטי העסק באתר.

על מנת לפתח מודל למידת מכונה שיחליט אילו בעלי עסק זכאים להלוואה, ישתמש משרד הכלכלה בחלק מה-dataset של בקשות שהוגשו בעבר ובהחלטות שהתקבלו לגביהן.

מבנה הנתונים של ה- dataset מוצג בטבלה שלהלן ונתוניו מצורפים בקובץ ששמו Loan_dataset.csv :

Variable	Description
Request_Number	Unique request ID
Employees	Number of employees
Monthly_Profit	Business average monthly profit from last year (US$)
Credit_History	Credit history positive (Boolean)
Customers	Business customers size (Large / Medium / Small)
Export_Abroad	Export out of USA (Y/N)
Loan_Amount	Loan amount in thousands (US$)
Payment_Terms	Number of months for paying back the loan
Gender	Applicant gender (Male / Female)
Education	Applicant education (Graduate / Under Graduate)
Married	Applicant married (Y/N)
Spouse_Income	Spouse monthly income (US$)
Request_Approved	Request decision (Y/N)

שימו לב מהם מטרתם וסוגם (נומרי או קטגוריאלי) של הנתונים. כמו כן, יש לשים לב מהו משתנה המטרה (class).

בתרגיל זה עליכם לבנות מודל מסוג עץ החלטה בשפת R.

תחילה, עליכם לבצע data cleaning ו- data preparation לפני בניית המודל. לאחר מכן, עליכם לבנות עץ החלטה ולבדוק את ביצועיו ע"י שינוי פרמטרים במודל. לבסוף, עליכם להציג בצורה ויזואלית את עץ ההחלטה שקיבלתם ולהעריך את ביצועיו.

II. משימות התרגיל

1. הכנת הנתונים:

1.1 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.
- יש לאפשר קריאה של קובץ הנתונים מנתיב שיבחר המשתמש.

1.2 יש להשלים ערכים חסרים בקובץ:

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

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

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

1.3 יש לבצע דיסקרטיזציה למשתנים הנומריים הבאים:

- Monthly_Profit
- Spouse_Income
- Loan_Amount

את מספר ה-bins יש לקבוע בצורה הגיונית לפי שיקולכם. לשם כך, בדקו מה משמעות המשתנה וחלקו את הטווח ל-3 עד 5 אינטרוולים. ניתן לחלק על פי Equal-frequency discretization או על פי Equal-width discretization. אתם אמורים להפעיל שיקול דעת בבחירת שיטת הדיסקרטיזציה המתאימה מבין שתי השיטות שהוצעו לכם על פי היתרונות שלהן לדטהסט שברשותכם.

- ציינו בדוח איזה סוג דיסקרטיזציה ביצעתם ולכמה bins.

1.4 ניתן לשנות טיפוסי משתנים לסוג Factor ע"פ שיקול דעתכם במידה ונדרש. יש לרשום הסבר על אילו משתנים עשיתם זאת ומהי הסיבה.

1.5 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 30% מהקובץ עבור סט הבדיקה והיתר עבור סט האימון.

2. בניית המודל:

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 במדדgini split , כדי לבחור את משתנה הפיצול הבא בקודקוד. נסו להגדיר פעם את **קריטריון הפיצול** gini ופעם את קריטריון ה-information gain. השוו את ביצועי המודל בין שני הקריטריונים.
- הגדירו את פרמטר ה- minsplit המציין מה **מספר הרשומות המינימלי** 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).

1.2. 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.
- יש להכריע בין Gini ל- IG עם אותו minsplit ואז לבדוק את מי שבחרנו מביניהם עם ערכי minsplit שונים.

3. הערכת ביצועי המודל:

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

3.1.1 קריטריון הפיצול gini או information gain שבחרתם בסעיף הקודם.

3.1.2 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) ע"י שימוש בשיטות שלמדתם.