1 מבוא והבנת הבעיה

הבעיה המחקרית בה עוסק הפרויקט הינה הסיווג של אנשים לפי הכנסתם השנתית. כלומר, סיווג האם פרט מרוויח מעל 50k דולר בשנה (1) או פחות מכך (0). בעזרת נתונים דמוגרפיים של הפרט, כמו גיל, מספר שעות עבודה שבועיות, סקטור עבודה אליו משתייך, ארץ ועוד, נרצה לבנות מודל שיכלול בתוכו את המאפיינים המשפיעים על הסיווג של ההכנסה השנתית.

מטרתם של מחקרים רבים הינה לבנות מודל אשר מנבא האם אדם מרוויח יותר מ50k דולר בשנה. משימה מסוג זה יכולה להתעורר בסביבת ארגונים ללא מטרות רווח, אשר בדרך כלל מתקיימים מתרומות. הבנת ההכנסה של הפרט יכולה לעזור לארגון ללא כוונות רווח להבין טוב יותר האם הפרט יכול לתרום לו, ומהו גודל התרומה הצפוי מאותו אדם. בזמן שהבנת ההכנסה השנתית הכוללת של אדם יכולה להיות קשה כאשר מנסים לשלוף נתון זה ממקורות ציבוריים, או אפילו על ידי שאילת האדם עצמו, ניתן לנסות לבנות מודל אשר לוקח בחשבון מאפיינים דמוגרפיים אחרים אשר זמינים לניתוח והערכה.

Chet Lemon et al. וגם Karan Bhanot בחנו נתונים סטטיסטיים של מפקד האוכלוסין, במטרה לזהות האם אדם מרוויח מעל\מתחת ל50k דולר בשנה. החוקרים עשו שימוש במספר מערכות לומדות כמו Logistic Regression, Decision Tree, Random Forest. נמצא כי המאפיינים הנחוצים ביותר לביצוע התחזית היו 'age', 'education', 'hours per week', 'occupation', and 'sex'. בנוסף נבחנה עקומת ROC להשוואה בין המודלים.

2 הבנת הנתונים

2.1 תיעוד מקורות הנתונים ומשמעותם

המידע נאסף ממאגר המידע של מפקד האוכלוסין בארצות הברית בשנת 1994. המידע מכיל 14 מאפיינים (רציפים וקטגוריאליים), ברובם מכילים מידע דמוגרפי אודות אדם מסוים. המטרה היא לחזות האם אותו אדם מרוויח יותר (או פחות) מ $ 50k בשנה. סט הנתונים שברשותנו מכיל 32,561 רשומות.

<u>משמעות המאפיינים:</u>

Name	Description	Type
age	גיל	Numeric
workclass	סוג עבודה	Categorical
fnlwgt	משקל – מספר אנשים שמפקד האוכלוסין מאמין פרט זה מייצג	Integer
education	רמת השכלה גבוה ביותר שיש לפרט	Categorical
education-num	רמת השכלה גבוהה ביותר בצורה נומרית	Numeric
marital	מצב משפחתי	Categorical
occupation	עיסוק מרכזי של הפרט	Categorical
relationship	קשר של הפרט לאחרים	Categorical
race	גזע	Categorical
sex	המין הביולוגי של הפרט	Categorical
capital-gain	רווח הון של הפרט	Numeric
capital-loss	הפסד הון לפרט	Numeric
hours-per-week	מספר שעות העבודה בשבוע שהפרט דיווח	Numeric
native-country	ארץ	Categorical
y	האם הלקוח מרוויח יותר או פחות מ50,000 דולר לשנה.	binary

2.2 הסתברויות אפריוריות וקשרים בין מאפיינים

- **הסתברויות אפריוריות**

<u>משתנה המטרה Y</u>

עבור משתנה המטרה נבדוק מה ההסתברות האפריורית לכך שלקוח מרוויח מעל 50k דולר בשנה $(over\ 50K)$ ואת ההסתברות האפריורית לכך שלקוח מרוויח פחות מכך $(under\ 50K)$. את ההסתברויות נחשב בעזרת השכיחויות של כל אחת מהאופציות להחלטה על סמך הנתונים שברשותנו.

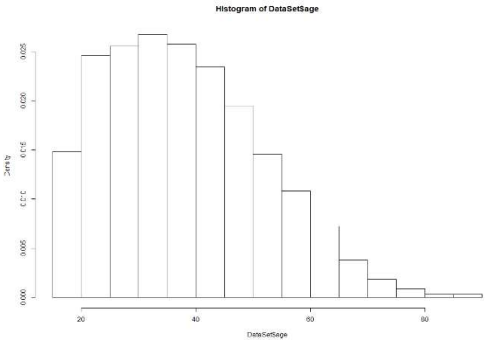
$$P(over\ 50K) = 0.24 \qquad P(under\ 50K) = 0.76$$

<u>מאפיין marital-status</u>

משתנה קטגוריאלי בעל 7 ערכים אפשריים. נציג את ההסתברויות האפריוריות הגבוהות ביותר (המתקבלות עבור 3 קטגוריות). עבור שאר הערכים ההסתברויות אפריוריות נמוכות במיוחד (פחות מ5%).

$$P(Divorced) = 0.1365,\quad P(Married\text{-}civ\text{-}spouse) = 0.4599, P(Never\text{-}married) = 0.328$$

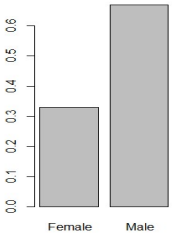
מאפיין age



משתנה רציף המתאר את גיל הנבדק. בכדי לנתח בצורה יעילה את ההסתברויות האפריוריות נחלק את הנתונים לדליים לפי ההיסטוגרמה המתקבלת למאפיין זה.

$$P(25-) = 0.1969\,,\ \ P(26-30) = 0.1278$$
$$P(31-40) = 0.2624, P(41-55) = 0.2873,\ \ P(56+) = 0.1255$$

מאפיין sex



משתנה קטגוריאלי המציג את מין הפרט.

$$P(Male) = 0.6692$$
$$P(Female) = 0.3308$$

מאפיין hours-per-week

משתנה רציף, לכן נרצה לאחד לדליים בכדי להציג את ההסתברויות האפריוריות.

$$P(<35) = 0.2113,\ \ P(35 < hours < 45) = 0.5695,\ \ P(>46) = 0.2192$$

מאפיין native-country

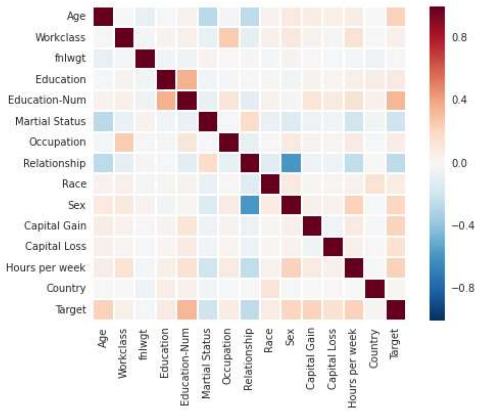
משתנה קטגוריאלי בעל 41 קטגוריות. ההסתברות האפריורית הגבוהה ביותר (בפער) מתקבלת עבור ארצות הברית $P(United\check{S}tates) = 0.8959$, עבור שאר המדינות ההסתברות האפריורית המתקבלת נעה סביב 5%.

- ## האם סט הנתונים מאוזן והאם מייצג את המציאות

לפי The Wall Street Journal (WSJ) כ73% מהאוכלוסייה מרוויחה פחות מ50K דולר בשנה, ורק השאר, כלומר כ27% מרוויחים מעל סכום זה. כלומר, הנתונים שבידנו אכן מציגים בצורה טובה את המציאות, למרות שסט זה אינו מאוזן.

- ## קשרים מעניינים בין המאפיינים

בכדי להבין לעומק את הנתונים, נרצה לבחון קשר בין מאפיינים ובין מאפיינים למשתנה המטרה Y.
בכדי לבחון את הקשרים בין המאפיינים השונים, נסתכל על מטריצת הקורלציות.



ניתן לראות כי קיימת קורלציה גבוהה בין Education וEducation Num , בין Workclass וOccupation, וקשר בין Sex, ו Marital Status עם Relationship. נשקול בהמשך הסרת חלק מהמאפיינים שעבורם זיהינו קורלציה גבוהה.

- ## מאפיינים חשודים כמשפיעים על משתנה המטרה

ישנם מספר מאפיינים שניתן לחשוד בהם בתור משפיעים על משתנה המטרה Y. בחלק זה נתייחס להשערות וידע על סמך מחקרים קיימים בנושא. כחלק מחקר סט הנתונים שקיבלנו, בחנו את הקשרים בין המאפיינים השונים למשתנה המטרה, בכדי לבחון את ההשערות ולבדוק אילו מאפיינים הם בעלי השפעה משמעותית על משתנה המטרה.

קשר בין מאפיין Age למשתנה המוסבר

ניתן לראות שככל שהגיל עולה, ההסתברות להכנסה של מעל 50k דולר בשנה הולכת וגדלה, עד לקבוצת הגיל האחרונה. ניתן להסביר זאת בעזרת ניסיון ומעמד שמגיעים אליו ככל שמתקדמים עם הגיל. כאשר עובדים מספר שנים צוברים ניסיון, ותק, ובדרך כלל מגיעים

למשרות בכירות יותר, ועל כן, ההכנסה השנתית גדלה. עם זאת, בגיל מסוים, ההכנסה מתחילה לקטון. ניתן להסביר זאת בעזרת פרישה לפנסיה למשל. (נספח 1)

קשר בין מאפיין Education Num למשתנה המוסבר

חשדנו במאפיין זה בגלל ההבנה שיש קשר בין רמת השכלה לבין הכנסה. נצפה שאנשים להם השכלה גבוהה יותר, תהיה גם הכנסה גבוהה יותר ולכן ההסתברות להרוויח מעל 50k ככל ההשכלה גבוהה יותר. ניתן לראות כי ככל שרמת ההשכלה אדם עולה, כך ההסתברות שהוא מרוויח מעל 50k בשנה תהיה גבוהה יותר. כלומר, קיים מתאם חיובי בין רמת ההשכלה לבין הכנסה שנתית. (נספח 1)

קשר בין משתנה Hours per week למשתנה המוסבר

אנחנו משערים, שככל שאדם עובד יותר, כך ככל הנראה המשכורת שלו עולה בהתאם. ואכן ניתן לראות שככל שמספר שעות העבודה השבועיות עולה, כך גם הסיכוי להרוויח מעל 50k דולר בשנה עולה. (נספח 1)

2.3 איכות הנתונים

- **נתונים חסרים**

עבור סט הנתונים הנתון ישנם מספר מאפיינים בעלי נתונים חסרים.

מאפיין	הסתברות לנתונים חסרים	כמות נתונים חסרים	משמעות
Workclass	5.64%	1836	ניתן לראות כי מדובר בכמות לא גדולה של נתונים חסרים. בגלל שמאפיין זה מוסיף מידע רלוונטי למשתנה המטרה, לא נרצה להסיר אותו ונשקול השלמת נתונים חסרים.
Occupation	5.66%	1843	ניתן לראות כי מדובר בכמות לא גדולה של נתונים חסרים. בגלל שמאפיין זה מוסיף מידע רלוונטי למשתנה המטרה, לא נרצה להסיר אותו ונשקול השלמת נתונים חסרים.
Native Country	1.79%	583	עבור מאפיין זה ניתן לראות כי אחוז הנתונים החסרים הוא יחסית קטן. בנוסף, ראינו כי הוא בעל השפעה על משתנה המטרה ולכן נשקול להשלים אותו.

את המשתנים להם יש נתונים חסרים, נבצע השלמה בעזרת אלגוריתם MICE. האלגוריתם מבצע השלמה לנתונים חסרים. השיטה מתבססת על מפרט מותנה לחלוטין, כאשר כל משתנה חסר מושלם על ידי מודל נפרד. האלגוריתם יכול להתמודד עם השלמת נתונים קטגוריאליים רציפים, בינאריים, לא מסודרים ועוד. (נספח 2)

3 הכנת הנתונים

3.1 בחירת מאפיינים

הסרת מאפיינים מסט הנתונים

מאפיין	תובנות	דרך פעולה
Education	עבור מאפיין זה, ראינו כי קיימת קורלציה גבוהה עם המאפיין Education Num. היות ומאפיין Education Num נותן מידע נוסף של סדר בין הערכים, נרצה להשאיר אותו.	משתנה זה אינו מוסיף מידע, מכיוון שזהה במשמעותו למאפיין אחר, ולכן נבחר להשמיט משתנה זה.
Relationship	עבור מאפיין זה, ראינו כי ישנם שני מאפיינים אשר בעזרתם ניתן להגדיר את ערך מאפיין זה.	מאפיין זה ניתן לייצוג על ידי שני מאפיינים אחרים (Sex, Marital Status) ולכן נבחר להסיר מאפיין זה.

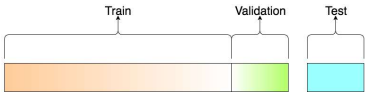
3.2 טיפול פרטני במאפיינים

מאפיין	הסבר	טיפול
Y	משתנה זה קיבל את הערכים 50> או 50=<. בכדי שנוכל לעבוד איתו בצורה נוחה, נרצה להציג בצורה בינארית. כלומר כעת המשתנה יענה על השאלה – האם הפרט מרוויח מעל 50k דולר בשנה?	המרה למשתנה בינארי – (0,1)
Native country	עבור מאפיין זה קיימים מספר נתונים חסרים (1.79%). אך עם זאת, המאפיין אכן משפיע על משתנה המטרה, וראינו כי קיים שוני בין ערך משתנה המטרה המתקבל עבור "ארצות הברית" לבין מדינות אחרות. בנוסף, ראינו כי קטגוריית "ארצות הברית" היא הנפוצה ביותר (90% מסך הנתונים).	נשנה את הקטגוריות ל2 ערכים בלבד – "ארצות הברית" ו"אחר".
Race	עבור מאפיין זה קיימות 2 קבוצות עיקריות – "Black" ו"White" . שאר הקטגוריות בעלות הסתברות אפריוריות השואפת לאפס.	נבחר לאחד לשלוש קטגוריות – Black, White, Other.
Capital loss	מאפיין זה מתאר את ההפסד הון שנתי. ההסתברות האפריוריות לערך השונה מ0 היא מאוד קטנה, ולכן נבחר להסתכל על המאפיין בצורה בינארית ולא לפי ערך ההפסד הממשי.	המרה לשתי קטגוריות – גדול מ0 \ שווה ל0. (משתנה בינארי)

Capital gain	מאפיין זה מתאר רווח הון שנתי. ההסתברות האפריורית לערך השונה מ0 היא מאוד קטנה, ולכן נבחר להסתכל על המאפיין בצורה בינארית ולא לפי ערך הרווח הממשי.	המרה לשתי קטגוריות – גדול מ0 \ שווה ל0. (משתנה בינארי)
Marital	עבור מאפיין זה ישנן 6 קטגוריות. ראינו כי עבור קטגוריות המתארות אנשים שלא חיים בזוגיות (גרושים, לעולם לא התחתנו, פרודים, אלמנים) ההתנהגות יחסית דומה. (השפעה זהה על משתנה המטרה).	נבחר לאחד את קטגוריות אלו לערך "לא נשואים", ולהישאר עבור מאפיין זה עם 3 קטגוריות.
Hours per week	משתנה זה הינו משתנה נומרי, המקבל ערכים בין 1-99, אנחנו משערים שעדיף להסתכל על מספר השעות ביחס לממוצע.	נחלק ל3 קבוצות – Low 0-35 Avg 36-45 High 46+

3.3 הכנת נתונים לאימון ובחינת מערכת לומדת

את סט הנתונים יש לחלק לשלושה חלקים – סט אימון (training), סט אימות (validation), ובחינה (test). את החלוקה ביצענו בשלבים – קודם כל ביצענו חלוקה לסט אימון training וסט בחינה test. בחרנו להקצות 20% מסט הנתונים לסט בחינה(6512 רשומות), אחרי סקירה לגבי החלוקה הנהוגה לסדר גודל נתונים כמו שלנו. בכדי לאמן את המודל כפי שצריך, רצינו להקצות את מירב הנתונים לסט האימון (80%) בכדי לאפשר למודל להיחשף לרשומות שונות ומגוונות. (26049 רשומות בסט האימון).



לאחר מכן, יש לחלק את סט האימון לסט אימון וסט אימות. את החלוקה נבצע בעזרת שיטת cross validation, הנקראת k-fold. בחרנו להשתמש בk=5. כלומר, לפי ערך זה, מתקבל כי מתוך 80% המהווים 26049 נתונים, 20% המהווים כ- 5210 נתונים יהיו סט האימות, וכל השאר –20839, יהיו סט האימון.

4 מידול

4.1 יער אקראי

מציאת יער אקראי עם ערכי ברירת המחדל של r

ערכי ברירת המחדל המתקבלים: **מספר העצים:** 500, **מספר משתנים מועמדים בכל פיצול:** 3. בנוסף, ניתן לראות לפי מטריצת המבוכה כי אחוז השגיאה הוא כ15%, כלומר אחוזי הדיוק על סט האימון הם 85%. אחוזי הדיוק המתקבלים לערכי ברירת המחדל:

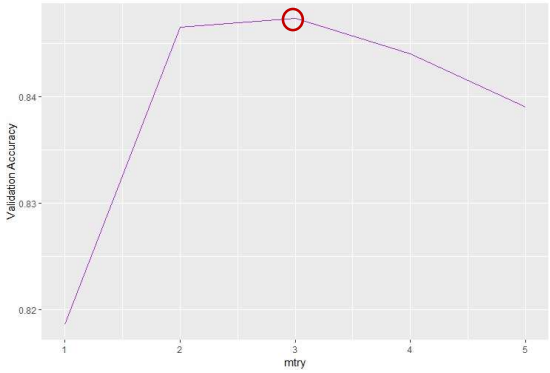
test	train
0.8499	0.8459

כיוונון פרמטרים

ישנן פרמטרים רבים שניתן לקבע בעת הרצת פקודת עץ אקראי. בחרנו לכוונן מספר פרמטרים אשר נלמדו בכיתה ולפי הנלמד בעלי ההשפעה המרבית על דיוק ואיכות היער האקראי.

Ntry - מספר הפרמטרים המועמדים בכל פיצול

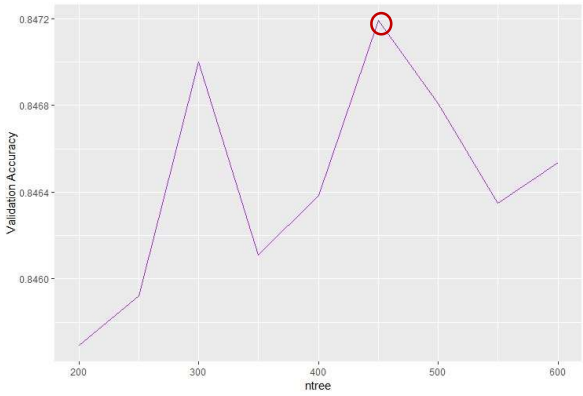
בכדי למצוא את מספר הפרמטרים האופטימאלי, נשתמש בשיטת k-fold ונחשב את אחוזי הדיוק של סט הווליד‌ציה עבור כל ntry מ1 ועד 10.



בהתחלה ישנה עליה עד הערך האופטימאלי, ולאחר מכן ירידה באחוזי הדיוק. ניתן לראות על פי הגרף של אחוזי הדיוק על סט הווליידציה שאכן נמצא שמספר הפרמטרים האופטימאלי לפיצול הוא ערך ברירת המחדל שנבחר – 3 פרמטרים.

Ntree - מספר העצים ביער

לאחר מציאת מספר פרמטרים אופטימאלי ntry=3, נבצע כעת כיוונון של מספר העצים ביער. נבצע תהליך איטרטיבי בו נבדוק את אחוזי הדיוק על סט הווליידציה עבור קפיצות במספר העצים מ200 ועד 600 בקפיצות של 50. לאחר ההרצה, ניתן לראות כי מספר העצים המיטבי הוא 450.



מודל יער אקראי נבחר

מאפייני היער הסופי המתקבלים לאחר כיוונון פרמטרים: **Ntry** = 3, **Ntree** = 450, **Maxnodes** = 1060. עבור יער זה מתקבלים אחוזי הדיוק הבאים:

test	train
0.85	0.8469

4.2 Neural Networks – רשת נוירונים

לפני תחילת האימון בעזרת רשת נוירונים, יש צורך להתמודד עם המשתנים הקטגוריאליים. את כל משתנים אלו הגדרנו בעזרת משתני דמה (one hot encoding) (נספח 3) והפעלנו את פונקציה as.factor() בכדי להגדיר אותם כמשתנים קטגוריאליים ולא כערך נומרי. בנוסף, הפעלנו את פונקציית scale() על המשתנים הרציפים (age, fnlwgt, educationNum) בכדי לנרמל אותם לאותה סקלת ערכים וסדר גודל.

הרצת הרשת עם ערכי ברירת המחדל

ערכי ברירת המחדל הם 28 נוירונים בשכבת הכניסה (כמספר המאפיינים), שכבה חבויה אחת בעלת נוירון אחד ושכבת יציאה בעלת 2 נוירונים (נספח 4). למעשה שכבת הכניסה והיציאה הינן קבועות, ואילו מספר השכבות החבויות, מספר הנוירונים בכל שכבה כזאת הוא היפר-פרמטר שיש לקבוע את ערכו (בחבילת nnet קיימת שכבה חבויה אחת). ישנן 33 משקולות שהרשת צריכה לקבוע בהינתן ערכי קונפיגורציה זו: $28_{\text{בכניסה}} + 1_{\text{בחבויה}} + 2_{\text{ביציאה}} + 2 \text{ biases} = 33\ weights$

חישבנו את אחוזי הדיוק בשלב האימון והבחינה בכדי לראות האם קיים מצב של over/under fitting.

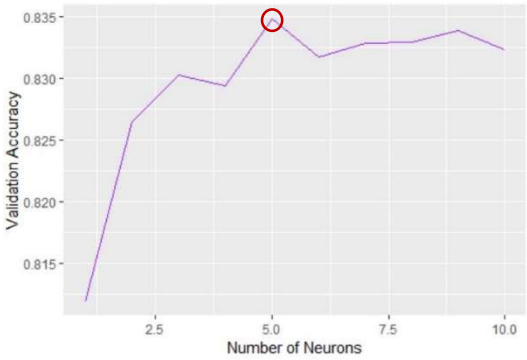
test	train
0.8268	0.8255

ניתן לראות כי קיבלנו הפרשים קטנים מאוד בין אחוזי הדיוק. מדד הדיוק עבור סט האימון נמוך 1.3×10^{-3} אחוזים מאשר הדיוק עבור סט הבחינה. כלומר, אנו לא נמצאים במצב של over fitting. באופן כללי, אחוזי הדיוק יחסית גבוהים הן עבור סט האימון והן עבור סט הבחינה.

כוונון פרמטרים

מספר נוירונים בשכבה החבויה

ביצענו תהליך של כיוונון פרמטרים – מספר הנוירונים בשכבה החבויה, בניסיון להעלות את אחוזי הדיוק של הרשת. נרצה למצוא את מספר הנוירונים בשכבה החבויה אשר ימקסמו את אחוזי הדיוק של סט האימות. נציג גרף של אחוזי הדיוק כפונקציה של מספר הנוירונים בשכבה החבויה. השתמשנו בשיטת k-fold עם k=5 על מנת לבחון את ערכי הפרמטרים שיביאו למקסום דיוק של סט האימות.



ניתן לראות כי בהתחלה ישנה עליה באחוזי הדיוק, כי הוספת הנוירונים בשכבה החבויה הופכת את המודל למורכב יותר, וכך המודל מבצע חיזוי בצורה טובה יותר גם על סט האימות. אך ברגע שעוברים את מספר הנוירונים האופטימאלי, מתקבל מצב של over fitting. נמצא כי מספר הנוירונים האופטימאלי הוא 5 (ערך זה הביא לאחוזי דיוק על סט הוולידציה הגבוהים ביותר – 83.48%). לאחר קיבוע ערך זה נקבל את אחוזי הדיוק הבאים –

test	train
0.832	0.8373

ערך weigh decay

Weight decay הינו פרמטר רשת אשר מנסה לכווץ את המשקלים ברשת, במטרה למנוע מצב של over fitting. נרצה למצוא את הערך המיטבי עבור פרמטר זה שיגדיל את אחוזי הדיוק על סט האימות (נבדוק זאת שוב בעזרת שיטת K-fold).

נתוני הקונפיגורציה: 28 נוירונים בשכבת הכניסה, 5 נוירונים בשכבה החבויה (ערך אופטימאלי שנמצא בחלק הקודם), מספר האיטרציות הוא 100, ערך הdecay האופטימאלי – 0.2 התקבל על ידי מקסום דיוק סט האימות (0.8347). ערך זה נמצא בעזרת אלגוריתם איטרטיבי שבחן את ערכי הפרמטר בקפיצות של 0.05 (נספח 5). כמות המשקולות הנלמדות – 157 (נספח 6).

test	train
0.8317	0.8369

הגבלת מספר איטרציות

מטרת קונפיגורציה זו היא מניעת מצב של over fitting. **נתוני הקונפיגורציה:** 30 נוירונים בשכבת הכניסה, 5 נוירונים בשכבה החבויה (ערך מקסימאלי שנמצא בסעיף הקודם), ערך weight decay = 0.2 האופטימאלי שנמצא בחלק הקודם. מספר האיטרציות המיטבי הנמצא הוא 200 אשר מקסם את דיוק סט האימות (0.8354). ערך זה נמצא בעזרת אלגוריתם איטרטיבי שבחן את ערכי הפרמטר בקפיצות של 10 בטווח הערכים 10-300 (נספח 7). עבור רשת זו, ניתן לראות כי אחוזי הדיוק על סט הבחינה הם הגבוהים ביותר. ולכן היא תהיה הרשת הנבחרת בתור הרשת הנוירונלית.

test	train
0.8378	0.8337

4.3 אשכול K-means

k-means

הוצאת משתנה המטרה מתוך הinput

אלגוריתם k-means הוא שיטה לניתוח אשכולות (clustering), אשר מבצע למידה לא מונחת, כלומר בקלט נכניס את מספר המחלקות הקיימות אך לא את החלוקה למחלקות בפועל. לכן, נרצה להוציא את משתנה המטרה מהקלט בכדי לאפשר את הלמידה הלא מונחת.

הרצת מודל עם ערכי ברירת מחדל

לפני הרצת המודל, נצטרך להכין את הנתונים להרצה של k-means. את המאפיינים הקטגוריאליים נציג בעזרת משתני דמה (כפי שעשינו ברשת הנוירונים), ונריץ את הפקודה ()scale בכדי לנרמל וליצור התאמה בסדרי הגודל של המשתנים. ביצענו הרצה של המודל עם ערכי ברירת המחדל, ועם k=2 כי בבעיה שלנו קיימות 2 מחלקות – רווח שנתי מתחת ל50k דולר בשנה (0) ומעל (1).

בכדי לחשב את אחוזי הדיוק של סט האימון, ביצענו השוואה בין הסיווג לקבוצות לבין הקבוצות בפועל (0\1).

אחוזי הדיוק של סט האימון חושבו לפי $\frac{sum(y=clust)}{number\ of\ rows} = 0.7024$. כלומר, 70.24% מהרשומות סווגו בצורה נכונה בהתאם למחלקה בפועל. הסיווג שקיבלנו לפי k means היה 1,2 ונדרשנו להבין מה סביר יותר שכל אשכול מייצג (1,0 בבעיה שלנו). בדקנו את שתי האפשרויות ובחרנו את זו שהניבה אחוזי דיוק גבוהים יותר (כלומר סבירה יותר).

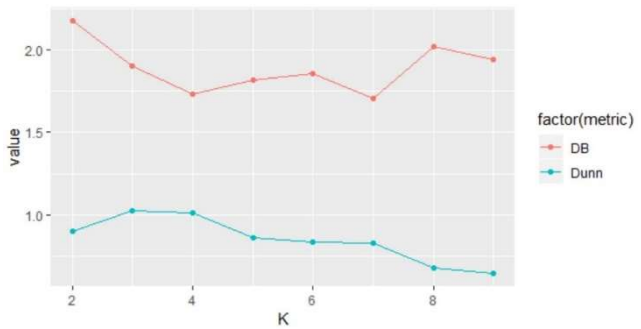
<u>טיב התאמה בין אשכולות למחלקות</u>

ראינו כי אחוזי הדיוק של האימון אינם גבוהים במיוחד. בבחינת אלגוריתם אשכול, נשאף להגיע למצב שבו המרחקים בין המחלקות השונות הם הגדולים ביותר (between), ולעומת זאת, בתוך המחלקה (within) – קטנים ביותר. נבחן 2 מדדים אשר לוקחים בחשבון את שני מרחקים אלו, כאשר המטרה שלנו היא למזער את DB (Davis Bouldin) המתייחס למרחק בתוך המחלקה, ולמקסם את dunn המתייחס למרחק בין המחלקות.

לאחר הרצת אלגוריתם k-means על סט האימון, קיבלנו את המדדים הבאים:

Value (train)	parameter
0.839	dunn
2.37	DB

<u>ערך k מיטבי</u>



בכדי למצוא ערך k מיטבי (כלומר מספר המחלקות המיטבי לסט הנתון), נבצע אימון על ערכי k שונים. עבור כל k נציג 2 מדדים – DB ו-dunn שני מדדים אלו מתארים את טיב המודל כפי שפורט מעלה. ערך k אופטימאלי יבחר לפי ערכי המדדים, כאשר ברצוננו למקסם את dunn (המתייחס למרחק בין המחלקות) ולמזער את DB (המתייחס למרחק בתוך המחלקה). בחנו 8 ערכי k, מ2 ועד 9.

לפי הגרף, ניתן לראות כי נקודת המינימום של DB ונקודת המקסימום של dunn מתקבלות עבור אותו הk, (k=4) ולכן ערך זה יהיה k המיטבי. ערך זה גבוה יותר מהמחלקות הקיימות בסיפור המקרה (2 מחלקות – "כן"\"לא"). ניתן להסביר זאת על ידי ההבנה שכך הנראה ישנה חלוקה יותר מתאימה לרווח של אנשים, ולא מספיק להסתכל על מעל\מתחת ל50k דולר בשנה (למשל חלוקה ל4 קבוצות לפי קפיצות של 25k).

k	DB	dunn	
2	2.180321	0.9005975	1.2797233
3	1.903292	1.0224134	0.8808791
4	1.730559	1.0110906	0.7194686
5	1.818096	0.8583873	0.9597088
6	1.859211	0.8357080	1.0235026
7	1.707347	0.8273776	0.8799697
8	2.022408	0.6752791	1.3471288
9	1.943186	0.6426970	1.3004891

4.4 מודל SVM

בחרנו בנוסף לבחון את מודל SVM שלא נלמד במסגרת הקורס. מודל זה מתאים גם הוא לבעיית סיווג כמו הבעיה שלנו ולכן בחרנו לבחון גם אותו. ביצועי המודל על סט האימון 0.8477 ועל סט הבחינה 0.847.

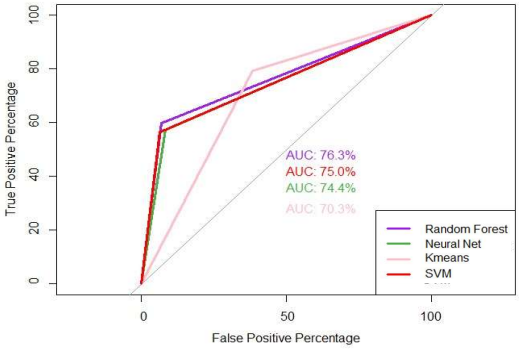
5 הערכה והשוואה בין מודלים

במהלך העבודה בדקנו 4 מודלים, וכעת נרצה להשוות ביניהם ולבחור את הטוב ביותר. שלושה מודלים הם מודלים מונחים – רשת נוירונים, יער אקראי וSVM והמודל הנוסף הוא מודל לא מונחה – אלגוריתם k means. בכדי להיות מסוגלים להשתמש באלגוריתם האשכול לביצוע תחזיות על סט הבחינה, יש להשתמש במרכזי המחלקות שנלמדו בסט האימון, ולבחון עבור הנתונים החדשים – לאיזה מחלקה הם משויכים לפי קרבה למחלקות השונות. השתמשנו בפונקציית הpredict של חבילת cclust, אשר עושה שימוש במרכזי המחלקות ומסווגת כל תצפית למחלקה הדומה לה ביותר.

Test	Chosen Model	
0.85	**ntree** = 450 ,**mtry** = 3	**Random Forest**
0.8378	**Opt decay**= 0.2 ,**Maxit** = 200 **Size** = 5	**Neural Network**
0.657	k = 2	**K - means**
0.847		**SVM**

עקומת ROC

בנוסף, נציג את עקומת ROC ונבחן את ה - AUC עבור כל מודל. עקומה זו מציגה את ביצועי המסווגים הדו ערכיים.



לסיכום, המודל הטוב ביותר לפי אחוזי הדיוק ולפי שטח מתחת לעקומת הROC יער אקראי.

6 שיפור המודל

בחרנו להשתמש בשיטת הצבעת הרוב (Majority Voting) לפי מודל אשר עושה שימוש במספר מודלים (Ensemble), ובוחרת את החלטת הרוב. כאשר שילבנו בין שלושת המודלים – יער אקראי, רשת נוירונלית ומודל אשכול, קיבלנו שהם מסווגים לאותה המחלקה ב87% את סט האימון וב90% את סט הבחינה.

אחוזי הדיוק של המודל המשולב (הסיווג מתבצע לפי החלטת הרוב) הם 87.3% על סט האימון ו84.9% על סט הבחינה. כלומר, שימוש במודל משולב נותן תוצאות דומות מאוד ואינו משפר את החלטת היער האקראי שנבחר כמודל הטוב ביותר בשלב הקודם.

7 סיכום ומסקנות

בעבודה זו ניתחנו את השאלה – האם אדם מרוויח מעל או מתחת ל50k דולר בשנה. עבדנו לפי כל שלבי מתודולוגיית CRISP DM, החל מהכנת נתונים ועד מידול הבעיה. בחרנו לבחון ארבעה מודלים שונים, שלושה מהם נלמדו במהלך הסמסטר (יער אקראי, רשת נוירונלית ואשכול k means), ומודל נוסף SVM אשר לא נלמד במסגרת הקורס. עבור כל אחד מהמודלים ביצענו תהליך של כיוונון מספר פרמטרים חשובים, במטרה לשפר את אחוזי הדיוק של המודל. בכדי לבחור את ערכי הפרמטרים הטובים ביותר השתמשנו בשיטת k fold ובחרנו בצורה איטרטיבית את ערך הפרמטר אשר ממקסם את אחוזי הדיוק על סט הולדיציה.

לאחר כיוונון פרמטרים ומציאת מודל טוב ביותר לכל סוג מידול, השווינו את המודלים השונים בהסתכלות על אחוזי הדיוק על סט הבחינה ובעזרת הסתכלות על עקומת ROC ושטח AUC. נמצא כי מודל היער האקראי הוא הטוב ביותר גם לפי אחוזי הדיוק וגם לפי עקומת ROC.

בנוסף, בדקנו אופציה של שילוב המודלים וביצוע סיווג לפי החלטת הרוב Majority Voting, אך מצאנו כי אחוזי הדיוק דומים מאוד לביצוע החלטה לפי יער אקראי בלבד, ולכן לא ראינו סיבה לבצע עבודה מיותרת כי אין ערך מוסף להחלטת הרוב.

לסיכום, מצאנו שהמודל הטוב ביותר לביצוע סיווג לבעיה המתוארת הוא יער אקראי עם 450 עצים, ו3 פרמטרים מועמדים בכל פיצול. אחוזי הדיוק על סט הבחינה הם 0.85.

8 ביבליוגרפיה

1. Kohavi, R. (1996, August). *Scaling up the accuracy of naive-bayes classifiers: A decision-tree hybrid*. In *Kdd* (Vol. 96, pp. 202-207).

2. Lemon C. , Zelazo C. , Mulakaluri K. *Predicting if income exceeds $50,000 per year based on 1994 US Census Data with Simple Classification Techniques.*

3. Valentini, G., & Masulli, F. (2002). *Ensembles of learning machines. In Italian workshop on neural nets* (pp. 3-20). Springer.

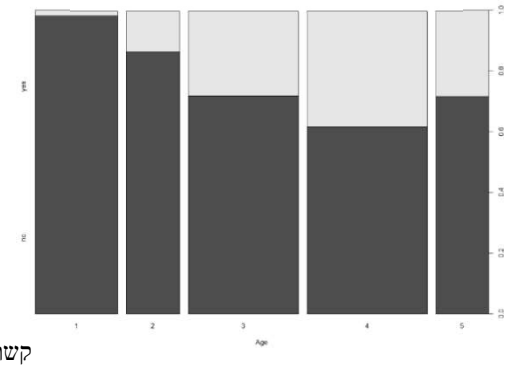
4. *Finding Donors for Charity using Machine Learning*, Github, https://sajalsharma.com/portfolio/finding_donors

5. *Will your income be more than $50K/yr? Machine Learning can tell,* https://towardsdatascience.com/will-your-income-be-more-than-50k-yr-machine-learning-can-tell-92138745fa24

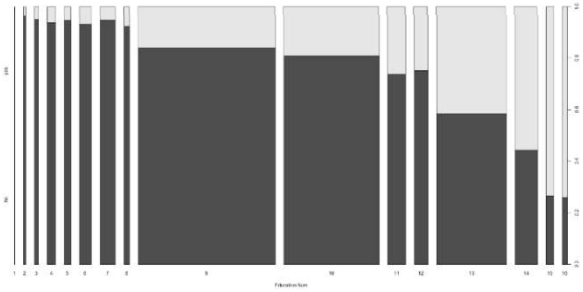
9 נספחים

נספח 1 – קשר בין המאפיינים למשתנה מטרה

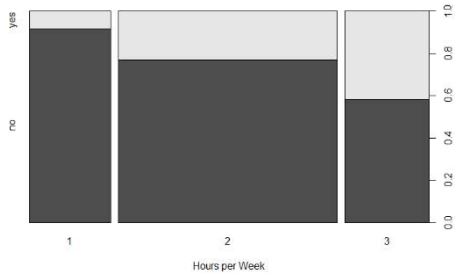
קשר בין גיל למשתנה המטרה



קשר בין שנות לימוד למשתנה המטרה



קשר בין שעות עבודה שבועיות למשתנה המטרה



נספח 2 - השלמה בעזרת אלגוריתם Mice

- השלמת עמודת Native Country – האלגוריתם לא הצליח להשלים מהסיבה שקיים חוסר איזון בין הקטגוריות השונות (90% מסך הרשומות שייכות לקטגוריה "ארצות הברית"). לכן, את הנתונים החסרים סיפחנו לקבוצת "אחר".
- השלמת עמודת Workclass – גם כאן קיים חוסר איזון, ולכן השמטנו 2 קטגוריות אשר לשתיהן הסתברויות אפריוריות השואפות ל0. לאחר מכן, ביצענו השלמה בעזרת אלגוריתם Mice.

נספח 3 - הכנות לרשת נוירונלית

מספר דוגמאות להמרה למשתני דמה.

#משתנה Workclass

תחילה איחדנו ל3 קטגוריות – private, self-emp, gov. רמת הבסיס היא – private. הגדרנו בעזרת 2 משתני דמה את 3 הקטגוריות

$$workclass\_D1 = \begin{cases} 1 & \text{self} - \text{emp} \\ -1 & \text{else} \end{cases} \quad ; \; workclass\_D2 = \begin{cases} 1 & \text{goverment} \\ -1 & \text{else} \end{cases}$$

#משתנה sex

הגדרנו בעזרת משתנה דמה יחיד

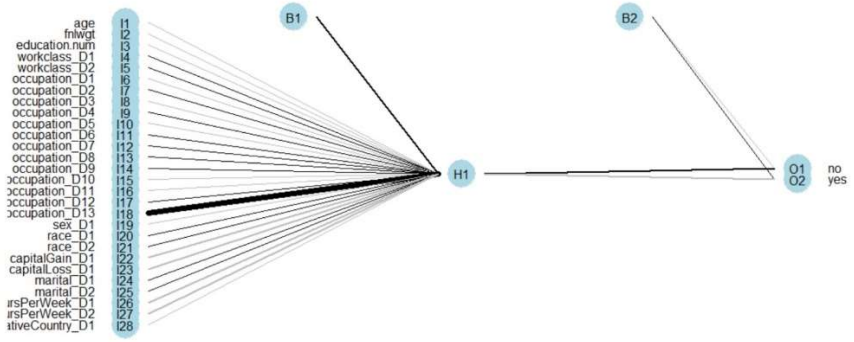
$$Sex\_D1 = \begin{cases} 1 & \text{Male} \\ -1 & \text{Female} \end{cases}$$

#משתנה Race

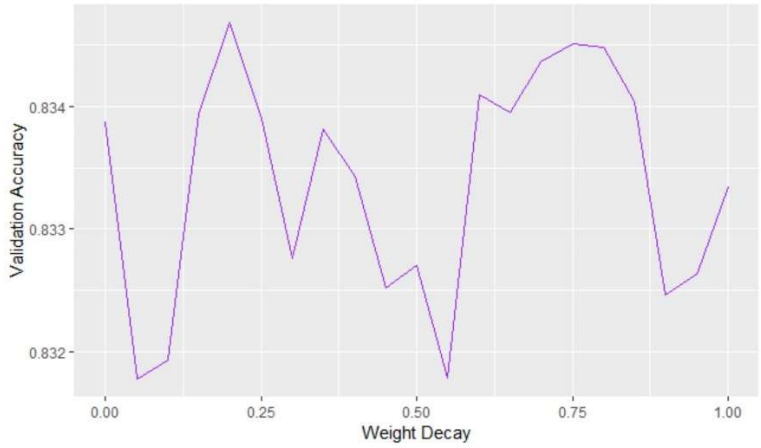
רמת הבסיס היא – White. הגדרנו בעזרת 2 משתני דמה את 3 הקטגוריות

$$Race\_D1 = \begin{cases} 1 & \text{Blace} \\ -1 & \text{else} \end{cases} \quad ; \; Race\_D2 = \begin{cases} 1 & \text{Other} \\ -1 & \text{else} \end{cases}$$

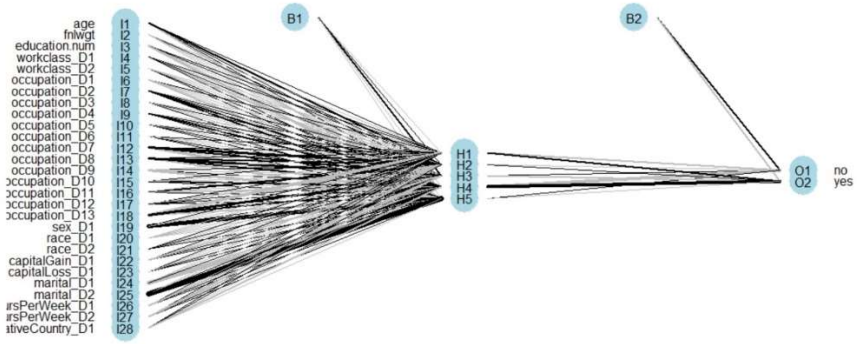
נספח 4 - רשת ברירת מחדל



נספח 5 - כיוונון פרמטר weight decay



נספח 6 – מספר משקולות עבור כיוונון weight decay



$$5 \times 28_{\text{בכניסה}} + 5 \times 2_{\text{בחבויה}} + 1 \times 5 \text{ biases} + 1 \times 2 \text{ biases} = 157\ weight$$

נספח 7 - כיוונון מספר איטרציות