פרויקט סיום

שם הקורס: מבוא לניתוח נתונים בפייתון

מספר הקורס: 094202

מגישים:
דביר טויטו 324270883
רועי מסקליק 212234637

תאריך הגשה: 02.07.2020

תוכן עניינים

תודות

ראשית, אנו רוצים להודות לכל מי שעזר לנו ללמוד את הקורס במהלך כל הסמסטר.

לעופרה אמיר המרצה, שהעבירה את כל ההרצאות באופן מעניין ופעיל, והצליחה להתגבר על הקושי של הוראה מקוונת בעזרת סרטונים מקדימים, עבודות עצמאיות, וסקרים על החומר שהכינה עבורנו.

לאלכס טואיסוב ורפאל שללה המתרגלים, שידעו לענות לנו על כל שאלה, וגם אם לא ידעו, הראו אכפתיות וניסו לעזור לנו לפתור את הבעיות אשר נתקלנו בהן.

לזהר גלעד בודק התרגילים, אשר בדק בקפידה את תרגילי הבית, ואף דאג לפרגן בבונוס כשהיה צריך.

תודה רבה לכולם, לכולכם, חלק בפרויקט זה!

פרק ראשון: תיאור מסד הנתונים

1.1 תיאור עמודות מסד הנתונים

מסד הנתונים שבו השתמשנו כולל בתוכו 48,842 רשומות כאשר כל רשומה מתארת את פרטיו של אדם מבוגר ואת משכורתו השנתית, במסד הנתונים כלול מידע נוסף אשר עשוי לעזור לנו במהלך כל תהליך ניתוח הנתונים.

1.2 תיאור עמודות מסד הנתונים

כאשר כל שורה במסד הנתונים מייצגת בן אדם, העמודות מייצגות:

שם העמודה	סוג משתנה העמודה	תיאור העמודה
age	מספר שלם	גילו של בן אדם, צריך להיות גדול מ- 0.
workclass	מחרוזת	תאגיד שבו בן אדם עובד
fnlwgt	מספר שלם	מתאר את מספר האנשים שהדירקטוריון מאמין כי האדם מייצג, כלומר אם ערך הfnlwgt שווה ל-2, מספר האנשים בעולם אשר מאפייניהם זהים למאפייניו של האדם הנוכחי, הוא 2.
education	מחרוזת	מוסד ההשכלה הגבוה ביותר שבו האדם למד
education-num	מספר שלם	מספר שנות ההשכלה שלמד האדם
marital-status	מחרוזת	מצבו המשפחתי של בן אדם
occupation	מחרוזת	משלח ידו של בן אדם
relationship	מחרוזת	מערכת היחסים בה נמצא בן אדם
race	מחרוזת	גזעו של בן האדם
gender	מחרוזת	מינו של בן אדם
capital-gain	מספר שלם	רווחי ההון של בן אדם
capital-loss	מספר שלם	הפסד ההון של בן אדם
hours-per-week	מספר שלם	שעות העבודה השבועיות של בן אדם
native-country	מחרוזת	המדינה שבה נולד בן אדם
income	מחרוזת	משכורתו השנתית של בן אדם, המחרוזת מתארת האם הבן אדם קיבל מעל 50 אלף בשנה או מתחת ל50 אלף בשנה

1.3 שינוים שביצענו במסד הנתונים:

בכדי להקל על חישובינו ביצענו שינוים במסד הנתונים:

מחיקת העמודה education:
החלטנו למחוק את העמודה הזו מכיוון שעמודה זו והעמודה **education-num** מייצגות את ההשכלה של בן אדם, העמודה **education-num** מתארת את השכלתו בדיוק רב יותר ותועיל לנו בכדי ליצור התפלגויות שקשורות בהשכלתו של בן אדם, לכן החלטנו להוריד את העמודה **education**

מחיקת העמודה capital-gain:
עמודה זו מכילה בתוכה 44,807 רשומות מתוך 48,842 שערכה בהן הוא אפס, העמודה הזו לא תוכל לשמש לנו תועלת, ואף היא עשויה להטעות את המסווג שיצרנו. מכיוון שכמות הרשומות בהן הערך שונה מאפס הוא זעיר לעומת כלל הרשומות, עשוי להיווצר מצב שבו המסווג בוחר לסט האימון רק רשומות בהן ערכי העמודה הם אפס, ולסט הבדיקה רק רשומות בהן הערכים שונים מאפס. דבר זה עשוי להקטין משמעותית את רמת הדיוק של המסווג. בנוסף לכך, עמודה זו לא תועיל לנו בתיאור ההתפלגויות והקורלציות שקשורת בערך זה, מכיוון שהוא ברובו אפסים, ולכן לא נוכל להבחין בשינוי שלו לצורכי התפלגויות וקורלציות.

מחיקת העמודה capital-loss
עמודה זו מכילה בתוכה 46,560 רשומות מתוך 48,842 שערכה בהן הוא אפס. נמחק את עמודה זו מאותן הסיבות שבגללן מחקנו את העמודה **capital-gain.**

בסך הכל, שתי העמודות האחרונות כוללות בתוכן 6317 ערכים ששונים מ - 0, כמות אשר אנו סבורים כי נוכל להוריד ממסד הנתונים מבלי לפגוע באופן משמעותי באמינותו. לכן, החלטנו להוריד את שתי העמודות.

הוספת העמודה binary_income
הוספנו לנתונים עמודה בשם binary_income. עמודה זו ממירה את תוצאות העמודה income לערכים בינארים, כאשר הערך הבינארי יהיה 1 אם הערך בעמודה **income** יהיה '>50k', הערך הבינארי יהיה 0 אם הערך בעמודה **income** יהיה '<=50k'. המרה זו עזרה לנו בחישובים לאורך הפרויקט. למשל, בחלק 3 רצינו לחשב את ההסתברות שלמבוגר תהיה הכנסה גבוה מחמישים אלף במסד הנתונים שלנו, בעקבות ההמרה שלנו, ממוצע על העמודה **binary_income** יתן לנו את התוצאה הזו. בנוסף לכך, על מנת להשתמש במסווג knn אנו זקוקים למשתנה מטרה אשר ערכיו הם מספריים, ולא קטגוריאליים, לכן צורך זה עודד אותנו יותר להוסיף את עמודה זו.

מחיקת העמודה fnlwgt
עמודה זו מתארת את מספר האנשים כי הדירקטוריון סבור כי כל עמודה מייצגת, כדי לבדוק את אמינות הערכים אשר מתוארים על ידי העמודה, סכמנו את מספר האנשים הכולל שקיים בעולם לפי עמודה זו ויצא לנו כי המספר גבוה יותר מ- 9 מיליארד, החלטנו להוריד את העמודה עקב חוסר אמינות של הנתונים המופיעים בה.

5 השורות הראשונות של מסד הנתונים לפני השינוים שביצענו:

	age	workclass	fnlwgt	education	educational-num	marital-status	occupation	relationship	race	gender	capital-gain	capital-loss	hours-per-week	native-country	income
0	25	Private	226802	11th	7	Never-married	Machine-op-inspct	Own-child	Black	Male	0	0	40	United-States	<=50K
1	38	Private	89814	HS-grad	9	Married-civ-spouse	Farming-fishing	Husband	White	Male	0	0	50	United-States	<=50K
2	28	Local-gov	336951	Assoc-acdm	12	Married-civ-spouse	Protective-serv	Husband	White	Male	0	0	40	United-States	>50K
3	44	Private	160323	Some-college	10	Married-civ-spouse	Machine-op-inspct	Husband	Black	Male	7688	0	40	United-States	>50K
4	18	?	103497	Some-college	10	Never-married	?	Own-child	White	Female	0	0	30	United-States	<=50K

5 השורות הראשונות של מסד הנתונים לאחר השינוים שביצענו:

	age	workclass	educational-num	marital-status	occupation	relationship	race	gender	capital-gain	capital-loss	hours-per-week	native-country	income	binary_income
0	25	Private	7	Never-married	Machine-op-inspct	Own-child	Black	Male	0	0	40	United-States	<=50K	0
1	38	Private	9	Married-civ-spouse	Farming-fishing	Husband	White	Male	0	0	50	United-States	<=50K	0
2	28	Local-gov	12	Married-civ-spouse	Protective-serv	Husband	White	Male	0	0	40	United-States	>50K	1
3	44	Private	10	Married-civ-spouse	Machine-op-inspct	Husband	Black	Male	7688	0	40	United-States	>50K	1
4	18	?	10	Never-married	?	Own-child	White	Female	0	0	30	United-States	<=50K	0

1.4 השאלות שרצינו לחקור

רצינו לחקור בפרויקט שלנו בעיקר שאלות אשר יעזרו לנו בסיווג ובהשערה שביצענו. לכן, שאלותינו היו:

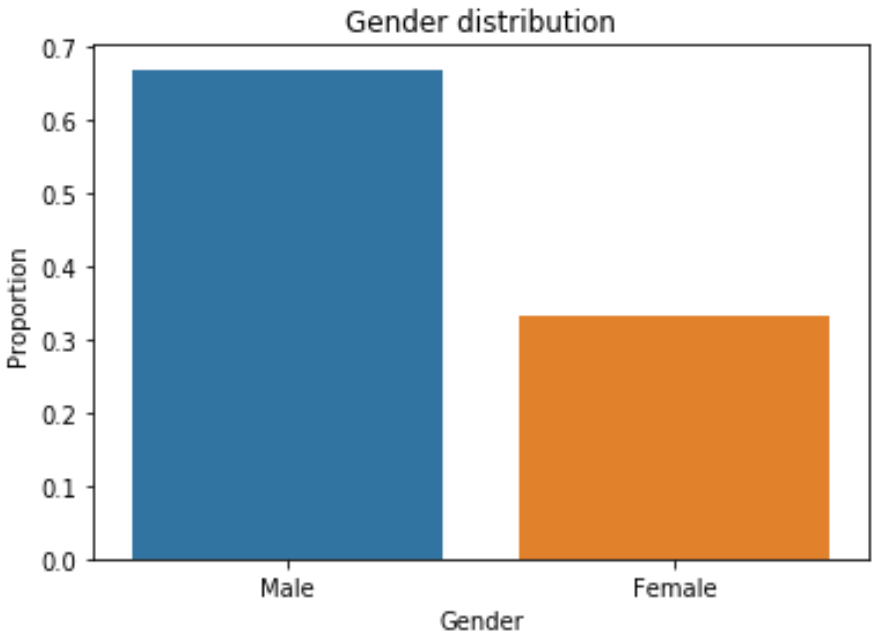
1. כיצד מתפלגים הנשים והגברים במסד הנתונים שלנו?
שאלה זו עניינה אותנו מכיוון שידענו שיש פוטנציאל במסד הנתונים שלנו לאפליה נגד נשים, לכן ידענו שהתפלגות המינים תעזור לנו בחישובים יותר מתקדמים בהשערתנו. בנוסף, שאלה זו עזרה לנו בהעלאת הטיה במסד הנתונים שלנו.
2. כיצד מתפלג מספר שנות הלימוד שכל אדם למד?
שאלה זו עניינה אותנו מהיותנו סטודנטים, רצינו לדעת בעיקר מהי כמות האנשים מתוך מסד הנתונים שלמדו מעבר ל-12 שנות לימוד.
3. כיצד מתפלג השכר בין האנשים במסד הנתונים?
שאלה זו עניינה אותנו עקב חשיבה על המשפט "80-20" (20 אחוז מהאנשים מחזיקים את 80 אחוז מן הכסף בעולם), ורצינו לדעת בעקבות משפט זה את אחוז האנשים בחברה ששכרם גבוה.
4. האם מספר שנות הלימוד משפיע על שעות העבודה השבועיות ועל השכר של האנשים במסד הנתונים.
שאלה זאת עניינה אותנו כי רצינו לדעת האם השקעה של אנשים בלימודים אכן משפיעה על עתידם הכלכלי.
5. השאלה המרכזית שלנו הייתה: האם ישנה אפליה בין גברים לבין נשים? כלומר, האם אחוז הגברים שמקבלים שכר שנתי שגבוה מחמישים אלף גבוה מאחוז הנשים שמקבלות שכר שנתי שגבוה מחמישים אלף.
שאלה זאת עניינה אותנו כי בימינו נעשים מאמצים רבים על מנת לתקן את הפער שנוצר אצל נשים, ובפרט בשכרן, ורצינו לדעת האם הנתונים שבידינו אכן תומכים במאמצים אלה.

פרק שני: ניתוח הנתונים

2.1 התפלגויות

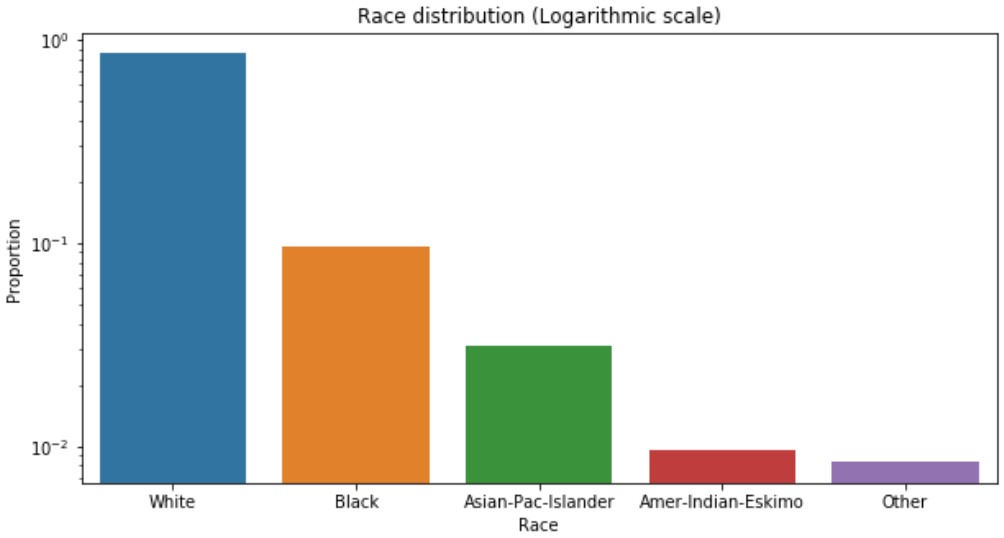
2.1.1 התפלגות המגדרים

במסד הנתונים מספר הגברים גדול בערך פי שניים מאשר מספר הנשים. השתמשנו בנתון זה כאשר ניסחנו את ההשערה שרצינו לחקור.



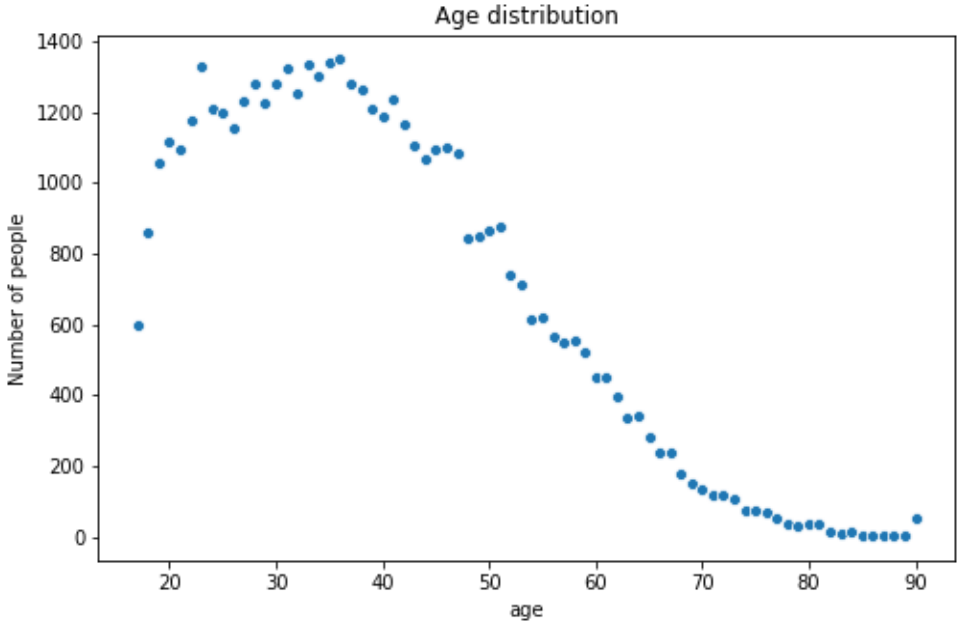
2.1.2 התפלגות הגזעים

ניתן לראות כי הרוב המוחלט במדגם שייך לגזע האנשים לבנים, ובמקום השני גזע האנשים השחורים.



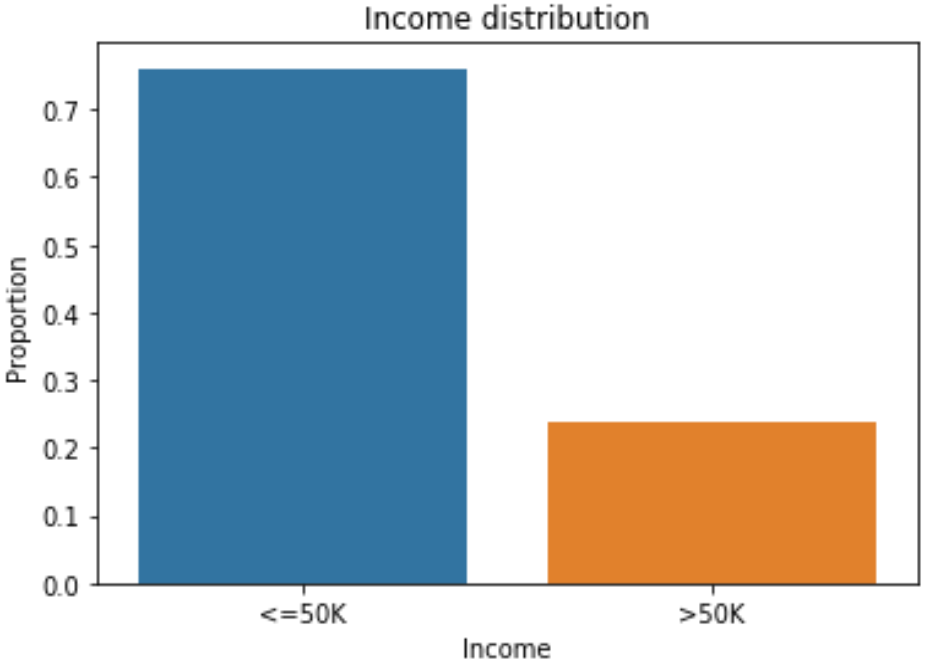
2.1.3 התפלגות הגילאים

ניתן לראות כי הגיל של המשתתפים במדגם נע בין עשרים לתשעים, כאשר הרוב הם בטווח הגילאים של עשרים עד חמישים.



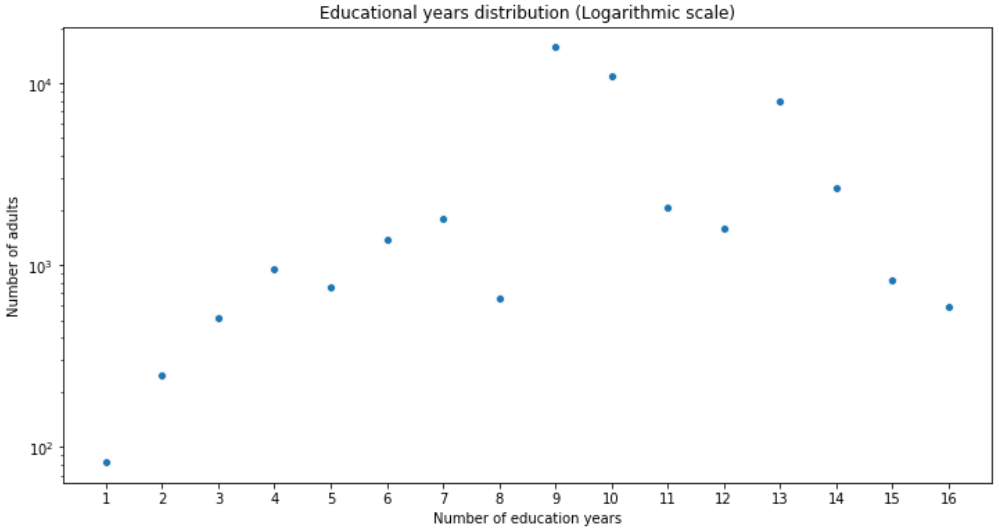
2.1.4 התפלגות ההכנסות

ניתן לראות כי הרוב המוחץ של המשתתפים במדגם משתכרים בשכר הנמוך מחמישים אלף או שווה לחמישים אלף, ואילו רק מעט מאוד אנשים משתכרים ביותר מכך. דבר זה עונה לנו על השאלה, ומראה לנו כי המשפט שהוביל אותנו לחקור את התפלגות זה כנראה נכון.



2.1.5 התפלגות שנות ההשכלה

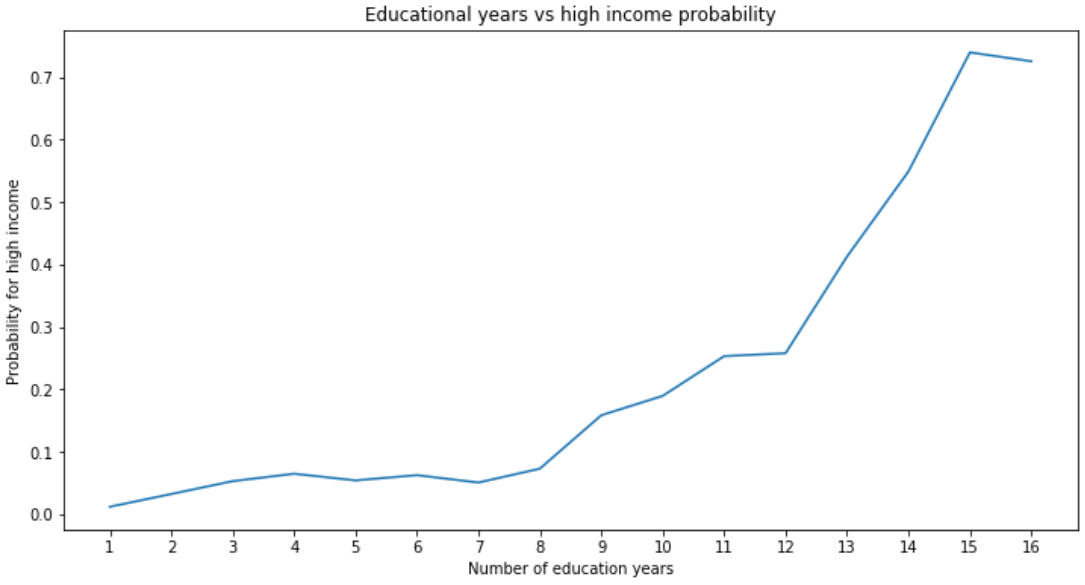
ניתן לראות כי חלק ניכר מהמשתתפים במדגם למדו בין תשע לעשר שנים. בנוסף, ניתן לראות כי רק מספר יחסית נמוך של משתתפים סיימו חמש עשרה או שש עשרה שנות לימוד. על כן, ניתן להסיק כי מרבית המשתתפים במדגם אינם סיימו (וכנראה לא התחילו בכלל) לימודים לתואר.



2.2 קורלציות

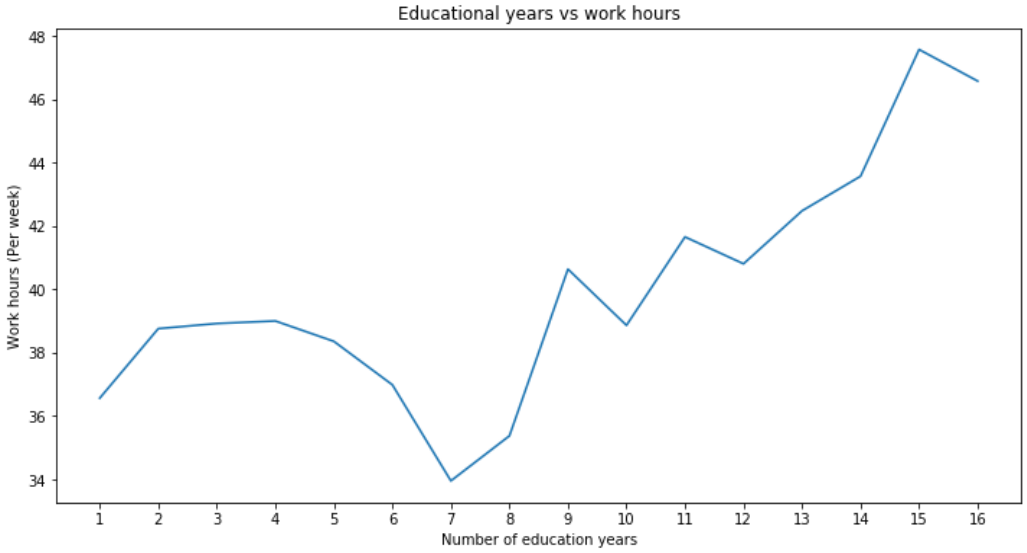
2.2.1 קורלציה בין מספר שנות ההשכלה לבין ההכנסה

ניתן לראות כי קיימת קורלציה חיובית גבוהה במיוחד בין מספר שנות הלימוד של המשתתפים במדגם לבין האם הכנסתם גבוהה מחמישים אלף. כלומר, אחוז האנשים שמשכורתם גבוהה מחמישים אלף עולה ככל שמספר שנות ההשכלה גדל. תוצאה זאת מאששת את השערתינו בנושא, שהייתה שאכן קיימת קורלציה חיובית בין השניים, ושככל שלומדים במשך יותר שנים, גדל הסיכוי להרוויח משכורת גבוהה.



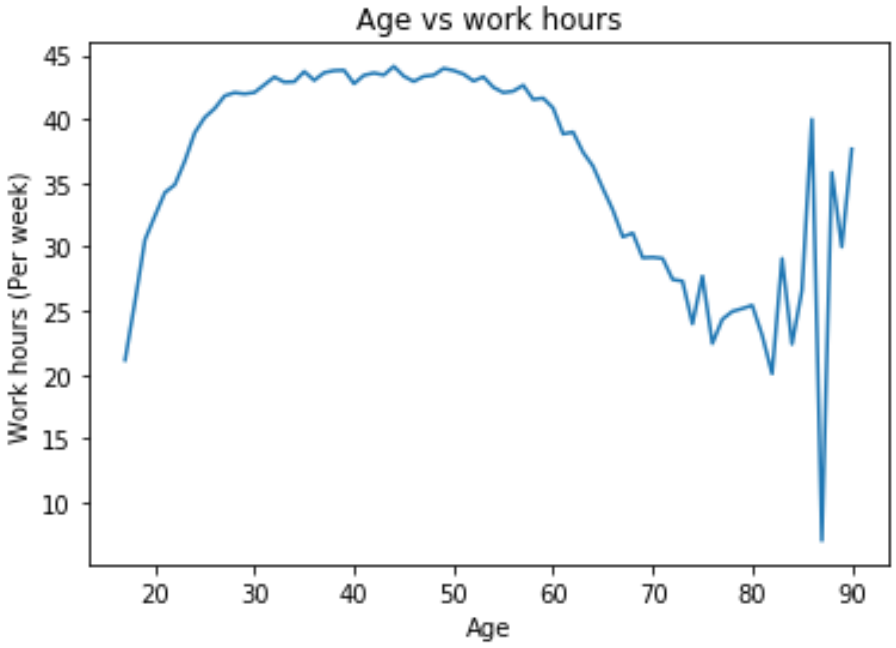
2.2.2 קורלציה בין מספר שנות ההשכלה לבין מספר שעות העבודה השבועיות

ניתן לראות החל משבע שנות השכלה ישנה קורלציה חיובית יחסית גבוהה בין מספר שנות ההשכלה ובין מספר שעות העבודה השבועיות. דבר זה מפריך את השערתינו בנושא, שהייתה שיש קורלציה **שלילית** בין השניים, ושככל שמספר שנות ההשכלה של בן אדם גבוה יותר כך הוא יהיה זקוק לפחות שעות עבודה.



2.2.3 קורלציה בין הגילאים לבין ההשכלה לבין מספר שעות העבודה השבועיות

ניתן לראות כי עבור הגילאים שבין עשרים לשלושים קיימת קורלציה חיובית בין הגיל לבין מספר שעות העבודה השבועיות, וכי עבור הגילאים שבין שישים לשמונים ישנה דווקא קורלציה שלילית. על כן, ניתן להסיק כי בסך הכל, אין קורלציה בין השניים.



פרק שלישי: העלאת השערות ובדיקתן

3.1 העלאת ההשערה

כפי שרשמנו כבר, אנו מתעניינים בשאלת האפליה המגדרית, ובגדר קיום אפליה נגד נשים, והאם הן מקבלות משכורת נמוכה יותר, רק מכיוון שהן נשים.

בעזרת ניתוח הנתונים גילינו שכ-30% מהגברים היו בעלי משכורת אשר גבוהה מחמישים אלף, ואילו רק כ-11% מהנשים היו בעלי משכורת אשר גבוהה מחמישים אלף

נתון זה הוביל להשערתנו שקיימת אפליה בין נשים לגברים.

3.2 ניסוח ההשערה

3.2.1 השערת האפס

השערת האפס שלנו הינה שאחוז הגברים שמקבלים משכורת הגבוהה מחמישים אלף (מתוך קבוצת הגברים) שווה לאחוז הנשים המקבלות משכורת הגבוהה מחמישים אלף (מתוך קבוצת הנשים).

3.2.2 השערת חלופית

השערת האפס שלנו הינה שאחוז הגברים שמקבלים משכורת הגבוהה מחמישים אלף (מתוך קבוצת הגברים) שונה מאחוז הנשים המקבלות משכורת הגבוהה מחמישים אלף (מתוך קבוצת הנשים).

3.2.3 סטטיסטי המבחן

סטטיסטי המבחן הוא ההפרש בין אחוז הגברים במדגם המקבלים משכורת הגבוהה מחמישים אלף (מתוך קבוצת הגברים) לבין אחוז הנשים המקבלות משכורת הגבוהה מחמישים אלף (מתוך קבוצת הנשים).

3.3 בדיקת ההשערה

3.3.1 בחירת שיטת הבדיקה

על מנת לבחון את ההשערה, השתמשנו בשיטת הבוטסטראפ. בחרנו את שיטה זאת מכיוון שלא יכלנו לערוך סימולציות בנוגע להשערה שהעלינו. ביצוע סימולציות בנושא זה אינו מתאפשר, כי אין ביכולתינו האפשרות לקבוע "אחוזים שווים" בין גברים ונשים. כלומר, לא נוכל להגדיר באופן מפורש ולסמלץ אחוזים שווים של גברים ונשים המקבלים משכורת גבוהה, וזאת כי אנו לא יודעים מהו אותו האחוז השווה שעלינו לסמלץ. על כן, בחרנו להשתמש בשיטת הבוטסטראפ.

3.3.2 ביצוע הבדיקה

על מנת לבחון את ההשערה חישבנו את התפלגות ההפרש בין אחוז הנשים אשר הכנסתן גבוהה מחמישים אלף לבין אחוז הגברים אשר הכנסתם גבוהה מחמישים אלף, בעזרת שיטת הבוטסטראפ.

בחרנו באקראיות בכל פעם אנשים מתוך המדגם, עם החזרה, כאשר מספר האנשים שבחרנו בכל פעם זהה לגודל המדגם. עבור כל דגימה, חישבנו את אחוז הגברים אשר הכנסתם גבוהה מחמישים אלף (מתוך קבוצת הגברים בדגימה), ואת אחוז הנשים אשר הכנסתן גבוהה מחמישים אלף (מתוך קבוצת הנשים בדגימה), ומצאנו את ההפרש בין האחוזים. את ההפרשים הכנסנו למערך שגודלו זהה למספר תהליכי הדגימות שביצענו.

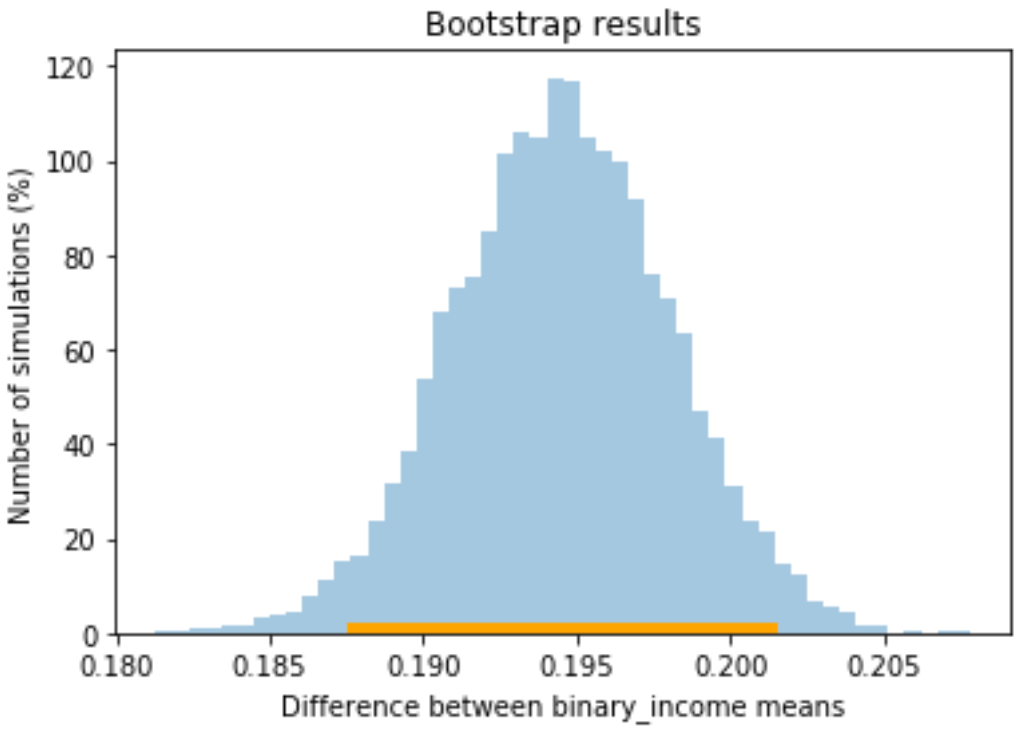
מספר החזרות שבחרנו לבצע את הדגימות הוא עשרת אלפים, מכיוון שזהו מספר הדגימות הסטנדרטי שעבדנו איתו עד כה בהרצאות והתרגולים.

לאחר מכן, יצרנו גרף המציג את התפלגות ההפרשים שחושבו, והוספנו לו רווח סמך של תשעים וחמישה אחוזים.

3.3.3 תוצאות הבדיקה ומסקנות

קיבלנו שרווח הסמך שלנו להפרש בין האחוזים נע בין 18.75 לבין 20.15. תוצאה זו רחוקה מהשערת האפס, שכן הפרש אפס (כלומר, שהאחוזים שווים) אינו נכלל הרווח הסמך.

לכן נוכל להסיק בביטחון של תשעים וחמישה אחוזים שאחוז הגברים אשר הכנסתם גבוהה מחמישים אלף גדול מאחוז הנשים שהכנסתן גבוה מחמישים אלף. כלומר, האפליה נגד נשים מובהקת סטטיסית.



3.4 הגבלות והטיות בבחינת ההשערה

כמו בכל ניתוח נתונים, ישנו הסיכוי אשר המידע שבידנו מוטה, מה שעלול לגרור לתוצאות שגויות ולמסקנות שגויות, אשר מבוססות על אותן תוצאות מוטות.

לדוגמה, יכולה להיות הטיית בחירה במדגם, שנובעת מכך שהוא לא מורכב מאנשים שנדגמו באופן אקראי, ומייצגים את האוכלוסייה. אנו חושדים בכך שמדגם זה אכן טומן בחובו הטיה שכזאת, שכן, מבדיקה שלנו באינטרנט, אחוז הגברים והנשים בארצות כמעט שווה, ואף יש יותר נשים מאשר גברים. עם זאת, ראינו שבמדגם שלנו, אחוז הגברים גדול כמעט פי שניים מאשר אחוז הנשים (המדגם מכיל 66.85% גברים, ואילו רק 33.15% נשים). הטיה זו משמעותית, מכיוון ששיטת הבוטסטראפ שביצענו מסתמכת באופן מפורש על ההנחה כי המדגם מייצג את האוכלוסיה. על כן, יכול להיות שאין באמת אפליה נגד נשים, אך ההטיה של המדגם גרמה להטיה בבוטסטראפ, ובכך היטתה את מסקנתנו.

דבר נוסף אשר עשוי לגרום למסקנתנו להיות שגויה הוא הצורך להוריד מן מסד הנתונים את העמודות: capital-gain, capital-loss, עקב מידע פגום אשר הן הכילו. במידה והן לא היו פגומות, יכול להיות שהיינו מוצאים להם שימוש בבדיקת ההשערה, שכן גם הן עוסקות ברווח של הנדגמים.

פרק רביעי: סיווג

4.1 בחירה בסיווג ובמשתנה המטרה

אנו בחרנו לסווג את האנשים במסד הנתונים, ולא לבצע ניתוח אשכולות. בחרנו כך מכיוון שהנתונים מכילים משתנה מטרה מאוד "ברור", שהוא האם ההכנסה של אדם גדולה מחמישים אלף, או לא. על כן, בחרנו לסווג את האנשים, ושההכנסה תהיה משתנה המטרה שלנו.

למעשה, משתנה המטרה שלנו הוא ה-binary_income, אשר יצרנו בתחילת העבודה מתוך משתנה ה-income.

4.2 ביצוע הסיווג

4.2.1 נקיון מסד הנתונים

הבחנו בכך שהיו כ-6000 רשומות אשר הכילו סימן שאלה בחלק מן העמודות הקטגוריאליות. התלבטנו בין האפשרויות הבאות:

- מחיקת כל עמודה בה יש לפחות ערך פגום אחד
- התייחסות לסימן השאלה כקטגוריה בפני עצמה
- מחיקת הרשומות בהן יש ערך פגום באחת מן העמודות

לא רצינו למחוק עמודות שלמות, מכיוון שכל העמודות הפגומות הכילו ברובן ערכים תקינים. כלומר, לא היו עמודות שרוב הנתונים בהם חסרים, ולכן נרתענו מלמחוק מידע של עמודות שלמות, בנוסף לכך, הורדת העמודות הייתה יכולה לפגוע באחוז הדיוק של המסווג, שכן יכול להיות שדווקא שלושת עמודות אלה תרמו למסווג בצורה הטובה ביותר.

בנוסף, לא רצינו להתייחס לסימני השאלה כקטגוריה בפני עצמה, כי התייחסות שכזאת גוררת דמיון בין אנשים שונים אשר המידע עליהם חסר. כלומר, אנו משייכים את האנשים הללו לאותה קטגוריה (קטגוריית "מידע חסר"), ודבר זה גורר דמיון של אנשים אשר לאו דווקא דומים, ורק איחדנו אותם תחת אותה קטגוריה "פיקטיבית".

לבסוף, החלטנו להסיר רק את הרשומות הפגומות. אנו סבורים כי להחלטה זו יהיו פחות חסרונות מאשר שאר האפשרויות, ולכן בחרנו בה.

4.2.2 קידוד משתנים קטגוריאליים

הבחנו בכך שקיימים משתנים קטגוריאליים בין כלל הנתונים, ושכולם הם נומינליים. על כן, קידדנו אותם בעזרת הפקודה cat.codes, על מנת שנוכל להשתמש בהם בתהליך הסיווג.

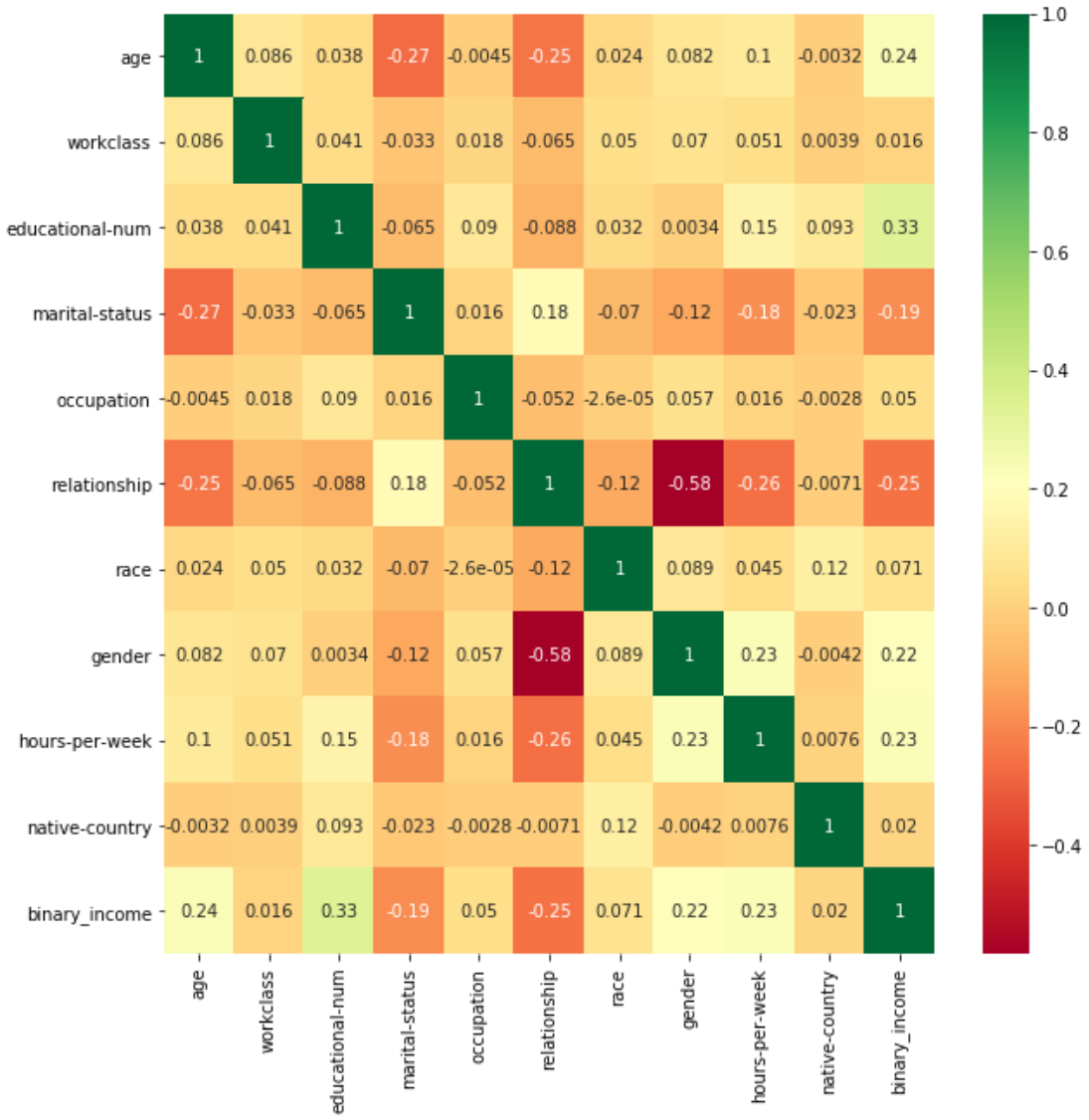
4.2.3 נרמול הנתונים

אלגוריתם knn רגיש למרחקים, ולכן היינו צריכים לדאוג לכך שקנה המידה של כל המשתנה יהיה זהה על כן, נרמלנו את הנתונים בעזרת אלגוריתם MinMax, על מנת שלא לתת לחלק מהם חשיבות גדולה יותר בחישוב המרחקים.

בחרנו באלגוריתם זה מכיוון שזהו האלגוריתם הסטנדרטי בו נתקלנו בתרגולים ואיתו עבדנו בתרגילי הבית.

4.2.4 מפת חום - Heatmap

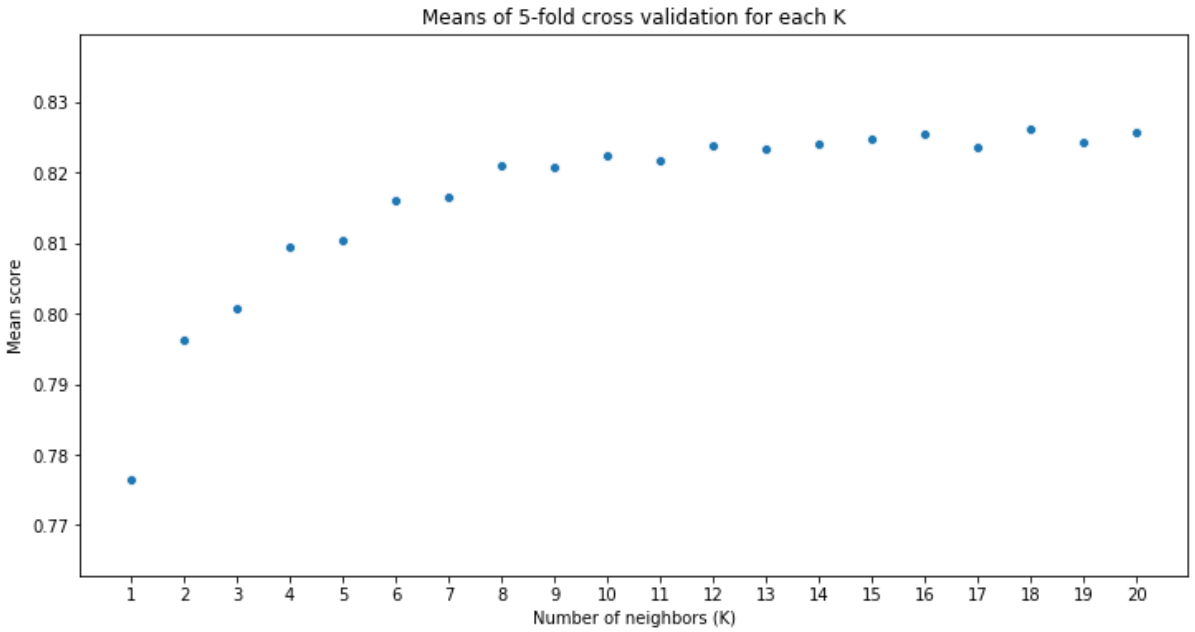
לאחר הנרמול, בדקנו את הקורלציה של כלל המשתנים אחד עם השני, ויצרנו מפת חום. הבחנו בה שאין קורלציה גבוהה בין אף משתנה למשתנה המטרה, ובכללי בין אף משתנה למשתנה אחר. על כן, החלטנו לסווג בעזרת כל המשתנים שיש לנו, כי אמנם כל אחד מהם בפני עצמו לא נמצא בקורלציה גבוהה עם משתנה המטרה, אך יכול להיות ששילוב ביניהם יוצר קורלציה גבוהה עמו. לכן,על מנת לא לפגוע בקשר כזה, אם קיים, החלטנו להשתמש בכלל המשתנים.



4.2.5 אימות צולב - Cross Validation

ביצענו תהליך של אימות צולב, על מנת למצוא את מספר השכנים האופטימלי לסיווג. חילקנו את הנתונים לסט אימון וסט בדיקה, ביחס של 20-80, והרצנו בדיקות למספרי שכנים אפשריים, החל מאחד ועד עשרים, כאשר כל בדיקה כללה בתוכה חמש fold-ים (המספר הסטנדרטי בו נתקלנו בתרגולים ובהרצאות, ואיתו עבדנו בתרגילי הבית).

בשלב זה, מצאנו כי מספר השכנים האופטימלי הינו 18. על כן, השתמשנו במספר זה בסיווג שערכנו על סט הבדיקה.



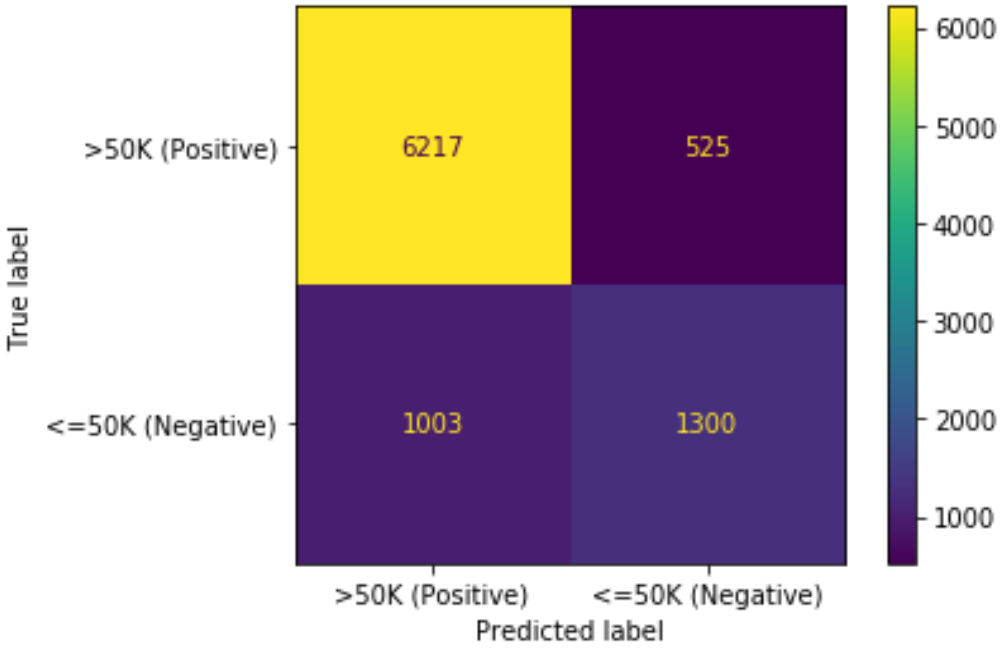
4.3 תוצאות הסיווג

4.3.1 אחוזי הצלחה

מצאנו כי אחוזי ההצלחה של המסווג שלנו עומדים על 83.11%.

4.3.2 מטריצת בלבול - Confusion matrix

השתמשנו בפקודה plot_confusion_matrix על מנת ליצור ולהציג את מטריצת הבלבול של המסווג שלנו.



4.3.3 ציון F1

מצאנו כי ציון ה-F1 של המסווג שלנו הינו 0.63.

4.4 מסקנות הסיווג

לאור תוצאות ההצלחה ומטריצת הבלבול, אנו סבורים כי המסווג שלנו הוא טוב בסך הכל, אם כי מספר הטעויות שעשה אינו קטן.

4.5 הגבלות בסיווג

גם בסיווג היו לנו הגבלות. הן נבעו בעיקר מהמידע שהיה חסר, אשר הוביל אותנו למחיקה של רשומות מן הנתונים. במידה ומידע זה לא היה חסר, אז היו בידינו עוד אלפי רשומות, בעזרתן יכלנו לאמן את המסווג בצורה יותר טובה, דבר שיכל להוביל לדיוק גבוה יותר בסיווג עצמו.

פרק חמישי: מבט לעתיד

5.1 שאלות שנותרו ללא מענה

לאחר בחינת ההשערות והסקת מסקנות, הבחנו כי יתכן והנתונים מוטים, עקב אחוז הגברים הגבוה במדגם. על כן, כפי שרשמנו, ייתכן שמסקנתנו בדבר האפליה שגויה, ואין כלל אפליה, או לפחות לא כזאת שמובהקת סטטיסטית. לפיכך, שאלה זו נותרה פתוחה בפנינו, כי אנו לא בטוחים שהמסקנה שלנו בנושא נכונה, וכי אכן יש אפליה.

בנוסף, ההטיה של הנתונים משאירה בפנינו שאלה נוספת פתוחה, והיא השאלה בגין קורלציה בין מספר שנות ההשכלה ובין מספר שעות העבודה השבועיות. כפי שרשמנו, חשבנו כי תהיה קורלציה שלישית בין השניים (למשל, כמו בחברות הייטק, שאנשים שלמדו הרבה, עובדים כיום מעט), והנתונים הפריכו את השערה זו. גם הפרכה זו עשויה להתערער עקב הטיה בנתונים, ולכן, גם היא נותרה פתוחה.

5.2 שאלות שעלו במהלך המחקר

במהלך המחקר וניתוח הנתונים, עלו בפנינו מספר שאלות, אשר לא יכלנו לענות עליהן רק מהנתונים אשר יש לנו.

למשל, שאלה אחת שעלתה לנו, היא מה ממוצע ההכנסות הכולל של גברים ונשים ביחד? התעניינו בשאלה זו מכיוון שכל החלק של ההשערה עסק סביב ערך שהוא מספרי (גודל ההכנסה), אך כל מה שידענו עליו זה האם הוא גדול מחמישים אלף או לא. כלומר, העיסוק סביב נושא זה הוביל אותנו לתהיה על ערכי ההכנסות.

שאלה זו הובילה אותנו לשאלה נוספת, שהיא, מה היה ממוצע ההכנסות או חציון ההכנסות של נשים וגברים בנפרד? התשובה לשאלה זו הייתה יכולה לספק לנו דרך נוספת לבדוק את השאלה העיקרית שלנו, בדבר קיום אפליה נגד נשים.

שאלה נוספת שעלתה במהלך הכנת הפרויקט היא מה היה קורה אילו היינו בוחרים בקובץ נתונים אחר? שאלה זאת מעניינת אותנו כי כל הפרויקט עוסק סביב הקובץ שבחרנו, ואנו סקרנים לדעת מה היה קורה אם היינו בוחרים בקובץ אחר, כי זה היה משנה לנו רבות את הליך העבודה והניתוח.