דו"ח – אומר זמיר 208552620 ואביחי מרדכי 318341278  
  
**תיאור הנתונים**

הקובץ שבחרנו לעבודה זו הוא מערך הנתונים "Census Income", הידוע גם בשם "Adult Dataset" מערך נתונים זה משמש לעתים קרובות במשימות של סיווג בינארי, וכולל נתונים שמקורם במפקד האוכלוסין האמריקאי. מטרת מערך הנתונים היא לחזות האם הכנסתו השנתית של אדם עולה על 50,000 דולר, בהתבסס על מספר מאפיינים אישיים ודמוגרפיים. הנתונים כוללים 15 עמודות, וכל שורה מייצגת אדם שונה.

**פירוט המידע**

1. **גיל**: משתנה מספרי רציף המציין את גילו של האדם בשנים.
2. **סוג עבודה (Workclass)**: משתנה קטגורי המציין את סוג התעסוקה של האדם או סוג המעסיק. לדוגמה, תעסוקה במגזר הפרטי, ממשלתי, או עבודה עצמאית.
3. **משקל סופי (Fnlwgt)**: משתנה מספרי רציף המייצג את משקלו הסופי של האדם במפקד, כלומר מספר האנשים שהאדם מייצג באוכלוסייה הכללית.
4. **השכלה**: משתנה קטגורי המתאר את רמת ההשכלה הגבוהה ביותר שהושגה על ידי האדם, כגון תואר ראשון, תעודת בגרות, או השכלה חלקית.
5. **מספר שנות השכלה (Education-num)**: משתנה מספרי המייצג את מספר השנים שהאדם למד, בהתאמה לרמת ההשכלה שצוינה.
6. **מצב משפחתי**: משתנה קטגורי המתאר את המצב המשפחתי של האדם, כגון רווק, נשוי, גרוש, או פרוד.
7. **מקצוע (Occupation)**: משתנה קטגורי המתאר את סוג המקצוע שבו עוסק האדם, כגון עבודה משרדתית, ניהולית, או מקצועות ידניים.
8. **יחס משפחתי (Relationship)**: משתנה קטגורי המתאר את הקשר המשפחתי של האדם ביחס לשאר בני הבית, כגון בן זוג, ילד, או קרוב משפחה אחר.
9. **גזע (Race)**: משתנה קטגורי המציין את הגזע של האדם, כגון לבן, שחור, אסייתי, או אמריקאי ילידי.
10. **מין (Sex)**: משתנה קטגורי המציין את מינו של האדם, זכר או נקבה.
11. **רווחי הון (Capital-gain)**: משתנה מספרי רציף המייצג רווחים פיננסיים שהושגו ממכירת נכסים או השקעות.
12. **הפסדי הון (Capital-loss)**: משתנה מספרי רציף המייצג הפסדים פיננסיים שהושגו ממכירת נכסים או השקעות.
13. **שעות עבודה בשבוע (Hours-per-week)**: משתנה מספרי רציף המייצג את מספר שעות העבודה שהאדם עובד בשבוע.
14. **מדינת מוצא (Native-country)**: משתנה קטגורי המציין את מדינת המוצא או המגורים של האדם, כגון ארצות הברית, מקסיקו, קובה, או מדינה אחרת.
15. **הכנסה**: משתנה קטגורי המייצג את רמת ההכנסה של האדם, והוא מחולק לשתי קטגוריות: הכנסה נמוכה או שווה ל-50,000 דולר בשנה, או הכנסה גבוהה מ-50,000 דולר בשנה.

**תיאור הבעיה**

בעבודה זו, אנו עוסקים במשימה של חיזוי הכנסתו של אדם ממוצע בארה"ב, האם היא עולה על 50,000 דולר לשנה או לא. משימה זו היא בעיית סיווג בינארית, שבה אנו מנסים לנבא ערך משתנה תלוי (הכנסה) בהתבסס על מספר משתנים (כגון גיל, השכלה, מקצוע ועוד). מערך הנתונים שנבחר למשימה מבוסס על נתוני מפקד האוכלוסין האמריקאי ומספק מגוון רחב של מאפיינים דמוגרפיים ותעסוקתיים של כל נבדק.

כדי לפתור את הבעיה, אנו משתמשים באלגוריתם רגרסיה לוגיסטית (Logistic Regression), שהוא מודל סטטיסטי שנועד להעריך את ההסתברות שמשתנה בינארי יקבל ערך מסוים. במקרה שלנו, המטרה היא לחזות אם ההכנסה של אדם תהיה מעל או מתחת ל-50,000 דולר בשנה. רגרסיה לוגיסטית מאפשרת לנו לבצע את הסיווג הזה בצורה יחסית פשוטה, תוך שהיא מתחשבת בקשר הלינארי בין המאפיינים לבין ההסתברות לתוצאה.

תהליך פתרון הבעיה מתחיל בהכנסת הנתונים והכנתם לצורך עיבוד. מאחר והנתונים המקוריים כוללים משתנים מסוג מחרוזות (כגון מצב משפחתי, מקצוע, ועוד), יש צורך להמיר משתנים אלו לייצוג מספרי באמצעות קידוד (Label Encoding), כדי שהמודל יוכל לעבוד עליהם. כמו כן, הנתונים מנורמלים כדי להבטיח שהערכים המספריים יהיו בקנה מידה אחיד, דבר שחשוב במיוחד ברגרסיה לוגיסטית.

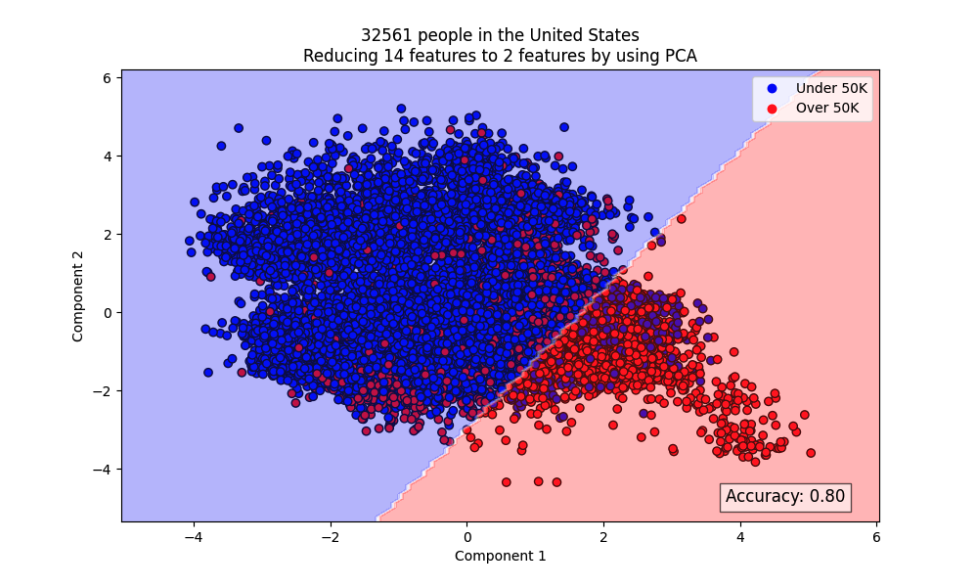
בכדי לבדוק את ביצועי המודל, אנו משתמשים בשתי גישות עיקריות:

1. **קרוס-וולידציה (Cross-Validation)**: שיטה זו מחלקת את מערך הנתונים למספר חלקים (folds), ובכל שלב המודל מתאמן על רוב החלקים ומנבא את החלק הנותר. התוצאה היא ממוצע של הביצועים בכל החלקים, מה שמבטיח שהמודל אינו תלוי בחלוקה ספציפית של הנתונים.
2. **חלוקה אקראית (Random Split)**: שיטה נוספת לבדיקת ביצועי המודל היא חלוקה אקראית של הנתונים ל-50% לאימון ו-50% לבדיקת המודל. גישה זו מאפשרת לבדוק את המודל במספר חלוקות אקראיות של הנתונים ולקבל ממוצע של הדיוק.

בנוסף לרגרסיה הלוגיסטית הבסיסית, בעבודה זו אנו מיישמים גם טכניקה מתקדמת יותר של AdaBoost בשילוב עם רגרסיה לוגיסטית. בעזרת AdaBoost אנו משפרים את יכולות הסיווג של המודל על ידי מתן משקל שונה לדגימות לפי מידת הקושי שלהן בסיווג נכון. כל שלב של אימון לוגיסטי מתבצע עם משקלים מתעדכנים לדגימות, ובסוף התהליך, מתקבל סיווג סופי המשלב את כל המודלים שנוצרו במהלך האימון.

המדד העיקרי שבו אנו משתמשים להערכת הביצועים של המודלים הוא דיוק (Accuracy), שהוא היחס בין מספר הדגימות שסווגו נכון למספר הכולל של הדגימות.

**תוצאות**

במסגרת העבודה, ביצענו ניתוחים סטטיסטיים באמצעות רגרסיה לוגיסטית בשלוש גישות שונות: קרוס-וולידציה, חלוקה אקראית (Random Split), ו-Adaboost. כל אחת מהגישות מאפשרת להעריך את ביצועי המודל בדרכים שונות ולספק מדדים חשובים בנוגע לדיוק הסיווג. להלן פירוט התוצאות שהתקבלו מכל אחת מהגישות.  
  
ראשית לפני תחילת העבודה הורדנו את הדאטה ל2 ממדים עם PCA והפעלנו LR על הדאטה בשביל לראות איך מתמודד האלגוריתם עם הדאטה ומה המפריד הלינארי שניתן לקבל מLR, לכן שרטטנו את הגרף הבא:  


**1. קרוס-וולידציה (Cross-Validation)**

קרוס-וולידציה היא טכניקה הנפוצה להערכת מודלים, במיוחד במערכי נתונים קטנים או כאשר רוצים למנוע תלות יתר בקבוצת אימון אחת. השתמשנו בקרוס-וולידציה עם 5 קיפולים (folds), כלומר, חילקנו את הנתונים ל-5 חלקים שווים ובכל פעם השתמשנו ב-4 חלקים לאימון ובחלק אחד לבדיקה. התהליך חזר על עצמו 5 פעמים כך שכל חלק שימש פעם אחת לבדיקה. תוצאות הדיוק שהתקבלו בכל אחד מהקיפולים הן:

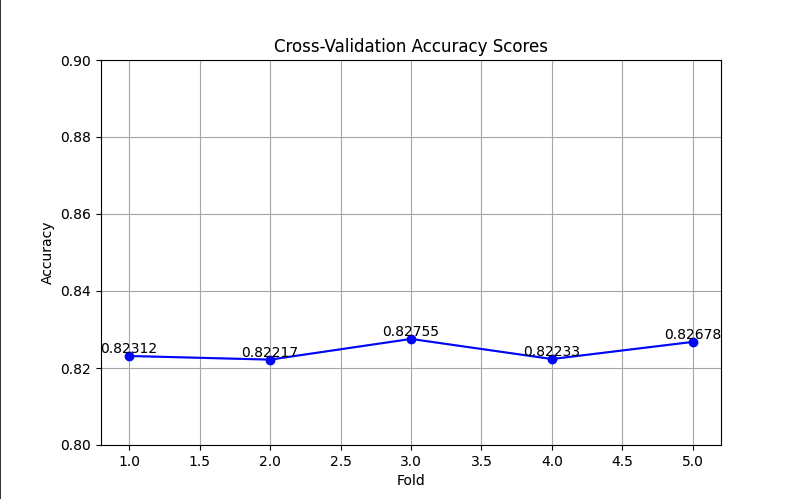
- קיפול 1: 82.31%

- קיפול 2: 82.22%

- קיפול 3: 82.75%

- קיפול 4: 82.23%

- קיפול 5: 82.68%

הממוצע של תוצאות הקרוס-וולידציה הוא 82.44%, מה שמספק לנו אינדיקציה טובה לביצועי המודל על פני כל מערך הנתונים, ללא תלות בחלוקה מסוימת. 

**2. חלוקה אקראית (Random Split)**

בגישת החלוקה האקראית, חילקנו את מערך הנתונים ל-50% נתוני אימון ו-50% נתוני בדיקה, כשהחלוקה בוצעה באופן אקראי בכל פעם. חזרנו על התהליך 5 פעמים כדי לקבל ממוצע דיוק מהימן עבור המודל. התוצאות שהתקבלו בכל אחת מהחלוקות הן:

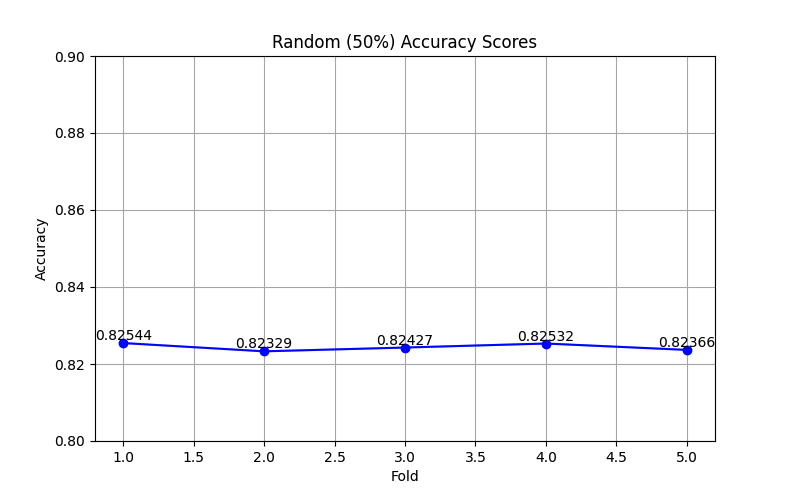
- חלוקה 1: 82.54%

- חלוקה 2: 82.33%

- חלוקה 3: 82.43%

- חלוקה 4: 82.53%

- חלוקה 5: 82.37%

הממוצע של תוצאות החלוקה האקראית הוא 82.44%, בדומה לממוצע הקרוס-וולידציה, מה שמעיד על יציבות המודל בחלוקות שונות של הנתונים.

**3. Adaboost עם רגרסיה לוגיסטית**

לבסוף, יישמנו טכניקת Adaboost בשילוב עם רגרסיה לוגיסטית כדי לשפר את ביצועי הסיווג. Adaboost היא שיטה שמבצעת אימון על מספר מודלים ומשקללת את תוצאותיהם כך שכל מודל מתמקד יותר בדגימות שבהן נכשלו המודלים הקודמים. במקרה זה, השתמשנו ב-5 מודלים של רגרסיה לוגיסטית, והתוצאה הסופית של הדיוק שהתקבלה היא 81.25%.

תוצאה זו נמוכה מעט מהתוצאות שהתקבלו בקרוס-וולידציה ובחלוקה האקראית, אך עדיין מספקת ביצועים טובים בהתחשב בכך ש-Adaboost מתמקד במשקול דגימות קשות יותר לסיווג.  
  
  
**מסקנות:**  
  
**ביצועי המודל**

1. **קרוס-וולידציה**: התוצאות מהקרוס-וולידציה הראו דיוק ממוצע של 82.44%. גישה זו מדגימה יציבות גבוהה בביצועים, מה שמעיד על כך שהמודל מצליח לייצר תוצאות עקביות גם כאשר הנתונים מחולקים לכמה חלקים שונים. התוצאה מצביעה על כך שהמודל מצליח להתמודד היטב עם הנתונים המגוונים, והדיוק המרבי שהוא מצליח להשיג הוא יציב יחסית בכל חלוקה.
2. **חלוקה אקראית**: גם כשביצענו חלוקה אקראית של הנתונים ל-50% אימון ו-50% בדיקה, הדיוק הממוצע היה 82.44%. תוצאה זו תומכת במסקנה שהמודל מצליח לספק תוצאות טובות גם בגישה זו, מה שמעיד על יציבות טובה בביצועים גם במצבים פחות מבוקרים.
3. **Adaboost עם רגרסיה לוגיסטית**: השימוש בAdaboost יחד עם רגרסיה לוגיסטית הניב דיוק של 81.25%. ירידה קלה בביצועים בהשוואה לשתי השיטות הקודמות מעידה על כך ש Adaboost למרות יתרונותיו בתיקון דגימות קשות לסיווג, לא תמיד מביא לשיפור משמעותי בתוצאות הכלליות. עם זאת, טכניקת Adaboost הוכיחה את יכולתה לשפר את ההתמודדות עם דגימות קשות על ידי מתן משקל גבוה יותר לדגימות שמסובכות לסיווג.

**סיכום**

בבחינה כללית, התוצאות מעידות על כך שהמודל של רגרסיה לוגיסטית פועל בצורה עקבית וטובה בשיטות הקרוס-וולידציה והחלוקה האקראית. השימוש ב-Adaboost מציע יתרון בתיקון דגימות קשות, אך עם ירידה קלה בדיוק הממוצע. המסקנה היא ש LRבשילוב עם גישות כמו קרוס-וולידציה, מספקת פתרון יציב לסיווג ההכנסה, בעוד ש-Adaboost מספק ערך מוסף במצבים של דגימות קשות, אם כי לא תמיד משפר את התוצאות הכלליות באופן משמעותי.