**עבודה סיום לקורס למידה חישובית**

* **תיאור הבעיה והדאטה:**
* הבעיה היא זיהוי הונאת כרטיסי אשראי באמצעות תהליך של למידת מכונה. הנתונים מכילים מאגר נתונים של עסקאות כרטיסי אשראי. המסגרת של הנתונים היא קובץ בשם "creditcard.csv" שכולל מידע על פעולות כרטיס אשראי ומזיקות. הקובץ מכיל 31 עמודות הכוללות את המאפיינים הבאים:

1. Time - זמן השורה כמספר שניות מאז הפעולה הראשונה בקובץ.

2. V1, V2, ..., V28 - מאפייני הסודיות של הכרטיס אשראי

3. Amount - סכום העסקה.

4. Class - מציין האם העסקה היא תקינה או פוגעת. ( הערכים האפשריים הם 0 או 1, כאשר 0 מייצג עסקה רגילה ו-1 מייצג עסקה שהיא הונאת כרטיסי אשראי.)

המטרה היא לבנות מודלי חיזוי המסוגל לזהות עסקאות פוגעות באמצעות אלגוריתמי Logistic Regression ו-Adaboost, ולבצע השוואה ביניהם.

* לניתוח ובניית מודלים נעשית על בסיס כמה מדדים שהם

1. יישום בעולם האמיתי: קבוצת הנתונים מייצגת בעיה בעולם האמיתי עם הישגים פרקטיים ישירים.

2. גודל הנתונים: קבוצת הנתונים נגודרת בגודל גדול יחסית.

3. זמינות: קבוצת הנתונים זמינה לציבור בקאגל, ובכך ניתן לגשת אליה בקלות לצורכי למידה ומחקר.

כתוצאה מהפקטורים שצוינו לעיל, קבוצת הנתונים "זיהוי הונאת כרטיסי אשראי" מהווה משאב יקר ללימוד ופיתוח של שיטות יעילות בתחום מדעי הנתונים ולמידה חישובית.

**הקדמה**:

במחקר זה, אנו שואפים להשוות את הביצועים של שני מודלים של למידת מכונה, כלומר **Logistic Regression ו-Adaboost,** עבור משימת הסיווג הבינארי של זיהוי הונאה בכרטיסי אשראי.

**. תיאורי דגם:**

**רגרסיה לוגיסטית** היא אלגוריתם סיווג ליניארי בשימוש נרחב המאומד את ההסתברות לקלט השייך למחלקה מסוימת. הוא יוצר מודלים של הקשר בין תכונות הקלט לבין הסיכויים ביומן של מחלקת היעד, ומאפשר החלטות סיווג בינארי המבוססות על סף.

**Adaboost**, היא טכניקת למידה אנסמבלית המשלבת לומדים חלשים (במקרה זה, עצי החלטה בעלי עומק מוגבל) ליצירת מסווג חזק. בכל איטרציה, Adaboost מקצה משקלים גבוהים יותר לדגימות שסווגו בצורה שגויה ומתאימה את המודל להתמקדות במופעים שקשה יותר לסווג אותם.

**. הגדרה נסיונית:**

עבור מחקר זה, חילקנו את מערך הנתונים ל-80% נתוני אימון ו-20% נתוני בדיקה. עץ ההחלטות בשימוש ב-Adaboost הוגבל לעומק מרבי של 1, מה שהופך אותו ללומד חלש. הגדרנו את מספר האיטרציות של Adaboost (num\_estimators) ל-50 כדי להשיג איזון בין ביצועי המודל וזמן החישוב.

**תוצאות ומסקנות**

**. השוואת ביצועי המודלים:**

עשינו שוואת ביצועים בין מודל אדבוסט לביצועי המודל הלוגיסטי הביצועים של שני המודלים הוערכו באמצעות ארבעה מדדי מפתח: דיוק, recall, Precisionוציון F1. מדדים אלו מספקים מבט מקיף על יעילותו של כל מודל באיתור עסקאות הונאה והימנעות מתוצאות שווא.

1. מודל הלוגיסטי מציג ביצועים טובים ברמה גבוהה:

- (Accuracy) מוערך כ-99.89%, כלומר, המודל מצליח לנבא כמעט ב-100% את התוויות הנכונות של המודל.

- (Precision) בגובה של כ-82.76%, כלומר, המודל קובע באמינות בגבול 82.76% מהעסקאות שהוא מסווה כחשודות בהונאה כמו עסקאות בהונאה.

- (Recall) נמוך יחסית, בגובה של כ-48.98%. זה אומר שהמודל לא מזהה באמינות גבוהה את כל העסקאות הממש מחוץ למקטע הספציפי שהוא זיהה כחשודות.

- F1 Score) יחסית נמוך בגובה של כ-61.54%, מה שמראה על יכולת נמוכה לבצע את שני המשימות באופן יחד: זיהוי מדויק של עסקאות בהונאה והמערכתית מחזרה ברמה נמוכה של המודל.

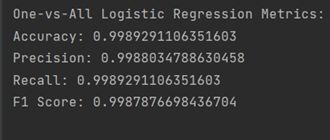
2. מודל ה-Adaboost מציג ביצועים מעולים:

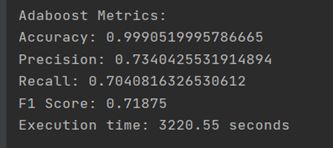
- (Accuracy) מוערך כ-99.91%, כלומר המודל מצליח לנבא כמעט ב-100% את התוויות הנכונות של המודל.

- (Precision) בגובה של כ-73.40%, כלומר, המודל קובע באמינות בגבול 73.40% מהעסקאות שהוא מסווה כחשודות בהונאה כמו עסקאות בהונאה.

- (Recall) יחסית גבוה בגובה של כ-70.41%, זה אומר שהמודל מזהה באמינות גבוהה רבה יותר מעסקאות חשודות בהונאה בהשוואה למודל הלוגיסטי.

- F1 Score) יחסית גבוה בגובה של כ-71.88%, זה מראה על יכולת טובה של המודל לבצע את שני המשימות באופן יחד: זיהוי מדויק של עסקאות בהונאה והמערכתית מחזרה ברמה טובה יותר מהמודל הלוגיסטי.

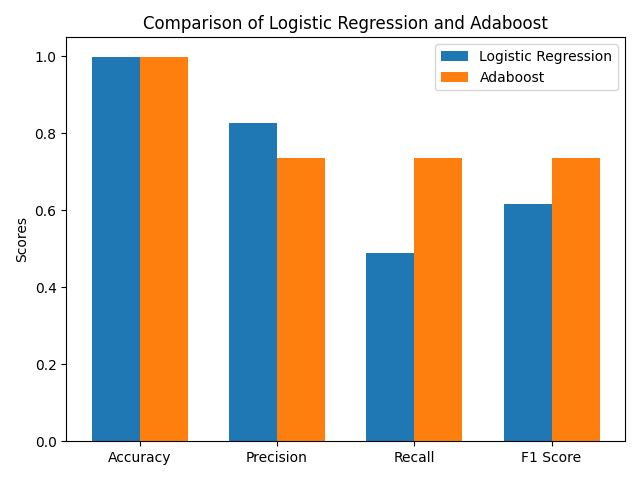




**. זמן ביצוע:**

- זמן הביצוע של התסריט כולו מחושב ושווה ל 3220.55 שניות

הערה: זמן הביצוע די גבוה (3220.55 שניות) בשל הגודל הגדול של מערך הנתונים והמורכבות של אלגוריתם Adaboost.



. **דיון**:

התוצאות מצביעות על כך שגם רגרסיה לוגיסטית וגם Adaboost השיגו דיוק גבוה בזיהוי הונאה בכרטיסי אשראי. עם זאת, Adaboost הפגינה recall מעולה, מה שמרמז על יכולת טובה יותר לזהות עסקאות הונאה בפועל. מצד שני, רגרסיה לוגיסטית השיגה דיוק גבוה יותר (ערך חיזוי חיובי), מה שמעיד על שיעור חיובי כוזב נמוך יותר.

ניתן לייחס את ההצלחה של Adaboost ליכולת שלה להתמקד בדגימות שסווגות בצורה שגויה בכל איטרציה, ולמעשה להתאים את המודל להתמודדות עם מקרים מאתגרים. לעומת זאת, גבול ההחלטה הליניארי של רגרסיה לוגיסטית עשוי להתקשה להפריד בין אזורי נתונים מורכבים וחופפים.

. **מסקנה**:

בהתבסס על הניתוח שלנו, הגענו למסקנה ש-Adaboost השיגה ביצועים טובים יותר על רגרסיה לוגיסטית עבור זיהוי הונאה בכרטיסי אשראי. הריקול הגבוה יותר שהשיגה Adaboost הוא בעל ערך במיוחד באיתור מקרים נוספים של הונאה, ומפחית את הסיכון לפעילויות הונאה שלא אותרו.

. **מגבלות**:

למרות התוצאות המבטיחות, ישנן מגבלות שיש לקחת בחשבון. מערך הנתונים עלול לסבול מחוסר איזון בכיתה, שעלול להשפיע על ביצועי המודל. בנוסף, הרגישות של Adaboost לחריגות עלולה להוביל להתאמת יתר על נתונים רועשים. מחקר עתידי צריך לחקור טכניקות לטיפול במגבלות אלו ולשפר עוד יותר את ביצועי המודלים.

. **סיכום**:

מחקר זה מדגיש את החשיבות של בחירת מודלי סיווג מתאימים למשימות גילוי הונאה. גם רגרסיה לוגיסטית וגם Adaboost הפגינו ביצועים חזקים, אבל היכולת של Adaboost להתמקד במקרים מאתגרים מספקת יתרון תחרותי. הבנת החוזקות והמגבלות של כל מודל חיונית ביישומים בעולם האמיתי כדי לקבל החלטות מושכלות למניעת הונאה בכרטיסי אשראי.