דוח פרויקט - חיזוי נטישת לקוחות בבנק

1. תיאור הפרויקט ומטרתו

הפרויקט עוסק בזיהוי נטישת לקוחות בבנק באמצעות כלי למידת מכונה. מטרת הפרויקט היא לנתח את מאגר הנתונים המכיל מידע פיננסי ודמוגרפי של לקוחות הבנק, ולחזות האם לקוח ינטוש את שירותי הבנק בעתיד. במסגרת המחקר נבחן ביצועם של מספר אלגוריתמים – Decision Tree, SVM, AdaBoost ו-k-NN – תוך התמקדות בהשוואת ביצועי המודלים על בסיס מדדים כגון דיוק (Accuracy), Precision, Recall ו-F1-Score. בנוסף, נבדוק השפעת פעולות הנדסת תכונות (כגון הורדת מימד, נרמול, ויצירת מאפיינים חדשים) על ביצועי המודלים.

קישור לגיט המכיל את הקוד ואת מאגר הנתונים -
https://github.com/MichaelM1997/Machine_Lerning_Project.git

2. תאור המאגר

המאגר מכיל 10,000 רשומות של לקוחות הבנק וכולל את העמודות הבאות:

- RowNumber: מספר סידורי של הרשומה.
- CustomerId: מזהה לקוח.
- Surname: שם המשפחה (עמודה זו הוסרה בשלב העיבוד מאחר ואינה תורמת לחיזוי).
- CreditScore: דירוג אשראי.
- Geography: מדינת מגורים.
- Gender: מין הלקוח.
- Age: גיל.
- Tenure: משך התקשרות עם הבנק (בשנים).
- Balance: יתרת החשבון.
- NumOfProducts: מספר המוצרים הפיננסיים שבבעלות הלקוח.
- HasCrCard: האם הלקוח מחזיק בכרטיס אשראי.
- IsActiveMember: האם הלקוח פעיל.
- EstimatedSalary: משכורת חודשית מוערכת.
- Exited: משתנה המטרה, המציין האם הלקוח עזב את הבנק (1 - עזב, 0 - נשאר).

במהלך עיבוד הנתונים בוצע ניקוי נתונים הכולל טיפול בערכים חסרים, הסרת עמודות לא רלוונטיות (כגון RowNumber, CustomerId, Surname) והמרת משתנים קטגוריים באמצעות קידוד One-Hot.

3. תאור המודלים

עץ החלטה (Decision Tree)

- **עיקרון הפעולה:**
  בניית מבנה עץ המחליף את הנתונים לשאלות בינאריות עד לקבלת החלטה סופית.
- **יתרונות:**
  קל להבנה, מספק גרף חשיבות תכונות המאפשר זיהוי משתנים מרכזיים.
- **חסרונות:**
  נוטה לאוברפיטינג אם אין הגבלות על עומק העץ.

SVM

- **עיקרון הפעולה:**
  מציאת היפר-מישור אופטימלי שמפריד בין שתי הקבוצות, תוך שימוש בליבתיות במקרים לא לינאריים.
- **יתרונות:**
  דיוק גבוה במקרים עם ממדיות גבוהה, מתאים לבעיות עם גבולות החלטה מורכבים.
- **חסרונות:**
  עלול להיות איטי בריצה וקושי בפרשנות התוצאות לעומת מודלים אחרים.

AdaBoost

- **עיקרון הפעולה:**
  שילוב של מספר מודלים חלשים (למשל עצי החלטה קטנים) ליצירת מודל חזק יותר באמצעות מתן משקל גבוה יותר לדוגמאות קשות.
- **יתרונות:**
  משפר ביצועים על מודלים בסיסיים ומפחית טעויות.
- **חסרונות:**
  רגיש לרעש בנתונים ועלול להוביל לאוברפיטינג במקרים מסוימים.

(k-NN (k-Nearest Neighbors

- **עיקרון הפעולה:**
  סיווג בהתבסס על קרבת הדוגמאות במרחב התכונות – סיווג המבוסס על רוב שכנים קרובים.
- **יתרונות:**
  פשוט ליישום, אינו מניח מודלים סטטיסטיים מוקדמים.
- **חסרונות:**
  רגיש לנורמליזציה של הנתונים ועלול להיות איטי בריצה בנתונים גדולים.

4. שאלות הפרויקט

הפרויקט מתמקד במענה על השאלות הבאות:

1. **מהם המשתנים המשפיעים ביותר על נטישת לקוחות בבנק?**
לדוגמה: האם גיל, יתרת החשבון, משך התקשרות והכנסה חודשית הם גורמים מרכזיים בנטישת לקוחות.
2. **איזה אלגוריתם מספק את הדיוק הגבוה ביותר בזיהוי לקוחות שעומדים לנטוש את שירותי הבנק?**
השוואת ביצועי המודלים על בסיס Accuracy ומדדים נוספים.
3. **כיצד פעולות הנדסת תכונות (כגון הורדת מימד, נרמול, יצירת מאפיינים חדשים) משפיעות על דיוק המודלים?**
ניתוח ההשפעה של עיבוד הנתונים על איכות ההחלטות של המודלים.

5. מימוש הקוד ויישום המודלים בפועל

המימוש נעשה בשפת Python עם שימוש בספריות Pandas, NumPy, Matplotlib, Seaborn scikit-learn. הקוד כולל פונקציות לפעולות הבאות:

- **טעינת הנתונים:**
קריאה של קובץ CSV והצגת הדוגמאות הראשונות ומידע כללי על המבנה.
- **ניקוי ועיבוד הנתונים:**
טיפול בערכים חסרים, הסרת עמודות לא רלוונטיות והמרת משתנים קטגוריים באמצעות קידוד One-Hot.
- **EDA (Exploratory Data Analysis):**
הצגת גרפים שונים – כגון גרף התפלגות המשתנה "Exited" ומפת מתאם – המאפשרים זיהוי ראשוני של תבניות וקשרים בנתונים.
- **נרמול תכונות:**
שימוש ב-StandardScaler לייעול תהליך האימון, במיוחד במודלים רגישים כמו SVM ו-k-NN.
- **חלוקת הנתונים:**
חלוקה לסטי אימון ובדיקה לשם אימון המודלים ובחינת ביצועיהם.
- **אימון מודלים:**
יישום ארבעת המודלים – Decision Tree, SVM, AdaBoost ו-k-NN – באמצעות פונקציות ייעודיות לכל מודל, כאשר בכל אחד מתבצעת הערכה על בסיס מדדים כמו Accuracy, Precision, Recall ו-F1-Score.
- **השוואת ביצועים:**
איסוף תוצאות המודלים בטבלה להשוואה מהירה והצגה נוחה של הביצועים.
- **יצירת גרפים:**
שמירה אוטומטית של גרפים לתצוגה בדוח, כגון גרף חשיבות התכונות מעץ ההחלטה, מפת המתאם וגרף ההתפלגות.

השיפורים בייעול זמן הריצה של הקוד נעשו באמצעות מספר התאמות אסטרטגיות:

1. **הסרת עמודות לא רלוונטיות:**
הסרנו את העמודות "CustomerId", "RowNumber" ו-"Surname" בשלב הניקוי. עמודות אלה אינן תורמות לחיזוי והן עלולות לגרום ליצירת מספר רב של עמודות מיותרות בעת קידוד One-Hot, מה שמגדיל את ממדיות הנתונים ומאט את תהליך האימון.

2. **הפחתת עומס בעיבוד גרפי:**
בעת יצירת מפת המתאם (heatmap), ביטלנו את ההצגה של הערכים (annotations) על כל תא על ידי הגדרת הפרמטר `annot=False`. פעולה זו מפחיתה משמעותית את העומס החישובי של יצירת הגרף, במיוחד כאשר מדובר במטריצה גדולה.

6. סיכום האלגוריתמים והתוצאות

לאחר ביצוע ניסויים והשוואת ביצועי המודלים, התקבלו התוצאות הבאות:

מודל	Accuracy	Precision	Recall	F1-Score
Decision Tree	79.97%	48.60%	50.51%	49.54%
SVM	86.10%	79.72%	38.36%	51.79%
AdaBoost	86.07%	73.06%	45.03%	55.72%
k-NN	83.37%	62.76%	35.79%	45.58%

פלט התוצאות מהרצת הקוד:

```
Comparison of Model Performance:
               Accuracy  Precision    Recall  F1-Score
Decision Tree  0.799667   0.485997  0.505137  0.495382
SVM            0.861000   0.797153  0.383562  0.517919
AdaBoost       0.860667   0.730556  0.450342  0.557203
k-NN           0.833667   0.627628  0.357877  0.455834
```

ניתוח תוצאות:

- **דיוק:** מודל ה-SVM השיג את הדיוק הגבוה ביותר (86.10%).
- **מדדים משלימים:** יש לשים לב שמדדים כגון Precision ו-F1-Score מצביעים על איזון טוב יותר במודל AdaBoost.
- **משתנים משפיעים:** גרף חשיבות התכונות מעץ ההחלטה מצביע על כך שמשתנים כמו יתרת החשבון, גיל, הכנסה חודשית ומשך התקשרות הם בעלות השפעה משמעותית על נטישת לקוחות.

השפעת הנדסת תכונות:
פעולות הסרת עמודות לא רלוונטיות, נרמול ויצירת מאפיינים חדשים סייעו לייעל את זמן הריצה ולשפר את ביצועי המודלים על ידי צמצום רעש והפניית המודל למאפיינים בעלי ערך חיזוי אמיתי.