**תרגיל – ניקוי מידע**

בתרגיל זה ובתרגיל הבא תעבדו על אותם מקורות נתונים העוסקים בבעיית חיזוי של טרוריסטים.

**הבעיה**

השב"כ אוסף רשימת חשודים בביצוע טרור מקרב תושבי עזה על ידי הפעלת רשת מלשנים ושימוש באמצעים מודיעיניים.

הבעיה היא שמדובר רק בחשודים לפיגוע בטרור ובפועל רק כ-10% מהחשודים אכן מתכננים לבצע פיגוע טרור.

המצב כיום הוא שאין שום מנגנון דירוג של החשודים לפי ההסתברות שלהם לבצע פיגוע טרור כך שעל השב"כ לבצע מעצרים ולהפעיל אמצעי מעקב ומניעה יקרים שדורשים משאבים רבים וזמן רב.

כיוון שהמשאבים של השב"כ אינם בלתי מוגבלים הוא אינו יכול לבצע מעקבים ומניעה על כל החשודים בו זמנית, ובזמן בו הוא משקיע על תחקור חשודים לא רלוונטיים, חשודים אחרים שהינם טרוריסטים עלולים לבצע פיגועים. בפועל כיום, לשב"כ יש אמצעים לבצע מעקב רק אחרי 20% מהחשודים וכיוון שאין שום מנגנון דירוג בין חשודים אלו אז רק 20% מהטרוריסטים נעצרים בעוד ש-80% מפיגועי הטרור מתממשים ללא שום יכולת למנוע אותם.

עליכם להציע לשב"כ דירוג של החשודים (כך שבראש הרשימה שתגישו לשב"כ יופיעו החשודים עם הסיכוי הגבוה ביותר לביצוע פיגוע טרור ובתחתית הרשימה אלו שהסיכויים שלהם הנמוכים ביותר), ובכך בתקווה תצליחו להקטין את כמות הפיגועים בצורה משמעותית.

**המידע**

1. **Db1.csv**: מאגר של כל תושבי עזה עד גיל 80. עבור כל תושב עזה, רשום במאגר ת.ז. שלו, של אביו ושל אימו. בנוסף רשום שם גם תאריך הלידה של התושב.
2. **Db2.csv**: רשימת כל המעצרים של טרוריסטים מהרצועה. כל רשומה במאגר מתייחסת למעצר ורשומים בה: ת.ז. של העצור, ומיקום המעצר (קורדינטות אורך ורוחב).
3. **Db3.csv**: רשימת הכנסות של כל החשודים. מאגר מידע זה מכיל נתונים על החשודים בלבד (שאותם עליכם לדרג). עבור כל חשוד שעבד בשנה האחרונה רשום שם ת.ז. שלו והשכר השנתי שאותו הוא דיווח לרשויות המס בשקלים.
4. **Db4.csv**: רשימה של מספרי פלאפון של כל תושבי הרצועה כפי שהתקבלו מהמפעילות הסלולאריות.
5. **Db5**: זוהי תיקייה שמכילה קובץ csv עבור כל חשוד. בקובץ זה יש טבלה של רשימת השיחות אותם החשוד ביצע בחודש האחרון. התחילית של שם כל קובץ הוא ת.ז. של החשוד שאת השיחות שלו הקובץ מפרט. כל רשומה בקובץ מתארת שיחה ומכילה את הפרטים הבאים על השיחה: מספר הפלאפון של הצד השני בשיחה, סוג השיחה (נכנסת/יוצאת), זמן השיחה בשניות, והמיקום ממנו החשוד ביצע את השיחה (קורדינטות אורך ורוחב).
6. **Train\_data.csv** – רשימה המכילה ת.ז. של חשודים ותיוג האם הם התגלו כטרוריסטים או לא (מידע מהעבר). מדגם זה ישמש ללמידת מודל.
7. **Test\_data.csv** – רשימה המכילה ת.ז. של חשודים ותיוג האם הם התגלו כטרוריסטים או לא (מידע מהעבר). מדגם זה ישמש לצורך הערכת המודל שלמדתם.

**הידע העסקי**

לצורך ביצוע הפרויקט והבנת עולם התוכן בצורה טובה יותר נפגשתם עם בכיר בשב"כ העוסק בתחום המודיעין, שמסר לכם את המידע הבא הנוגע לטרוריסטים מתוך הניסיון אותו הוא צבר:

1. המניעים של טרוריסטים צעירים שונים ממניעים של טרוריסטים מבוגרים. הצעירים לרוב מבצעים פיגועים בשל השפעה מתוך המשפחה שלהם (אב שעסק בטרור, אח בכור (שבחברה הערבית הרבה פעמים מהווה ממלא מקום דמות האב ודמות לחיקוי) שעסק בטרור.
2. המבוגרים לעומת זאת, מבצעים פיגועים לרוב ממניעים כלכליים: מדובר לרוב בבעלי משפחות בעלי הכנסה נמוכה במיוחד שהקצבה אותה מציעים ארגוני הטרור יכולה לאפשר למשפחתם להתקיים בכבוד אחרי מותם.
3. יש מספר מועט של מקומות מפגש ברצועה בה נפגשים הטרוריסטים לצורך תכנון פיגועים. מהניסיון, מפגעים לרוב נוטים להסתובב בקרבת מקומות אלו כאשר הם מתכננים לבצע פיגוע. גם המעצרים לרוב מתבצעים במקומות אלו כי שם הכי קל למצוא את הטרוריסטים.
4. פעילי טרור לרוב נשארים במעגל הטרור גם לאחר שחרור ממעצר.
5. בנים יחידים פחות נוטים לעסוק בטרור.

**המשימה שלכם**

כמו כל פרויקט data science, הכל מתחיל באיכות טובה של הנתונים איתם משתמשים.

במשימה הבאה עליכם לאמוד בצורה כמה שיותר עמוקה את איכות המידע ולגלות בעיות אפשריות באיכותו. נסו לאפיין את הבעיות בצורה טובה (למשל במקום להגיד שחסר מידע להסביר מה נפח המידע החסר (אם ניתן) ולנסות לבדוק האם יש מאפיינם משותפים לפיסות המידע החסרות (אם ניתן)). ככל שתתנו מידע יותר מפורט ומוסבר על בעיות במידע יהיה ללקוח (ולכם) יותר קל להבין כיצד להתגבר על הבעיות השונות.

במיוחד שימו דגש על מידע חסר וערכים שגויים של שדות במידע (לרוב בעיות אלו מצריכות השגת מידע מחדש מהלקוח, תהליך ארוך וסיזיפי שעדיף שיתחיל כמה שיותר מוקדם בכדי לא לעקוב את הפרויקט).

אתם מוזמנים לשחק עם המידע, לשאול עליו שאלות שמתקפות את ההיגיון הסביר והידע העסקי ולהצליב בין מקורות מידע שונים בכדי להשיג את מטרה זו.

בהצלחה!