# **פרויקט – כריית נתונים**

# **חיזוי שימוש בקוקאין - Cocaine use prediction**

# חברי הצוות:

# **אור בן טובים - 313373854**

# **נועה מור - 205530991**

# **אוראל נגר – 206261786**

**תקציר:**

נכון להיום, למעלה ממיליון אמריקאים מכורים לקוקאין - אחד החומרים הממכרים ביותר הנחשב קשה מאוד לטיפול. בישראל, ובעיקר בתל אביב, נפוץ מאוד השימוש בקוקאין. ישראל מדורגת במקום ה-20 בעולם בהיקף השימוש בקוקאין, ביחס לגודל האוכלוסייה.

כמו כן, בפסיכולוגיה, מתארים את אישיותו של האדם באמצעות 'חמש התכונות הגדולות' אשר בעלות השפעה ניכרת עליו, שהינן: מוחצנות-מופנות (Extraversion), נעימות (Agreeableness), מצפוניות (Conscientiousness), יציבות רגשית-נוירוטיות (Neuroticism) ופתיחות מחשבתית (Openness to experience).

בפרויקט זה, אנו מעוניינים לבחון את הקשר בין 'חמשת התכונות הגדולות' וכן עוד משתנים נוספים (מין, גיל, מוצא אתני, השכלה וכו'), לבין התנסות בקוקאין – השתמש/לא השתמש. הפרויקט יוכל לשמש ככלי עזר למוסדות ביטחוניים/ציבוריים בכך שיוכלו לאמת את נטייתם של המועמדים הנבחנים למשרות רלוונטיות. פרויקט זה הינו פרויקט יישומי - אנו צופים שמוסדות רבים יוכלו להיעזר בכלי זה, ויראו בו כלי שימושי לזיהוי מקדים של המעומדים אשר אינם מתאימים למוסד.

לבסוף, הטמענו ממשק אינטרנטי לחיזוי התנסות בקוקאין, ע"י מודל ANN בעל רמת דיוק של 0.77 במדד ה-AUC ROC. מעבר לכך, מצאנו ש"חמשת התכונות הגדולות" מהוות גורמים משמעותיים ומשפיעים בחיזוי זה.

**שיטה :**

אנו נשתמש בבסיס נתונים המכיל מידע על הערכים השונים שתוארו קודם, אשר מתבסס על 1,885 רשומות ו-32 פי'צרים. אנו נבחן רק האם אדם השתמש בקוקאין או לא, בעוד בסיס הנתונים הנ"ל מספק חומרים ממכרים נוספים אחרים. יש לציין, שאנו נבצע שלבי מתודולוגיה קלאסיים של למידת מכונה, משלב חקר הנתונים ו-EDA עד לשלב החיזוי עצמו ובחינת המודל. שלבים אלו יכולים להתאים לחיזוי שימוש בכל אחד מהחומרים הממכרים, אך אנו בחרנו בפרויקט זה להתמקד אך ורק בשימוש בקוקאין. בכדי לבצע את מטרת הפרויקט אנו ניישם מספר טכניקות של למידת מכונה וכריית נתונים. ראשית, אנו נבצע תהליך של EDA על בסיס הנתונים בכדי להבין את התפלגות המשתנים, חשיבותם, מציאת דפוסים שונים, אנומליות וכו', אשר יבוצע באמצעות וויזואליזציות שונות. לאחר שלב זה, נבצע שלב של Data Cleaning במטרה לקבל בסיס נתונים תקין שנוכל להפיק ממנו תוצאות נכונות ומהימנות. לאחר מכן, נבצע מספר טכניקות של כריית נתונים (אשר נלמד במהלך הסמסטר), בכדי לקבל תבניות ודפוסים שונים של הנתונים. לבסוף, באמצעות הבנת הנתונים ויצירת התבניות השונות בשלבים הקודמים, ניצור מודל חיזוי אשר יחזה האם האדם השתמש בקוקאין בעבר או לא. המודל יוצג באמצעות GUI של ספריית Gradio.

כפי שציינו, אנו נבצע מספר מודלים של למידת מכונה אשר יתאמנו על מערך אימונים וייבחנו באמצעות מערך המבחן. חלוקת מערך האימונים ומערך המבחן, יחולקו בהתאם להתפלגות הנתונים וכן ביצוע אימות צולב (Cross validation) לבחינת המודלים השונים. יתר על כן, אנו מתכננים להעריך את תוצאות המודל באמצעות המדדים: AUC ROC ו-F1\_score. בחרנו במדדים הנ"ל היות ולאחר בדיקת התפלגות הדאטה, קיים איזון בין שתי הקבוצות. בנוסף, בחרנו להשתמש במדד F1\_score בכדי לקבל ממוצע (הרמוני) של ה-Recall וה-Precision לטובת החיזוי.

קישור לבסיס נתונים:

[https://archive.ics.uci.edu/ml/datasets/Drug+consumption+%28quantified%29#](https://archive.ics.uci.edu/ml/datasets/Drug+consumption+%28quantified%29)

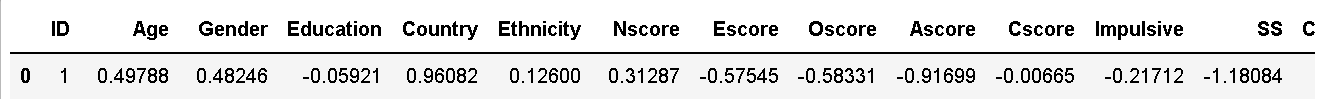
**ניסויים שבוצעו:**

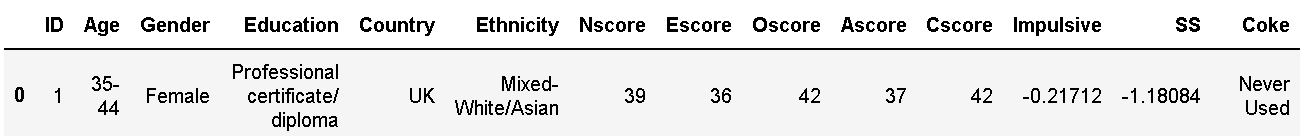
**קריאת הנתונים ותהליך ה-Pre-Processing**

A picture containing chart

Description automatically generatedבתחילה קלטנו את הנתונים והתחלנו בסידור שלהם. הנתונים הגיעו בצורה מספרית אשר לא מובנת לקריאה, אז התחלנו בהמרת כל הערכים המספריים לערכים המקבילים שלהם המילוליים. לחלק מהפיצ'רים היו הרבה סוגים של ערכים שונים, ולחלק סט ערכים קטן יותר. מטעמי נוחות המרנו את הפיצ'רים עם כמות הערכים המצומצמת ע"י טעינתם ב-JSON, ובפיצ'רים עם ערכים מרובים, טענו את ה"מילון" שלהם ב-CSV.

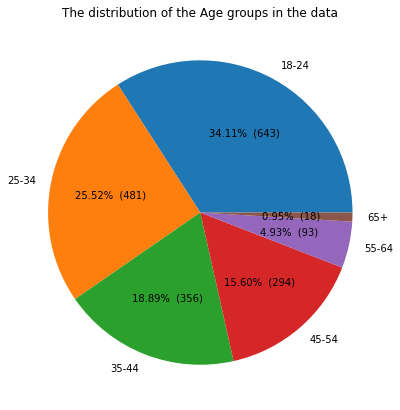
דוגמא לקובץ JSON :

לפני:

אחרי:

ישנם פיצ'רים אשר חסרים להם הערכים המילוליים המתאימים. לכן, השארנו אותם כמו שהם, במטרה לבחון האם קיים קשר בין הפיצ'רים הנ"ל בצורתם הנוכחית לשימוש בקוקאין, או לפיצ'רים אחרים.

בנוסף, נמצא שאין ערכים חסרים וחריגים (בבחינת הממוצע, החציון וסטיות התקן)



**Exploratory Data Analysis**

מבחינת גילאים נדגמו בעיקר אנשים בעלי גיל הנמוך מ-55 והגדול מ-18, שזו האוכלוסייה הרלוונטית יותר למטרות המחקר שלנו. מבחינת מין יש 50 אחוז גברים , ו-50 אחוז נשים. בנוסף, כ-60 אחוז בעלי השכלה (חלק בעלי דוקטורט וחלק בעלי תואר ראשון) , וכ-40 אחוז אינם בעלי השכלה אקדמאית (חלק סיימו תיכון וחלק נשרו מבית הספר בגילאים שונים). עבור פיצ'רים אלו המידע יחסית מאוזן. וכן המשתנה התלוי שלנו מאוזן גם הוא, כאשר 55% לא התנסו, ו45% התנסו בקוקאין.

מבחינת מיקום המגורים והמוצא של הנשאלים, בסיס הנתונים משקף בעיקר את אוכלוסיית UK ו-US (85% מהמדגם) עם נטייה מובהקת לנשאלים בעלי צבע עור לבן (91% מהמדגם).

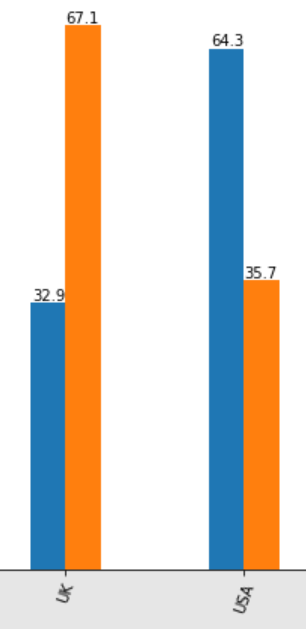
Application, Teams

Description automatically generated with medium confidenceכעת, ניסינו לראות האם קיימת קורלציה בין המשתנים. בHeat-Map המוצגת ניתן לראות קשרי קורלציה חלשים בין המשתנים. קשר אחד אשר מעל 0.5 ואולי יכול לתת לנו אינפורמציה הוא הקשר בין האימפולסיביות למדד התחושתיות העומד על 0.62.

Chart, bar chart

Description automatically generatedחשדנו שיהיה קשר בין גיל הנשאל להתנסותו בקוקאין, קשר שלא ניתן לראות ב- Heat-Mapלעיל בגלל שמשתנה הגיל הינו קטגוריאלי. אכן, ניתן לראות שעד גיל 44 מידת ההתנסות של הנשאלים עומדת על כ50% פחות או יותר. לעומת זאת מעל גיל 45 אחוז המשתמשים יורד לאזור ה37%, ולבסוף לנשאלים היותר מבוגרים ניתן לראות מידת התנסות נמוכה משמעותית. הסבר אפשרי הוא שבעבר הייתה פתיחות נמוכה יותר להתנסות בחומר זה.

ביצענו גרפים דומים על מנת לראות קשרים בין שאר המשתנים הקטגוריאליים להתנסות. ומצאנו ש-50% מהגברים התנסו לעומת 40% מהנשים.



בנוסף, אם נסתכל על US ו-UK, שהם מיקומי המגורים העיקריים שלנו, נראה הבדלים משמעותיים באחוז המשתמשים שהתנסו.

לאחר מכן, בדקנו בגרפים את הקשר בין מידת החוזק של כל אחת מ"חמשת התכונות הגדולות" להתנסות. הגרף מתאר את כמות המשתמשים ולכן נבחן אותו ע"י הסתכלות על שינוי המגמה. ברוב הגרפים לא ניתן לראות הבדלים משמעותיים, אך בהסתכלות על תכונת ה-Openness, שזה מידת הפתיחות לחוויות. היינו מצפים שככל שמדד זה יהיה גבוה יותר מידת ההתנסות תהיה גבוהה יותר. ואכן נראה שהיחס בין המתנסים ללא מתנסים משתנה מערך 48, ויותר אנשים התנסו בקוקאין לעומת אותם שלא התנסו: Chart, line chart

Description automatically generated

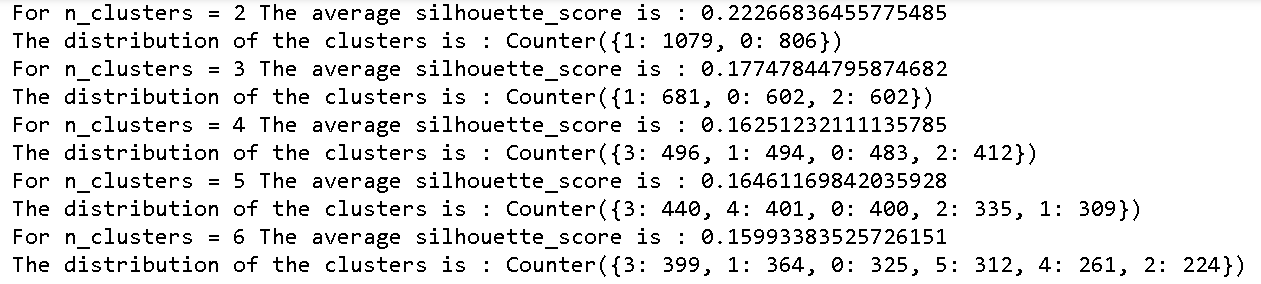
Chart, line chart

Description automatically generated  
  
כמו כן, אם נסתכל גם על Conscientiousness , המתארת את הנטייה לוויסות עצמי, התמקדות במשימות ובמטרות, חשיבה לפני מעשה ודחיית סיפוקים, היינו מצפים לראות ירידה ביחס ההתנסות ככל שמדד זה עולה. ואכן כך ניתן לראות בגרף זה:

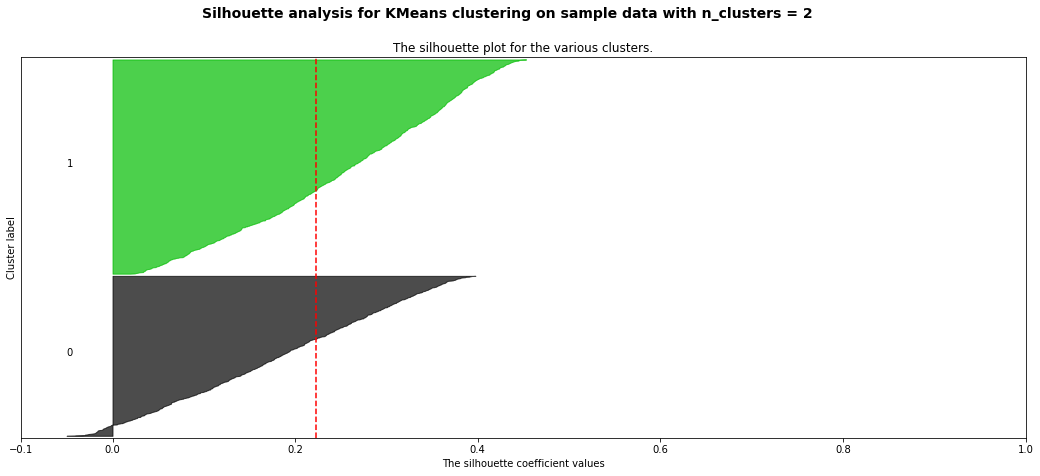
**Clustering and Binning**

בנוסף לפיצ'רים הקיימים שציינו, רצינו לבחון יצירת פיצ'רים חדשים ע"י שימוש באלגוריתמי Clustering וביצירת Binning. בחנו 2 אלגוריתמי Clustering עבור "חמשת התכונות הגדולות": K-means ו- DBSCAN. יש לציין שבשני האלגוריתמים ביצענו Scaling בכדי לוודא שכל המשתנים באותו הטווח.

1. K-means – נעזרנו בדוגמא לקוד מהאתר של scikit-learn ובחנו טווח של 2-6 קלאסטרים. לאחר הצגה ויזואלית של ציון ה- silhouette בחרנו חלוקה ל2 קלאסטרים אשר הניבה את הציון הטוב ביותר, 0.222. ניתן לראות בתוצאות מטה, שככל שעולים במספר הקלסטרים ממוצע המדד יורד בכל קבוצה.



להלן מספר גרפים של מדד ה-Silhouette ל-3 חלוקות של קלאסטרים:

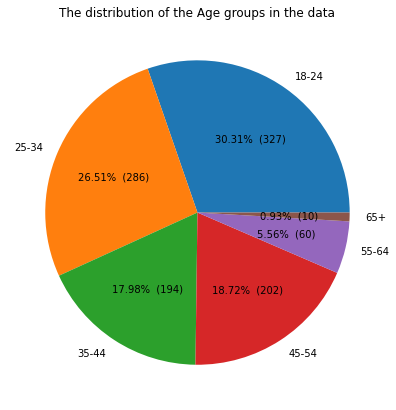
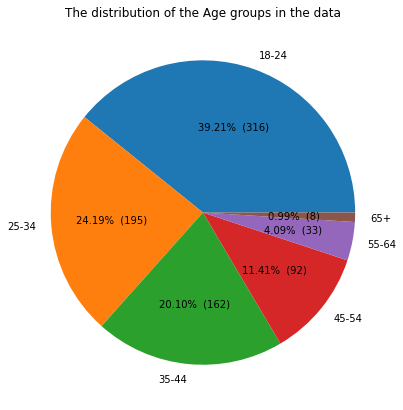


רצינו לבחון מהם המאפיינים של כל קבוצה באמצעות גרפי הפאי הבאים:

גיל:

Cluster 1

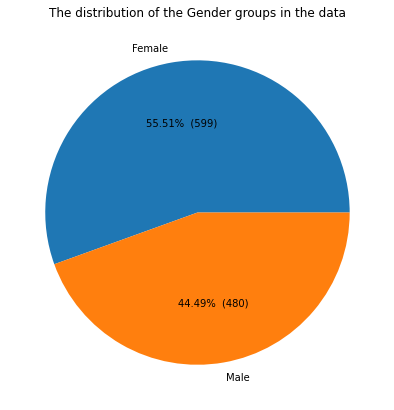
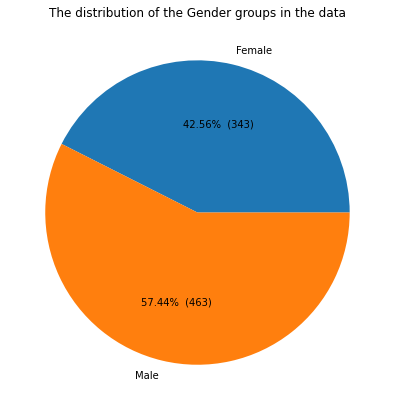
Cluster 0



ניתן לראות שבקבוצה 0, האלגוריתם סיווג יותר צעירים (בגילאים 18-24). בנוסף, ניתן לראות שקיים שוני באחוז האנשים בגילאים 45-54. כפי שציינו קודם לכן, קבוצת הגילאים הצעירה (18-24) התנסה יותר בקוקאין לעומת קבוצת הגילאים 45-54.

מגדר:

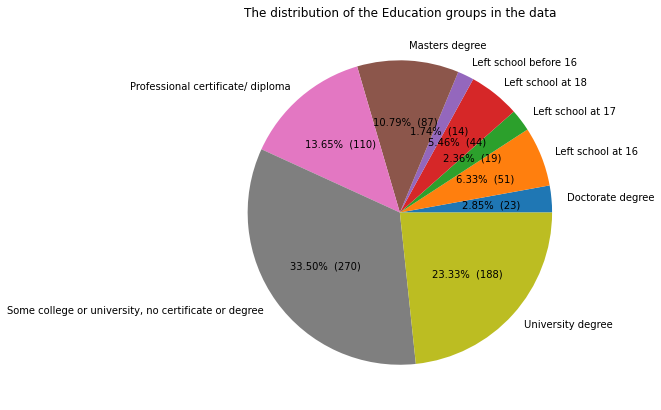
Cluster 0



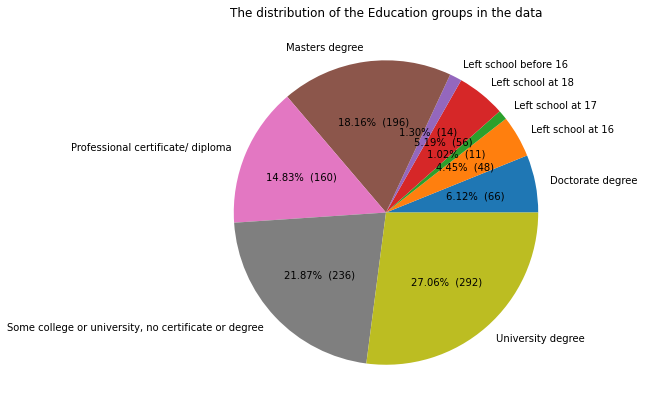
Cluster 1

ניתן לראות שבקבוצה 0 אחוז הגברים גדול יותר מאחוז הנשים ובקבוצה 1 המצב הפוך. כפי שציינו קודם, גם פה ניתן לראות שקבוצה 0 מכילה את האוכלוסייה (במקרה הנ"ל מגדר) שהתנסה יותר בקוקאין.

השכלה:



Cluster 1

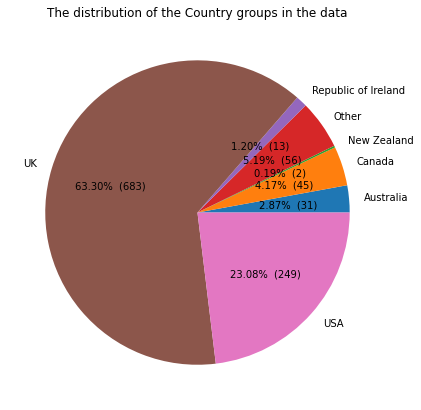


Cluster 0

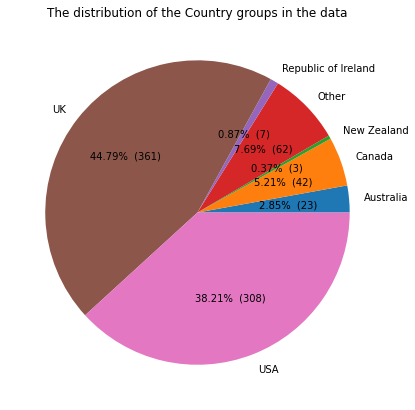
ניתן לראות שבקבוצה 0 יש פחות אנשים בעלי השכלה גבוהה בהשוואה לקבוצה 1. בהמשך הדו"ח מוצג גרף המראה שיותר אנשים שאין להם השכלה גבוהה התנסו יותר בקוקאין מאשר אנשים בעלי השכלה גבוהה. כמו כן, ניתן לראות שקיים הבדל משמעותי באוכלוסיית האנשים ללא תעודה/תואר (הבדל של כ-12 אחוזים).

מדינות:

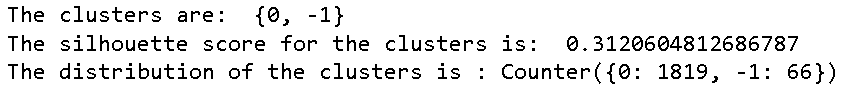
Cluster 1



Cluster 0



ניתן לראות שההבדל המשמעותי בין 2 הקבוצות, עבור משתנה זה, הינו החלוקה בין האוכלוסייה שגרה ב-UK וב-USA. בקבוצה 0 נמצאים פחות אנשים הגרים ב-UK מאשר בקבוצה 1, ויותר אנשים הגרים ב-USA נמצאים בקבוצה 0 לעומת בקבוצה 1. כפי שהצגנו קודם לכן, ב-USA אחוז האנשים שהתנסה בקוקאין גדול הרבה יותר מהאנשים שלא התנסו, וב-UK ההיפך. גם פה נראה שבקבוצה 0 קיימים יותר אנשים שהתנסו בקוקאין (בהתייחס ל-2 המדינות הנ"ל).

1. DBSCAN – השתמשנו באלגוריתם זה כדי לבחון צורת חלוקה שונה מהאלגוריתם הקודם. לאחר מספר ניסיונות, בשינוי 2 הפרמטרים של האלגוריתם, מצאנו שגם אלגוריתם זה מביא את התוצאות הטובות ביותר בחלוקה ל2 קלאסטרים. חלוקות אלגוריתם זה, הבעלות ערך silhouette הגבוה מ0.22, היו מוטות במעל 95% לכיוון אחת הקבוצות. חלוקה זאת הייתה נראית לנו לא משמעותית אשר לא תביא לנו ערך מוסף, ולכן בחרנו לממש את האלגוריתם הראשון.

כפי שציינו, דרך נוספת ליצירת פיצ'רים נוספים הינה באמצעות Binning . ראינו לנכון לאחד מספר קטגוריות בפיצ'רים נבחרים אשר התנהגו בצורה דומה, לשם יצירת פיצ'ר חדש שיוכל לתרום למודל החיזוי. יש לציין שבשלב זה איננו יודעים אם פיצ'רים אלו יתרמו בצורה טובה בהמשך למודל החיזוי. שלב הEDA נתן לנו את היכולת להעריך אילו Binning עלינו לבצע. הפיצ'רים שבחרנו ליצור עבורם Binning הם: גיל, השכלה, מדינה (ליבשת ומדינות עיקריות) ומוצא.

Bar chart

Description automatically generated with low confidenceלדוגמא, האיחוד שנעשה בהשכלה הפריד בין נשאלים בעלי השכלה גבוה או לא.

**ML Prediction Models**

בפרויקט זה בחנו 3 מודלי חיזוי: Random Forest, Adaboost ו- ANN. עבור המשתנים הקטגוריאליים ביצענו תהליך של One Hot Encoding על מנת שנוכל לבחון את ההשפעה של כל קטגוריה על מודל החיזוי. את טיב החיזוי בחנו באמצעות 2 מדדים: F1-Score ו-AUC ROC.

ניסינו לעשות Scaling למודל הANN אך תוצאות החיזוי היו פחות טובות ולכן ויתרנו על כך.

בשני המודלים הראשונים שצוינו לעיל, השתמשנו בHyperparameter על מנת למצוא את הפרמטרים אשר יתנו את החיזוי הטוב ביותר על פי מדד ה-AUC ROC.

A picture containing chart

Description automatically generatedציוני המודלים:

Text

Description automatically generated

Text

Description automatically generated with low confidence

כמו כן, עבור שני המודלים הראשונים ביצענו Feature Importance במטרה לבחון מה הם הפיצ'רים המשפיעים ביותר על מודל החיזוי.

Chart, bar chart

Description automatically generatedChart, bar chart

Description automatically generatedRandom Forest Adaboost

ניתן לראות ש"חמשת התכונות הגדולות" הן מהוות כפיצ'רים המשפיעים ביותר על החלטת מודלי החיזוי.

בחרנו במודל הANN מכיוון שהתוצאות שלו במדד ה-AUC ROC היו הכי טובות, וגם ה-F1Score שלו היה גבוה כמו המודלים האחרים.

**Web Interface – Gradio**

לבסוף, החלטנו להטמיע את מודל ה-ANN בתוך ממשק אינטרנטי ידידותי למשתמש, על מנת להקל את השימוש לכל אדם, גם למשתמשים אשר אין להם ידע בתכנות או בלמידת מכונה.

על המשתמש יהיה להזין את כל הפיצ'רים שהמודל שלנו משתמש בהם, חלקם מספריים וחלקם קטגוריאליים. את המשתנים המספריים המשתמש פשוט יזין כטקסט חופשי, ואת המשתנים הקטגוריאליים המשתמש יבחר מתוך סט קבוע של ערכים, כל זאת על מנת להקל על המשתמש וליצור את הממשק אינטואיטיבי ופשוט, מה שימנע טעויות הזנה.

את הערכים המספריים שהמשתמש מזין, אנו מכניסים ישירות למודל. את הערכים הקטגוריאליים הנבחרים, אנחנו מגדירים כ- "1" , ואת שאר הקטגוריות שלא נבחרו מאותו הפיצ'ר נגדיר כ"0".

Graphical user interface, application

Description automatically generated

**חלקות משימות בעבודה:**

יש לציין **שכל אחד** לקח חלק **בכל שלבי המטלה** ורוב הפעמים נפגשנו **ועבדנו ביחד**. אך להלן חלוקה גסה.

אור –קליטה ועיבוד הנתונים ו-Gradio.

אוראל – EDA, מודלים ועריכת הדוח.

נועה – Binning, Clustering ועריכת המצגת.