פרויקט מסכם

מסמכים להגשה:	1. קובץ קוד מוכן בפורמט Jupyter notebook. 2. מסמך PDF לדו"ח בשם report_{group_number}.pdf 3. קובץ csv בשם results_{group_number}.csv אשר כולל את תחזיות הקלסיפיקציה (Prediction **Probabilities** - מצ"ב דוגמא להגשה בקובץ example_results.csv) **שימו לב לשמות העמודות!**
תוכנות דרושות:	Anaconda for python 3
אמצעי הגשה:	קובץ zip שיכלול את שלושת הקבצים הנ"ל – שם הקובץ בפורמט {group_number}.zip
דוגמת הגשה (הגשה שחורגת מהפורמט תגרור הורדה אוטומטית של 5 נקודות)	עבור קבוצה מספר 70 יוגש קובץ 70.zip המכיל את הקבצים: 1. notebook_70.ipynb 2. report_70.pdf 3. results_70.csv
יצירת קשר:	ilanv@mail.tau.ac.il
מועד אחרון להגשה:	30.06.2023 בשעה 23:59 (היום האחרון של הסמסטר)

מבוא

בפרויקט זה ינתנו לכם מספר פיצ'רים אודות קבצי הרצה (.exe) והמשימה שלכם היא להחליט אם הקובץ זדוני (malicious) או לא (benign), באמצעות ניתוח סטטי של הקבצים – כלומר ניתוח המידע אודות הקבצים מבלי להריץ אותם.

בפרויקט נעסוק בבעיית Binary Classification (כלומר – משני קלאסים) בה עליכם לסווג רשומות לשתי קטגוריות – האם הקובץ זדוני (1) או לא (0), על סמך מספר פיצ'רים בדאטה סט. חלק מן הפיצ'רים ידועים וחלקם אנונימיים.

המטרה בפרויקט המסכם הינה לחשוף אתכם לעבודה מעשית בה תוכלו להתנסות בחומר הנלמד בקורס, תוך כדי יישום הכלים בסביבת נתונים אמתית לחלוטין.

אין בכוונת הצוות להגביל אתכם בצורת החשיבה והעבודה, אולם קיימות מספר הנחיות בסיסיות אשר עליכם לעמוד בהן.

הנחיות כלליות

- אין להכין ולעצב את סט הנתונים הגולמי בעזרת אקסל (אלא בחבילות python בלבד).
- מימוש הקוד ייעשה במחברת ג'ופיטר בסביבת Anaconda, ויכלול הסברים מלאים על אופי המימוש בקוד עצמו (בעזרת markdowns והערות בתוך הקוד)
- מותר להשתמש בכל החבילות שמגיעות עם סביבת ה Anaconda. שימוש בחבילות נוספות הדורשות התקנה אפשרי באישור של מתרגל הקורס. במידה ויתקבל אישור כזה יש לכלול במחברת תא ייעודי להתקנת החבילות הנוספות (`!pip install...`)
- הימנעו מהצגת אזהרות במחברת הסופית – ניתן לעשות זאת באמצעות :

```
import warnings
warnings.filterwarnings('ignore')
```

- משך הרצת הקוד מתחילתו ועד סופו לא יעלה על שעה.
- אין להשתמש בקבצים חיצוניים פרט ל train.csv, test.csv כאשר בקובץ הtest אין את ה labels.
- כפי שתלמדו במהלך הסמסטר, אין חובה להשתמש בכל הפיצ'רים של סט הנתונים. תוכלו להנדס ולעצב את הפיצ'רים כרצונכם.
- טיב הביצועים של המודל שלכם ייעשה על ידי מטריקת **AUC**
- שימושים בפונקציות וטכניקות שלא נלמדו בקורס הם מבורכים, אולם הם אינם מהווים תחליף לשיטות המסורתיות.
- בונוס יחולק לסטודנטים לפי ביצועי המודל שלהם – 7 נקודות למקום הראשון, 4 נקודות למקום השני, נקודה למקום השלישי.
- קנס גדול יינתן על קוד שלא רץ ועל קובץ csv שלא מוגש בפורמט המבוקש.
- יש לפרט את ההנחות שנלקחו בכל שלב של הפרויקט. הנחות שלא יפורטו יחשב כאילו לא נלקחו בחשבון ומצב זה עלול להוביל להורדה בציון.
- עבור כל ויזואליזציה שאתם מציגים, יש לכתוב הסבר מה הסקתם ממנה ו/או איזה חלק מעניין בה. ויזואליזציה ללא הסבר תחשב כלא קיימת.
- גם אם ניסתם "לפצח" את הבעיה בדרך מסוימת ואין שיפור בתוצאות: אל תסירו את הניסיון מהקוד. רק חשוב שתדגישו שמדובר בניסיון לא מוצלח ואין לו מקום ב-work flow הסופי.
- המחברת צריכה לספר סיפור – כיצד חקרתם את הנתונים ואיך שיפרתם את המודל שלכם כשניסיתם פתרונות מכיוונים שונים (גם אם הם נכשלו).

משימת התכנות (הניקוד עבור הסעיפים השונים בסוגריים)

חלק ראשון – אקספלורציה (8 נקודות)

- עליכם לחקור את הנתונים בכל אופן שבו עולה על רוחכם: האופי שבו כל פיצ'ר מתפלג, התנהגות קורלטיבית בין הפיצ'רים, נתונים סטטיסטיים על הפיצ'רים. בשלב זה של הפרויקט יש המון מקום לוויזואליזציה! נצלו זאת. ציון ינתן בעיקר על המסקנות המתקבלות משלב זה.

חלק שני – עיבוד מקדים (35 נקודות)

עבור השאלות אשר מופיעות בסעיפים יש לענות בגוף המחברת (**Markdown**) בצמוד לחלקי הקוד הרלוונטיים

- האם קיימים נתונים חריגים (Outliers) בדאטה? אם כן, עליכם להסירם או לפחות לתת עליהם את הדעת **(3)**
- האם הנתונים מנורמלים? אם לא- האם צריך לנרמל אותם? מה החשיבות של נרמול הנתונים בבעיה? **(3)**
- האם ישנם נתונים חסרים? כיצד בחרתם לטפל בהם ומדוע באופן זה? **(3)**
- התמודדות עם משתנים קטגוריאליים **(4)**
- האם המימדיות של הבעיה גדולה מדי? למה מימדיות גדולה עלולה ליצור בעיה? איך נזהה כי מימדיות הבעיה גדולה מדי? **(6)**
- הקטנת המימדיות על ידי טכניקה אחת שנלמדה בכיתה – PCA, ו/או על ידי בחירת תת קבוצה של פיצ'רים קיימים (feature selection). כיצד הקטנת המימדיות השפיעה על המודל? **(6)**
- בניית פיצ'רים חדשים ו/או מניפולציה מתמטית על פיצ'רים קיימים **(2)**
- החלת העיבוד המקדים על סט ה-Test **(8)**
- ניתן לבצע ניסיונות נוספים אשר לא נלמדו בקורס על מנת לעבד את הפיצ'רים הניתנים לכם (בונוס)

חלק שלישי – הרצת המודלים (20 נקודות)

- בניית שני מודלים ראשוניים מבין השלושה הבאים והחלתם על סט הנתונים: **(10)**
  - Naïve Bayes Classifier
  - KNN
  - Logistic Regression
- בחירת שני מודלים מתקדמים מבין הארבעה הבאים והחלתם על סט הנתונים: **(10)**
  - Multi-Layer Perceptron (ANN)
  - Decision Tree
  - Random Forest or Adaptive Boosting
  - Support Vectors Machine

- יש לתת הסבר על משמעות ההיפר פרמטרים שבחרתם לשנות וכיצד הם משפיעים על המודל מבחינת שונות והטייה (יכול להופיע בנספחים).

- נסו במידת האפשר להסביר את התרומה (החשיבות) של כל פיצ'ר להצלחת המודל. (בונוס יינתן לניתוח מרשים בהיבט זה).

חלק רביעי – הערכת המודלים (23 נקודות)

- בניית Confusion Matrix (מדגמית) על אחד המודלים, עליכם להסביר מה אומרים התאים בתוך המטריצה בהקשר למודל שבחרתם, כלומר מה ניתן להסיק על ביצועי המודל בהקשר זה. **(10)**
- הערכת המודל באמצעות K-Fold Cross Validation, ובניית פלט ROC על כל K-Fold עבור כל אחד מהמודלים שהורצו (רצוי באותו התרשים). **(8)**
- פערי ביצועים בין הרצת המודל על ה-Train או על ה-Validation, האם המודל שלכם הוא Overfitted? מה עשיתם / עליכם לעשות על מנת להגדיל את יכולת ההכללה שלו? **(5)**

חלק חמישי – ביצוע פרדיקציה (9 נקודות)

- לאחר בחירת המודל, עליכם לבצע תחזית על נתוני קובץ "test.csv", ולשמור אותה בקובץ results_{group_number}.csv (החליפו את group_number למס' הקבוצה) אשר כולל את תחזיות הסתברות הסיווג (**Prediction <u>Probabilities</u>** - מצ"ב דוגמא).
- יש לבצע תחזית על **כל אחת** מהתצפיות המופיעות בקובץ test.csv
- **חשוב מאוד** – עליכם ליצור בסוף המחברת pipeline להרצת המודל הסופי.

כלומר, חלק במחברת שבו יש את התהליך מתחילתו עד סופו (מטעינת הנתונים וביצוע עיבוד מקדים ועד ביצוע חיזוי) עם הדרך שמצאתם לנכון להשתמש בה לצורך עיבוד וחיזוי (המודל). החלק הזה מיועד עבור שחזור תוצאות זריז ומהימן, שבו תשתמשו בפונקציות השונות שפיתחתם לאורך המחברת. (5)

- מינימום Test AUC: 0.9 (4)

חלק שישי – שימוש בכלים שלא נלמדו בקורס (5 נקודות)

- יש לתאר לפחות כלי אחד שיישמתם בפרויקט ולא למדנו עליו בקורס.
- ניתן לבחור בכל חלק בפרויקט ליישום חלק זה – אקספלורציה, עיבוד מקדים, למידה, הערכה.
- הסבירו מדוע בחרתם להשתמש בו, איך הוא עובד וכיצד הוא השפיע על המודל.

<u>הערות:</u>

1. מיותר לציין שאת המודלים תצטרכו לבחון על סט **Validation** ולא על ה-**Train** עצמו (בחירת מודל לפי ביצועיו על ה-**Train** עלול להביא לתוצאות מאוד נמוכות ולהורדת ציון משמעותית!). הרצת הפרידקציות על ה-**Train** יכולה לסייע במציאת **Overfitting** אבל לא מהווה אינדיקטור לטיב המודל.
2. עליכם לכתוב מפורשות את ההיפר פרמטרים של המודלים הנבחרים כפי שנלמדו בכיתה, גם אם הוחלט להשתמש בערכי ברירת המחדל שלהם.
3. סדר מימוש השלבים אינו מחייב, במסגרת הפרויקט סביר מאוד להניח שתצטרכו לחזור אחורה אל שלבים מוקדמים יותר (בדומה לכל פרויקט Data Science).

<u>- (הנחיות לדו"ח המסכם בעמוד הבא) -</u>

הדוח המסכם

הדו"ח יכלול לכל היותר 5 עמודים (לא כולל שער ונספחים) אשר בו יוסברו כלל השלבים שננקטו במהלך ניתוח הנתונים. נדרש הסבר מפורט על מענה לשאלות הנשאלו לעיל, הרציונל מאחורי בחירת כל אחת מהשיטות שצוינו בשלבים לעיל, של ההיפר-פרמטרים שנבחרו, ותוצאות המודלים השונים. אין צורך להרחיב במילים ואין צורך לצטט את חומר הקורס בפרויקט.

בהיבט פחות פורמלי, זה המקום להסביר את כלל התהליך שביצעתם. דילמות, החלטות וכו' ("סיפור").

הדו"ח ייכתב בעברית או באנגלית בגופן Calibri, עם רווח שורות של 1.15.

- יש לצרף נספח בדו"ח הכולל הסבר על אחראיות כל שותף ותרומתו לעבודה. השמטת חלק זה תוביל להורדה משמעותית בציון.
- יש לפרט את ההנחות שנלקחו בכל שלב של הפרויקט. הנחות שלא יפורטו יחשבו כאילו לא נלקחו בחשבון ומצב זה עלול להוביל להורדה בציון.
- ניתן להוסיף נספחים ככל העולה על רוחכם (ולהפנות אליהם בקוד). ויזואליזציה תופיע בנספחים (וגם בגוף הקוד).
- יש לציין בראש הדו"ח את שמות המגישים + ת"ז. את הדו"ח יש להגיש במערכת ה-Moodle עד התאריך 30.06.2023 בשעה 23:59.
- הדו"ח מהווה חלק בלתי נפרד מהפרויקט המסכם. הניקוד אשר הוגדר בחלקי הפרויקט השונים נשען גם על טיב תיאורם בדו"ח.

שימו לב: מועד ההגשה הינו סוף הסמסטר. לא יינתנו הארכות אז אנא תכננו את זמנכם בהתאם. בנוסף, יש להקפיד הקפדה יתרה על פורמט ההגשה.

בהצלחה!!!