

## עבודת סוף קורס – למידת מכונה וסטטיסטיקה ביישומים רפואיים

### העבודה בזוגות.

ביחרו דאטא סט רפואי עליו תבצעו אנליזה לפי ההוראות בהמשך (ניתן להיעזר במאגרי המידע שהעלתי לתאר המודל, למשל דאטא של מיקרוביום, דאטא של גנים וכדומה) ועדכנו בקובץ האקסל שהועלה למודל את שמות חברי הקבוצה ואת הדאטא סט שבחרתם.

עבודה זו מורכבת מ2 שלבים:

#### - **שלב 1:**

הצגת עיקרי העבודה, שאלת החקר, שיטות החקר, האתגר בדאטא איתו עליכם להתמודד, תוצאות ראשוניות, מסקנות ראשוניות ומה מתכוונים להרחיב בהמשך.  
**תאריך הגשה: XX** נא להשתבץ באקסל שהעליתי למודל.  
ההצגות יועברו פרונטלית בכיתה על ידי הזוג בליווי מצגת ובגרפים תומכים. משך זמן ההצגה כ-5 עד 7 דקות. אין לחרוג מעבר ל10 דקות.  
יש להגיע פרונטלית למפגש המסכם של הקורס, אנא כבדו את החברים שלכם.

#### - **שלב 2:**

הגשה כוללת של העבודה במבנה של מאמר (יוגש בשפה האנגלית pdf) הכולל את ראשי הפרקים הבאים (שימו לב אורך העבודה המקסימלי הוא **8 עמודים**):

- Abstract – main background to the question, main results and conclusion
- Introduction- literature review on the subject
- Results-presenting the main hypothesis and answering them using the tools of the course and new tools. You should support your claims with tables, figures, visualizations.
- Methods – describing the tools and calculations used during the project, such as- algorithms, statistical tests, etc, there is no need for explaining the theoretical methods you used, for example, if you used a specific feature selection, there is no need to define what is feature selection that report the details of the specific feature selection you used (dimension, filtering criteria, etc.)
- Discussion-the conclusions and limitations of your analysis

בנוסף, יש לצרף קישור לGitHub מסודר עם Readme מפורט והוראות הפעלה מדוייקות כך שיהיה ניתן לשחזר את האנליזה המלאה שנעשתה בעבודה. יש להעלות אף את הדאטא לגיטהב.

**שימו לב שאתם מספרים סיפור ולא רק מבצעים גיבוב של מבחנים סטטיסטיים בלי קו מקשר ביניהם.**

### **הבהרה:**

בזמן ניתוח הפרוייקט יש להתייחס למושגים הבאים (במידה ורלוונטי לשאלת המחקר), נא דאגו שהשאלה תהיה מקיפה ולהשתמש בכמה שיותר כלים ומושגים מהקורס כדי לענות עליה:

מה הדאטא איתו אתם עובדים, מה מאפייני ההתפלגות שלו, מה האתגרים שנובעים מסוג הדאטא הזה, מה האתגרים שנובעים מהמשימה הספציפית שאתם רוצים לבצע, האם הדאטא טבלאי, תמונותי, טקסטי, האם הפיצ'רים תלויים, בלתי תלויים, האם processing שאתם מבצעים לדאטא מתאים להנחות של המודלים שאתם משתמשים בהם (נירמול, סקלות ערכים, אי תלות, מולטי קולינאריות, רגישות למרחקים). בנוסף – יש להתייחס לשיטות ה processing שנלמדו בקורס, מודלי הלמידה שנלמדו בקורס (מודלים לדאטא טבלאי פשוט, מודלים לתמונות CNN, מודלים לגרפים GAT, GCN), משמעות המבנה של הדאטא אל מול הערכים אם רלוונטי. אימון מודלים, לוס, רגולריזציה, overfitting, חלוקה נכונה של הדאטא לקבוצות אימון, ולידציה,

מבחן, שיטות הערכה של המודל תוך שימת דגש על מטריקות מדידה שונות כמו ( r squared, correlation, mistakes, confusion matix, accuracy, AUC, Precision, Recall, F1 score) מודלים תוך שימוש במבחנים סטטיסטיים מתאימים וויזואליזציות מתאימות. בדיקת השערות, היפותיזה, רווח סמך, רמת מובהקות, p-value, טעות סוג ראשון, טעות סוג שני והיחסים ביניהן, עוצמת המבחן, מבחן דו צדדי, מבחן חד צדדי, מבחן למדגם אחד, מבחן לזוג מדגמים בלתי תלויים, מבחן למדגמים מזווגים, התפלגויות ובחינתם, מבחן חי בריבוע, מבחן F, מבחנים פרמטריים, מבחנים אפרמטריים, מתאמים, תיקוני למידות מרובות, השוואות, וכן כל מושג שהזכרנו במהלך ההרצאות או התרגולים ולא צויין מפורשות.

**חשוב מאוד – שהפרויקט לא יוצג כאוסף מבחנים ובדיקות בלי קו מקשר, אלא כסיפור. אתם באים לספר את הסיפור של הדאטה שלכם תוך חיקוי התהליך שמחשנו בקורס של בחירת הכלים המיטביים להראות זו ושיכלול כלים קיימים במידה ויש בכך צורך.**

בנוסף – על הגרפים להיות קריאים, כתב קריא, צבעים משמעותיים, כותרות לצירים, figure caption מובנים.

**תאריך הגשה של השלב השני של העבודה: XX לא יינתנו הארכות מעבר.**

### **חלוקת הציון בפרויקט הסופי:**

30% - שלב א (פרזנטציה)

70% - שלב ב (העבודה הסופית) + מבחן הגנה על העבודה (הגנה בבירור, אעדכן)

שימו לב- שיינתנו נקודות (עד 10) גם ליצירתיות והתרשמות כללית, מעבר לנכונות המתמטית של העבודה.

סומכת עליכם 😊 שיהיה מעניין...

**בהצלחה!**