

## "מבוא למדעי הנתונים" – עבודת גמר

**הקדמה:**

מצ"ב פרטים על עבודת הגמר אשר יש להגיש. מטרת העבודה היא להכיר לסטודנט בצורה אישית ובלתי אמצעית את מגוון הנושאים אשר נלמדו בקורס. בין היתר, יש להבין את הנושאים שעברנו עליהם בקורס בצורה משמעותית וטובה. יש להגיש קישור לריפוזטורי בגיטהאב המכיל 2 מחברות גופיטר. מחברת אחת שנקראת על שם הדאטה שתבחרו, ומחברת נוספת שתקרא Apple\_Visualization עבור דאטה מצורף שנקרא AppleStore.csv. כמו כן, יש לכתוב קובץ readme או של גיטהאב או בוורד כפי שפורט בתרגול.

**2 משימות על:**

במסגרת עבודת הגמר, על כל סטודנט לבחור DATASET אחד מתוך האתר KAGGLE. כאשר הבעיה שתחקר היא בעיית סיווג (קלסיפקציה). כמו כן, מצורף לעבודה זו DATASET שישמש אותנו לדימות נתונים בלבד.

הערה: אם ישנם סטודנטים שהגישו במהלך הסמסטר את ארבעת המטלות בזמן(!), והיו מעוניינים שמטלות אלו לא תבדקנה ומבקום ינתן תוספת משקל לעבודה זו, נא לעדכן במייל "אישור הדאטה" עד התאריך 25.1.20 (מפורט בפסקה הבאה).

**עבור המחברת שלכם:**

במסגרת עבודת הגמר, על כל סטודנט לבחור DATASET אחוד מתוך האתר KAGGLE. כאשר הבעיה העיקרית שתחקר היא בעיית סיווג (קלסיפקציה), יש לאשר את הdata הנבחר מול המתרגל במייל, שימו לב שהמידע חייב להיות גדול, לא פחות מ 10,000 רשומות כמו גם לא פחות מ 10 פיצ'רים מסוגים שונים

(types of data ). הקובץ הוא קובץ csv בלבד (אלא אם אישרנו אחרת). אין להשתמש בדאטה אותו פגשתם במהלך הסמסטר. אין להשתמש בדאטה שמוגדר "לימודי" כמו "אירוסים" או "טיפים".

מטרת המחברת הזו היא להביא את כל ידיעותיכם בקורס לידי ביטוי. עליכם להציג את כלל תהליך עיבוד הדאטה עד הרצת המודלים וחיזוי. תזכורת, תהליך למידת מכונה הוא כדלקמן : 1.התאמת הדאטה 2. ניקיון הדאטה 3. חלוקת הדאטה לאימון וקבוצת מבחן. 4. אימון המודל והצגת ביצועי מודל (הערכת המודל)

בעת הרצת המודל יש להשתמש במשתנים טפשיים dummies , ויש להמיר כל משתנה מסוג object למספרי. יש לעשות את תהליך הסקיילינג scaling בכל מקרה .

עליכם לדעת להסביר, מדוע מחקתם פיצ'רים מסויימים מהדאטה המקורי, כאשר ההמלצה היא לא לעשות זאת, ובאופן כללי יותר, על כל מניפולציה שתעשו על הדאטה המקורי, חייב להיות הסבר מניח את הדעת. עבור הערכת ביצועי מודל, יש להבין בצורה מושלמת (!) האם המודל שלכם טוב או לא ולמה לא. יש להציג את מטריצת השגיאות בסופו של דבר ומומלץ לעשות זאת גם בצורת מפת חום עבור סיווג של שלושה או יותר משתנים דיכוטומיים.

**עבור מחברת חנות אפל:**

מצורף לכם דאטה עבור חנות האפליקציות של אפל לאחר שניקינו לכם את חלקו.

לדאטה זה ישנם כמה פיצ'רים כאשר שלושת הסוגים נמצאים בו (קטגוריאלי, אורדינלי, ונומרי)

חידוד על פיצ'רים שאולי לא ברורים :

סך כל המדרגים rating\_count\_tot

סך כל המדרגים בגרסה האחרונהrating\_count\_ver

הדירוג עבור הגרסה האחרונה בלבדuser\_rating\_ver

דירוג כללי עבור כל הגרסאותuser\_rating

מספר השפות הנתמכותlang.num

מספר המכשירים הנתמכים sup\_devices.num

יש לשרטט לפחות 12 גרפים משמעותיים (שמספרים סיפור) כמו גם דקורטיביים, לפי ההנחיות הבאות:

שישה גרפים במטפלוטליב ושישה גרפים בסיבורן.

1.חייבים להופיע גרפים אותם שרטטתם בגישה מונחית עצמים, וגם בגישה הפשוטה.

2.יש לשרטט לפחות שישה גרפים שונים (עדיפות לכך שכל 12 הגרפים יהיו מסוגים שונים).

3. פעם אחת עליכם לצייר על אותו הקנבס לפחות שני גרידים.

4.עליכם להציג את 15 האפליקציות המדורגות ביותר.

5.עליכם להציג את 20 האפליקציות עם הדירוג הגבוה ביותר.

עבור קטגוריית משחקים – 1. חינמיים. 2. בתשלום. Pairplot 6.מתוך 12 הגרפים, עליכם להציג 2

שתי הפעמים הללו הן היחידות בהן תוכלו להציג את סוג הגרף הזה.

כל השאר לבחירתכם.

את כלל הגרפים שהצגתם במחברת יש לייצא כקובץ ולהגיע עם 12 קבצים לפרזנטציה ולא רק עם המחברת.

הערה: גרפים שלטעמנו לא יהיו מספיק דקורטיביים יגררו הורדה בציון, להקפיד ביותר על "מקרא" , שנתות, ערכים, שמות לצירים, כותרת לגרף וכו'. גרפים אשר לא נלמדו הן בתרגול והן בהרצאה מבורכים ותינתן תוספת ניקוד על גרפים מהסוג הזה, כמו גם גרף בתלת מימד או גרפים אינטרקטיביים.

לצורך ביצוע משימת הדימות, מומלץ מאוד להכיר קצת יותר טוב את הדאטה, ולבצע פעולות הכנה וניקיון פשוטות באם תראו צורך, עם זאת, הדבר היחיד שנבדוק במחברת יהיה הדימות.

**פירוט המשימות:**

* נא לעבור על הקורס של git באתר udacity. . בנוסף, יש לשנות את המחברות בגיטהב שלכם דרך commits שונים.
* חשוב להציג את הנתונים **בצורה גרפית טובה שניתן ללמוד ממנה תובנות לגבי ה-DATA.** **זה חלק משמעותי מהציון.**
* יש להשתמש בין היתר באלגוריתמים שלמדנו במהלך הסמסטר. בנוסף כדאי לנסות להשתמש גם בעוד אלג' שלא נלמדו בין היתר: NAÏVE BAYES, , ADABOOST , עצי החלטה, Logistic Regression וכו'. שימוש בלפחות שנים נוספים הוא בגדר חובה, יש ללמוד מודל אחד נוסף לבד ומה הוא עושה ולהגיע מוכנים איתו לפרזנטציה עבור נשאל שאלות.
* מה מידת הדיוק של המודל שלכם יחסית למנצחים בתחרות KAGGLE? במה הוא יותר טוב ממודל 0?
* מומלץ להעזר באינטרנט, במחברות דומות מקאגל ובסטודנטים להתייעצות, אך בשום אופן אין להעתיק מהנ"ל וגם לא חלקי קוד. נבדוק זאת אל מול שאר המחברות בקאגל, ומול שאר עבודות הסטודנטים .

בהצלחה לכולם, ניתן לשאול שאלות או לבקש הבהרות בפורום התרגול.

רועי ועמית.