הפקולטה למנהל עסקים - מסלול למערכות מידע

ניתוח נתונים באמצעות PYTHON – תרגיל סיום

הוראות כלליות:

יש להגיש לאתר הקורס קובץ ipynb מלא על בסיס הקובץ Python_data_analysis_final.ipynb בו מופיעות שאלות תרגיל זה.

ניתן לבצע את התרגיל בקבוצות של עד 3 תלמידים.

שימו לב! לתרגיל שאלות מילוליות בנוסף למקטעי הקוד שיש למלא- הקפידו לענות על כל הסעיפים.

את התשובות המילוליות יש לכתוב באנגלית, במידה וקיימת בעיה עם האנגלית אנא צרפו תשובות מילוליות בעברית בקובץ word נפרד תוך הפניה לשאלה/סעיף המתאים בצורה ברורה.

את התרגיל יש להגיש לא יאוחר מה-12/2/2022 בשעה 23:59.

בכל מקום בו יש למלא מקטע קוד תופיע ההערה:

```
#Your code here:
```

השאלות במטלה מבוססות באופן ישיר על ההדגמות שראינו ותרגלנו בכיתה - העזרו בהן לצורך חזרה והדגמה.

לפני תחילת המטלה יש למצוא טבלת נתונים מסוג csv מאחד המקורות הבאים או מכל מקור אינטרנטי אחר. על הטבלה לכלול לפחות 5 משתנים (עמודות) כמותיים (מספריים) מסוג int או float, 2 עמודות קטגוריאליות (המורכבות ממחרוזות) ולפחות 1000 שורות (תצפיות או datapoints).

מוזמנים/ות לפנות אלי בכל בעיה של מציאת נתונים לעזרה.

מקורות אפשריים לנתונים:

- kaggle.com/datasets
- datasetsearch.google.com
- data.cdc.gov
- opendata.cityofnewyork.us/data

ניקוד השאלות: שאלות 1-2 10 נק', שאלות 3-6 20 נק'.

חלק א'- עיבוד נתונים (10 נק')

1. טענו את כל החבילות הרלוונטיות לפתרון התרגיל. הנחיה: לעיתים קשה לחזות מראש מהם כלל הכלים שנצטרך לטעון מתוך חבילות ומודולים ולכן נוסיף לטעינה במקטע הקוד הראשון חבילות/כלים/מודולים ככל שנצטרך אותם לאורך הקוד.
2. באמצעות מקטע הקוד המוגדר בקובץ פרויקט הסיום טענו את טבלת הנתונים שמצאתם/ן באינטרנט.

3. בעזרת הפונקציות/מתודות describe ו-shape הציגו סיכום של הנתונים. וודאו שכל העמודות מוצגות כולל הקטגוריאליות.

4. באמצעות הפונקציה pd.drop הסירו עמודות לא רלוונטיות (קטגוריאליות עם הרבה קטגוריות/ מזהה ייחודי כמו מספר זיהוי- הכל בתלות בקובץ הנתונים אותו בחרתם/ן). **שימו לב!** עליכם/ן להשאיר לפחות עמודה אחת קטגוריאלית (המורכבת ממחרוזות- מומלץ שלעמודה זו יהיו שתי קטגוריות בלבד). זכרו: ללא השמה – הורדת עמודה לא תישמר בטבלה. כלומר, עליכם ליצור השמה מחדש לטבלת הנתונים כדי לשמור את הטבלה הערוכה.

חלק ב': Matplotlib – גרפים בפיתון (עבור השאלות הבאות מומלץ להיעזר בקובץ ההדגמה: matplotlib_demo_and_exerecise.ipynb) (30 נק')

5. מן הטבלה החדשה שנוצרה, בחרו משתנה מספרי ותיצרו היסטוגרמה המציגה את התפלגות המשתנה. בהיסטוגרמה שנו את צבע "המקלות" לצבע אחר מהכחול של ברירת המחדל והגדירו שקיפות מחדש (פרמטר alpha בפונקציית הגרף).
6. בחרו שני משתנים רציפים והציגו גרף נקודות (כפי שמודגם מתחת לשאלה 3 בקובץ ההדגמה) או קווי מגמה (כמו בדוגמת הקוסינוס והסינוס) כפי שהדגם בשיעור.
במידה ובחרתם/ן גרף נקודות (פיזור XY), שנו את צבע הנקודות כך שיציגו משתנה שלישי והוסיפו כותרות לציר ה-X, ציר ה-Y וכותרת המתארת את הגרף.
במידה ובחרתם/ן גרף קווי מגמה, הציגו שני קווי מגמה לפחות (כמו בדוגמת הקוסינוס והסינוס) והוסיפו מקרא (legend) כותרות לציר ה-X, ציר ה-Y וכותרת המתארת את הגרף.

חלק ג'- למידת מכונה - machine learning: קיבוץ אשכולות/ clustering (עבור השאלות הבאות מומלץ להיעזר בקובץ ההדגמה: Kmeans_demo_mall.ipynb) (30 נק')

7. צרו dataframe חדש בשם df_numeric המכיל אך ורק משתנים מספריים (ראו דוגמא בתרגיל בית 1).
8. באמצעות df_numeric בצעו ניתוח אשכולות (קיבוץ לקבוצות) באמצעות אלגוריתם הקיבוץ kmeans כפי שהודגם בכיתה עבור K =5 אשכולות.
9. לאחר מכן, עבור אותה טבלת נתונים ובאמצעות הכלים שהודגמו בכיתה בצעו את שיטת המרפק (elbow) עבור טווח האשכולות (K) 2-12.
10. באופן מילולי באנגלית (1-2 משפטים) בגוף הקוד (או בעברית בקובץ וורד נפרד) ענו על השאלות הבאות:
מהו ה-K הטוב ביותר? כיצד קבעתם זאת? מהם המאפיינים של קיבוץ טוב לאשכולות?
11. עתה בצעו על אותה טבלה ובעזרת ההדגמה שבכיתה הצעו ניתוח סילואט (Silhouette) למציאת K החלוקות לאשכולות הטוב ביותר) עבור טווח האשכולות (K) 2-12. תזכורת: סילואט הוא מדד למרחק של נקודה משאר הנקודות במקבץ/אשכול (cluster) לעומת המרחק במקבץ/אשכול הסמוך. במדד זה 1 היא תוצאה מושלמת (הפרדה מלאה בין מקבצים, לעומת זאת, -1 היא הפרדה רעה מאוד. עליכם/לחפש את מספר המקבצים (K) שממוצע מדד הסילואט שלו יהיה גבוהה ככל הניתן.
12. ענו בקצרה (משפט אחד): האם שני המדדים מסכימים על מספר המקבצים (K) הטוב ביותר? אם כן, מהו? אם לא, האם יש לכם רעיונות כיצד לבחור את ה-K הטוב ביותר?

חלק ד'- למידת מכונה - machine learning: סיווג/ classification (עבור השאלות הבאות מומלץ להיעזר בקובץ ההדגמה: churn_logistic_regression.ipynb) (30 נק')

13. באמצעות טבלת הנתונים המקורית בה השתמשתם/ן, בצעו את השלבים הבאים:
   I. הסירו את עמודות קטגוריאליות בהן יותר מ-3 קטגוריות שונות.

   II. בדקו האם ישנם ערכים חסרים בטבלת הנתונים באמצעות הפונקציה שהודגמה בכיתה:
   df.isnull().sum()

   III. במידה וישנם הסירו אותם באמצעות הפונקציה
   df.dropna()
   IV. המירו את כל המשתנים הקטגוריאליים למשתנים מספריים כפי שהודגם בכיתה.

14. לצורך השלבים הבאים יש לבחור עמודה בה שני ערכים בלבד – היא עמודת החיזוי (Y). אם הנתונים שלכם מכילים עמודה עם שני ערכים (בדומה לעמודת ה-churn בדוגמא בכיתה) אז העבירו אותה להיות העמודה האחרונה כפי שמודגם בקוד. במידה ולא, צרו עמודה חדשה בשם Y ממשתנה מספרי קיים כך שתכיל שני ערכים, כל ערך מעל לחציון יקבל 1 וכל ערך מתחת לחציון יקבל 0 (כפי שביצעתם בתרגיל הבית 1 שאלה 4 סעיף B). לאחר מכן הסירו את העמודה המספרית המקורית ממנה יצרתם/ן את העמודה החדשה.
15. עתה צבעו את השלבים הבאים:
   a) חלקו את הטבלה לסט אימון וסט מבחן כפי שהודגם בכיתה (בעזרת הפונקציה: train_test_split) בשיעור של 80% אימון- 20% מבחן.
   b) אמנו מודל של רגרסיה לוגיסטית על סט האימון בלבד (בעזרת הפונקציה: LogisticRegression ועל גביה שימוש בפונקציה fit שמאמנת את המודל).
   c) חזו את ערכי משתנה החיזוי (Y) על סט המבחן (בעזרת הפונקציה: predict) ושמרו את תוצאת החיזוי תחת השם predictions.
   d) בעזרת הפונקציה score חשבו והדפיסו את דיוק המודל על סט המבחן.
   e) צרו את מטריצת הטעות (confusion matrix) באמצעות הפונקציה confusion_matrix, את המטריצה שתתקבל שמרו בשם cm.
   f) הדפיסו גרף המציג את מטריצת הטעות כפי שהודגם בכיתה.
16. ענו מילולית על השאלות הבאות: מהו הדיוק (accuracy) שהתקבל? כמה פעמים מודל החיזוי צדק כשחזה את קטגוריה 0 וכמה פעמים צדק כשחזה את קטגוריה 1?

בהצלחה!!