פרויקט קורס – 13401 ניטור סטטיסטי וזיהוי אנומליות

תשפג

הנחיות כלליות:

- קבוצות: העבודה היא קבוצתית – בקבוצות עד 3.
- מועד הגשה: כשבועיים אחרי סוף הסמסטר (מועד מדויק יפורסם במודל).
- תוצרים להגשה: קוד + מצגת (עד 20 שקפים).

תקציר הפרויקט:

מטרת הפרויקט היא יישום השיטות שנלמדו בקורס – הסטודנטים מתבקשים לפתח ולהציג מדיניות שנועדה לענות על דרישה מוגדרת הייטב בכדי לקבל החלטה עסקית או תפעולית. הפרויקט (ותוצריו) יורכבו מ4 מרכיבים: בחירת מערך נתונים, הצבת דרישה (שניתן לענות עליה בעזרת הכלים שנלמדו בקורס), הצגת 3-2 תוכניות אלטרנטיביות שיתנו מענה לדרישה (ויכלו לתת מענה לאותה דרישה גם בעתיד), והשוואת התוכניות (בעזרת מערך הנתונים שנבחר) בכדי לנמק יישום של אחת מהן. נושא הפרויקט פתוח במהותו – על הסטודנטים להפגין יצירתיות בהתאמת השיטות שנלמדו לעולמות תוכן ונתוני אמת. הציון הסופי יינתן על סמך התאמת התוכנית לעולם התוכן של הנתונים, בהירות התקשורת בתוצרים (במצגת ובהערות בקוד), והפגנת חשיבה עמוקה ויצירתית.

פירוט על מרכיבי הפרויקט:

בחירת מערך נתונים

יש לבחור אחת מהאופציות:

1. מכירות ברשת חנויות
   - https://www.kaggle.com/competitions/rossmann-store-sales/data
   - יש להשתמש בקובץ TRAIN ובקובץ STORES (בשניהם) ולהתמקד בערכי המכירות וגם בכמות הצרכנים.
   - יש לייצר תכנית שתתאים לכל סוג אגרגציה (לדוגמא, לפי גודל חנות, היקף מכירות, מיקום גאוגרפי וכו), אך מספיק להציג יישום ממוקד במצגת.
   - ניתן לפצל את הנתונים למערך TRAIN וTEST (לדוגמא, לפי קבוצת אגרגציה שונה).
   - הקשר לשאלת הפרויקט: המלצה למנהלי הרשת בנושא אסטרטגיית השקעות ותמחור.
2. פופולריות של פוסטים
   - http://archive.ics.uci.edu/ml/datasets/News+Popularity+in+Multiple+Social+Media+Platforms
   - יש לייצר תכנית שתתאים לכלל הפלטפורמות, אך מספיק להציג יישום ממוקד במצגת.
   - ניתן לפצל את הנתונים למערך TRAIN וTEST (לדוגמא, לפי פלטפורמות).

- הקשר לשאלת הפרויקט: זיהוי מקדים של פוסטים שעתידים להפוך ויראליים, למטרת REPOST ופרסום.

3. שינוי מצב רוח
   - https://ubicomp.eti.uni-siegen.de/home/datasets/icmi18/
   - יש לייצר תכנית שתתאים לכל זוג רגשות, אך מספיק להציג יישום ממוקד במצגת.
   - ניתן לפצל את הנתונים למערך TRAIN וTEST (לדוגמא, לפי סוג רגש).
   - הקשר לשאלת הפרויקט: זיהוי מהיר של שינויי מצב רוח, למטרת הצלבה עם נתונים חיצוניים.
4. אחר
   - ניתן להציע מערך נתונים אחר (אם זמין פומבית באינטרנט = +5 נק בונוס) – יש להגיש בקשה מיוחדת למתרגל הקורס לאישור.
   - הבקשה המיוחדת צריכה לכלול: מערך הנתונים (או קישור אליו), הקשר לשאלת הפרויקט, והצדקה לבחירה במערך הנתונים האחר.
   - לא ניתן לעבוד עם מערך נתונים שלא אושר!

<u>הצבת דרישה</u>

לאחר שנבחר מערך נתונים, יש להגדיר דרישה שניתן לענות עליה בעזרת השיטות שנלמדו בקורס. הדרישה צריכה להתייחס הן לעבר (שניתן לראותו בנתונים) והן לעתיד (בהנחה שנתונים זהים יאספו בצורה רציפה בעתיד):

- הדרישה צריכה להיות מנוסחת כמטרה!
- אם יש צורך בהגדרות חדשות (של מונחים או פרמטרים), יש לנמקן.
- יש להצדיק את חשיבות הדרישה מבחינה כלכלית.
- ניתן להיעזר במקורות מידע חיצוניים לביסוס טענותיכם.
- ההיו יצירתיים – נסו "למכור" את חשיבות המענה על הדרישה המוצגת.

<u>הצגת תוכניות אלטרנטיביות</u>

עליכם לפתח 2-3 תוכניות עבודה שיתנו מענה לדרישה בהווה (בעזרת הנתונים), ולהסביר את ישומן בעתיד (כאשר נתונים חדשים יהיו זמינים). התוכניות המוצגות חייבות להתבסס על שיטות מ-2 יחידות לימוד שונות בקורס. לדוגמא: תכנית אחת יכול להיות תכנית דגימה לפי תכונות, ותוכנית אחרת לפי שיטות זיהוי אנומליות. ניתן לשלב (לשרשר) שיטות כנדרש. לגבי כל תכנית יש להציג את:

- האינפורמציה הנדרשת.
- פירוט שלבי תפעול התוכנית וקבלת המסקנות.
- הצגת יישום התוכנית בעזרת מערך הנתונים (כלל החישובים הנדרשים).

<u>השוואה ומסקנה</u>

יש להשוות בין התוכניות האלטרנטיביות המוצגות תוך התייחסות למאפייני התוכניות והמדדים שהוצגו בקורס. מטרת ההשוואה היא הצגת יתרונות וחסרונות של כל אחת מהתוכניות המוצעות. לאחר ההשוואה, יש לבחור התוכנית המועדפת ולנמק את הבחירה. ניתן להשתמש בנתונים חיצוניים כנדרש (לייצור הנימוקים) ו\או להתייחס לתרחישים שונים (לדוגמא, תרחיש בו עלות דגימה נמוכה מול תרחיש בו עלות דגימה היא גבוהה). חשוב להתייחס לתפעול התוכנית בהקשר מציאותי.

תוצרים להגשה:

קוד:

- הקוד ייבדק בעזרת הרצה מלאה עם מערך הנתונים הנבחר.
- יש לכתוב הערות בקוד (כך שיהיו מובנות לכל משתמש).
- יש לציין פלטים רלוונטיים בבירור.

מצגת:

- חייבת לכלול התייחסות לכל מרכיבי הפרויקט.
- יש להניח שקהל היעד מכיר את הקשר הנתונים אז לא בקיא בתוכנם.
- יש להניח שקהל היעד אינו מכיר את השיטות המוצעות – יש להסביר כל שלב בבהירות.

הערות:

מיקוד הפרויקט הוא יישום הנושאים שנלמדו בקורס – לא יכולת תכנות או הנדסת נתונים.

מערכי הנתונים שהוצאו גדולים מאד – ניתן לבחור חלק מצומצם מהנתונים ולהתמקד בהם. מערכי הנתונים גדולים דיו בכדי שלא יהיו קבוצות שעובדות על נתונים זהים – יש להימנע מתיאום עבודה בין קבוצות.

ניתן לשלב ידע מקורסים קודמים – אך להתמקד בניתוח בעזרת השיטות שנלמדו בקורס זה.

נא להפנות שאלות בנושא הפרויקט למתרגל הקורס.