

המכללה האקדמית תל אביב יפו
בית הספר לכלכלה וניהול
החוג למערכות מידע

274077 ,Big Data

ד"ר גלית גורדוני

תשע"ח, סמסטר א', מועד ב', 29/03/2018

--	--	--	--	--	--	--	--	--

מספר זהות

משך הבחינה: 3 שעות

ללא חומר עזר

בבחינה 20 שאלות, יש לענות על כולן. משקל השאלות זהה - 5 נקודות לשאלה.
בכל שאלה יש לסמן את התשובה הנכונה (אחת בלבד), באופן ברור, בדף התשובות בלבד. אם סימנת יותר מתשובה אחת - השאלה תיפסל!
יתקבלו תשובות שנכתבו על גבי דף התשובות בלבד. לא ייבדקו תשובות שלא ייכתבו בדף המיועד לכך.

סימון תשובות בטופס הבחינה לא יהווה בסיס לערעור.

יש להחזיר את דף התשובות ביחד עם טופס הבחינה, אחרת הבחינה לא תיבדק.

**נא לשים לב כי דף התשובות מוכן לטופס בחינה של 6 תשובות. בבחינה זו יש 5 תשובות לכל היותר.
יש להקפיד למלא את התשובות במקום הנכון!**

בהצלחה!

לפניך נתונים לחמש השאלות הבאות:

פרס אוסקר הנו פרס המוענק על ידי האקדמיה האמריקאית לקולנוע ליוצרים ושחקנים המועמדים בקטגוריות שונות. באקדמיה יש כ-6000 חברים בעלי זכות הצבעה ורק הם יכולים לבחור את הזוכים בפרסים בכל קטגוריה. תוצאות ההצבעה האנונימית מפורסמות בטקס האוסקר השנתי. לקראת אחד מטקסי האוסקר שהתקיים בשנים האחרונות ניסו שתי חברות טכנולוגיה ישראליות לנחש, באמצעות ניתוח ביג דאטה, מי יזכה בפרסים הבולטים.

שאלה מספר 1:

החברה הראשונה ביססה את התחזית שלה על הנוכחות החזקה ביותר ברשת, מבחינה ויזואלית, בחמש הקטגוריות העיקריות באוסקר: הסרט הטוב ביותר, השחקן הטוב ביותר, השחקנית הטובה ביותר, שחקן המשנה הטוב ביותר ושחקנית המשנה הטובה ביותר. החברה ביצעה ניתוח סימולטני של מאות מיליוני תמונות באתרים, בלוגים, אתרי נישה, תרבות, חדשות, רשתות חברתיות וכדומה. הממצאים הראו לדוגמא, שלשחקן אחד הייתה פי-3 נוכחות ברשת לעומת שחקן אחר.

מהי הפרשנות המתאימה ביותר לתחזיות של חברת הטכנולוגיה לגבי הזוכים העתידיים בפרסי האוסקר?

- א. מודל החיזוי לא מתבסס על נתונים שנאספים מהאוקלוסייה המתאימה ולכן יש כאן בעיה של סלקטיביות
- ב. התחזית אמורה להיות מדויקת מכיוון שהיא מבוססת על נתונים הנאספים בזמן-אמת
- ג. התחזית אמורה להיות מדויקת מכיוון שנעשה שימוש בביג דאטה ולא בסקרים
- ד. התחזית אמורה להיות מדויקת מכיוון שהטכנולוגיה מודדת התנהגות בעבר לשם ניבוי התנהגות עתידית
- ה. התחזית אמורה להיות מדויקת מכיוון שהיא מבוססת על ניתוח תמונות ולא על ניתוח טקסטים

שאלה מספר 2:

החברה השנייה שניסתה לחזות את זהות הזוכים ביססה את התחזית על טכנולוגיית בינה מלאכותית שפיתחה. הטכנולוגיה מאפשרת למדוד את סוג ועוצמת הרגש שגולשים מביעים בעת שיח על המועמדים לזכייה. החברה בחנה אילו שחקנים מעוררים הכי הרבה רגשות חיוביים. עבור כל שחקן ושחקנית חושב אחוז התגובות המכילות רגשות חיוביים כלפיהם מתוך כלל התגובות כלפיהם. התחזית לגבי הזוכה בקטגוריה מסוימת התבססה על דירוג כל מועמד או מועמדת בהתאם לאחוז התגובות עם רגשות חיוביים שקיבלו ברשת בהשוואה לשאר המועמדים.

מהי הפרשנות המתאימה ביותר לתחזיות של חברת הטכנולוגיה לגבי הזוכים העתידיים בפרסי האוסקר?

- א. התחזית לא מתבססת על נתונים שנאספים מהאוקלוסייה המתאימה ולכן יש כאן בעיה של סלקטיביות
- ב. התחזית אמורה להיות מדויקת מכיוון שהיא מבוססת על נתונים הנאספים בזמן-אמת
- ג. התחזית אמורה להיות מדויקת מכיוון שהיא מבוססת על סקרים המודדים את העדפות הצופים בסרטים השונים
- ד. התחזית אמורה להיות מדויקת מכיוון שהטכנולוגיה מודדת התנהגות ולא דעות אישיות
- ה. התחזית אמורה להיות מדויקת מכיוון שהיא מבוססת על ניתוח סנטימנט ולא על ניתוח תמונות

שאלה מספר 3:

האם התחזיות של שתי חברות הטכנולוגיה לגבי הזוכים בפרסי האוסקר אמורות להיות זהות?

- א. כן, מכיוון שהן מנסות לחזות את אותה תופעה
- ב. לא, מכיוון שהן עושות שימוש בסוג אחר של נתונים
- ג. לא, מכיוון שלא ניתן להשוות בין ניתוח תמונות לניתוח טקסטים
- ד. לא, מכיוון שהתחזיות מתבססות על גודל מדגם שונה
- ה. אין מספיק מידע לגבי תהליך יצירת המידע ותהליכי הבקרה על איכות הנתונים כדי לקבוע אם התחזיות אמורות להיות זהות

שאלה מספר 4:

מדדי הרייטינג של חברת נילסן, המבוססים על ניתוח שיח ברשתות החברתיות, מוצגים כחלופה לשימוש בסקרים. ספקות רבים מועלים בנוגע לביסוס החלטות, כמו זמני שידור של תכניות טלוויזיה והשקעת תקציבים בהפקת תכניות שונות, על השיח ברשתות החברתיות במקום על סקרים.

מבין הספקות הבאים, איזה מתייחס לחסרון האפשרי של שימוש בנתונים מהרשתות החברתיות?

- א. האם הגולשים ברשתות החברתיות הם מדגם מייצג של כלל האוכלוסייה הרלבנטית?
- ב. כיצד מושפע מדד הרייטינג של חברת נילסן מההבדלים הקיימים ברשת בכמות התגובות לתוכניות טלוויזיה שונות?
- ג. מהי יחידת הניתוח במאגר המשמש לניתוח השיח ברשתות החברתיות?
- ד. עד כמה הדעות שגולשים מעלים לרשתות החברתיות משקפות את הדעות האמתיות שלהם?
- ה. עד כמה הגולשים ברשתות החברתיות מאמינים לדעות גולשים אחרים ברשת על תכניות טלוויזיה?

שאלה מספר 5:

בדיון שנערך על איכות התחזיות של שתי חברות הטכנולוגיה, לגבי הזוכים בפרסי האוסקר, בהשוואה לאיכות מדד הרייטינג של חברת נילסן, הועלו מספר טענות. מהי הטענה המתאימה ביותר?

- א. כל הטענות נכונות
- ב. בכדי להעריך את איכות התחזיות ומדד הרייטינג נדרש מידע מפורט על תהליכי איסוף המידע, מקורות המידע, שלב ה-ETL והנתונים שבהם נעשה שימוש במודלים לחיזוי
- ג. מדד הרייטינג של נילסן עלול לסבול מסלקטיביות, מכיוון שלא כל צופי הטלוויזיה פעילים ברשתות החברתיות
- ד. בניתוח סנטימנט קיימות טעויות שמקורן בסיווג לא מדויק של השיח
- ה. התחזיות של הזוכים בפרסי האוסקר מבוססות על נתונים שלא נוצרו לצרכי מחקר

לפניך נתונים לחמש השאלות הבאות:

אחת המגמות המאפיינות את השימוש בביג דאטה, הנה של חברות אשראי, בנקים ובנקים למשכנתאות העושים שימוש ב- predictive analytics כדי לקבל החלטות בנוגע ללקוחות ספציפיים. לדוגמא, השוואה בין הלקוח ללקוחות אחרים, בעלי מאפיין זהה, כדי לחזות האם אותו לקוח יצליח לעמוד בהתחייבויות.

שאלה מספר 6:

בבנק למשכנתאות מתכננים לעשות שימוש בנתונים פנימיים של הבנק על פעילות עבר של הלקוחות ומאפייניהם לשם קבלת החלטות על מתן הלוואות לרכישת דירה ללקוחות ספציפיים. המודל כלל השוואה בין הלקוח שביקש הלוואה לבין לקוחות אחרים בעלי מאפייני אשראי זהים כמו למשל, ממוצע סכום הרכישות החודשי ומידע היסטורי על מידת העמידה לאורך זמן של הלקוחות בהתחייבויות האשראי שלהם. המודל שימש לחיזוי הסיכוי שהלקוח יעמוד בהתחייבויות העתידיות של המשכנתא.

במחלקת המחקר התלבטו לגבי כדאיות הוספת נתונים חיצוניים למודל החיזוי, כדוגמת מידע על תחביבים, השכלה, ניסיון בשוק העבודה ותחומי עניין של הלקוח מתוך הפרופיל של הלקוח ברשת החברתית. מהו המידע שאמור להופיע בשורות בקובץ הנתונים?

- הלקוח הוא יחידת הניתוח ולכן בכל שורה אמור להופיע לקוח אחד
- בשורות אמור להופיע ישוב המגורים של הלקוח
- בשורות אמור להופיע מידע על התנהגות עבר של הלקוח. בכל שורה אמור להיות מאפיין אחר כמו ממוצע סכום הרכישות החודשי
- הסיכוי שהלקוח יעמוד בהתחייבויות המשכנתא העתידיות הוא יחידת הניתוח ולכן בכל שורה יופיע סכום ההחזר החודשי העתידי
- בשורות יופיעו רק לקוחות שקיבלו הלוואות בעבר. הניתוח מתייחס למידע היסטורי ולכן בכל שורה אמור להופיע לקוח שקיבל בעבר הלוואה

שאלה מספר 7:

מהו המידע הנדרש כדי למזג בין הנתונים הקיימים בבנק לנתונים חיצוניים מהרשת החברתית?

- בנתונים הקיימים ובנתונים החיצוניים אמור להופיע לפחות משתנה זהה אחד שיאפשר לזהות את הלקוח הספציפי
- בנתונים הקיימים ובנתונים החיצוניים אמור להופיע רק מידע משלים שאינו חופף
- בנתונים הקיימים ובנתונים החיצוניים אמור להופיע מידע מפורט וזהה על התנהגות הלקוחות שיאפשר להשוות את מידת הדיוק של הנתונים בין מקורות המידע השונים
- בנתונים הקיימים ובנתונים החיצוניים אמור להופיע מידע ברמת המקור
- המידע הנדרש למיזוג הנתונים הוא שם סניף הבנק של הלקוח. המידע אמור להופיע גם ברשת החברתית וגם בנתוני הבנק

שאלה מספר 8:

המתנגדים להוספת נתונים חיצוניים למאגר הנתונים הפנימי של הבנק טענו שהעובדה שרק לחלק מהלקוחות יש פרופיל ברשתות חברתיות תגרום לפגיעה באיכות הנתונים. מהו מרכיב הטעות בביג דאטה שבו מתמקדת הביקורת?

- טעות בתאים הנובעת מנתונים חסרים (cell errors-missing data)
- טעות בתאים הנובעת מטעות אפיון (cell errors-specification error)
- טעות בשורות הנובעת מכפילות (raw error-duplication error)
- טעות בשורות הנובעת מסלקטיביות (raw error-selectivity)
- טעות בטורים הנובעת מטעות בתיוג (column error-erroneous labeling)

שאלה מספר 9:

ביקורת נוספת התייחסה לכך שגם אם ללקוחות יש חשבון פעיל ברשתות חברתיות יתכן והמידע המופיע בו לא עדכני או לא אמין. מהו מרכיב הטעות בביג דאטה שבו מתמקדת הביקורת?

- א. טעות בתאים הנובעת מטעות בתוכן (cell errors-content error data)
- ב. טעות בתאים הנובעת מטעות אפיון (cell errors-specification error)
- ג. טעות בתאים הנובעת מנתונים חסרים (cell errors-missing data)
- ד. טעות בטורים הנובעת מטעות בתיוג (column error-erroneous labeling)
- ה. טעות בשורות הנובעת מסלקטיביות (raw error-selectivity)

שאלה מספר 10:

לו הייתם חלק מצוות המחקר, האם הייתם ממליצים לבנק לאסוף מטוויטר נתוני ציוצים של לקוחות הבנק, על דפוס הרכישות שלהם באשראי, ולכלול את המידע במודל החיזוי?

- א. לא, מכיוון שצפויה טעות בתאים הנובעת מנתונים חסרים (cell errors-missing data)
- ב. לא, מכיוון שצפויה בטורים טעות הנובעת מטעות בתיוג (column error-erroneous labeling)
- ג. כן, מכיוון שציוצים בטוויטר משקפים התנהגויות בזמן אמת ומעודכנים יותר מהנתונים הפנימיים של החברה
- ד. כן, מכיוון ששימוש בנתונים מטוויטר יאפשר לחברת האשראי לצמצם טעות בתאים הנובעת מנתונים חסרים (cell errors-missing data)
- ה. כל התשובות נכונות

לפניך נתונים לחמש השאלות הבאות:

בינואר 2016 פורסם על ידי נציבות הסחר הפדרלית בארצות הברית (Federal Trade Commission) דו"ח המתמקד ביתרונות הרבים הגלומים בשימוש בביג דאטה על ידי חברות עסקיות ובהשלכות שליליות אפשריות. בכדי למקסם את היתרונות של שימוש בביג דאטה ולצמצם את ההשלכות השליליות, נוסחו מספר נושאים שחברות עסקיות אמורות לבחון לפני שהן עושות שימוש בביג דאטה.

שאלה מספר 11:

- הנושא הראשון התמקד בשיקול של עד כמה מאגר הנתונים מייצג. מהי הטענה המתאימה ביותר לתיאור הנושא?**
- ניתוח נתונים המבוסס על מאגר מידע שחסרים בו נתונים על אוכלוסיות מסוימות, כמו למשל אנשים השומרים על פרטיותם, אנשים מעוטי הכנסה הצורכים פחות ומעורבים פחות בפעילויות כלכליות, אנשים שאין להם גישה לטכנולוגיה או שיש להם פחות יכולות טכנולוגיות, עלול להוביל לחישוב אומדנים מוטים
 - מאגרי נתונים המוגדרים כביג דאטה כוללים את כל האוכלוסייה ולכן שאלת הייצוג לא רלבנטית
 - מאגרי נתונים המוגדרים כביג דאטה כוללים עקבות התנהגויות דיגיטליות ולא דיווח עצמי על התנהגות ולכן נושא הייצוגיות לא רלבנטי
 - מאגרי נתונים המוגדרים כביג דאטה כוללים נתונים הנוצרים כשלעצמם לכן הנתונים לא אמורים לייצג את האוכלוסייה
 - אף תשובה לא נכונה

שאלה מספר 12:

הנושא השני התמקד בשיקול של עד כמה מדויק צפוי להיות הסבר המבוסס על ביג דאטה. מהי הטענה המתאימה ביותר לתיאור הנושא?

- השימוש בביג דאטה מתאים מאוד לזיהוי מתאמים (קורלציות בין שני משתנים) אך לא מאפשר לבדוק סיבתיות
- ביג דאטה הם נתונים הנוצרים לכשעצמם ולכן הם פחות מדויקים מנתונים הנוצרים לצרכי מחקר
- ביג דאטה הם נתונים העוברים תהליך ארוך של כרייה, בניית משתנים חדשים וסיון נתונים לא רלבנטיים מתוך נתונים רלבנטיים. התהליך כרוך בטעויות רבות ולכן לא ניתן לבצע חיזוי מדויק
- הסבר המבוסס על ביג דאטה הוא סיבתי מכיוון שביג דאטה כוללת את כל המידע הקיים ולא מדגם
- כל התשובות נכונות

שאלה מספר 13:

הנושא השלישי התמקד בשיקול של אפשרות ליצירת אפליה בתהליך סינון מועמדים לעבודה המבוסס על ניתוח נתונים היסטוריים. מהי הטענה המתאימה ביותר לתיאור הנושא?

- כל התשובות נכונות
- השימוש באלגוריתמים המבוססים על נתונים היסטוריים כמו למשל, נתוני העובדים המועסקים בחברה עלול לגרום לאפליה בקבלה לעבודה. לדוגמא, אם לפי מודל החיזוי, מוסד הלימודים האקדמי שבו למד העובד נמצא במתאם גבוה עם משך ההישארות שלו במקום העבודה, השימוש באלגוריתם עלול לגרום להעדפת מועמדים שלמדו באותם מוסדות אקדמיים בהם למדו העובדים המועסקים בחברה. מועמדים מתאימים עלולים להידחות רק משום שלמדו במוסדות אקדמיים אחרים מאלו של העובדים המועסקים בחברה
- היכולת למזג נתונים על המועמד מהרשתות החברתיות יחד עם קורות החיים ששלח ולסנן מועמדים מתאימים על בסיס מאפיינים אישיים שלא רלבנטיים לקבלה לעבודה, היא אחד מהסיכונים הקיימים בשימוש בביג דאטה
- השימוש בביג דאטה עלול לגרום לאי שוויון בקבלת עובדים. לדוגמא, אם האלגוריתם לחיזוי מידת ההתאמה של מועמדים לעבודה מבוסס על נתונים של עובדים המועסקים בחברה, ניתן יהיה לחשב לכל מועמד את סיכויי ההצלחה שלו בתפקיד לפי מידת הדמיון בינו לבין פרופיל העובדים המועסק בחברה, ולסנן מועמדים על סמך שיקולים הידועים רק למעסיק
- שימוש באלגוריתמים לקבלת החלטות על סינון מועמדים עלול לגרום לאפליה בקבלה לעבודה של עובדים מתאימים מכיוון שהמודלים לחיזוי כוללים רק את הנתונים הנגישים למעסיקים, כמו למשל מאפייני העובדים המועסקים בחברה, שיתכן ולא מייצגים את כלל המועמדים המתאימים לעבודה

שאלה מספר 14:

אחת הדוגמאות שניתנה בדו"ח של נציבות הסחר הפדרלי בארצות הברית, התמקדה בחיזוי של גוגל את ממדי התפרצות מחלת השפעת בארצות הברית. מהי הטענה המתאימה ביותר לתיאור האתגרים בחיזוי ממדי התפרצות המחלה לאורך זמן?

- א. האלגוריתם לחיזוי היה מדויק, אולם המלצות לחיפוש של גוגל, בעקבות החיפוש של מילות מפתח הקשורות לשפעת, גרמו לכך שלאורך זמן האלגוריתם הצביע על אומדנים מוגזמים
- ב. האלגוריתם לחיזוי ממדי התפרצות מחלת השפעת בארצות הברית לא היה מדויק מכיוון שהיה מבוסס על מילות חיפוש שהקלידו גולשים בגוגל, הכוללות טעויות איות רבות
- ג. האלגוריתם לחיזוי ממדי התפרצות מחלת השפעת בארצות הברית הפך משנה לשנה למדויק יותר, מכיוון שמספר האנשים שחיפשו מידע בגוגל על שפעת עלה
- ד. האלגוריתם לחיזוי ממדי התפרצות מחלת השפעת בארצות הברית כלל טעויות בחישוב האומדנים והגדרות לא נכונות של משתנים ולכן התחזיות לא היו מדויקות
- ה. האלגוריתם היה מבוסס על נתונים בזמן אמת על כל האנשים שחלו בשפעת, כך שניתן היה לחזות במדויק את התפרצות המגפה בקרב קבוצות סיכון באוכלוסייה כמו ילדים וקשישים

שאלה מספר 15:

בעיריית בוסטון פותחה אפליקציה המכונה Street Bump. האפליקציה חנימית וניתנת להורדה מאתר העיריה לטלפון חכם. האפליקציה עושה שימוש בפונקציות קיימות של הטלפון החכם, כדוגמת GPS, לשם איסוף נתונים על מפגעי תשתית בשכונות מגורים בעיר. עם תחילת השימוש באפליקציה התברר כי מרבית הפניות מגיעות משכונות שבהן מתגוררת אוכלוסייה במעמד כלכלי גבוה. בעיריית בוסטון הוחלט לצייד פקחים של העירייה באפליקציה ולשלוח אותם לסייר באופן שוטף, באזורי מגורים פחות מבוססים, כדי לדווח על בעיות בתשתית. כיצד התמודדות העירייה עם בעיית האי שוויון בנגישות לאפליקציה אפשרה לשפר את איכות הנתונים המופקים באמצעותה?

- א. צמצום טעות בשורות הנובעת מסלקטיביות (raw error-selectivity)
- ב. צמצום טעות בתאים הנובעת מטעות בתוכן (cell errors-content error data)
- ג. צמצום טעות בתאים הנובעת מטעות אפיון (cell errors-specification error)
- ד. צמצום טעות בתאים הנובעת מנתונים חסרים (cell errors-missing data)
- ה. צמצום טעות בטורים הנובעת מטעות בתיג (column error-erroneous labeling)

לפניך נתונים לחמש השאלות הבאות:

בין האפשרויות הרבות לשימוש בביג דאטה מושם כיום על פיתוח מערכות להמלצה אישית בתחומים שונים. בשאלות הבאות מופיעים נתוני ביג דאטה בהם נעשה שימוש במערכות להמלצה אישית.

שאלה מספר 16:

במערכת להמלצה אישית על מסלולי רכיבה, הליכה וריצה נכללו באלגוריתם נתונים כמו גיל האדם, מגדר ומשקל, רמת זיהום האוויר ונפח התחבורה בשעות שונות. כיצד ניתן למזג בין הנתונים ברמות הקבוצה השונות לשם הרצת האלגוריתם ומתן המלצה אישית לאדם?

- א. נתונים על זמן ומיקום גאוגרפי הם נתונים המאפשרים זיהוי של מידע גם ברמת המקור וגם ברמת המיקור. נתוני המיקום האישי מהטלפון החכם בנקודת זמן ספציפית ניתנים למיזוג עם נתוני מקור בזמן ספציפי על אזור ההמצאות של האדם.
- ב. לא ניתן למזג בין נתוני מקור ומיקור
- ג. ניתן למזג בין נתוני מיקור לבין נתוני מקור רק אם מקבצים את הנתונים האישיים של אנשים לנתוני מקור ויחידת הניתוח היא ברמת המקור. המגבלה היא שלא ניתן יהיה להגיע להמלצה אישית לאדם ספציפי
- ד. ניתן למזג נתוני מיקור רק עם נתוני מיקור ולכן לא ניתן יהיה להריץ את האלגוריתם
- ה. ניתן למזג נתוני מקור רק עם נתוני מקור ולכן לא ניתן יהיה להריץ את האלגוריתם

שאלה מספר 17:

במערכת להמלצה אישית בתחום רכישות בגדים נכללו באלגוריתם נתונים כמו גיל האדם, מגדר ומשקל, העדפות אישיות לסגנון לבוש ודפוסי רכישה מקוונים באתרי בגדים. מהי יחידת הניתוח שאמורה להופיע בשורות?

- א. לקוח
- ב. דגם של בגד ספציפי
- ג. אתר מקוון לרכישת בגדים
- ד. מדינת המוצא של הלקוח אליה הוא נוהג להזמין משלוחי בגדים שרכש
- ה. סכום הבגד

שאלה מספר 18:

במערכת המלצה אישית על אירועי תרבות נכללו באלגוריתם נתונים על הטעם המוזיקלי של הלקוח, תחומי עניין, גיל, מזג האוויר, תגובות ברשתות החברתיות אחוז מכירת הכרטיסים לאירוע, מיקום האירוע ומיקום הלקוח. כיצד תושפע יעילות מערכת ההמלצות במצב שבו אין מידע על המיקום הגאוגרפי של חלק גדול מהלקוחות?

- א. היעילות תפגע מכיוון שרק הלקוחות שניתן לקבל עליהם נתוני מיקום ספציפי יוכלו לקבל את ההמלצה האישית
- ב. המידע על מיקום לא אמור להשפיע על האלגוריתם שעליו מבוססת מערכת ההמלצות האישית
- ג. היעילות לא תושפע. המיקום הגאוגרפי של הלקוח הוא מידע ברמת המיקור ולכן הוא לא רלבנטי להמלצה על אירועי תרבות
- ד. היעילות לא תושפע כל עוד יש נתונים על מיקום האירוע
- ה. אף אחת מהתשובות לא נכונה

שאלה מספר 19:

בבניית מאגר הנתונים עבור מערכת ההמלצות האישית נעשה שימוש בשלושה ממדים: זמן, מקום ונושא. מבין התהליכים הבאים איזה תהליך מתייחס לשלב ה-ETL?

- א. בניית משתנים ברמת המקור באמצעות נתונים על ציורים בטוויטר
- ב. חישוב אלגוריתם להמלצה אישית
- ג. ביצוע ניתוח היסטורי לאיתור דפוסי קשר בין משתנים
- ד. חישוב מדדים תיאוריים
- ה. כריית הנתונים

שאלה מספר 20:

מערכת המלצות אישית לחולי אסטמה כללה אפליקציה לטלפון הנייד השולחת התראות אישיות למשתמשים בהתאם למאפייניהם האישיים. בזמן מצבי סיכון, האפליקציה שולחת למשתמשים התראה להישאר בבית או להימנע ממאמץ גופני ובימים אחרים לקבל המלצות לצאת לפעילות גופנית מחוץ לבית. מה נדרש כדי לשלוח התראה או המלצה אישית מדויקת למשתמשי האפליקציה?

- א. מיזוג בין נתוני מיקרו של הלקוח עם נתוני מקרו שונים כמו רמת הסיכון בטווח זמן מסוים באזור גאוגרפי ספציפי לאלרגיה נשימתית
- ב. מידע ברמת המקרו על כל הישובים שבהם התגורר הלקוח במהלך חייו
- ג. מידע על תדירות הגלישה של הלקוח ברשתות החברתיות
- ד. נתונים ברמת המקרו ומידע על המיקום המדויק בקואורדינטות של בית הלקוח
- ה. כל התשובות נכונות

--- סוף המבחן ---