בינה עסקית

מבחן סופי - מועד ג'

סמסטר א' תשע"ח

מרצה: פרופסור אילן שמשוני

מתרגל: מר אבנר אינוז

משך הבחינה שלוש שעות.

ניתן להשתמש בכל חומר עזר כתוב ומחשב כיס.

בבחינה **ארבע** שאלות.

יש לענות על השאלות **במחברת המצורפת**.

שים לב כי בידך 5 דפים כולל דף זה.

הבחינה מיועדת לגברים ונשים כאחד ומנוסחת בלשון זכר מטעמי נוחות בלבד.

שאלה	ניקוד מקסימאלי	ציון
1	40	
2	25	
3	30	
4	15	
סה"כ	110	

בהצלחה!

1) Data Warehouse, MDX, ETL (40 נקודות)

חברת "רפרפת" מחזיקה ברשותה מחלבות ברחבי העולם לצורך ייצור ומכירת בקבוקי חלב. כל מחלבה מחזיקה ברשותה פרות בהתאם לשטח האכלוס שברשותה ובהתאם לסוג הפרות שבסביבתה. פרה יכולה להיות מסוג אחד או יותר (במקרה בו מדובר בפרה מעורבת), אך בכל מקרה כל פרה נמצאת במחלבה אחת בלבד.

כדי לייצר את בקבוקי החלב, כל מחלבה שואבת חלב מהפרות שברשותה כאשר כל שאיבה מתבצעת מפרה אחת בלבד, מעבירה אותו תהליך פסטור (במהלכו החלב מאבד נוזלים) וממלאת בקבוקים, כך שכל בקבוק ממולא משאיבה אחת בלבד אך ייתכן כי שאיבה אחת תתחלק בין מספר בקבוקים.

בסיום התהליך, עבור כל בקבוק המערכת שומרת את עלות הייצור שלו, עלות המכירה שלו, תכולתו (לדוגמא, ליטר אחד) ותאריך מכירת הבקבוק. בנוסף, לכל שאיבה נשמר תאריך השאיבת החלב.

לצורך בקרה וייעול תהליכים מנהל החברה ביקש להפיק את הדוחות הבאים:

א- סך הרווח ממכירת הבקבוקים בחציון השני של שנת 2015.

ב- רשימת הערים בהן נשאב לפחות 10 ליטר חלב לפני תהליך הפסטור במהלך חודש אוגוסט 2014.

ג- חמש המדינות המרוויחות ביותר.

ד- רווחי המחלבות לפי השנים באופן הבא:

	שנה א	*שנה ב*	....	*שנה N*
מחלבה א	....			
...				
מחלבה M				

ה- הפרות מהן נשאב חלב "טהור" בשנת 2015. חלב "טהור" הינו חלב שאיבד עד 5% בתהליך הפסטור.

ו- שלושת הפרות שמחליבתן נגרע הכי מעט חלב. כלומר, הפרות מהן כמות החלב המופסדת בתהליך הפסטור הנמוכה ביותר.

ז- עבור כל סוג פרה בה כמות השאיבה לאחר הפסטור בQ2-2014 נמוך מכמות השאיבה לאחר הפסטור השאיבה בQ3-2014, הצג את כמות השאיבה בפועל לQ4-2014.

א) **(10 נקודות)** שרטט סכמת פתית השלג לפי הסיפור העונה על הדוחות הנדרשים.

ב) רשום שאילתת MDX המציגה את הדוחות הנ"ל:

a. **(5 נקודות)** דו"ח ד'

b. **(5 נקודות)** דו"ח ו'

c. **(7 נקודות)** דו"ח ז'

ג) **(8 נקודות)** חלק מהפרות עברו בין המחלבות והחברה דאגה לעדכן את השינויים. איך היית מטפל בבעיה זו? הצע שני פתרונות אפשריים.

ד) **(5 נקודות)** לפנייך שתי טבלאות ממסד הנתונים המקורי בחברה, תאר במילים כיצד יתבצע תהליך הETL ממסד הנתונים התפעולי לסכמה אותה יצרת בסעיף א'. בתיאורך יש להתייחס לפעולות אגרגאציה (ממוצע, סכימה, מינימום וכו'), פעולות מתמטיות בסיסיות, נתונים לא רלוונטיים ונתונים שאין באפשרותך לדעת על בסיס **כל השדות שבטבלאות אלו**.

טבלה שאיבה:

זיהוי שאיבה	זיהוי פרה	זיהוי מחלבה	תאריך שאיבה	כמות שאיבה לפני פסטור

טבלת בקבוק:

זיהוי בקבוק	זיהוי שאיבה	מחיר ייצור	מחיר מכירה	תכולה

2) מאפיינים features (20 נקודות)

נתון dataset של N נקודות. כל נקודה בעלת 5 מאפיינים. הנקודות מסווגות לשתי מחלקות.

הפעלת על ה dataset את שלושת האלגוריתמים הבאים: עץ החלטה, SVM לינארי ו kMeans (ב kMeans התעלמת כמובן מהמחלקה אליה שייכת הנקודה). בשלושת האלגוריתמים קיבלת תוצאות טובות. שני האלגוריתמים הראשונים סיווגו את נקודות טוב ובאלגוריתם השלישי החלוקה ל 2 אשכולות הייתה קונסיסטנטית עם המחלקות שלהן שייכות הנקודות.

כשלא שמתם לב האויב שלכם הוסיף ל dataset מאפיין נוסף לכל נקודה (המאפיין ה 6). ערך המאפיין הוגרל אקראית ללא שום קשר למחלקות שאליהן שייכות הנקודות.

הפעלתם את האלגוריתמים מחדש על הנתונים החדשים.

א) האם ישתנו תוצאות הסיווג (או ב kMeans התאמת החלוקה לאשכולות למחלקות) של האלגוריתמים יחסית לתוצאות שקיבלתם לפני השינוי. אם השתנו אז עד כמה ובעיקר למה.

ב) בהנחה שהמאפיין החדש נמצא בטור לא ידוע והנתונים המקוריים לא בידיכם, אם התבקשתם לזהות את המאפיין האקראי איך הייתם עושים את זה.

ג) לאיזו בעיה אמיתית ניתן להשתמש בשיטה שפיתחתם בסעיף ב.

3) SVM (30 נקודות)

נתון לך dataset של נקודות מסווגות. כל נקודה שייכת לאחד מ 6 מחלקות. אתה מתבקש לבנות מסווג מסוג multiclass SVM כדי לסווג נקודות כאלה.

א) הוחלט שכל מסווג SVM במערכת שתבנה יחלק את הנקודות לשתי קבוצות בצורה הבאה: נקודות משלוש מחלקות לא ישתתפו במסווג ושלושת המחלקות האחרות יתחלקו לשתי מחלקות מול המחלקה אחת.

מה תהיה גודלה של המטריצה M שמתארת את זה (שורות , עמודות)?

ב) כתוב את המטריצה M.

ג) אחרי שהמטריצה הוכנה איך יתבצע תהליך הלימוד. ניתן להניח שיש לך פונקציה שיודעת ללמוד מסווג SVM סטנדרטי שאתה יכול להשתמש בו.

ד) אם ידוע לך שחלק (לא ידוע) מנקודות הלימוד מסווגות לא נכון מה תעשה כדי לנסות לפתור את הבעיה?

ה) בהינתן נקודה לא מסווגת, איך המערכת תסווג אותה. הצע שתי שיטות לסווג הנקודה. הסבר בפירוט את התהליך.

4) Agglomerative Clustering (15 נקודות):

א) תאר את אלגוריתם ה agglomerative clustering.

ב) מהם הקריטריונים השונים לאיחוד אשכולות (Clusters) באלגוריתם.

ג) תן דוגמא גרפית (צייר את הנקודות) שבו אם בוחרים בקריטריון אחד מקבלים פתרון אחר מאשר אם בוחרים קריטריון אחר. הסבר.

ד) תן דוגמא גרפית שבה זה לא משנה. הסבר.