מספר התלמיד הנבחן
רשום את כל תשע הספרות

הדבק כאן את מדבקת הנבחן

האוניברסיטה הפתוחה

י"ב בתמוז תשע"ו

18
ביולי 2016

סמסטר 2016ב

4 / 20595

מס' שאלון - 535

מס' מועד 86

שאלון בחינת גמר

20595 - כריית מידע

משך בחינה: 3 שעות

בשאלון זה **5** עמודים

מבנה הבחינה:

בבחינה שני חלקים.

יש לענות על כל השאלות.

חלק א:
ענו על ארבע השאלות שבחלק זה.
90 נקודות לחלק זה.

חלק ב:
ענו על השאלה שבחלק זה.
10 נקודות לחלק זה.

משקל כל שאלה מפורט בגוף השאלון.
אפשר להשתמש בהנחות במהלך הפתרון, אם תזדקקו לכך.
פרטו את הנחותיכם!

חומר עזר:
כל חומר עזר מותר בשימוש.
אסור בשימוש כל מכשיר אלקטרוני שבאמצעותו ניתן לאצור מידע לרבות מכשיר טלפון נייד, מחשב נישא, שעון חכם וכד'.

בהצלחה !!!

החזירו
למשגיח את השאלון
וכל עזר אחר שקיבלתם בתוך מחברת התשובות

חלק א (90 נקודות)

בחלק זה **ארבע** שאלות. ענו במפורט על **ארבע** השאלות.

שאלה 1 (30 נקודות) – עצי החלטה (decision tree)

נתונה טבלת נתוני אימון:

מס' נבדק	מדד מסת גוף	לחץ דם	סכרת	פרופיל שומנים	סוג טיפול
1	27	תקין	לא	תקין	דיאטה
2	22	גבוה	לא	לא תקין	ללא טיפול
3	34	תקין	לא	תקין	תרופתי
4	25	גבוה מאוד	כן	תקין	ללא טיפול
5	29	תקין	כן	לא תקין	תרופתי
6	32	תקין	לא	תקין	תרופתי
7	31	גבוה	לא	תקין	תרופתי
8	30	גבוה מאוד	לא	תקין	דיאטה
9	26	תקין	כן	תקין	דיאטה
10	33	גבוה	כן	לא תקין	ניתוחי
11	32	גבוה מאוד	לא	תקין	תרופתי
12	18	גבוה	כן	תקין	ללא טיפול
13	28	תקין	לא	תקין	דיאטה
14	30	גבוה מאוד	כן	לא תקין	דיאטה

א. בנו עץ החלטה חלקי, הכולל את רמת השורש ורמה אחת נוספת בלבד, עבור נתוני האימון שבטבלה לחיזוי החלטה מהו הטיפול שעל הנבדק לקבל לאור מדד מסת הגוף שלו. בתשובתכם הדגימו את שלבי בחירת התכונה המפצלת בעץ.

הערה: יש לכלול חישוב של **אחד** המדדים כדוגמת אנטרופיה, Gain ratio, מדד גיני.

ב. האם ניתן לגזום את עץ ההחלטה שבניתם בסעיף א? נמקו תשובתכם.

המשך הבחינה בעמוד הבא

שאלה 2 (15 נקודות) – חוקי הקשר (association rules)

נתון בסיס הנתונים :

תנאי מזג אויר	נהיגה בשכרות	עבירות תנועה	חגירת חגורה	דרגת חומרת התאונה
טוב	כן	נהיגה במהירות מופרזת	לא	חמורה
גרוע	לא	אין	כן	קלה
טוב	לא	אי ציות לתמרור עצור	כן	קלה
טוב	לא	נהיגה במהירות מופרזת	כן	חמורה
גרוע	לא	נהיגה ברמזור אדום	לא	חמורה
טוב	כן	אי ציות לתמרור עצור	כן	קלה
גרוע	כן	אין	כן	חמורה
טוב	לא	נהיגה ברמזור אדום	כן	חמורה
טוב	כן	אין	לא	חמורה
גרוע	לא	נהיגה ברמזור אדום	לא	חמורה
טוב	כן	נהיגה במהירות מופרזת	כן	חמורה
גרוע	לא	אי ציות לתמרור עצור	כן	קלה

א. בצעו בינאריזציה לבסיס הנתונים והציגו אותו.

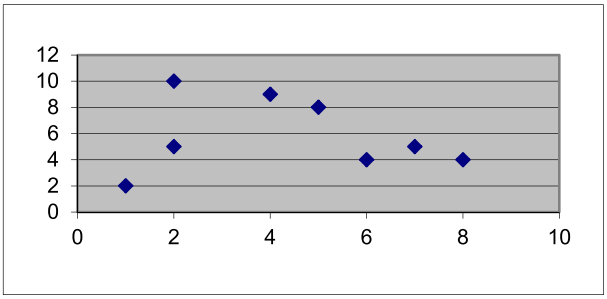
ב. הגדירו מהו הרוחב המקסימלי עבור כל אחת מהטרנזקציות בבסיס הנתונים הבינארי ?

ג. בהנחה min_support= 30% מצאו את כל הקבוצות התדירות.

המשך הבחינה בעמוד הבא

שאלה 3 (25 נקודות) – ניתוח אשכולות (clustering)

בצעו אשכול לשמונה הנקודות המתוארות באיור הבא תוך שימוש באלגוריתם k-means.



בתשובתכם הניחו:

א. k = 3

ב. מרכזי האשכולות ההתחלתיים הם: (1,2) (5,8) (2,10)

ג. השתמשו במרחק אוקלידי

ד. תנאי סיום: לא חלים שינויים באשכול

שימו לב,
בתשובתכם הדגימו את שלושת האשכולות ההתחלתיים וכן את כל השלבים והגדירו את שלושת האשכולות הסופיים.

שאלה 4 (20 נקודות) – למידה בייאסינית

במחשב הביתי התקבלו הודעות המייל הבאות:

D1:"Send us your password"	**spam**
D2:"Send us your review"	**ham**
D3:"Review your password"	**ham**
D4:"Review us"	**spam**
D5:"Send your password"	**spam**
D6:"Send us your account"	**spam**

יש לזהות את התולעת (spam) בעזרת המסווג naïve Bayes .

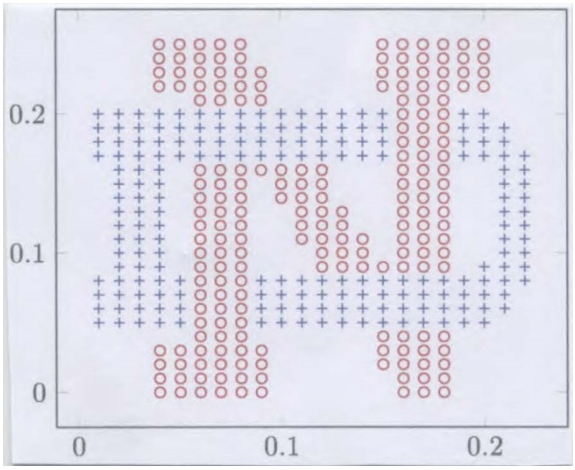
המשך הבחינה בעמוד הבא

חלק ב (10 נקודות)

בחלק זה שאלה **אחת**. ענו עליה במפורט.

שאלה 5 (10 נקודות)

באיור הבא מודגם בסיס נתונים שיש לסווג את נתוניו . הסטודנטים בקורס כריית מידע הציעו את המסווגים הבאים: עץ החלטה, naïve Bayes , k-Nearest Neighbor.



בחנו את ביצועי שלושת המסווגים וציינו מיהו המסווג בעל הביצועים הגבוהים ביותר.

ב ה צ ל ח ה !