אוניברסיטה הפתוחה

20595

כריית מידע

חוברת הקורס אביב 2019ב

כתבה: ד"ר מיה הרמן

פברואר 2019 - סמסטר אביב – תשע"ט

פנימי – לא להפצה.

תוכן העניינים

אל הסטודנט

אנו מקדמים את פניך בברכה עם הצטרפותך אל הלומדים בקורס "כריית מידע".

בחוברת זו תמצא את "לוח הזמנים ופעילויות", תנאים לקבלת נקודות זכות ומטלות הקורס.

לקורס קיים אתר באינטרנט בו תמצאו חומרי למידה נוספים, אותם מפרסם/מת מרכז/ת ההוראה.
בנוסף, האתר מהווה עבורכם ערוץ תקשורת עם צוות ההוראה ועם סטודנטים אחרים בקורס.
פרטים על למידה מתוקשבת ואתר הקורס, תמצאו באתר שה"ם בכתובת:
http://telem.openu.ac.il
מידע על שירותי ספרייה ומקורות מידע שהאוניברסיטה מעמידה לרשותכם, תמצאו באתר הספריה באינטרנט www.openu.ac.il/Library.

ייעוץ ינתן ביום ד' בין השעות 9:30-11:30 בטלפון 09-7781260. פגישה יש לתאם מראש.
ניתן לפנות גם בדואר אלקטרוני maya@openu.ac.il

לתשומת לב הסטודנטים הלומדים בחו"ל:

למרות הריחוק הפיסי הגדול, נשתדל לשמור אתכם על קשרים הדוקים ולעמוד לרשותכם ככל האפשר.
הפרטים החיוניים על הקורס נכללים בחוברת הקורס וכן באתר הקורס.
מומלץ מאד להשתמש באתר הקורס ובכל אמצעי העזר שבו וכמובן לפנות אלינו במידת הצורך.

אני מאחלת לך לימוד פורה ומהנה.

בברכה,

ד"ר מיה הרמן
מרכזת הקורס

1. לוח זמנים ופעילויות (20595 / ב2019)

שבוע לימוד	תאריכי שבוע הלימוד	יחידת הלימוד המומלצת	מפגשי ההנחיה*	תאריך אחרון למשלוח ממ״ן (למנחה)
1	1.3.2019-24.2.2019	יחידה 1 מבוא		
2	8.3.2019-3.3.2019	יחידה 2 תורת המידע	מפגש ראשון	
3	15.3.2019-10.3.2019	יחידה 3 הכנת נתונים		
4	22.3.2019-17.3.2019 (ה-ו פורים)	יחידה 4 סיווג וחיזוי	מפגש שני	
5	29.3.2019-24.3.2019	יחידה 5 עצי החלטה- חלק א		
6	5.4.2019-31.3.2019	יחידה 6 עצי החלטה- חלק ב	מפגש שלישי	ממ״ן 11 5.4.2019
7	12.4.2019-7.4.2019	יחידה 7 למידה בייסיאנית ולמידה מבוססת תצפיות		
8	19.4.2019-14.4.2019 (ו ערב פסח)	יחידה 8 חוקי הקשר – חלק א		ממ״ן 21 16.4.2019
9	26.4.2019-21.4.2019 (א-ו פסח)	יחידה 9 חוקי הקשר – חלק ב		

** התאריכים המדויקים של המפגשים הקבוצתיים מופיעים ב״לוח מפגשים ומנחים״.*

לוח זמנים ופעילויות - המשך

שבוע הלימוד	תאריכי שבוע הלימוד	יחידת הלימוד המומלצת	מפגשי ההנחיה*	תאריך אחרון למשלוח הממ"ן (למנחה)
10	3.5.2019-28.4.2019 (ה יום הזכרון לשואה)	יחידה 10 ניתוח אשכולות- חלק א	מפגש רביעי	
11	10.5.2019-5.5.2019 (ד יום הזיכרון) (ה יום העצמאות)	יחידה 11 ניתוח אשכולות- חלק ב		
12	17.5.2019-12.5.2019	יחידה 12- *פשלת* רשתות אינפו-עמומות	מפגש חמישי	ממ"ן 12 17.5.2019
13	24.5.2019-19.5.2019 (ה ל"ג בעומר)	יחידה13 *פשלת* בחירת מאפיינים	מפגש שישי	
14	31.5.2019-26.5.2019	יחידה 14 *פשלת* נושאים מתקדמים בכריית מידע		ממ"ן 22 31.5.2019
15	7.6.2019-2.6.2019	חזרה	מפגש שביעי	
16	14.6.2019-9.6.2019 (א שבועות)			

מועדי בחינות הגמר יפורסמו בנפרד

** התאריכים המדויקים של המפגשים הקבוצתיים מופיעים ב"לוח מפגשים ומנחים".*

2. תיאור המטלות בקורס

קרא היטב עמודים אלו לפני שתתחיל לענות על השאלות

פתרון המטלות הוא חלק בלתי נפרד מלימוד הקורס – הבנה מעמיקה של חומר הלימוד דורשת תרגול רב. המטלות תבדקנה ותוחזרנה לך בצירוף הערות המתייחסות לתשובות.

2.1 מבנה המטלות

כל מטלה מורכבת מכמה שאלות.
את הפתרונות למטלה עליך להדפיס. רצוי להשאיר שוליים רחבים להערות המנחה.

אם השאלה בממ"ן אינה ברורה לך, אל תהסס להתקשר אל המנחה בשעות הייעוץ הטלפוני בלבד לצורך קבלת הסבר.

המטלות מלוות את יחידות הלימוד בקורס. להלן פירוט המטלות והיחידות שאליהן מתייחסת כל מטלה.

2.2 חומר הלימוד הדרוש לפתרון המטלות

ממ"ן 11 – יחידות לימוד 1-6 – רשות 2 נקודות.

ממ"ן 21 – יחידות לימוד 1-6 – **חובה** – 13 נקודות
(פרויקט – שלב א).

ממ"ן 12 – יחידות לימוד 7-8 – רשות 2 נקודות.

ממ"ן 22 – יחידות לימוד 7-11 – **חובה** – 13 נקודות
(פרויקט – שלב ב).

ממ"נים 21 ו-22 (פרויקט):
מכיוון ואלו **מטלות חובה** ומהווה משקל רב בציון הסופי, אין להגישן באיחור ללא קבלת אישור מראש. על – כן הקפד לשלוח את המטלות במועד.

2.3 ניקוד המטלות

סה"כ ניתן לצבור 26 - 30 נקודות במטלות.
מטלות החובה בקורס כוללות פרויקט המוגש בשני שלבים ומהוות יחד 26 נקודות.
מומלץ להגיש את כל המטלות

3. התנאים לקבלת נקודות זכות בקורס

א) הגשת מטלות מנחה 21 ו- 22 (פרויקט חובה).

ב) ציון של לפחות 60 נקודות בפרויקט.

ג) ציון של לפחות 60 נקודות בבחינת הגמר.

ד) ציון סופי בקורס של 60 נקודות לפחות.

לבחינת הגמר רשאי לגשת רק סטודנט שצבר 26 נקודות לפחות.

לתשומת לבכם!

כדי לעודדכם להגיש לבדיקה מספר רב של מטלות הנהגנו את ההקלה שלהלן:

אם הגשתם מטלות מעל למשקל המינימלי הנדרש בקורס, **המטלות** בציון הנמוך ביותר, שציוניהן נמוכים מציון הבחינה (**עד שתי מטלות**), לא יילקחו בחשבון בעת שקלול הציון הסופי.

זאת בתנאי שמטלות אלה **אינן חלק מדרישות החובה בקורס** ושהמשקל הצבור של המטלות האחרות שהוגשו, מגיע למינימום הנדרש.

זכרו! ציון סופי מחושב רק לסטודנטים שעברו את בחינת הגמר בציון 60 ומעלה והגישו מטלות כנדרש באותו קורס.

הכנת המטלות חייבת להעשות על-ידי כל סטודנט בנפרד.

מטלות שלא יבוצעו באופן עצמאי – יפסלו!!!

מטלת מנחה (ממ"ן) 11

הקורס: 20595 - כריית מידע

חומר הלימוד למטלה: יחידות 1-6

מספר השאלות: 2 **משקל המטלה:** 2 נקודות

סמסטר: ב2019 **מועד אחרון להגשה:** 5.4.2019

אנא שים לב:
מלא בדייקנות את הטופס המלווה לממ"ן בהתאם לדוגמה שלפני המטלות.
העתק את מספר הקורס ומספר המטלה הרשומים לעיל.

בממ"ן זה **שתי** שאלות המתייחסות לטבלת נתוני האימון הרצ"ב.

ניתן להשתמש בתוכנת WEKA וכן בגיליון אלקטרוני EXCEL . ענו במפורט על **שתי** השאלות.

בתשובתכם, ציינו כל הנחה שבצעתם וצרפו את התוכנות והקבצים שערכתם בהם שימוש בחישובים.

בעמוד הבא נתונה טבלת נתוני האימון ---->

מס'	גיל נבדק	משקל	תזונה	ספורט	מצב סוציואקונומי
1	58	גבוה	רגילה	ל	נמוך
2	60	גבוה	רגילה	Y	גבוה
3	32	גבוה	צמחונית	ל	נמוך
4	58	תקין	רגילה	כ	גבוה
5	663	נמוך	צמחונית	כ	גבוה
6	39	נמוך	רגילה	ל	נמוך
7	70	נמוך	צמחונית	כ	גבוה
8	27	72	צמחונית	ל	נמוך
9	45	נמוך	רגילה	ל	גבוה
10	64	תקין	רגילה	כ	גבוה
11	48	תקין	רגילה	כ	גבוה
12	62	תקין	צמחונית	כ	גבוה
13	50	גבוה	רגילה	ל	גבוה
14	24	תקין	צמחונית	????	נמוך

שאלה 1 (25 נקודות)

ציינו והדגימו את שלבי הכנת הנתונים לביצוע כריית מידע כדוגמת טיפול בערכים חסרים, ערכים שגויים ועוד. בסיום, בנו בסיס נתונים מטוייב.

שאלה 2 (75 נקודות)

א. בנו עץ החלטה עבור נתוני האימון שבטבלה לחיזוי מצב סוציואקונומי. בתשובתכם הדגימו את שלבי בחירת התכונה המפצלת בעץ.

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

ב. איזה מבין התכונה/תכונות ניתן להסיר ומדוע? אם אין תכונה הניתנת להסרה, יש לציין זאת מפורשות.

מטלת מנחה (ממ"ן) 21 - פרויקט גמר

הקורס: 20595 - כריית מידע

חומר הלימוד למטלה: יחידות 1-6

מספר השאלות: 2 **משקל המטלה:** 13 נקודות

סמסטר: ב2019 **מועד אחרון להגשה:** 16.4.2019

אנא שים לב:
מלא בדייקנות את הטופס המלווה לממ"ן בהתאם לדוגמה שלפני המטלות.
העתק את מספר הקורס ומספר המטלה הרשומים לעיל.

הנחיות

נתון בסיס נתונים מתחום תעשיית היין הלבן המצוי בכתובת:

http://archive.ics.uci.edu/ml/datasets/Wine+Quality

הסברים אודות נתוני הקובץ תוכלו למצוא בכתובת:

http://archive.ics.uci.edu/ml/machine-learning-databases/wine-quality/winequality.namess

מטרת הפרויקט:

לחזות את איכות היין הלבן מתוך סט התכונות הנתונות.

הפרויקט כולל שימוש בחבילת תוכנה WEKA לכריית מידע.

הפרויקט יבוצע בשני שלבים:

א. ממ"ן 21 – בשלב הראשון תידרשו להגדיר את הבעיה, להכין את הנתונים ולפתור את הבעיה בעזרת שיטות סיווג וחיזוי.

ב. ממ"ן 22 – בשלב השני תידרשו לפתור את הבעיה בעזרת חוקי הקשר וניתוח אשכולות. כמו כן, יהיה עליכם לסכם את עיקרי התוצאות שקיבלתם בממ"ן 21 וכן בממ"ן 22 והמסקנות .

על מנת לסייע לכם בפתרון הפרויקט מורכב הממ"ן ממספר שאלות הנחיה. **הינכם נדרשים לענות על כל השאלות לפי סדר הופעתן.** לצורך פתרון השאלות הינכם רשאים להניח כל הנחה ו/או הפשטה סבירה שתידרשו לה. יש לציין במפורש בתחילת הפרויקט את ההנחות ו/או ההפשטות בהן הנכם משתמשים. כמו כן, יש לציין כל הנחה ו/או הפשטה עליה אתם מתבססים במהלך הפרויקט.

במסגרת הפרויקט יש להשתמש בחבילת תוכנה חופשית WEKA המצויה בכתובת:

WEKA: http://www.cs.waikato.ac.nz/ml/weka/index.html

צורת הגשה:

יש להגיש את הפרויקט מודפס. הקפידו על כתיבה בהירה ומאורגנת וכן על תרשימים ברורים וקריאים. יש להקפיד על תיעוד מפורט של כל שלבי הפרויקט. **אין צורך** לצרף לפרויקט נתונים טכניים של חבילת התוכנה WEKA.

1. הגדרת הבעיה והכנת הנתונים (50%)

(7%) א. הגדירו את מטרות כריית המידע. ציינו את ההנחות וההפשטות בהן השתמשתם.

(6%) ב. הגדירו את הנתונים בהם השתמשתם בפרויקט כדוגמת: תכונות, סוג הנתונים, נתונים חסרים, תחומי ערכים ועוד.

(7%) ג. בהמשך לסעיפים א ו-ב , הגדירו ותארו את שלבי ה- KDD עבור הבעיה הנתונה.

(15%) ד. בהמשך לסעיפים א ו-ב ערכו סקירה השוואתית לכלל החלופות האפשריות (לפחות 4 חלופות) לביצוע כריית מידע. בתשובתכם יש להתייחס ליתרונות/חסרונות כל אחת מהחלופות בהקשר לבעיה הנתונה.

(15%) ה. תארו את שלבי הכנת הנתונים. בתשובתכם יש להתייחס לבעיות באיכות הנתונים כדוגמת טיפול בערכים חסרים, תצוגה גרפית של הנתונים, ניקוי הנתונים, שילוב והמרה של נתונים ועוד.

הערה:

בשלב זה ניתן להשתמש בתוכנת Excel.

2. סיווג וחיזוי (50%)

(5%) א. בחרו שתי שיטות לחיזוי הנתונים.
הסבירו את השיטות ונמקו את בחירתכם.

(15%) ב. תארו את שלבי השיטות שבחרתם בסעיף א.

(8%) ג. עבור כל שיטה דווחו את תוצאות הניתוחים.

(7%) ד. העריכו את מידת הדיוק של כל שיטה.

(15%) ה. נתחו השוואתית את התוצאות והסיקו מסקנות כולל הצעות לשיפורים.

מטלת מנחה (ממ"ן) 12

הקורס: 20595 - כריית מידע

חומר הלימוד למטלה: יחידות 7-11

מספר השאלות: 3 **משקל המטלה:** 2 נקודות

סמסטר: ב2019 **מועד אחרון להגשה:** 17.5.2019

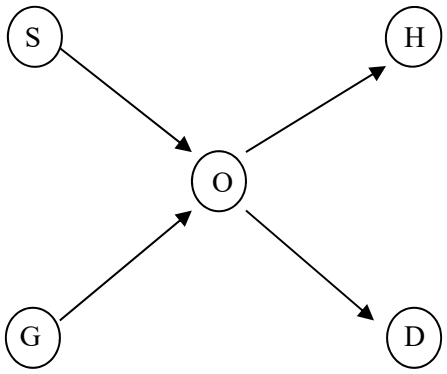
אנא שים לב:
מלא בדייקנות את הטופס המלווה לממ"ן בהתאם לדוגמה שלפני המטלות.
העתק את מספר הקורס ומספר המטלה הרשומים לעיל.

שאלה 1 (30%):

באיור הבא נתונה רשת בייסיאנית המתארת את העובדות הבאות:

השמנה בטנית (יסומן O) מתפתחת כתוצאה מאכילה מופרזת של סוכרים (יסומן S) או/ו כתוצאה מתורשה (יסומן G).

השמנה בטנית גורמת ליתר לחץ דם (יסומן H) או/ו למחלת סוכרת (יסומן D).



בהנחה ש: P(S)=0.3 P(G)= 0.4

נגדיר:

$$P(O \mid S \wedge G) = 0.8$$
$$P(O \mid S \wedge \overline{G}) = 0.4$$
$$P(O \mid \overline{S} \wedge G) = 0.6$$
$$P(O \mid \overline{S} \wedge \overline{G}) = 0.2$$

א. חשבו את ההסתברות של נבדק ללקות בהשמנה בטנית –P(O)

ב. חשבו את ההסתברות שנבדק בעל גורם תורשתי ילקה בהשמנה בטנית P(O|G)

שאלה 2 (30%):

בצעו אשכול היררכי אגלומרטיבי לנתוני האימון המטוייבים בממ"ן 11 .
בתשובתכם יש לכלול:

- אופן הגדרת המרחקים
- קריטריון לעצירת בניית הדנדרוגרם

שימו לב,
בתשובתכם הסופית יש להדגים את כל שלבי בניית הדנדרוגרם .

שאלה 3 (40%):

נתונה הטבלה הבאה:

מספר הזמנה	פריטים
1001	{i1,i4,i5}
1024	{i1,i2,i3,i5}
1012	{i1,i2,i4,i5}
1031	{i1,i3,i4,i5}
1015	{i2,i3,i5}
1022	{i2,i4,i5}
1029	{i3,i4}
1040	{i1,i2,i3}
1033	{i1,i4,i5}
1038	{i1,i2,i5}

בהנחה:

Min_support =60%

Min_confidence = 80%

א. מצאו את כל הקבוצות התדירות תוך שימוש באלגוריתם FP-growth . הדגימו את שלבי בניית עץ FP .

ב. הציגו את חוקי ההקשר החזקים.

מטלת מנחה (ממ"ן) 22 - פרויקט גמר

הקורס: 20595 - כריית מידע

חומר הלימוד למטלה: יחידות 7-11

מספר השאלות: 3 **משקל המטלה:** 13 נקודות

סמסטר: ב2019 **מועד אחרון להגשה:** 31.5.2019

אנא שים לב:
מלא בדייקנות את הטופס המלווה לממ"ן בהתאם לדוגמה שלפני המטלות.
העתק את מספר הקורס ומספר המטלה הרשומים לעיל.

הנחיות:

בהמשך לממ"ן 21 השלימו במסגרת מטלה זו את הפרויקט בקורס . מומלץ לשוב ולעיין בהנחיות שניתנו בממ"ן 21 .

ענו על השאלות הבאות:

1. חוקי הקשר (45%)

(10%) א. בחרו שני אלגוריתמים של חוקי הקשר.
תארו ונתחו את האלגוריתמים שבחרתם. נמקו את בחירתכם.

(15%) ב. בהנחה:

Min_confidence = 60% ,Min_support =40%

מצאו את כל הקבוצות התדירות תוך שימוש בשני האלגוריתמים שבחרתם בסעיף א.

(8%) ג. הציגו את חוקי ההקשר החזקים.

(4%) ד. הריצו ודווחו את התוצאות של שני האלגוריתמים.

(8%) ה. נתחו השוואתית את התוצאות של שני האלגוריתמים והסיקו מסקנות.

2. ניתוח אשכולות (45%)

(2%) א. הגדירו מהו ניתוח אשכולות.

(2%) ב. הגדירו מדדי איכות לאשכולות.

(12%) ג. בחרו שתי גישות לניתוח אשכולות.
בתשובתכם יש לכלול הסבר אודות כל גישה וכן נימוק לבחירתכם.

(12%) ד. תארו את שלבי ניתוח האשכולות עבור 2 הגישות שציינתם בסעיף ג.
בתשובתכם יש להתייחס בין היתר לאופן הכנת הנתונים, מה הם הפרמטרים, ערכי הפרמטרים ועוד.

(7%) ה. עבור כל גישה דווחו את תוצאות הניתוחים.

(10%) ו. נתחו השוואתית את התוצאות והסיקו מסקנות.

3. סיכום ומסקנות (10%)

סכמו בקצרה את עיקרי התוצאות שהתקבלו בממ״ן 21 וכן בסעיפים הקודמים בממ״ן הנוכחי ואת המסקנות שניתן לקבל מתוצאות אלה.