

אוניברסיטה הפתוחה

20595

כריית מידע

חוברת הקורס אביב 2016ב

כתבה: ד"ר מיה הרמן

מרץ 2016 - סמסטר אביב – תשע"ו

פנימי – לא להפצה.

© כל הזכויות שמורות לאוניברסיטה הפתוחה.

תוכן העניינים

1	אל הסטודנט
2	1. לוח זמנים ופעילויות
4	2. תיאור המטלות
4	2.1 מבנה המטלות
4	2.2 חומר הלימוד הדרוש לפתרון המטלות
4	2.3 ניקוד המטלות
5	3. התנאים לקבלת נקודות זכות בקורס
7	ממ"ן 11
9	ממ"ן 21 (פרויקט)
13	ממ"ן 12
15	ממ"ן 22 (פרויקט)

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 הספרייה באינטרנט www.openu.ac.il/Library.

ייעוץ ינתן ביום ד' בין השעות 11:30-9:30 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 מאוד 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 / ב2016)

שבוע לימוד	תאריכי שבוע הלימוד	יחידת הלימוד המומלצת	מפגשי ההנחיה*	תאריך אחרון למשלוח ממ"ן (למנחה)
1	11.3.2016-6.3.2016	יחידה 1 מבוא		
2	18.3.2016-13.3.2016	יחידה 2 תורת המידע	מפגש ראשון	
3	25.3.2016-20.3.2016 (ה-ו פורים)	יחידה 3 הכנת נתונים		
4	1.4.2016-27.3.2016	יחידה 4 סיווג וחיזוי	מפגש שני	
5	8.4.2016-3.4.2016	יחידה 5 עצי החלטה- חלק א		
6	15.4.2016-10.4.2016	יחידה 6 עצי החלטה- חלק ב	מפגש שלישי	
7	22.4.2016-17.4.2016 (ו ערב פסח)	יחידה 7 למידה בייסיאנית ולמידה מבוססת תצפיות		ממ"ן 11 22.4.2016
8	29.4.2016-24.4.2016 (א-ו פסח)	יחידה 8 חוקי הקשר – חלק א		
9	6.5.2016-1.5.2016 (ה יום הזכרון לשואה)	יחידה 9 חוקי הקשר – חלק ב	מפגש רביעי	

* 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	13.5.2016-8.5.2016 (ד יום הזיכרון) (ה יום העצמאות)	יחידה 10 ניתוח אשכולות- חלק א		ממ"ן 21 13.5.2016
11	20.5.2016-15.5.2016	יחידה 11 ניתוח אשכולות- חלק ב	מפגש חמישי	
12	27.5.2016-22.5.2016 (ה ל"ג בעומר)	יחידה 12- <i>ת/ש</i> רשתות אינפו-עמומות	מפגש שישי	ממ"ן 12 27.5.2016
13	3.6.2016-29.5.2016	יחידה 13 <i>ת/ש</i> בחירת מאפיינים		
14	10.6.2016-5.6.2016 (א יום ירושלים)	יחידה 14 <i>ת/ש</i> נושאים מתקדמים בכריית מידע		ממ"ן 22 10.6.2016
15	17.6.2016-12.6.2016 (א שבועות)	חזרה	מפגש שביעי	
16	24.6.2016-19.6.2016			

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

משקל המטלה: 2 נקודות

מספר השאלות: 2

מועד אחרון להגשה: 22.4.2016

סמסטר: 2016ב

אנא שים לב:

מלא בדיוקנות 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.

נתונה טבלת נתוני אימון בעמוד הבא:

ת.ז. שחקן	מגדר	גיל	ילדים	סוג רכב	סיכון לתאונה
11111	ז	38	כ	משפחתי	נמוך
22222	נ	24	ל	ספורט	גבוה
33333	ז	29	ל	ספורט	??
44444	נ	36	כ	משפחתי	נמוך
55555	ז	38	כ	ספורט	גבוה
66666	נ	28	כ	משפחתי	גבוה
77777	ז	39	כ	99999	גבוה
88888	ז	32	כ	משפחתי	נמוך
99999	ז	48	כ	ספורט	גבוה
10111	ז	??	ל	משפחתי	נמוך
11222	נ	37	כ	משפחתי	נמוך
3AB11	ז	26	כ	משפחתי	גבוה
13444	נ	31	0	ספורט	גבוה
14555	נ	22	ל	משפחתי	גבוה

שאלה 1 (25 נקודות)

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

שאלה 2 (75 נקודות)

א. בנו עץ החלטה חלקי, הכולל את רמת השורש ורמה אחת נוספת בלבד, עבור נתוני האימון שבטבלה לחיזוי רמת סיכון לתאונה. 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: 2016

מועד אחרון להגשה: 13.5.2016

אנא שים לב:

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

הנחיות

נתון קובץ נתוני פופולריות חדשות מקוונות המצוי בכתובת:

<http://archive.ics.uci.edu/ml/datasets/Online+News+Popularity>

נתוני הקובץ תוכלו למצוא בכתובת:

<http://archive.ics.uci.edu/ml/machine-learning-databases/00332>

מטרות הפרויקט:

-חיזוי מידת הפופולריות של החדשות המקוונות – ממ"ן 21.

-איפיון פרופיל החדשות המקוונות הפופולריות – ממ"ן 22.

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%)

- א. (6%) הגדירו את הנתונים בהם השתמשתם בפרויקט כדוגמת: תכונות, סוג הנתונים, נתונים חסרים, תחומי ערכים ועוד.
- ב. (7%) הגדירו את מטרות כריית המידע. ציינו את ההנחות וההפשטות בהן השתמשתם.
- ג. (7%) בהמשך לסעיפים א ו-ב, הגדירו ותארו את שלבי ה-KDD עבור הבעיה הנתונה.
- ד. (15%) 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.
- ה. (15%) תארו את שלבי הכנת הנתונים. בתשובתכם יש להתייחס לבעיות באיכות הנתונים כולל 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%)

- א. (5%) בחרו שתי שיטות לחיזוי הנתונים. הסבירו את השיטות ונמקו את בחירתכם.
- ב. (15%) תארו את שלבי השיטות שבחרתם בסעיף ב.
- ג. (8%) עבור כל שיטה דווחו את תוצאות הניתוחים.
- ד. (7%) העריכו את מידת הדיוק של כל שיטה.
- ה. (15%) 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: 2016

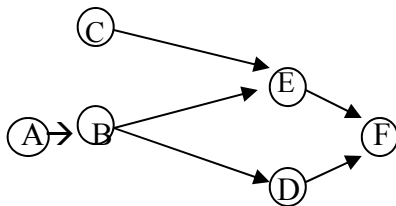
מועד אחרון להגשה: 27.5.2016

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



א. מהי ההסתברות המשותפת של A,B,C,D,E,F.

ב. עבור כל זוג צמתים בגרף קבעו האם הם תלויים או בלתי תלויים אחד בשני.

שאלה 2 (30%):

נתונה הטבלה:

מס' לקוח	מוצרים
1	{מחשב, דיסק און קי, מדפסת, עכבר}
2	{מחשב, דיסק און קי, מדפסת, אוזניות}
3	{מחשב, דיסק און קי, אוזניות, כונן חיצוני, משחק מחשב}
4	{מחשב, מדפסת, משחק מחשב}

בהנחה:

$$\text{Min_confidence} = 60\%, \text{Min_support} = 50\%$$

א. מצאו את כל הקבוצות התדירות תוך שימוש באלגוריתם א-פריורי.

ב. הציגו את חוקי ההקשר החזקים.

שאלה 3 (40%):

בטבלה ההבאה נתונים ארבעה סוגי תרופות. לכל תרופה שתי תכונות (pH- Index ו- weight).

Medicine	Weight	Ph- Index
A	1	1
B	2	1
C	4	3
D	5	4

בצעו אשכול לנתונים אלו תוך שימוש במטריקה של מרחק אוקלידי.

תנאי התחלה: $c1=A, c2=B$

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: 2016ב

מועד אחרון להגשה: 10.6.2016

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. הניחו מינימום confidence & support ובהתאמה, הגדירו חוקים המקיימים זאת. הסבירו ונמקו.
- (4%) ב. הגדירו את סט הפרמטרים. הסבירו ונמקו.
- (10%) ג. בחרו שני אלגוריתמים של חוקי הקשר. תארו את האלגוריתמים ונמקו את בחירתכם.
- (4%) ד. הריצו ודווחו את התוצאות של שני האלגוריתמים.
- (7%) ה. לאור סעיף ד הגדירו אילו מהחוקים שהגדרתם בסעיף א יש בהם עניין ועשויים להיות שימושיים.
- (10%) ו. נתחו השוואתית את התוצאות של שני האלגוריתמים והסיקו מסקנות.

2. ניתוח אשכולות (45%)

- (2%) א. הגדירו מהו ניתוח אשכולות.
- (2%) ב. תארו את אופן מדידת איכות האשכולות.
- (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 שצינתם 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.