א ו נ י ב ר ס י ט ה ה פ ת ו ח ה

20595

כריית מידע

חוברת הקורס אביב 2014ב

כתבה: ד"ר מיה הרמן

מרץ 2014 - סמסטר אביב – תשע"ד

פנימי – לא להפצה.

תוכן העניינים

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

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

אני מאחלת לך לימוד פורה ומהנה.

בברכה,

ד"ר מיה הרמן
מרכזת הקורס

1. לוח זמנים ופעילויות (20595 / ב2014)

שבוע לימוד	תאריכי שבוע הלימוד	יחידת הלימוד המומלצת	מפגשי ההנחיה*	תאריך אחרון למשלוח ממ"ן (למנחה)
1	7.3.2014-2.3.2014	יחידה 1 מבוא		
2	14.3.2014-9.3.2014	יחידה 2 תורת המידע	מפגש ראשון	
3	21.3.2014-16.3.2014 (א-ב פורים)	יחידה 3 הכנת נתונים		
4	28.3.2014-23.3.2014	יחידה 4 סיווג וחיזוי	מפגש שני	
5	4.4.2014-30.3.2014	יחידה 5 עצי החלטה- חלק א		
6	11.4.2014-6.4.2014	יחידה 6 עצי החלטה- חלק ב	מפגש שלישי	ממ"ן 11 11.4.2014
7	18.4.2014-13.4.2014 (ב ערב פסח) (ג-ו פסח)	יחידה 7 למידה בייסיאנית ולמידה מבוססת תצפיות		
8	25.4.2014-20.4.2014 (א-ב פסח)	יחידה 8 חוקי הקשר – חלק א		
9	2.5.2014-27.4.2014 (ב יום הזכרון לשואה)	יחידה 9 חוקי הקשר – חלק ב	מפגש רביעי	ממ"ן 21 2.5.2014

** 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	9.5.2014-4.5.2014 (ב יום הזכרון, ג יום העצמאות)	יחידה 10 ניתוח אשכולות- חלק א		
11	16.5.2014-11.5.2014	יחידה 11 ניתוח אשכולות- חלק ב	מפגש חמישי	ממ"ן 12 16.5.2014
12	23.5.2014-18.5.2014 (א ל"ג בעומר)	יחידה 12- משלים רשתות אינפו-עמומות	מפגש שישי	
13	30.5.2014-25.5.2014 (ד יום ירושלים)	יחידה13 משלים בחירת מאפיינים		ממ"ן 22 30.5.2014
14	6.6.2014-1.6.2014 (ג-ד שבועות)	יחידה 14 משלים נושאים מתקדמים בכריית מידע		
15	13.6.2014-8.6.2014	חזרה	מפגש שביעי	
16	20.6.2014-15.6.2014			

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: ב2014 **מועד אחרון להגשה:** 11.4.2014

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.

להלן טבלת נתוני אימון:

מס׳ ת.ז.	גיל	מגדר	תואר	שפת תכנות	ידע ב-WEKA	ניסיון בעבודה	ניסיון בהוראה	העסקה כמנחה בקורס כריית מידע
12345	25	נ	מדעי המחשב	Java	כן	לא	לא	לא
67891	30	נ	מדעי המחשב	C++	לא	כן	כן	כן
71294	28	ז	ניהול מערכות ויישומים	Java	??	כן	לא	כן
34568	35	נ	מדעי המחשב	C++	כן	כן	לא	לא
24680	29	ז	ניהול מערכות ויישומים	Java	לא	לא	כן	X
12578	38	נ	מדעי המחשב	C++	כן	כן	כן	כן
25804	33	ז	ניהול מערכות ויישומים	Java	כן	??	כן	כן
16923	40	ז	מדעי המחשב	C++	כן	כן	לא	לא
34890	24	נ	מדעי המחשב	Java	לא	לא	לא	לא

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 המטייבים שבטבלה לחיזוי העסקת מועמד כמנחה בקורס כריית מידע. בתשובתכם הדגימו את שלבי בחירת התכונה המפצלת בעץ.

שימו לב: יש לכלול חישובים כדוגמת אנטרופיה, Gain ratio, מדד גיני

ב. איזו 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: ב2014 **מועד אחרון להגשה:** 2.5.2014

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 בדיקות רפואיות לאבחון בעיות בלוטת התריס המצוי בכתובת:

http://archive.ics.uci.edu/ml/datasets/Thyroid+Disease

הסברים אודות נתוני הקובץ תוכלו למצוא בכתובת:

http://archive.ics.uci.edu/ml/machine-learning-databases/thyroid-disease/HELLO

מטרת הפרויקט:

לחזות האם הנבדק סובל מבעיה בבלוטת התריס כדוגמת פעילות יתר של בלוטת התריס (hypothyroid).

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. **<u>היכם נדרשים לענות על כל השאלות לפי סדר הופעתן.</u>** לצורך פתרון השאלות היכם 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%)

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

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

(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 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%)

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

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

(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

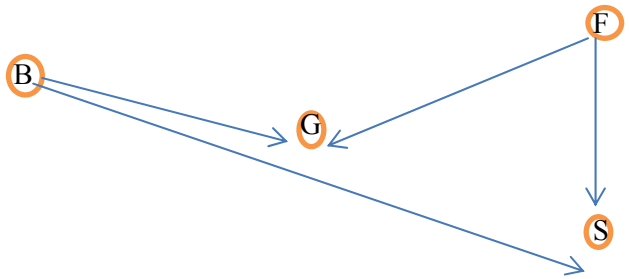
סמסטר: 2014ב **מועד אחרון להגשה:** 16.5.2014

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

שאלה 1 (25%):

באיור הבא נתונה רשת בייסיאנית המתארת את העובדות הבאות:

התנעת הרכב (יסומן S) נקבעת בהתאם לכמות הדלק במיכל הדלק (יסומן F) או/ו למצב הבטריה (יסומן B) המעודכנים במונה (יסומן G).



בהנחה ש: P(F = empty)=0.2 P(B=bad)=0.1

נגדיר:

P(G=empty|B=good, F=not empty)=0.1

P(G=empty|B=good, F=empty)=0.8

P(G=empty|B=bad, F=not empty)=0.2

P(G=empty|B=bad, F=empty)=0.9

P(S=no|B=good, F=not empty)=0.1

P(S=no|B=good, F= empty)=0.8

P(S=no|B=bad, F=not empty)=0.9

P(S=no|B=bad, F= empty)=1.0

א. בהנחה שהבטריה התרוקנה (bad)חשבו את ההסתברות שהרכב אכן יצליח להתניע.

ב. חשבו את ההסתברות P(B=good,F=empty,G=empty,S=yes)

שאלה 2 (45%):

נתונה הטבלה הבאה:

מספר תלמיד	רישום לחוגים
1	{כדורסל,כדורגל,שחמט}
2	{פסנתר,דרמה,כדורסל}
3	{כדורסל, שחמט, אומנות}
4	{פסנתר, דרמה}
5	{כדורגל, אומנות, פסנתר}
6	{כדורסל, שחמט, פסנתר, דרמה}
7	{פסנתר, אומנות, כדורסל}
8	{כדורגל, שחמט}
9	{כדורסל, שחמט, פסנתר, דרמה}
10	{כדורגל, אומנות}

בהנחה ש:

Min_support=50%

Min_confidence=70%

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

שאלה 3 (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 .

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: ב2014 **מועד אחרון להגשה:** 30.5.2014

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 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.