קווים מנחים לפרויקט המסכם בקורס "מערכות לומדות וכריית נתונים"

1. עבודת הגמר תעסוק בנושא כריית נתונים ותבוצע לפי קווי היסוד של מדריך CRISP-DM (מופיע באתר הקורס).
2. למידת המדריך CRISP-DM תתבצע כלמידה עצמית ויש לשלב בה את הלמידה והשימוש בתוכנת R.
3. עבודת הגמר תתבצע במהלך הסמסטר בצוותים של יחידים או זוגות לכל היותר לאחר אישור המרצה.
4. כל צוות יגיש לתיבת הדואר של מרצה הקורס (מספר 73) דו"ח פרויקט מודפס עד תאריך שיימסר לקראת סיום הסמסטר. סקריפטים של R יוגשו יחד עם עותק רך של הדו"ח במייל ל- boaz@bgu.ac.il.
5. בחיפוש נושא ו/או מאגר נתונים לפרויקט (אם לא בחרתם נושא/מאגר מהמחקר או העבודה שלכם), ניתן להשתמש במאגרי המידע הממוחשבים שבספרייה, הרשימות הביבליוגרפיות של מאמרים, הקישורים שניתנו בהרצאה הראשונה או לעשות שימוש במנועי החיפוש באינטרנט או באתר Machine Learning Repository UCI, ובבסיסי הנתונים של KD Nuggets (קישורים רלוונטיים הועלו לאתר). מומלץ מאד להשתמש במאגר נתונים המגיע מתחום המוכר לצוות הפרויקט, ושיש לצוות ידע מקצועי/אישי נוסף עליו.
6. בנוסף למאגר הנתונים, יש לבחור גם מספר מצומצם של מאמרים, ועל פי הצורך, בנושא הנבחר, שישמשו לסקר ספרות בסיסי וכקו מנחה. מומלץ לבחור מאמרים מהכנסים/כתבי-העת המובילים בתחום:
International Conference on Machine Learning (*ICML*), Advances in Neural Information Processing Systems (*NIPS*), Uncertainty in Artificial Intelligence (*UAI*), Artificial Intelligence and Statistics Conference (*AISTATS*), Journal of Machine Learning Research (*JMLR*), Machine Learning).
7. יש לשלוח סיכום של מאגר הנתונים ותמצית שאלת המחקר לאישור המרצה ע"י העלאתם לאתר הקורס ב- Moodle לקבוצת הדיון המתאימה עד לתאריך שיפורסם בקורס.
8. לכל צוות יוקצה משך שיפורסם לקראת מועד ההצגות לצורך הצגת הפרויקט מול הכתה, מתוכן יש להשאיר כ- 5 דקות לשאלות ודיון. יש לבדוק מראש עמידה בזמנים. המצגת תכלול את הרקע לפרויקט, את תמצית שיטת הניתוח, את המודלים שנבחרו, תוצאותיהם, מדדי הביצוע שלהם ולסיכום את תרומת השימוש במערכות לומדות וכריית נתונים לצורך פתרון הבעיה.
9. כדי לחזות/לסווג את ערכי המטרה השתמשו במודלים:
   - Neural Network (“nnet” package)
   - Random Forest (“randomForest” package)
   - Kmeans (“stats” default package)

לכל מודל, דונו במשמעות התיאורטית של הפרמטרים השונים של המודל (למשל, מספר שכבות חבויות, מספר עצים ביער, מספר אשכולות...), והשפעתם על תפקוד וביצועי המודל, והציגו תוצאות אמפיריות ע"ס בסיס הנתונים שלכם לחיזוי/סיווג עבור אותו מודל, העושה שימוש בערכי פרמטרים שונים (כל סט פרמטרים למודל מגדיר עבורו קונפיגורציה). יש להציג תוצאות (למשל, אחוזי דיוק, rmse, או מרחקים בתוך אשכול ובין אשכולות) עבור מספר קונפיגורציות של המודל, כל אחת מובילה לביצועים שונים, ולהסביר את הסיבות לביצועים השונים בקרב הקונפיגורציות השונות, היתרונות והחסרונות של כל קונפיגורציה ואת אופן קביעת הקונפיגורציה המיטבית (הפרמטרים המיטביים) לכל מודל. השוו תיאורטית בין המודלים השונים ואמפירית בין הקונפיגורציות המיטביות של המודלים עבור אותה משימה (נניח חיזוי/סיווג, לאחר שעבור ה-Kmeans השמטתם את משתנה המטרה) והסבירו את ההבדלים בין התוצאות וקיום פערים, אם קיימים, בביצועים. במידה וישנם מודלים נוספים שאתם חושבים שיתאימו טוב יותר לנתונים שלכם, הציגו אותם ודונו בהם ובמוטיבציה לשימוש בהם, וכן הציגו ניתוח אמפירי דומה עבורם בהתבסס על הנתונים.

10. הדו"ח יכלול את המרכיבים הבאים:

- שער – המוסד, הפרויקט, המבצעים, המרצה, תאריך הגשה.
- תקציר – עד חצי עמוד, תמצית של הדו"ח.
- תוכן עניינים – כולל עימוד.
- רשימת סימונים וקיצורים – בהתאם לצורך.
- חמישה פרקים בהתאם למבנה ולתוכן של מדריך CRISP-DM (ניקוד):
  i. מבוא והבנת התחום/בעיה – Business understanding (15)
  ii. הבנת הנתונים – Data understanding (15)
  iii. הכנת הנתונים – Data preparation (15)
  iv. מידול – Modeling (30)
  v. הערכה – Evaluation (15)
- סיכום, דיון, ומסקנות (10)

בהצלחה!