קווים מנחים לפרויקט הסיום בקורס "מכונות לומדות וכריית נתונים"

1. עבודת הגמר תעסוק בנושאי הקורס, ותבוצע לפי קווי היסוד של מדריך CRISP-DM (כמתואר בהרצאה מספר 1) ובסביבת R/מטלב/פייתון.
2. למידת מדריך ה-CRISP-DM והשימוש בתוכנת R/מטלב/פייתון תבוצענה עצמית.
3. עבודת הגמר תתבצע במהלך הסמסטר בצוותים של יחידים עד שלשות לכל היותר, לאחר אישור המרצה בפורום הקורס במודול. לא ניתן להתחיל, לבצע או להגיש העבודה ללא אישור המרצה בפורום, וזאת עד למועד שיימסר בשיעור.
4. בחיפוש נושא ו/או מאגר נתונים לפרויקט, ניתן להשתמש במאגרי המידע הממוחשבים שבספרייה, הרשימות הביבליוגרפיות של מאמרים, הקישורים שניתנו בהרצאה הראשונה או לעשות שימוש במנועי החיפוש באינטרנט או באתר UCI Machine Learning Repository, ובבסיסי הנתונים של Kaggle, (קישורים רלוונטיים הועלו לאתר). מומלץ מאד להשתמש במאגר נתונים המגיע מתחום המוכר לצוות הפרויקט, ושיש לצוות ידע מקצועי/אישי נוסף עליו. כמו כן, ניתן לעשות שימוש בבסיס נתונים עצמי מהמחקר או ממקום העבודה של הסטודנט. בסיס הנתונים יבטא בעיית סיווג או אישכול (במידה ומעוניינים בניתוח בעיה אחרת, יש לקבל מראש אישור מרצה לכך).
5. בנוסף למאגר הנתונים, יש לבחור גם מספר מצומצם של מאמרים, ועל פי הצורך, בנושא הנבחר, שישמשו לסקר ספרות בסיסי וכקו מנחה לעבודה. 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*), and Machine Learning).
6. ניתוח בסיס הנתונים ייעשה ע"י מערכ(ו)ת לומדות לבחירת צוות הפרויקט. הצוות ינמק בראשית העבודה את הבחירה שעשה במערכת שנבחרה לעבודה לאור אתגרי בסיס הנתונים/הבעיה/התחום. בפרק הדיון בעבודה, הצוות יבחן את נימוקיו אלה לאור התוצאות האמפיריות שהושגו ועד כמה הבחירה הראשונית שעשה הוכיחה עצמה, או מה היה עושה אחרת לו הפרויקט נמשך. מוצע, שאם המערכת שנבחרה אינה יער אקראי, רשת נוירונלית MLP, או אשכול kmeans, הרי הצוות ישווה – תיאורטית/אלגוריתמית (באופן איכותני) וניסיונית – את המערכת שבחר למערכות אלה. מצופה שבעבודה תכינו, במידת הצורך, את בסיס הנתונים לאחר שלמדתם אותו, תפעילו את המערכ(ו)ת שבחרתם (ואולי את זו אליה אתם משווים) על בסיס הנתונים לשם הסיווג/אישכול, תבצעו אופטימיזציה לביצועים עבור כל מערכת, תשוו את ביצועיהן, תשקלו שיפורים והרחבות של בסיס הנתונים ו/או המערכות (בונוס לעבודה, המזכה בהערכה נוספת) ותסכמו את העבודה.

7. את פירות העבודה תציגו בסמינר בסוף הקורס (סעיף 8) ובעבודה כתובה (סעיף 9), שתוגש במהלך תקופת המבחנים על פי תאריך שיימסר בסוף הקורס. כל חברי הצוות ישתתפו בצורה פעילה בשתי המטלות.
8. לכל צוות יוקצו כ-20 דקות (משך ההצגה ייקבע בכל שנה בהתאם למספר המציגים בקורס, ויימסר לקראת מועד ההצגות) לצורך הצגת הפרויקט מול הכתה, מתוכן יש להשאיר כ-5 דקות לשאלות ודיון. בהכנות להצגה, יש לבדוק מראש עמידה בזמנים. המצגת תכלול את הרקע לפרויקט, תמצית שיטת הניתוח, המודלים שנבחרו, תוצאותיהם, מדדי הביצוע שלהם ולסיכום את תרומת השימוש במערכות לומדות וכריית נתונים לצורך פתרון הבעיה.
9. היקף הדוח לא יעלה על 8 עמודים, לא כולל: שער (עמוד אחד), ביבליוגרפיה (עמוד אחד) ונספחים במידת הצורך (עד 2 עמודים). הפונט יהיה מסוג times new roman בגודל 11, עם מרווח בין השורות 1.15 ורוחב שוליים של 2.5 ס"מ מכל אחד מארבעת הצדדים. הדו"ח יכלול את המרכיבים הבאים (ההגשה היא לתיבת ההגשה במודל של דו"ח פרויקט, כולל סקריפטים של התוכנה, וזאת עד לתאריך שיימסר לקראת סוף הקורס):
   - 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:
     i. מבוא והבנת התחום/בעיה – Business understanding.
     ii. הבנת הנתונים – Data understanding.
     iii. הכנת הנתונים – Data preparation.
     iv. מידול – Modeling.
     v. הערכה – Evaluation.
   - סיכום, דיון, ומסקנות

בהצלחה!

בעז