שם הקורס בעברית: מבוא להנדסת נתונים

שם הקורס באנגלית: Introduction to Data Engineering

מספר הקורס: 382.1.1101

מבנה הקורס: הרצאה: 3, תרגיל 1, מעבדה 0, סה"כ: 4 ש"ש

נקודות זכות: 3.5

סמסטר א, תשפ"ב

תאור הקורס:

הנושאים הנלמדים בקורס כוללים: סקירה כללית של התחום, מסדי נתונים רלציוניים, מחסני נתונים ובינה עסקית, עיבוד נתונים, אמידה סטטיסטית, חיזוי, סיווג, ניתוח אשכולות, חוקי קשר, אתיקה ופרטיות נתונים.

בשיעורי המעבדה התלמידים יתנסו בשפות תכונות ותוכנה לעבודה נוחה עם נתונים. בפרט התלמידים ילמדו לבנות מסדי נתונים, לבצע ניתוחים סטטיסטיים באמצעות Python ו- SQL ותוכנת MS Excel והרצת אלגוריתמים של כריית מידע באמצעות שפת Python.

מטרות הקורס:

הקורס נועד לסקור את השיטות והכלים המובילים לאחסון, עיבוד, ניתוח ותצוגה של כמויות נתונים גדולות (Big Data) לצורך פתרונן של בעיות הנדסיות מורכבות. במהלך הקורס הסטודנטים יחשפו למקורות עשירים ומגוונים של נתונים, וכלים בסיסיים לניתוחם והבנתם, למשל ממערכות חברתיות, עסקיות, ייצור, רפואה, אינטרנט ועוד.

תנאי קדם:

1. מבוא לתכנות (קורס מקביל)
2. אלגברה לינארית (קורס מקביל)

דרישות הקורס והרכב הציון:

- התלמיד נדרש ללמוד את נושאי הקורס המפורטים להלן על פי ההרצאות בכיתה וחומר הקריאה, להשתתף בהרצאות ובמעבדות ולהגיש עבודות בית במועד ולהבחן בסוף הקורס על החומר.
- במשך הקורס יינתנו 9-10 עבודות בית מעשיות הקשורות לנושאי המעבדה (מתוכם 6-8 תרגילי חובה להגשה).
- משקל מטלות – כ-10% (תרגילי בית שבועיים קצרים)
- משקל הבחינה הסופית – כ-90% (חובת מעבר).
- במידה ולא תתאפשר בחינה פרונטלית הבחינה תתקיים באופן מקוון

רשימת קריאה (בבליוגרפיה):

1) Han, J., Kamber, M., and Pei, J., 2011, *Data Mining Concepts and Techniques*, 3rd Edition, Morgan Kaufmann.
2) Kelleher, J.D. and Tierney, B., 2018. *Data Science*. The MIT Press.
3) Guttag, J., 2016. *Introduction to Computation and Programming Using Python: with Application to Understanding Data*. MIT Press.
4) Bloomfield, V.A., 2014. *Using R for numerical analysis in science and engineering*. Chapman and Hall/CRC.
5) Python Data Science Handbook, Jake VanderPlas

תכנית המפגשים:

מס' השיעור	נושא השיעור	קריאה נדרשת
1	מבוא לקורס. סקירה כללית של תחום מדעי הנתונים כולל הרקע ההיסטורי והטכנולוגי. דוגמאות של פרויקטים מעשיים.	Han: פרק 1 Kelleher: פרק 1
2-3	נתונים, מאגרי נתונים, תכנון מסדי נתונים רלציוניים. שליפת מידע ממסדי נתונים רלציוניים. מבוא לשפת SQL. גישת NoSQL. מודל Entity-Relationship.	Han: פרק 1 Kelleher: פרק 1
4	עבודה עם קבצים עבודה עם DataFrames (Pandas)	DS Handbook פרק 2
5-6	עיבוד נתונים (data manipulation). ניקוי נתונים (data cleaning). זיהוי חריגים (outlier detection). המרת נתונים (data transformation). כמו כן נלמד לעבוד עם Regular Expressions	Han: פרק 3
7	מחסני נתונים ובינה עסקית (business intelligence). ויזואליזציה של נתונים (עבודה עם plotly express ו matplotlib)	Han: פרק 4

מס' השיעור	נושא השיעור	קריאה נדרשת
8	חיזוי ערכים רציפים. רגרסיה ליניארית פשוטה, רגרסיה לינארית רבת משתנים, ושימוש במודלי רגרסיה. היכרות עם חבילת scikit-learn	Han: פרק 3
9-10	סיווג (classification). גישת השכנים הקרובים ביותר (k-NN). עצי החלטה (decision trees). היכרות עם חבילת weka	Han: פרק 8 Kelleher: פרק 4 Guttag: פרק 24
11-12	ניתוח אשכולות (clustering). חוקי קשר (association rules).	Han: פרק 10 Han: פרק 6 Guttag: פרק 23
13	אמידה סטטיסטית ובדיקת השערות. אתיקה ופרטיות נתונים	Han: פרק 2 Bloomfield: פרק 3+ 10 Kelleher: פרק 6 Guttag: פרק 17+18