

תבנית מוצעת לפרשיית לימודים (הניתנת לסטודנט)
אוניברסיטת בן-גוריון
המחלקה להנדסת תעשייה וניהול

מדעי הנתונים לכלכלנים
(Data Science For Economists)
364-1-1391
רון סרפיאן

פרטי המרצה:

שעות קבלה: תפורסמה בתחילת הסמסטר
דואר אלקטרוני: ronsar@post.bgu.ac.il

תיאור הקורס:

מדעי הנתונים הוא תחום רחב העוסק בהפיכת נתונים גולמיים למידע המאפשר קבלת החלטות אופטימלית. מטרת קורס זה הנה להעניק לסטודנטים לכלכלה כלים לעיבוד וניתוח מגוון בסיסי נתונים וכן גישה ליישומים שונים של למידת מכונה. קורס זה יילמד באמצעות תוכנת R, אשר הופכת לאחת מסביבת ניתוח הנתונים המועדפת באקדמיה ובתעשייה. בקורס נלמד כיצד לנתח נתונים בשיטות סטטיסטיות קלאסיות כגון מודלים של רגרסיה וסיווג, וכן שיטות חדשות יותר ללמידת-מכונה, הורדת מימד ועוד. מלבד ניתוח הנתונים עצמם גם נסקור כלים אוטומטיים לדיווח ממצאים. דגש יינתן על שימוש בתוכנת R ובשפת SQL בכדי לממש שיטות אנליזה חדשות המתאימות לקבצי נתונים גדולים (Big data).

מטרות הקורס:

- התלמיד ידע לעבוד ולנתח מגוון בסיסי נתונים בסביבת R.
- התלמיד ידע ליישם אלגוריתמים ללמידת מכונה בסביבת R.

דרישות הקורס:

לתלמידי כלכלה: קורסי הבסיס בסטטיסטיקה ואקונומטריקה.
לתלמידי מחלקות אחרות: הקבלה באישור המרצה ותדרוש קורסי קדם ברמת מבוא בהסתברות, וסטטיסטיקה. אלגברה לינארית היא יתרון אבל אינה הכרח. למען הסר ספק, אין הכרח בידע מוקדם בחדו"א או תכנות.

מבנה ציון הקורס:

מבחן: 50%
מטלות: 30%
תרגיל מסכם: 20%

תהינה 2 או 3 עבודות בית, בזוגות, שמצריכות ליישם דברים דומים למה שראו בשיעור אבל מצריכות גם לחשוב וללמוד לבד (קצת). בנוסף, יהיה תרגיל מסכם בסגנון [kaggle תחרות](https://www.kaggle.com). התלמידים יקבלו הדרכה - כיצד בונים pipeline לבעיות למידה באופן כללי, ויצטרכו לייצר תחזיות באמצעות machine learning עבור בעיית נתונים כלשהי שאבנה. הציון של תרגיל זה יתבסס על איכות הניבוי שלהם.

שימו לב שהמטלות אינן רק יישום של מה שלומדים בכיתה ובספר הקורס, אלא, הן מצריכות גם יכולת למידה עצמית, כלומר, חיפוש פתרונות (גוגל, פורומים) לבעיות שנתקלים בהן בזמן כתיבת קוד או בניית מודל. יכולת זו חשובה מאוד בתחום ה data science, בעיקר בתעשייה, והיא חלק ממה שהקורס בא לפתח.

חובת קבלת ציון עובר בבחינה הסופית: כן

מפגש	נושאי המפגש
1	היכרות עם R
2	תכנות בסיסי ב-R
3	SQL
4	סטטיסטיקה תיאורית
5	תרשימים
6	מודלים ליניארים
7	מודלים ליניארים מוכללים
8	מודלים היררכיים
9	למידה מונחית (1)
10	למידה מונחית (2)
11	למידה בלתי מונחית
12	חישוב מקבילי, כתיבת דוחות
13	היכרות עם סביבות תוכנה נוספות (פייתון)

Topic description:

1. **Introduction:** The R eco-system (CRAN, BioConductor, Microsoft R, StackExchange, R-help, ...). The programming rational, object types.
2. **R basics:** Interacting with different objects, clean, merge and manipulate data, writing functions, loops and scoping rules.
3. **Introduction to SQL language:** Writing SQL queries, introducing SQL interfaces for R (sqldf, SQLite,...)
4. **Exploratory data analysis:** Summary statistics and data aggregation in R. Introducing the data.table package.
5. **Visualization:** Introducing graphics system in R and ggplot2 package.
6. **Linear models:** The lm() function. How to fit a linear regression and interpret output: the coefficients, goodness of fit, etc.
7. **Generalized linear models:** The glm() function. Logistic regression, Multinomial, Poisson, Exponential, etc.
8. **Hierarchical models:** The lmer4 and nlme packages for hierarchical models, mixed models, longitudinal data, repeated measures and panel data.
9. **Supervised Learning:** Introduction: Train/Test split, validation methods; Learning algorithms: Ridge regression, SVM, Neural Nets, Tree based modeling. Introducing the caret package.

10. **Unsupervised Learning:** Dimensionality reduction: PCA, FA, ICA; Clustering: K-means, and hierarchical.
11. **Reporting and reproducible research:** Sweave, knitr, bookdown, Shiny.
12. **Parallel computing for BigData:** Parallel computing infrastructures (parallel and foreach packages)
13. **Other software environments:** introducing Python.

רשימת ספרות:

הקורס נצמד לספר שיכתב עבור הקורס, ויהיה זמין באופן חופשי. הספר יבוסס ברובו על ספר הקורס "R-סביבת תוכנה לניתוח נתונים" (364.2.1201) אשר ניתן במחלקה להנדסת תעשייה וניהול באוניברסיטת בן גוריון לתלמידי תואר שני וזמין באתר: <http://www.john-ros.com/Rcourse>