

סילבוס - תוכנית הוראה לקורס

ניהול והכנת נתונים ליישומי AI

| המחלקה לניהול

Data Management and Preparation for AI Applications | 55123

הרצאה ותרגיל	סוג הקורס:
3_	היקף נ"ז:
ב'	שנת לימודים:
א'	סמסטר:
_____	יום ושעה
_____	שעת קבלה:
_____	מייל מרצה:
_____	קישור לאתר למדה:

תיאור הקורס ומטרות למידה

תקציר הקורס

בעולם הבינה המלאכותית, איכות התובנות הניהוליות תלויה באופן ישיר באיכות הנתונים המזינים את המודל ("Garbage In, Garbage Out"). קורס זה מתמקד בשלב הקריטי והתובעני ביותר בכל פרויקט דאטה (כ-80% מזמן הפרויקט): הכנת הנתונים, ניקויים וטיבם. הסטודנטים ירכשו מיומנויות טכניות מתקדמות בשימוש בספריית Pandas לטיפול בנתונים "מלוכלכים" ממקורות מגוונים: קבצים, מסדי נתונים ו-APIs. הקורס שם דגש על אוטומציה של תהליכי עיבוד ידניים באמצעות קוד קריא ויעיל, והפיכת "החומר הגולמי" הארגוני לתשתית אסטרטגית המוכנה לאימון מודלים של למידת מכונה.

מטרות הקורס

1. להקנות לסטודנטים שליטה טכנית עמוקה בכלים לעיבוד, ניקוי והכנת מערכי נתונים רחבי היקף.
2. לפתח הבנה של הקשר ההדוק בין איכות הנתונים לבין ביצועי מערכות ה-AI בארגון.
3. להכשיר את הסטודנטים בשילוב מקורות מידע שונים לכדי תמונה ניהולית אחת מבוססת נתונים.
4. להטמיע חשיבה ביקורתית בזיהוי הטיות (Biases) ועיוותים בנתונים כחלק מתהליך הניקוי.

תוצרי הלמידה

ידע

1. הלומדים יתארו את ארכיטקטורת ה-Modern Data Stack ואת השלבים השונים בצינור הנתונים מרגע האיסוף ועד להזנה למודל.
2. הלומדים יגדירו את המושגים המרכזיים בטיוב נתונים, לרבות ערכים חסרים, חריגים, ונתונים לא מאוזנים. FEDA
3. הלומדים יכירו את עקרונות העבודה עם נתונים בסביבת Python ואת השיטות המרכזיות להכנת נתונים ליישומי AI.

מיומנויות

1. הלומדים ינתחו מערכי נתונים גולמיים באמצעות חקירת נטונים ראשונית (EDA) כדי לזהות שגיאות, עונתיות וקשרים בין משתנים.
2. הלומדים יעריכו את איכות הנתונים, יזהו בעיות מרכזיות במבנה הנתונים, ויבחרו שיטות מתאימות לניקוי, לאיזון ולהנדסת משתנים.

תכנון מהלך השיעורים (כולל למידה פעילה):

מס' השיעור	נושא השיעור	למידה פעילה	קריאה/ צפיה נדרשת	הערכה תהליכית/מעצבת
1	Pandas Foundations	התקנת סביבת עבודה, טעינת קובצי Excel/CSV ופקודות חקירה ראשוניות (head, info).	1, 6	
2	סינון ובחירת נתונים	תרגול טכניקות חיתוך (Slicing) ואיתור חריגים (Outliers) באמצעות מדדים סטטיסטיים.	1, 6	
3	חקירת נתונים (EDA)	יצירת ויזואליזציות מהירות היסטוגרמות ו- Scatter plots לזיהוי שגיאות במבנה הדאטה.	1, 2	מטלה 1: טעינה, סינון וזיהוי חריגים במערך נתונים "מלוכלך".
4	Missing Data & Cleaning	מעבדה: יישום טכניקות מילוי ערכים מול הסרה, וניקוי שגיאות טקסטואליות.	1, 4	
5	Data Types & Dates	עבודה עם אובייקטי זמן, המרת טיפוסים ויצירת ציר זמן לצרכי ניהול.	1	
6	התמודדות עם נתונים לא מאוזנים	תרגול טכניקות Oversampling ו- Undersampling (שימוש בספריית imbalanced-learn).	7	מטלה 2: ניקוי דאטה וטיפול בערכים חסרים ובנתונים לא מאוזנים.
7	Aggregation & Pivot	בניית טבלאות ציר (Pivot Tables) וסיכומי נתונים מורכבים לפי פלחי שוק/זמן.	1, 6	
8	מיזוג מקורות מידע	תרגול מיזוג והצלבת נתונים ממספר טבלאות וקבצים ליצירת תמונה אחת.	1	
9	SQL to Pandas Bridge	חיבור לבסיס נתונים ארגוני, הרצת שאילתות SQL ומשיכת תוצאות ישירות ל- DataFrame.	8, 6	מטלה 3: מיזוג נתונים ממקורות שונים ושליפה מ-SQL.
10	עבודה עם APIs	שימוש בספריית requests למשיכת נתוני זמן אמת (מניות/מזג אוויר) ופענוח JSON.	1	
11	הנדסת משתנים (Feature Engineering)	יצירת וקטורים, קידוד משתנים One-Hot (Encoding) וביצוע Scaling לקראת אימון מודל.	2, 3	
12	בקרת איכות ואוטומציה	בניית בדיקות תקינות אוטומטיות (Sanity Checks) למערכי הנתונים למניעת הטיות.	4, 5	מטלה 4: הכנה מלאה של מערך נתונים (Features) למודל AI.
13	סיכום: מניחול נתונים להיערכות ארגונית ל- AI	חזרה לבחינה		

ציון סופי

מסקל בציון הסופי	תיאור התוצר
50 % מהציון הסופי	ביצוע מטלות בית
50 % מהציון הסופי	מבחן מסכם

דרישות הקורס

נכחות חובה במפגשי המעבדה (לפחות 80%) והגשת כל המטלות במועד_

דרישות קדם

שם הקורס	מס' הקורס
מבוא לתכנות בסביבת פייתון	55-011
ניהול מערכות מידע ומסדי נתונים	55-051

ביבליוגרפיה: תכנים לקריאה, צפיה והאזנה (עדכני)

1. McKinney, W. (2022). *Python for Data Analysis* (3rd ed.). O'Reilly Media.
2. Géron, A. (2022). *Hands-On Machine Learning with Scikit-Learn, Keras, and TensorFlow* (3rd ed.). O'Reilly Media.
3. Zheng, A., & Casari, A. (2018). *Feature Engineering for Machine Learning*. O'Reilly Media.
4. Redman, T. C. (2013). "Data's Bad Data Problem". *Sloan Management Review*.
5. Pearson, K. E., Saunders, C. S., & Galletta, D. F. (2024). *Managing and Using Information Systems: A Strategic Approach*. John Wiley & Sons.
6. Pandas Official Documentation. <https://pandas.pydata.org/docs/>
7. Imbalanced-learn Documentation. <https://imbalanced-learn.org/>
8. SQLAlchemy Documentation. <https://docs.sqlalchemy.org/>