מבוא לניתוח נתונים – תרגיל מספר 4

- תאריך הגשה: 18/5/2021
- ההגשה לבד או בזוגות. במקרה של הגשה בזוגות -רק אחד מבני הזוג מגיש, אך יש להקפיד לרשום את השם ות.ז. של שני הסטודנטים שהגישו. מי שלא רשום על אף עבודה לא יקבל ציון, ולא יוכל להצטרף לעבודה בדיעבד.
- במידה וסעיף מסוים לא רץ – הציון על סעיף זה הינו אפס.
- יש להראות את הקוד דרכו הגעתם לפתרון. סעיף עם תשובה נכונה אך ללא קוד יקבל ציון אפס.
- התרגיל הינו אישי וניתן לבצעו בזוגות בלבד. אין להעביר פתרונות או חלקי פתרונות בין סטודנטים בכיתה. העברת התרגיל על כל המשתמע מכך לאנשים אחרים שאינם רשומים לקורס אסורה. כל הנ"ל הינם עבירות משמעת אשר יועברו לבחינת ועדת המשמעת של האוניברסיטה.
- **הקפידו להגיש עבודה ברורה עם הסברים**. אל תשאירו אותנו במתח! כתבו באמצעות markdown על איזה סעיף אתם עונים. על סעיף ללא הסברים (או במידה ולא ברור באיזה סעיף מדובר) יורדו לפחות חצי מהנקודות.
- יש להגיש קובץ אחד בלבד, מסוג ipynb.
- איחורים בהגשה מכל סיבה שאינה נמצאת בתקנון הרשמי של האוניברסיטה: על כל יום איחור בהגשה יורדות 5 נקודות בציון התרגיל באופן אוטומטי.

הפעם תנתחו נתונים על שחקני NBA, את המידע קיראו מכאן:

https://raw.githubusercontent.com/ShaiYona/Data-Science2021B/main/Assignments/4/nba.csv

1. (5%) קראו את הקובץ, השמיטו ערכים חסרים והציגו את 3 השורות הראשונות עם העמודות ['Age',Height,'Weight','Salary'] בלבד.
2. (10%) המטרה היא לבנות מודל שמנבא את 'Salary' בהתבסס על העמודות 'Age',Height,'Weight'. אבל אחת העמודות שם בעייתית. תקנו אותה לפורמט שיאפשר ניתוח והסקת מסקנה, הציגו שוב את 3 השורות הראשונות של ['Age',Height,'Weight','Salary']
3. (20%) צרו countplot ו pieplot עבור העמודה position, וכיתבו איפה לדעתכם ניתן לראות את הדאטה בצורה ברורה יותר (הבהרה: זו דעתכם בלבד ולכן אין כאן תשובה שגויה).
4. (10%) פצלו את הדאטה כך ש 30% מהדאטה ישמש למבחן ו70% לאימון. הציגו את 5 השורות הראשונות של עמודות הfeatures שישמשו לאימון ואת 5 השורות הראשונות של הlabel\target שישמשו למבחן.
5. (10%) התאימו מודל לאימון, באיזה מודל השתמשתם ומדוע? (הבהרה: יש להשתמש באחד ממודלי עצי ההחלטה שנלמדו בשיעור ובתרגול).
6. (10%) בצעו תחזית למשכורת השחקנים בהתאם לfeatures בדאטה של המבחן וחשבו את הדיוק של המודל שיצרתם.
7. (20%) הוסיפו את העמודה position , הריצו את המודל מחדש וכיתבו האם המודל עכשו טוב יותר, גרוע יותר או אותו הדבר. רמז: יש לטפל בעמודה position לפני ההוספה.
8. (15%) שפרו את ביצועי המודל על ידי הגבלת עומק העץ (הבהרה: אין צורך לכתוב פונקציות, מספיק למצוא עומק אחד בו הביצועים משתפרים).
9. (10%) נקודות מתנה. יורדו במידה ויוגש ערעור על הציון.