היום נממש עץ החלטה לומד מאפס בפייתון.

נסתכל על ה-data set שנמצא בלינק הבא:
http://archive.ics.uci.edu/ml/datasets/Breast+Cancer+Wisconsin+%28Diagnostic%29

זהו data set שמתאר נתונים על גידולים והאם הגידול הוא שפיר (Benign) או סרטני (Malignant). הפיצ'רים שיש במערכת הם רדיוס הגידול, הקעירות שלו, חלקות, היקף, שטח, מרקם ועוד. הנתונים הינם labeled, כלומר על כל גידול, המספר השני ברשימת ה-features שלוו הינו ה-label (האם הוא שפיר או סרטני. B עבור שפיר,M עבור סרטני). המספר הראשון הינו ה-id.
הנתונים שאותם נחקור נמצאים בקובץ "wdbc.data" מתאריך 5 לפברואר.

הלינק הבא מכיל הסברים מלאים לפתרון של עץ החלטה לומד:
http://machinelearningmastery.com/implement-decision-tree-algorithm-scratch-python/

1. יש לבנות אלגוריתם רקורסיבי לבניית עץ החלטה לפי Gini Index. השתמש.י בשמונים אחוז מהנתונים לשם כך (training data). קודם יש לבנות קוד לפיצול של רמה אחת ואחרי זה נריץ את זה על עץ בצורה רקורסיבית.
   a. בנה פונקציה שמפצלת את ה-dataset לפי חוק.
      הפונקציה מקבלת:
      1. dataset
      2. אינדקס של פיצ'ר שמפצלים עליו
      3. ערך שמפצלים עליו
      הפונקציה מחזירה שני datasets שהם המקורי מפוצל לשניים
   b. בנה פונקציה (get_split) שמקבלת רשימה של נתונים כולל ה-label ובוחרת לפי כלל ג'יני מה הפיצ'ר המפצל ובאיזה ערך.
   c. בנה אובייקט עם מבנה של node. השדות שצריך שיהיו באובייקט:
      i. node ימני
      ii. node שמאלי
      iii. מספר סידורי של feature מפצל
      iv. ערך שבו ה-feature מפצל
      v. עומק שבו אנו נמצאים בעץ
      vi. כל ה-datasets שנמצאים בחלק זה של העץ
      vii. ערך תוצאה אם זה עלה
   d. הרץ את split_node באופן רקורסיבי על העץ. בכל שלב בודקים אם יש את תנאי העצירה. אם לא - מריצים את split_node על הימני ואז על השמאלי ושוב באופן רקורסיבי, כמו בלינק שלמטה. כלל עצירה לפי עומק או לפי שאין איך לפצל יותר, כלומר העלה טהור
2. יש כעת להריץ את העץ על עשרים אחוז הנותרים של הנתונים - test data

בהצלחה!