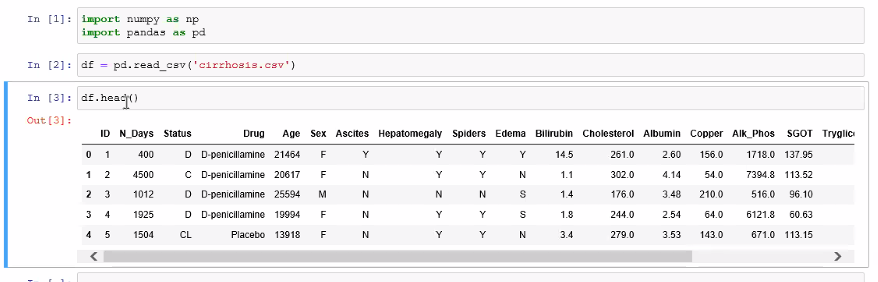
**תרגול 8**

הכנה לתרגיל בית 2



יבוא של ספריות רלוונטיות

קריאה של הקובץ נתונים

סקירה כללית של הdata

Describe עושה סטטיסטיקות של משתנים נומריים

Dtype אומר לכל עמודה מה הסוג שלה

Shape נותן כמה שורות ועמודות יש לנו

Table

Description automatically generated with medium confidence

היסטוגרמה- משתמשים למשתמשים נומרי רציף

Barplot- משתמשים למשתנה קטגורי

\*להוסיף את כל השלבים האלה של סקירת נתונים לתרגיל בית 1, כולל לעשות barplot לפי התכונה שאנחנו רוצים לנבא (area\_type)

נוסיף את matplotlib וseaborn לויזואליזציות של הנתונים:

Text

Description automatically generated

רוצים לעשות ויזואלציה למשתנה הtarget שלנו stage בbarplot כדי לדעת מה השכיחות של כל אחת מהקטגוריות בstage.

Chart, bar chart

Description automatically generated

נהפוך את שלבים 4 ל"יש מחלה" ו1-3 ל"אין מחלה"

נשנה את stage בנתונים ל1 או 0:

Graphical user interface, text, application, Word

Description automatically generated

נמשיך לpreprocessing

לשנות את המשתנה stage לקטגורי

Table

Description automatically generated with medium confidence

אם אנחנו רוצים רק את המשתנים הנומריים, או כל סוג אחר-נשתמש בselect\_dtypes(include=)

לדוגמה לקבל רק את המשתנים הנומריים:

Table

Description automatically generated with medium confidence

\***למשתנים קטגוריים לא עושים קורלציות**

מקדם מתאם זה covariance בין 2 משתנים חלקי מכפלת הסטיות תקן שלהם

לא ניתן לעשות זאת בין משתנים קטגוריים

מטריצת קורלציות בין המשתנים הנומריים השונים:

A screenshot of a computer

Description automatically generated with medium confidence

הערכים בכל מלבן הם מקדם המתאם בין התכונות השונות

אם הערך נמוך אז אין קשר לינארי ביניהם

אם יש קורלציה גודלה מאוד בין 2 תכונות אז אפשר לבדוק אם להוריד אחד מהם, כי הם בעצם מעידים על אותו הדבר, כך נשפר את זמן הריצה.

בדרך כלל משתמשים בקורלציה ועוד שיטה, כמו random forest ולהסתכל על feature importance כדי לדעת אם כדאי להוריד תכונה או לא.

Figsize נותן את הגודל של הplot

Annotation=true נותן את הערכים המספריים בheat map

Axis=0 שורות

Axis=1 עמודות

Inplace=true באמת מוריד את מה שעושים לו drop ולא רק מראה איך הdf יראה בלעדיהם



לבדוק כמה ערכי null יש בכל עמודה:

Graphical user interface, application

Description automatically generated

להרבה תכונות יש את אותם ערכים חסרים, זאת אומרת שיש סיכוי שיש הרבה שורות שרובן עם מידע חסר

Table

Description automatically generated

כיוון שיש הרבה שורות עם הרבה ערכי תכונות חסרות, נוריד את תצפיות אלו

נשתמש בdropna עם 9=threshold, זאת אומרת שאם יש שורה שיותר מ9 תכונות חסרות אז נוריד את שורה זו(9 כי זה יותר מחצי מהתכונות)

Graphical user interface, text, application, email

Description automatically generated

נבדוק עם כמה שורות עם na נשארנו עכשיו

Table

Description automatically generated

משתנים קטגוריים:

Table

Description automatically generated

נעביר את הfeatures הקטגורים ל1 ו0 כדי שנוכל לעשות knn:

exclude בselect dtypes אומר מה לא להכניס

pd.get\_dummies() –מבצע one hot encoding

one hot encoding- משתמשים למעבר של משתנים קטגוריים לנומריים

|  |  |  |  |
| --- | --- | --- | --- |
| ירוק | כחול | אדום | צבע |
| 0 | 0 | 1 | אדום |
| 0 | 1 | 0 | כחול |
| 1 | 0 | 0 | ירוק |
| 0 | 1 | 0 | כחול |

כל שורה מקבלת מערך של 0 ו1 שאומר איזה סוג היא

אם זה לא אדום ולא כחול אז זה בטוח ירוק, לכן אפשר לוותר על העמודה הראשונה של הירוק-drop\_first=true

Table

Description automatically generated



נוריד את המשתנים הקטגוריים:



נעשה concat כדי לחבר את כל המשתנים הנומריים וכל הקטגוריים אחרי encode

כדי לקבל df אחד עם כל השינויים שעשינו בתכונות:

Table

Description automatically generated

עכשיו נבצע KNN כדי למלא את הערכים החסרים:

נייבא מsklearn את knn imputer אשר ממלא ערכים חסרים



בKNNimputer נגדיר את מספר השכנים ובweight distance -ממוצע ממושקל על פי מרחקים

לאחר מכן נבצע fit\_transform כדי לקבל את מערך של מערכים שהוא הdf החדש עם הערכים המשלימים:

Text

Description automatically generated

נהפוך את המערך עם הערכים המשלמים לdf באמצעות דריסה של הdf הישן:

Text

Description automatically generated with medium confidence

תרגול 8: ML\_example כולל את ההמשך בו עושים ניבוי

**MSE**

הוא מושפע מיחידות המדידה שבודקים ( לכן בתרגיל יצא ערכים מאוד גדולים)

אם רוצים מדד בין 0 ל1 עד כמה טוב עץ רגרסיה עובד הוא R^2 .

תרגיל בית 2

נקבל 3 קבצים של נתונים על מכירות בwalmart

2012-2013 train

2014 test

בנוסף נקבל נתונים של תחנות מזג אוויר

הdata מאוד בעייתי וצריך לנקות אותו כדי שנוכל להשתמש בו לניבוי

50% מהציון הוא על לנקות את הdata

לעשות ויזואליזציות כדי להבין את הקשר בין התכונות

נציג את הdf החדש עם הערכים המשלימים:

Table

Description automatically generated with medium confidence