תרגול 2

Pre-processing- לנקות את הנתונים מדברים לא רלוונטיים, ערכים לא תקינים, ערכים חסרים

**Imputation**

התהליך של להחליף נתונים חסרים

אופציה 1: ניתן להוריד תצפיות או משתנים

צריך לחשוב טוב לפני שעושים זאת כדי לא לאבד מידע

אופציה 2: ניתן לתת לכל הערכים החסרים ערך חדש

אופציה 3: אמדים סטטיסטים(ממוצע, חציון, שכיח(mode) )

להחליף את הערכים החסרים באמדים סטטיסטיים, תהליך זה מתבצע עבור כל feature בנפרד

אופציה 4: KNN -אלגוריתם שמחפש את K התצפיות הכי קרובות לתצפית הנוכחית, כך האלגוריתם מנבא את הערך של הערכים החסרים

אופציה 5: MICE-:

1.עבור כל אחד מהfeatures תחילה מנסים להשלים את כל הערכים החסרים עם הממוצע

2. מחסירים מfeature אחד את הערך החסר

3. משתמשים ברגרסיה לינארית כדי להשלים את הערך החסר שהורדנו מהfeature.

4. מבצעים זאת עבור כל ערך חסר בכל feature.  
5.מחסירים מהתוצאה האחרונה שקיבלנו מתהליך זה, את הנתונים שהשלמנו עם הממוצע בשלב 1.

6.חוזרים על שלבים 2-5 עד שההפרשים הם מאוד קטנים או אחרי מספר איטרציות שקבענו מראש

**קורלציות**

התהליך של לזהות קשרים בין features. כאשר יש קורלציה יותר מדי גבוהה בין features אז נשקול להוריד אחד מהfeatures. קורלציה גבוהה מדי בין משתנים אומרת שהfeatures אומרים את אותו הדבר.

**קורלציית פירסון**- מעריך את היחס הלינארי בין 2 משתנים המשכיים. ככל שאנו קרובים יותר ל1 אז יש יותר קורלציה חיובית בין המשתנים( יחס ישר), ככל שמתקרבים ל-1 יש קורלציה שלילת יותר גדולה (יחס הפוך). ככל שמתקרבים ל0 אין קורלציה.

**קורלציית ספירמן-** בודק את היחס המונוטוני בין 2 משתנים

קשר מונוטוני עולה- ככל שxעולה אז y מעולם לא יורד

קשר מונוטני יורד- ככל שx עולה אז y מעולם לא עולה

קשר לא מונוטוני- ככל שx עולה אז y לפעמים עולה ולפעמים יורד

\*קורלציה אומרת שיש קשר בין המשתנים אבל לא אומר שמשתנה אחד גורם לשני (לא בהכרח קשר של סיבה ותוצאה בין המשתנים)

**מטריצת קורלציות-** לוקחים את המשתנים, ובודקים את הקורלציות בין כל המשתנים, כל תא במטריצה מראה את רמת הקורלציה בין המשתנים. ככל שהצבע יותר קרוב לכחול אז יש יותר קורלציה בין המשתנים. באלכסון המרכזי תמיד יהיה 1 כי זה יהיה הקורלציה בין משתנה לעצמו.

באופציית ויזואליזציה השנייה של המטריצה האלכסון המרכזי מראה בהיסטוגרמה את ההתפלגות של המשתנה.

**טרנספורמציה**

**סטנדרטיזציה-** שינו התפלגות הנתונים להתפלגות נורמלית

**נורמליזציה-** שינוי והגבלת ערכי הfeatures לטווח מסוים

**One hot encoding-**כאשר יש משתנה קטגורי, כמו צבעים, אז המודלים לא עובדים איתם טוב בדרך כלל, לכן נחליף את המשתנים הקטגורים לעמודות כאשר בכל שורה יש 1 בשורה אחת אשר מסמן את הערך הקטגורי הנוכחי ובכל שאר השורה יהיו אפסים.

Pandas

Head()- נותנת 5 שורות ראשונות

Tail()- נותנת 5 שורות אחרונות

Sample(x)- נותן מדגם אקראי של מסד הנתונים, מספר השורות הוא לפי הx שנותנים

Describe()- מתאר את המידע לפי חציון, מינימום,maximum.

Corr()- מבצע קורלציות, הדיפולט הוא פירסון

ניתן להציג בheat map או דרכים שונות.