למידה חישובית – תרגיל 2

מיכל אברמוב, 301834297
בוריס בורשבסקי, 311898746

מה מצורף

1. מחברת המפרטת בשלבים את מה שעשינו
2. קובץ main.py המכיל את כל הקוד של התרגיל
3. קובץ html שמאפשר צפיה נוחה יותר במחברת
4. קבצי הdata.

הFlow של התרגיל

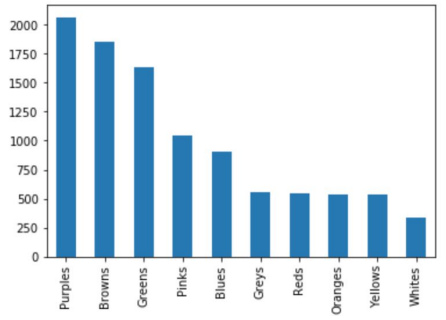
1. קריאת וחלוקת את המידע

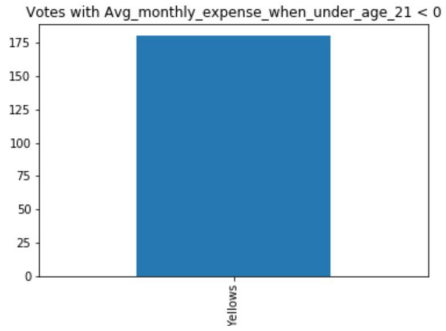
a. קראנו את המידע והוספנו עמודה בשם "Split" שתעזור לנו בחלוקה לDatasets.
b. חילקנו את המידע ל60%, 20%, 20%, וייצאנו את הraw data אחרי החלוקה (וללא עמודת הsplit) ל3 קבצים:
   i. Raw_data_train
   ii. Raw_data_validation
   iii. Raw_data_test
c. בהמשך נשתמש בעמודה הsplit כדי לוודא שחלוקת השורות על הmodified data לקבצים מתאימה לחלוקת השורות בraw data.

2. מיפוי הפיצ׳רים

a. מיפינו את הפיצ׳רים הקטגוריאליים והנומריים ע״מ לעבוד עליהם בהמשך
b. חיפשנו פיצ׳רים שבהם ישנם ערכים לא הגיוניים, מצאנו שישנם מספר פיצ׳רים עם ערכים שליליים שאינם מסתדרים, למשל:
   i. Avg_monthly_expense_when_under_age_21
   ii. AVG_lottary_expanses

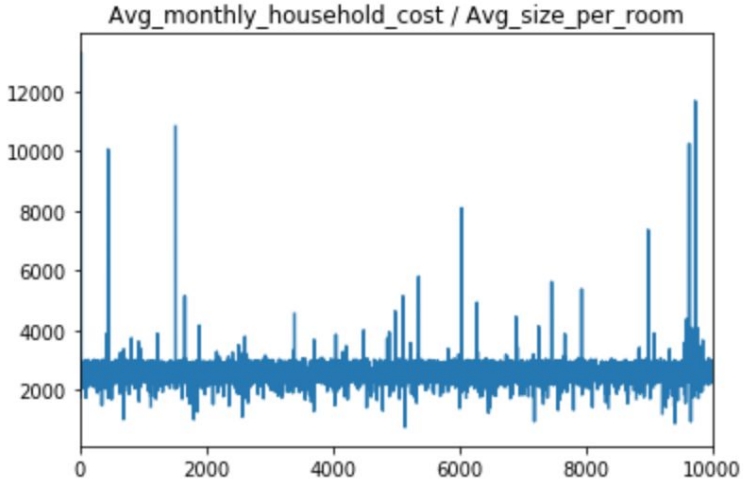
אבל ראינו שכל הערכים שקיימת קורלציה בינם לבין המטרה (כולם בחרו בצהוב) אז החלטנו להשאיר אותם במחשבה שהם ינורמלו לאחר מכן - ניתן לראות גרף.



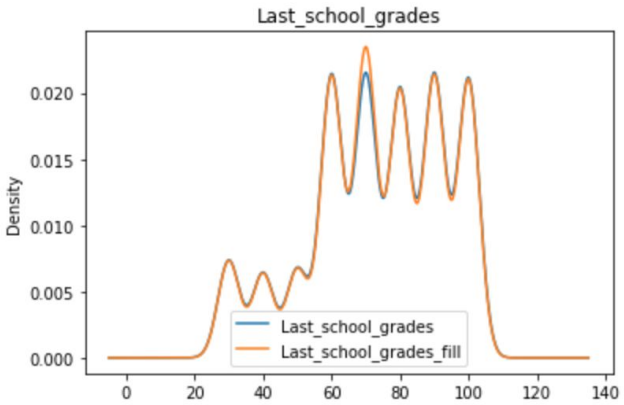


3. מילוי ערכים חסרים

a. בהתחלה בדקנו קורלציה בין הפיצ׳רים וחיפשנו את כל הפיצ׳רים הנומריים ביניהם ישנה קורלציה של לפחות 0.95 - במקרה זה השלמנו את כל הערכים החסרים בעזרת הקורלציה וסימנו את הפיצ׳ר שהשלמנו ממנו כלא רלוונטי
באיור: דוגמא לפיצ׳רים עם קורלציה:



b. לאחר מכן עברנו על כל הפיצ'רים עם קורלציה של 70% ואותם השלמנו גם בעזרתה אך ללא סימון הפיצרים
c. לבסוף השלמנו את הערכים שנותרו חסרים ע״י הערך החציוני ובדקנו שלא פגענו בהתפלגות כמו בגרף המצורף:



d. את הערכים הקטגוריאליים השלמנו ע״י הערך הנפוץ ביותר תחת אותו label

4. טרנספורמציה של המידע

a. ערכים קטגוריאליים:
   i. בערכים בהם יש חשיבות לסדר מיפינו את הערכים לפי הסדר
   ii. בערכים בוליאנים הגדרנו ש-NO מייצג -1, yes מייצג 1, ו maybe 0
   iii. בערכים ללא חשיבות לסדר יצרנו פיצ׳רים בוליאני עבור כל אחד מהאפשרויות הקיימות כך שנוצר בעצם עבור כל פיצ׳ר קבוצה נוספת של פיצ׳רים למשל: עבור Occupation יצרנו Is_occupation_hightech וכו׳
b. ערכים נומריים:
   i. בחרנו לעשות scale לינארי לכל הערכים כך שכולם יהיו בטווח בין 0 ל1. (בהנחה שכבר בדקנו קורלציה בין הפיצ׳רים)
c. שינינו את הvote לערך נומרי לפי כל צבע

5. ניקוי רעשים

a. ביצענו ניקוי רעשים בעזרת outlier detection ע״מ לנקות את כל המופעים שבהם ערך הפיצ׳ר רחוק בלפחות 3 סטיות תקן מהערך הממוצע של אותו פיצ׳ר.
b. הסתכלנו על כל השורות שבהם הורדנו פיצ׳ר אך לא מצאנו משהו חשוד.

6. בחירת פיצ׳רים

a. בשלב זה הורדנו את הפיצ׳רים שסימנו מיותרים בשלב הקורלציה והורדנו מופעים בהם יש עדיין ערכים ריקים
b. השתמשנו בVarianceThreshold על מנת להעיף את כל הפיצרים שאין להם שונות (לא הצלחנו להעיף כלום)
c. שלב ה Filter Method:
   i. הרצנו select Percentile עבור הפיצ׳רים בעזרת f-classif ובעזרת mutual-information בחרנו את הפיצ׳רים ב25% אחוז העליונים בשני המקרים
d. שלב בWrapper Method:
   i. השתמשנו בRFE בשני ריצות שונות, אחת עם cross validation = 3 ובשניה עם cross validation מסוג StratifiedKFold הוספנו את כל הפיצ׳רים שנוספו בכל אחת מהריצות.
e. השתמשנו ב tree based feature selection, גם שם בחרנו את 25% הפיצ׳רים עם הtree weight הגבוה ביותר והוספנו לרשימת הפיצ׳רים הנבחרים.

7. שמירת המידע

a. שמרנו את המידע המעובד לקבצים לפי עמודת split שהגדרנו בהתחלה לקבצים
   i. processed_test.csv
   ii. processed_train.csv
   iii. processed_validation.csv

8. הפיצ׳רים שנבחרו לבסוף:

a. פיצרים חדשים:
   i. Married'
   ii. 'Avg_monthly_income_all_years'
   iii. 'Is_Most_Important_Issue_Financial'
   iv. 'Yearly_IncomeK'
   v. 'Number_of_valued_Kneset_members'
   vi. 'Is_Most_Important_Issue_Environment'
   vii. 'Garden_sqr_meter_per_person_in_residancy_area'
   viii. 'Will_vote_only_large_party'
   ix. 'Avg_monthly_expense_when_under_age_21'
   x. 'Is_Most_Important_Issue_Military'
   xi. 'Is_Most_Important_Issue_Education'
   xii. 'Is_Most_Important_Issue_Foreign_Affairs'
   xiii. 'AVG_lottary_expanses'
   xiv. 'Is_Most_Important_Issue_Other’
   xv. 'Last_school_grades'
   xvi. 'Looking_at_poles_results'
   xvii. 'Weighted_education_rank'
   xviii. 'Overall_happiness_score'
   xix. 'Is_Most_Important_Issue_Social'

b. פיצרים מקוריים (שעליהם מבוססים הפיצ׳רים החדשים):
   i. Number_of_valued_Kneset_members'
   ii. 'Avg_monthly_expense_when_under_age_21'
   iii. 'Avg_monthly_income_all_years'
   iv. 'Yearly_IncomeK'
   v. 'Married'
   vi. 'Last_school_grades'
   vii. 'Weighted_education_rank'
   viii. 'Garden_sqr_meter_per_person_in_residancy_area'
   ix. 'Will_vote_only_large_party'
   x. 'Looking_at_poles_results'
   xi. 'Overall_happiness_score'
   xii. 'Most_Important_Issue'
   xiii. 'AVG_lottary_expanses'