עבודת בית מסכמת - למידה חישובית

מגישות : הלל דודיאן 318593720 יהודית פרל 311596852

<u>מסקנות על המשימות שאחרי ההגשה</u>

נזכיר -
הקובץ נתונים שאנחנו בחרנו נלקח מאתר UCI ,שבו מאגרי נתונים ללימוד למידה חישובית . הקובץ מכיל מאגר מידע על אנשים ממדינות שונות. כולל בתוכו 14 מאפיינים ועמודת תוצאה ו- 32562 שורות. עמודת התוצאה הינה המשכורת עבור כל אדם (אם המשכורת שלו מעל או מתחת ל50 אלף בשנה). נרצה לבנות מודל שיחזה האם אדם מסוים ירוויח יותר או פחות 50 אלף בשנה.

בחרנו בנושא זה כי נושא שמעניין את רובנו, מכיל מידע רב, שברובו נתונים מורכבים המאפשר ביצוע של פילוחים שונים ומגוונים.

לצורך כך בחרנו בשני אלגוריתמים Logistic Regression ו-Decision tree classification המבצעים

בנינו שני מודלים על סמך כל אלגוריתם והשוואנו בניהם על סמך הרצה כל פעם על מדינה אחת

<u>המשימות שקיבלנו אחרי הגשת הפרוייקט :</u>
להשוות את התוצאות השגיאות של המודלים על מדינה אחת ועל מספר מדינות(בכל מודל)
לבדוק האם יש overfiting לעץ במודל שלנו

תוצאות גרפיות על אלגוריתם logistic regression

נציג את תוצאות המודל logistic regression על מדינה אחת ועל מספר מדינות באמצעות הגרפים הבאים :

1. גרף המציג את השגיאה הממוצעת עבור כל למדה
2. שגיאות Test ו train

1. <u>גרף המציג את השגיאה הממוצעת עבור כל למדה</u>

בפרוייקט, חיפשנו את ערך ה λ האופטימלי, בעזרת k-fold cross validation.
והצגנו בגרף את ערך השגיאה הממוצעת לכל λ.

2. <u>שגיאות Test ו train</u>

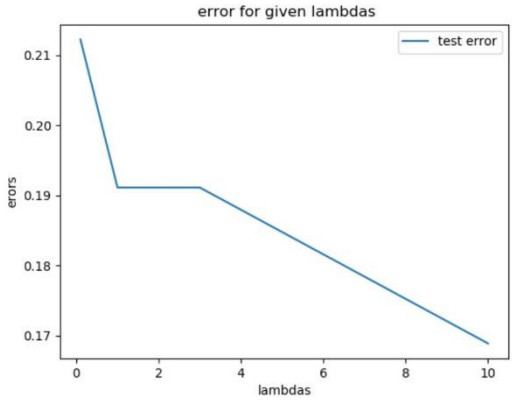
עבור ה λ הכי טוב ציירנו גרף שציר ה-x שלו זה מספר הדוגמאות וציר ה-y הוא השגיאה על קבוצת הלמידה, והשגיאה על קבוצת test קבועה

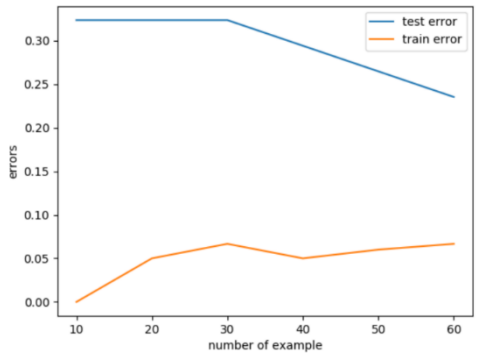
בהרצה על מדינה אחת – הוצגה בדוח הקודם
בפרוייקט הרצנו על מספר מדינות (כל פעם על מדינה אחרת)

בדוח הצגנו דוגמת הרצה עבור המדינה קובה המכילה 95 דוגמאות.

הגרף השמאלי מראה את השגיאות עבור כל למדה, הלמדה שהשגיאה בה הכי מינימלית היא λ =10 , אבל זה לא אומר שהיא האופטימלית ביותר, יכול להיות שעבור למדה גדולה יותר הגרף ימשיך לרדת.

הגרף מימין מראה שהשגיאות של המודל עבור הנתונים שהוא מכיר נמוכות בהרבה מהשגיאות על הנתונים שהוא לא מכיר, כלומר overfitting

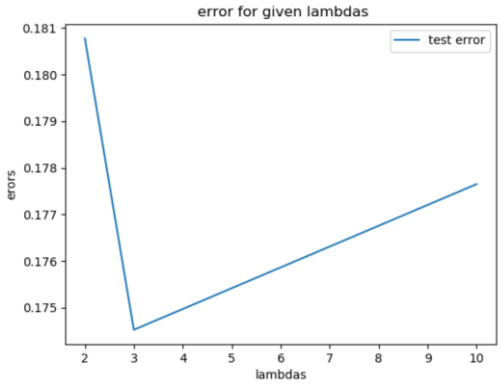


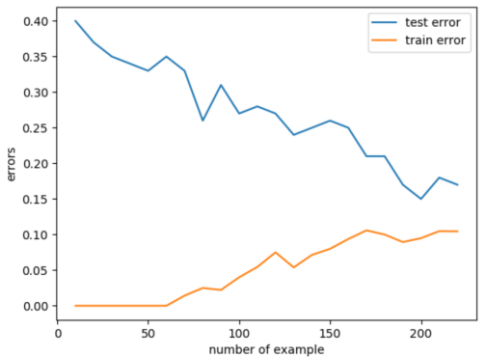


הרצה על מספר מדינות – בעקבות המשימות שאחרי ההגשה כעת, הרצנו על שלוש מדינות: קובה, פיליפין, אקוודור (מספר הדוגמאות 322)

<u>בגרף השמאלי</u> – הלמדה האופטימלית היא 3
ואפשר לדעת בוודאות שזו הלמדה האופטימלית, מכיוון שעבור למדה נמוכה יותר או גבוהה יותר – השגיאות גבוהות יותר.

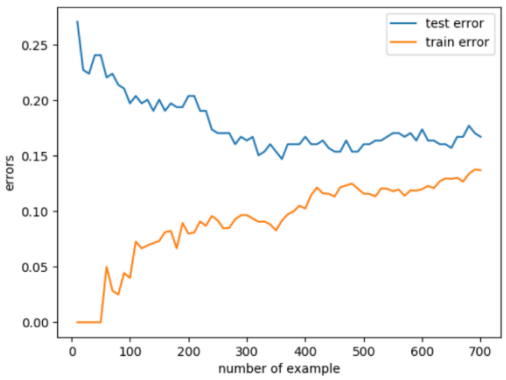
<u>בגרף הימני</u> – רואים שהפער בין השגיאות עבור הנתונים שהוא מכיר לבין השגיאות של הנתונים שהוא לא מכיר – הולך ומצטמצם ככל שמספר הדוגמאות עולה

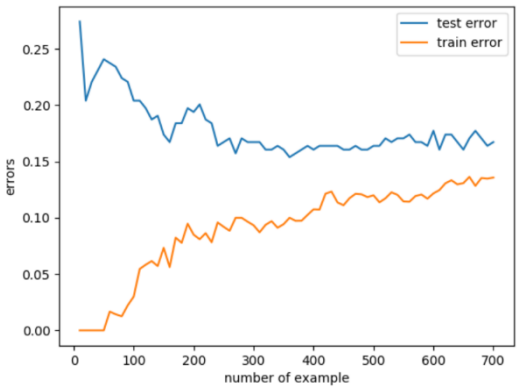




על מנת לבצע הרצה על מספר רב יותר של דוגמאות לקחנו את הנתונים על ארה"ב והרצנו על 1000 דוגמאות מתוכם (מעבר למחשב היה קשה להגיב)
בנוסף, כדי שההשוואה בין מדינה אחת לבין מספר מדינות תיהיה הוגנת הרצנו מספר דוגמאות שווה בין מדינה אחת לבין מספר מדינות.

הגרף הימני - מציג את השגיאות על הרצת מדינה אחת - אהר"ב
הגרף השמאלי - מציג את השגיאות על הרצת 3 מדינה אחת - אהר"ב, קובה וגרמניה





ניתן לראות שגם כאן, שבשתי הגרפים הפער בין השגיאות על הtest והשגיאות על ה train הולך וקטן, ככל שגדל מספר הדוגמאות עולה.

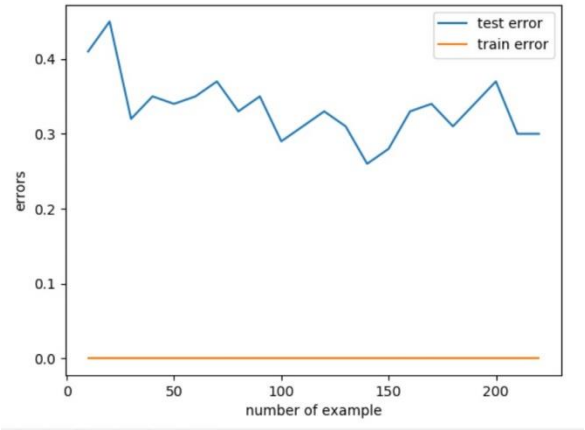
מסקנות–

- כאשר הרצנו רק על ארה"ב שמכילה לבדה אף מספר רב יותר של דוגמאות בהשוואה להרצה על השלוש מדינות ראינו שיפור בתוצאות (ניתן לראות זאת גם במודל העץ)
- הרצה לפי מדינות נראית לנו נכונה יותר, הקושי בנתונים שלנו שלרוב ,בכל מדינה היה יחסית קצת דוגמאות ולכן המודל הראה תוצאות טובות יותר על כמה מדינות .
- מתקבל מודל מדויק יותר ככל שמריצים על יותר דוגמאות.
- אין הבדל מהותי בין הרצה על מדינה אחת לבין הרצה על מספר מדינות (כאשר מספר הדוגמאות שווה)

תוצאות גרפיות על אלגוריתם Decision tree

נציג את תוצאות המודל Decision tree על מדינה אחת ועל מספר מדינות
באמצעות גרף שגיאות על קבוצת ה- Test ועל קבוצת ה-train
ונבדוק האם יש אובר פיטינג לעץ

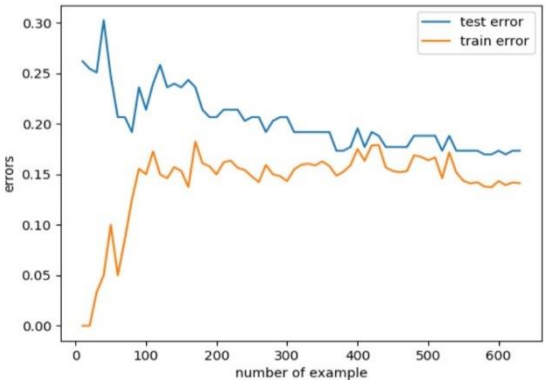
בהתחלה קיבלנו את הגרף הבא:

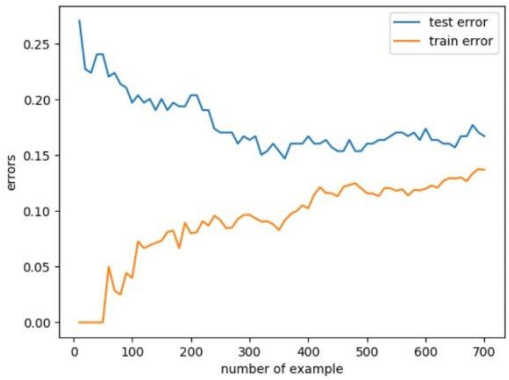


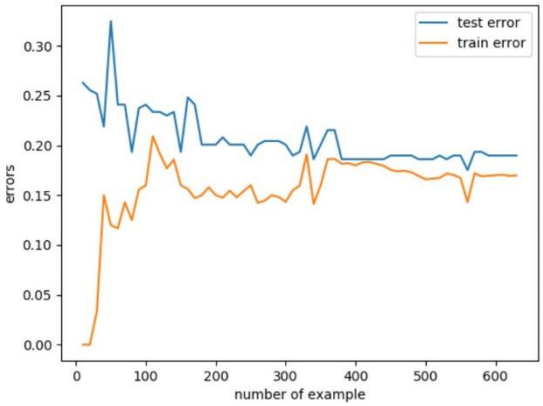
כלומר, המודל תמיד צודק על קבוצת trainה
יש למידה על קבוצת testה ככל שמספר הדוגמאות גדל – אבל היא מעטה מאוד.
והתקבל overfitting
חשבנו למה זה קרה,
והבנו שאם לא מגבילים את העץ – הוא יקבל החלטות רק אם הן נכונות במאה אחוז.
ולכן הוא יהיה מותאם באופן מוחלט לקלטים שהוא מכיר, ולא יחזה בצורה טובה קלטים שאיננו מכיר.

זוהי הסיבה ל overfiting שקבלנו ,ולכן הוספנו הגבלה של עומק העץ.

<u>בגרף משמאל -</u> סיננו את הנתונים **לפי ארצות הברית**, והרצנו על 1000 דוגמאות – עץ החלטה מסוג ג׳יני עם עומק של עד 3.
<u>בגרף מימין -</u> סיננו את הנתונים **לפי ארצות הברית, גרמניה וקובה** והרצנו על 1000 דוגמאות – עץ החלטה מסוג ג׳יני עם עומק של עד 3.







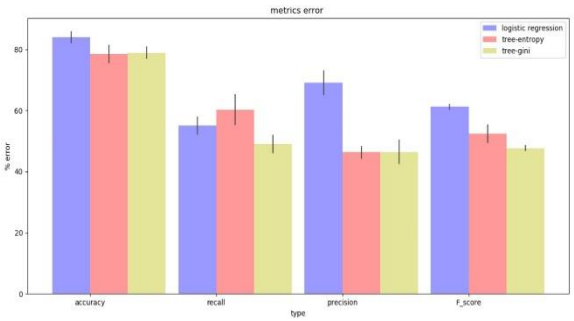
בגרף הבא,
סיננו את הנתונים **לפי ארצות הברית, גרמניה וקובה** והרצנו על 1000 דוגמאות – עץ החלטה מסוג **entropy** עם עומק של עד 3.

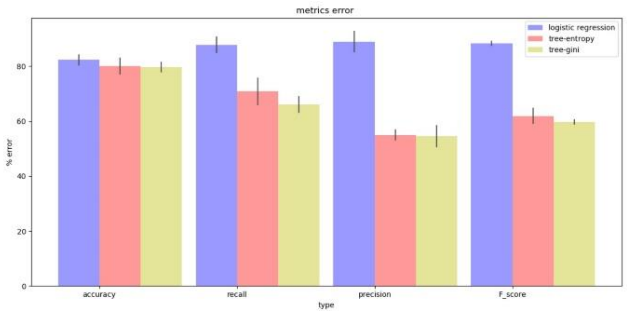
מסקנות –

- לאחר הגבלת עומק העץ ראינו שאין overfitting לעץ, כלומר ככל שמספר הדוגמאות הולך וגדל ככה הפער בין השגיאות על הנתונים שהוא מכיר ועל הנתונים שהוא לא מכיר, הולך וקטן.
- על מנת לקבל דיוק מריבי על מדינה ספציפית נעדיף להריץ את המודל על כל מדינה בנפרד

תוצאות גרפיות השוואה בין על אלגוריתם Decision tree ל logistic regression

הגרף מציג את מדדי הדיוק בlogistic regression מול מדדי הדיוק ב decision tree לפי מדד gini ו entropy בעץ





הגרף מימין - הרצה על שלוש מדינות ארצות הברית, גרמניה וקובה

גרף משמאל- על מדינה אחת (ארה״ב)

העמודות **הסגולות** מתארות את אחוזי השגיאה של logistic regression
העמודות **הורודות** מתארות את אחוזי השגיאה של עץ החלטה בעומק 3 מסוג entropy
העמודות הצהובות מתארות את אחוזי השגיאה של עץ החלטה בעומק 3 מסוג ג׳יני

לפי המדדים הבאים Accuracy,recall ,Precision ,F_score (בהתאמה)

מסקנות –

- ניתן לראות שהמודל מדויק יותר כאשר אנחנו מריצים logistic regression לעומת decision tree- **חשוב לציין** כי בדוח הקודם קבלנו תוצאה **הפוכה** כיוון שלא הגבלנו את עומק העץ ולכן הדיוק של מודל עץ ההחלטה יצא טוב יותר
- כשמריצים על שלוש מדינות יש אחוז שגיאה גבוה יותר מאשר הרצה על מדינה אחת , נכון ורצוי יותר להשוות על מדינה אחת מכיוון שלכל מדינה מאפיינים ייחודים לה
- כאשר אנחנו משווים בין המדדים השונים בעץ gini ו entropy קבלנו דיוק גבוהה יותר בממד ו entropy ונעדיף להשתמש בו.
אופן החישוב שונה המדדים gini ו entropy כיוון שממד entropy משתמש בחישוב ב- log ים ,העלות החישובית שלו גדולה יותר –ניתן לומר שעדיף להשתמש במדד גיני במודל עץ החלטה בהיבט חישובי.
ולכן, בעניין של העדפה של דיוק מול עלות חישובות במידה ונרצה דיוק גבוה יותר נמליץ להשתמש במדד entropy ואם נרצה להוריד עלות חישובית – מדד גיני.

תוצאות entropy מול logistic regression

```
--------DecisionTree type entropy ---------------
Results of DecisionTreeClassifier with about the country United-States
accuracy 78.59778597785979
recall 60.37735849056604
precision 46.3768115942029
F_score 52.459016393442624
TPR 60.37735849056604
FPR 16.972477064220186
-----------------------------------------------
--------------LogisticRegression----------------
Results of LogisticRegression about the country United-States
accuracy 84.0
recall 55.072463768115945
precision 69.0909090909091
F_score 61.29032258064515
TPR 55.072463768115945
FPR 7.35930735930736
```

תוצאות gini index מול logistic regression

--------DecisionTree type gini ----------
Results of DecisionTreeClassifier with about the country United-States
accuracy 78.96678966789668
recall 49.056603773584904
precision 46.42857142857143
F_score 47.70642201834863
TPR 49.056603773584904
FPR 13.761467889908257

----------LogisticRegression----------
Results of LogisticRegression about the country United-States
accuracy 84.0
recall 55.072463768115945
precision 69.0909090909091
F_score 61.29032258064515
TPR 55.072463768115945
FPR 7.35930735930736