למידת מכונה פרויקט

מאת :

אוריאל זקס

ליעד בן משה

תיאור המאגר

בפרויקט זה אנחנו משתמשים במאגרים https://www.kaggle.com/uciml/student-alcohol-consumption .

הנתונים התקבלו בסקר שנערך בקרב תלמידי תיכון ומחולק לשני מאגרים תלמידים שלקחו קורס במתמטיקה ותלמידים שלקחו קורס בפורטוגזית (משני בתי ספר שונים) . הוא מכיל הרבה מידע חברתי, מגדרי ולימודי מעניין על סטודנטים, ומשפחותיהם.

שני המאגרים מכילים אותם נתונים על תלמידים שונים.

חשוב לציין שהמאגר של תלמידי קורס המתמטיקה מכיל מידע על 395 סטודנטים . ושהמאגר של קורס הפורטוגזית מכיל מידע על *650* סטודנטים.

שאלות ותוצאות

בפרויקט שלנו הצגנו ארבע שאלות שאותם ניסנו לחזות בעזרת המאגרים באמצעות 4 שיטות למידת מכונה שונות. עבור כל שאלה חילקנו את המאגר באופן אקראי לשני חציים . חצי לאימון וחצי לבחינה\ בדיקה. בחלק זה של הפרויקט נציג את השאלות , את דיוק החיזוי של כל שיטת לימוד עבור כל מאגר בנפרד, (מתמטיקה ופורטוגזית). בעיות שנתקלנו בהם ודרך ההתמודדות שלנו איתם. וניתוח שלנו על השאלה ותוצאות הלמידה.

שאלה 1: על פי מאפיינים (ציונים,מצב סציואקונומי,גיל,בריאות וכו'), האם התלמיד יצרוך כמות גדולה של אלכוהול?

בשאלה זאת היה קשה לנו להגדיר במפורש מהי צריכת אלכוהול גובהה.

במאגר יש התייחסות לכמות שתיית אלכוהול בשני אופנים

שתיית אלכוהול בימי עבודה (1-5) , שתיית אלכוהול בסוף השבוע (1-5).

בלי התייחסות למשמעות הדירוג.

לכן בשאלה זאת נחזה שתי תוצאות , עבור שתייה בימי עבודה ועבור שתייה בסופי שבוע.

כאשר דירוג 4-5 יחשב שתייה מרובה , 1-3 יחשב שתייה לא מרובה.

דיוק

	מתמטיקה	**פורטוגזית**
ADABOOST	ימי שבוע: 0.92451178 סוף שבוע: 0.7908249158	ימי שבוע: 0.9252307692 סוף שבוע: 0.8032307692
SVM	ימי שבוע: 0.947306397 סוף שבוע: 0.83207070	ימי שבוע: 0.945538461 סוף שבוע: 0.832717
KNN	ימי שבוע: 0.948821 סוף שבוע: 0.79528619	ימי שבוע: 0.940923076 סוף שבוע: 0.791333333
NN	ימי שבוע: 0.95484848 סוף שבוע: 0.791245791	ימי שבוע: 0.948061538 סוף שבוע: 0.7987692

ניתוח תוצאות שאלה **1**

1) ניתן לראות שהדיוק יחסית קרוב בין שיטות הלימוד(עם חיסרון קל לADABOOST) .
2) אין הפרש חד משעמי באף אחד מהשיטות בין קורס המתמטיקה לקורס הפורטוגזית.

3)**בכל** השיטות יש הפרש משמעותי בדיוק בכמות השתייה בימי שבוע מול סופי שבוע.(לדוגמא בNN יש הפרש בדיוק מעל 0.16)

ההשערה שלנו להפרש הגדול הזה היא שקשה יותר לחזות את כמות צריכת האלכהול בסופי שבוע, כיוון שכמות גדולה יותר מהתלמידים שותים בסופש מכיוון שלדעתנו רובם חוגגים ומבלים בסופש בלי קשר לפרמטרים שלהם.

כדי לחזק את ההשערה שלנו. נראה את יחסי השתייה בקרוס פורטוגזית :

<u>ימי השבוע רגילים:</u> צורכים 34 לעומת 615 שלא צורכים הרבה.

<u>בסופי שבוע:</u> צורכים 132 תלמידים לעומת 517 .

אבל שאר הפרמטרים נשארים זהים ולכן התלות בהם חלשה יותר ומעט קשה יותר לחזות את צריכת האלכוהול.

שאלה 2: על פי כמות שתייה (באמצע שבוע וסופ"ש) ונוכחות בבית הספר , חיזוי ציוני התלמיד.

המאגרים מדרגים את ציוני התלמיד 0-20 , כדי לפשט את חיזוי הציונים נתנו אפשרות טווח בין הציון המקורי לתוצאת החיזוי (ברירת מחדל 3).

	מתמטיקה	**פורטוגזית**
ADABOOST	0.596363636	0.72141538
SVM	0.551515	0.75984
KNN	0.56272727	0.7331692
NN	0.5409090	0.7183999999

ניתוח תוצאות שאלה 2

1) ניתן לראות שאחוזי ההצלחה שלנו לחיזוי נמוכים יחסית לשאלה 1.
גם אחרי הגדלת הטווח לחיזוי הציון , עדיין 0-20 תוצאות זה הרבה אפשריות לחיזוי . ופוגע משמעותית בדיוק .
2) ניתן לראות הבדל משמעותי בין הדיוק לקורס פורטוגזית לעומת קורס המתמטיקה .

ההשערה הראשונית שלנו הייתה שכיוון שקורס המתמטיקה מכיל פחות סטודנטים אז הוא לומד פחות נתונים ולכן החיזוי פחות טוב.
כדי לתמוך בהשערה הזאת צמצמנו את גודל האימון בכל השיטות בקורס הפורטוגזית לסדר גודל של קרוס המתמטיקה. אכן הדיוק ירד בכל השיטות כבערך ב0.02 . אבל ההפרש עדיין נשאר משמעותי.

השערה שנייה

הסתכלנו על התפלגות ציוני התלמידים כדי לזהות דפוסים שעלולים לגרום להפרש בדיוק.

מתמטיקה		פורטוגזית	
0	38	0	15
4	1	1	1
5	7	5	1
6	15	6	3
7	9	7	10
8	32	8	35
9	28	9	35
10	56	10	97
11	47	11	104
12	31	12	72
13	31	13	82
14	27	14	63
15	33	15	49
16	16	16	36
17	6	17	29
18	12	18	15
19	5	19	2
20	1		

ההבדלים המשמעותיים בין הקורסים שבהם הבחנו :

בקרוס המתמטיקה המקבצים יותר קטנים. (המקבץ הכי גדול הוא 56, לעומת 104,97,82 ועוד... בפורטוגזית).

המקבצים הגדולים בקורס הפורטוגזית נמצאים במרכז 10-15 , ובגלל שהגדלנו את הטווח לחיזוי יהיה קל למסווגים לחזות טווח קרוב למרכז. לעומת המתמטיקה שהמקבצים מפוזרים.

שאלה 3: על פי נתוני הסטודנט כגון (צריכת שתייה זמן פנוי וכו') האם ההורים גרים ביחד או לא?

	מתמטיקה	פורטוגזית
ADABOOST	0.8719191919191915	0.8717538461538463
SVM	0.8944444444444442	0.877723076923077
KNN	0.8934343434343429	0.8681846153846154
NN	0.8692929292929289	0.8742769230769231

ניתוח תוצאות שאלה 3:

הופתענו לראות דמיון רב בין כל התוצאות . המסקנה הראשונית שלנו הייתה שיש קורלציה גובהה בין הפרמטרים שלנו למצב ההורים.

לאחר מכן ניסנו לשנות את הפרמטרים לפרמטרים אחרים וקיבלנו תוצאות זהות.

מכך הסקנו שהמסקנה הראשונה שלנו שגוייה.

ההשערה שנייה שלנו הייתה, שהאלגוריתם רואה אחוזים גבוהים בקרב הורים שגרים ביחד, ולכן כמעט תמיד יחזה את האפשרות הזאת.

כדי לחזק את הטענה

1) בדקנו את אחוזי ההורים הגרים ביחד

בקרב הקורס בפורטוגזית 569 . מתוך 650 יוצא 0.8753846

(קרוב מאוד לדיוק כל השיטות.)

שאלה 4: על פי נתוני הסטודנט כגון צריכת שתייה , זמן פנוי ,מצב ההורים , איזו השכלה ההורים קיבלו?

	פורטוגזית	מתמטיקה
ADABOOST	Mother education:0.34455 Father education:0.29455	Mother education:0.3773 Father education:0.312828
SVM	Mother education:0.39821 Father education:0.32821	Mother education:0.36717 Father education:0.279999
KNN	Mother education:0.38633 Father education:0.30633	Mother education:0.39505 Father education:0.29010
NN	Mother education:0.31633 Father education:0.28633	Mother education:0.32313 Father education:0.261010

ניתוח תוצאות שאלה 4:

ניתן לראות שכל שיטות החיזוי לא היו מדויקות. לפי דעתנו בגלל ש:

1) יש חמש תוצאות שונות, מה שמקשה מאוד על החזוי.
2)בפרמטרים שנתנו אין תלות למצב חינוך ההורים.

ניסנו לשנות את הפרמטרים כדי לראות אם נוכל לקבל חיזוי טוב יותר למצב חינוך ההורים.
אבל לא הצלחנו להגיע לתוצאה מרשימה.