למידה חישובית
ספטמבר 2021

אלעד דוד (206760274)
ענבר שמייה (208774026)

עבודת בית מסכמת - למידה חישובית/ אלעד דוד וענבר שמייה

מבוא

בעבודה זו מתבצעת השוואה בין שני אלגוריתמים מסוג רגרסיה, הן ברמות-הדיוק (באמצעות השוואה בין ה-y המוערך לבין ה-y האמיתי) והן בזמן-הריצה; תוך השתדלות על שימוש והתנסות במספר שדרוגים לאלגוריתמים. טרם הרצות אלה, נמצאו התכונות החשובות ביותר ע"י כמה אלגוריתמים שונים מתוך סך התכונות הנבדקות. לבסוף חושב האם ישנו over-fitting באלגוריתמי הרגרסיה שנבנו ע"ס רשומות ה-train.

הנתונים שנבחרו ללימוד ולבחינה הינם 29 ערכים מספריים (כולל הציון הסופי בלימודים) אודות 396 סטודנטים וסטודנטיות מאוניברסיטת מינהו, פורטוגל. מראש נבחרו נתונים מסוג למידה מכוונת Supervised-Data, משמע לצד פירוט ערכים טכניים על הסטודנטים ישנם גם ציונם הסופי בלימודים.
בעבודה זו הבאנו למבחן את **הקורלציה בין נתונים יבשים על הסטודנט/ית לבין ציונו/ה הסופי בלימודים**.

רקע תאורטי

אודות הנתונים

הפרמטרים שנבחנו עבור כל סטודנט מפורטים ב*טבלה 1*. מקור: kaggle: Student-grade-prediction

טבלה 1: מידע מסוכם על כלל התכונות שנבחנו עבור כל סטודנט מקובץ הנתונים המקורי (student-mat.csv).

שם התכונה	ערך מינימלי	ערך מקסימלי	הערות
המוסד האקדמי	0	1	0- גבריאל פריירה; 1- מוזיניו דה סילביירה
מין (זכר/ נקבה)	0	1	0- נקבה; 1- זכר
גיל (בשנים)	15	22	גיל הסטודנט/ית
מגורים עירוניים/ כפריים	0	1	0- כפרי; 1- עירוני
גודל המשפחה (מעל 3 נפשות)	0	1	0- מתחת ל-3 נפשות; 1- מעל
האם ההורים גרים יחד	0	1	0- לא; 1- כן
השכלת האם	0	4	0- ללא השכלה; 1- עד כיתה ד'; 2- כיתה ה'-ט'; 3- השכלה תיכונית; 4- השכלה גבוהה
השכלת האב	0	4	
זמן הגעה מהבית למוסד (בדקות)	1	4	1- פחות מ-15 דק'; 2- 15-30 דק'; 3- 30-60 דק'; 4- מעל לשעה
זמן למידה שבועי	1	4	1- פחות משעתיים; 2- 2-5 שעות; 3- 5-10 שעות; 4- מעל ל-10 שעות
מס' הקורסים בהם נכשל	1	4	1-3 – בהתאמה; 4- אחר
האם המוסד מעניק תמיכה לימודית	0	1	0 – לא; 1 – כן
האם המשפחה מעניקה תמיכה לימודית	0	1	
שיעורי-עזר	0	1	
פעילויות מחוץ לתכנית הלימודים	0	1	
בוגר/ת בי"ס לאחיות	0	1	
מעוניין/ת בלימודים גבוהים יותר	0	1	
גישה לאינטרנט בבית	0	1	
המצאות במערכת-יחסים רומנטית	0	1	
יחסים עם המשפחה	1	5	1 – גרוע; 5 – מצוין
זמן חופשי אחר שעות הלימודים	1	5	1 – נמוך; 5 – גבוה
תדירות יציאות עם חברים	1	5	
צריכת אלכוהול במהלך אמצ"ש	1	5	
צריכת אלכוהול במהלך סופשבוע	1	5	
מצב בריאותי	1	5	1 – גרוע; 5 – מצוין
חיסורים	0	93	מס' החיסורים משיעורים
ציון בסמסטר הראשון	0	20	ישנה קורלציה גבוהה בין השקלול של שתי תכונות אלה לבין הציון הסופי בקורס
ציון בסמסטר השני	0	20	
ציון סופי	**0**	**20**	**ציון סופי בקורס**

אלעד דוד (206760274)
ענבר שמייה (208774026)

כלים ושיטות עבודה

נרמול נתונים

נרמול הנתונים מתבצע באמצעות מעבר על כל תכונה בטבלה ומעבר לסקאלות דומות של ערכים. כפי שניתן לראות בטבלה 1: מידע מסוכם על כלל התכונות שנבחנו עבור כל סטודנט מקובץ הנתונים המקורי (student-mat.csv).*טבלה 1*, ערכי התכונות נעים בקשתות מספרים שאינן בהכרח חופפות; מצב כזה עלול להוביל לחוסר-פרופורציות בעת הרצת אלגוריתמים שונים. נרמול הנתונים מתבצע עבור כל עמודה בנפרד, באמצעות חישוב ההפרש בין ערך נתון לערך המינימום של התכונה, חלקי ההפרש בין ערך המינימום למקסימום של כל תכונה.

מציאת תכונות חשובות

נרצה למצוא תת-קבוצה של התכונות הנבדקות שייצגו את התכונות המשמעותיות ביותר מביניהן. שקלול של התכונות החשובות יניב קורלציה גבוהה בין ערכיהן לבין ערך ה-y המתקבל. קיום יתר תכונות לא רלוונטיות בנתונים עשוי להקטין את הדיוק של המודלים. שלושה יתרונות של ביצוע בחירת תכונות לפני עיבוד הנתונים:

1. מפחית overfitting: פחות נתונים מיותרים משמע פחות הזדמנות לקבל החלטות על סמך רעש.
2. משפר את הדיוק: נתונים פחות מטעים פירושם דיוק המודל משתפר.
3. מקצר את זמן האימון (train): פחות נתונים פירושם שאלגוריתמים לומדים מהר יותר.

ישנן מס' דרכים למצוא את התכונות החשובות, בעבודה זו בוצע שימוש בשלוש מהן (השוואת אלגוריתמים):

אלגוריתם ראשון הוא אלגוריתם חמדן, אשר בוחר תחילה את התכונה החשובה ביותר המהווה קורלציה גבוהה ביותר בין ערכיה לבין ערכי ה-y. בכל חזרור ועד הגעת ההרצה למס' התכונות הרצוי, האלגוריתם ישפר את הדיוק באמצעות הוספת התכונה שבאמצעות הוספתה- דיוק האלגוריתם משתפר באופן המשמעותי ביותר.
אלגוריתם שני דוגל בשיטת האלימינציה Recursive Feature Elimination (RFE), בכל חזרור רץ ברקורסיה על כלל התכונות ומסיר את התכונה שמוסיפה הכי מעט אינפורמציה ודיוק (ואף עשויה לגרוע מן הדיוק).
אלגוריתם שלישי מוצא את וקטור תטא (Θ) עבור האלגוריתם רגרסיה לינארית, בוחר את k התכונות שעבורן התטאות שהתקבלו בחישוב בערך מוחלט הן הגבוהות ביותר (רחוקות מ-0). ככל שהמקדם במשוואה הלינארית גבוה יותר, כך ישנה חשיבות גבוהה יותר ומשקל גבוה יותר לתכונה הנכפלת בו.

אלגוריתם K-NN

האלגוריתם מקבל רשומה (בתור test) ומספר שלם k, ומוצא את k הרשומות בקבוצת ה-train בעלות הערכים הדומים ביותר לרשומה הנבדקת; דמיון זה נבדק באמצעות ההפרש בין הרשומה הנבדקת לבין הרשומות בריבוע (לביטול הערכים השליליים). האלגוריתם קובע כי ערך ה-y של הרשומה הנבדקת שווה בקירוב לממוצע של ערכי ה-y של k הרשומות שנמצאו. ניתן לחשב באמצעות שיטת המרפק מהו המספר האידיאלי של שכנים.

Linear Regression

אלגוריתם זה מקבל את כל רשומות-האימון (train) ויוצר משוואה לינארית. משוואה זו יוצרת מודל המסביר את הקשר בין התכונות הנבדקות לבין ה-y המתקבל. לאחר מציאת משוואה לינארית שכזו ניתן להזין את תכונות הרשומה הנבדקת (test), להכפילן בווקטור תטא ולקבל חזרה את ה-y המוערך בעת הצבת הפרמטרים הנבדקים במשוואה.

מתודולוגיה

טרם הרצת אלגוריתמים מגוונים, נרמלנו את הנתונים כך שממוצע כל תכונה שווה ל-0 וסטיית התקן של כל תכונה שווה ל-1 (בקירוב, שהרי ייתכנו שגיאות כתוצאה מעיגול וקיצוץ של המחשב).

לקראת הרצת האלגוריתמים נחלקו הרשומות ל-2 קבוצות: קבוצת האימון (train) וקבוצת המבחן (test). גודל קבוצת האימון הוא כ-⅔ מסך הרשומות (נתוני הסטודנטים), וקבוצת המבחן גודלה כ-⅓ הרשומות הנותרות. החלוקה ל-2 תתי-קבוצות אלו הינה רנדומלית.

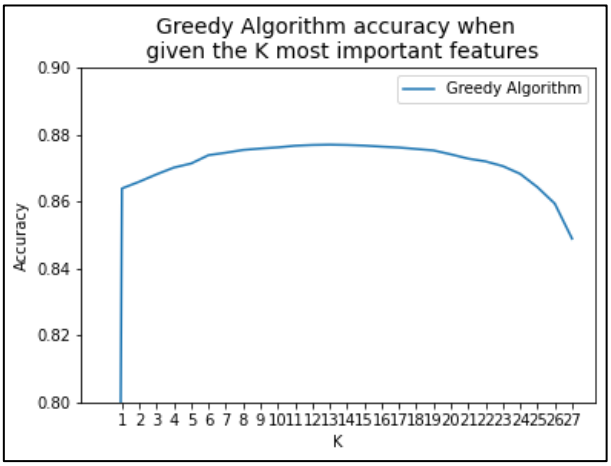
אז הורצו שלושה אלגוריתמים שונים במטרה לבחור את מספר התכונות החשובות ביותר. חיוני למצוא את המספר האידיאלי עבור התכונות החשובות ביותר, כאשר מספר זה מייצג את הרף הגבוה ביותר שמשפר באופן ניכר את המודל המחשב את הערך החזוי. לשם כך הורצו כל האלגוריתמים, המפורטים בתת-הפרק *מציאת תכונות חשובות*, על קבוצת האימון ונבדקה רמת השיפור מהוספת תכונה נוספת לסט התכונות שכבר נבחר על-ידי כל אלגוריתם. החסם הנבחר לשיפור רמת הדיוק של המודל באמצעות הוספה של תכונה נוספת

אלעד דוד (206760274)
ענבר שמייה (208774026)

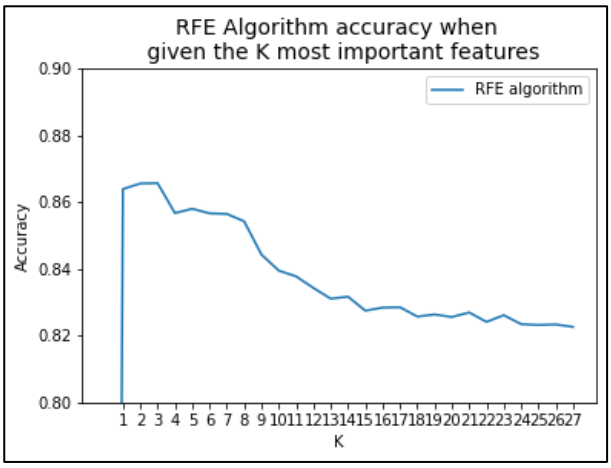
הוא 10^{-3}, כלומר- אם הוספת התכונה הבאה לא משפרת דיה את המודל, לא נוסיף אותה והלאה.
החמדן בחר <u>שש תכונות</u>, אלג' ה-RFE ואלג' התטאות הגבוהות בחרו <u>שתי תכונות</u> עד ההגעה לסף.

לצורך השוואה בין האלגוריתמים השונים בהיבט של בחירת מספר התכונות האידיאלי בכל אלגוריתם, להלן הגרפים השונים שהתקבלו מכל אלגוריתם:

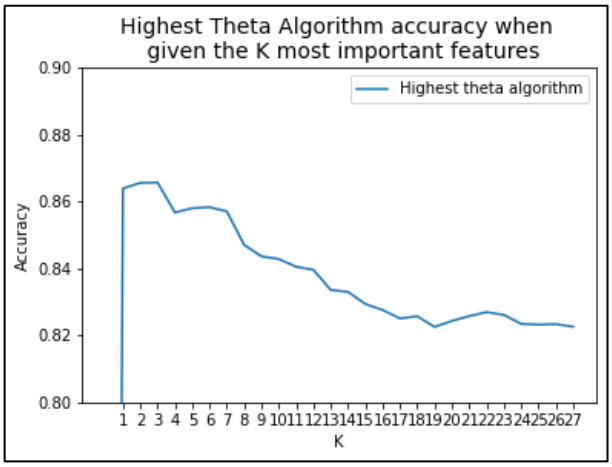
<u>גרף 3:</u> רמת הדיוק כתלות במס' התכונות החשובות ביותר בהרצת האלגוריתם החמדן לבחירת k תכונות חשובות ביותר.



<u>גרף 2:</u> רמת הדיוק כתלות במס' התכונות החשובות ביותר בהרצת האלגוריתם ה-RFE לבחירת k תכונות חשובות ביותר.



<u>גרף 1:</u> רמת הדיוק כתלות במס' התכונות החשובות ביותר בהרצת האלגוריתם למציאת התטאות הגבוהות ביותר לבחירת k תכונות חשובות ביותר.



אלעד דוד (206760274)
ענבר שמייה (208774026)

מגרפים אלו אנו למדים כי שלושת האלגוריתמים המפורטים תחת *מציאת תכונות חשובות* הניבו תוצאות שונות על הנתונים שהוזנו. העלייה בגרפים 1-3 נבחנת ומחפשים אחר ערך ה-k הקטן ביותר אשר התכונה הבאה (k+1) שמוספת לאימון המודל כבר לא מהווה שיפור משמעותי (ערך ה- y החזוי קרוב מדי לערך ה-y האמיתי); עבור *גרף 3* ערך k זה שווה 6 (אחריו ישנה ירידה חדה בגרף), עבור *גרף 2* ערכו הוא 2 ועבור *גרף 1* ערך זה הוא k=2. כאמור, האלגוריתמים הניבו מספרים שונים המייצגים את המספר האידיאלי של התכונות החשובות ביותר עבור נתוני האימון, אך גם ערכן שונה מאלגוריתם אחד למשנהו.

האלגוריתמים המדוברים, החזירו את התכונות המפורטות ב*טבלה 2* בתור התכונות החשובות ביותר (כאשר התכונות מסודרות לפי סדר הוספתן לסט התכונות החשובות ביותר מימין לשמאל, מלמעלה למטה):

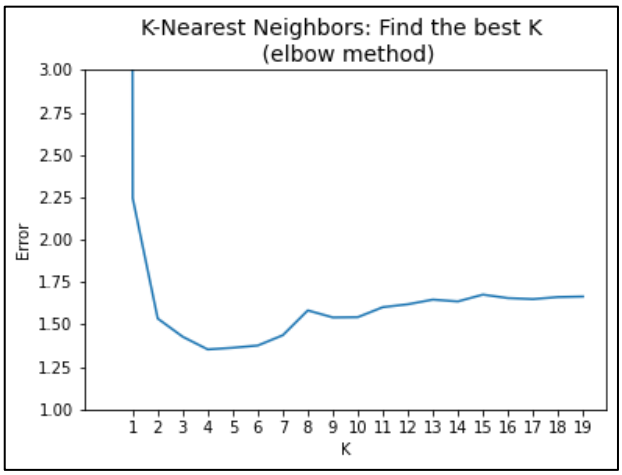
טבלה 2: מידע מסוכם על כלל התכונות שנבחנו ונמצאו כחשובות ביותר ע"י כל אחד מהאלגוריתם שהורצו.

Greedy Algorithm	ציון בסמסטר השני	יחסים עם המשפחה	חיסורים
	פעילויות מחוץ לתכנית הלימודים	האם המוסד מעניק תמיכה לימודית	ציון בסמסטר הראשון
RFE Algorithm	חיסורים	ציון בסמסטר השני	
Highest Theta	ציון בסמסטר השני	חיסורים	

להמשך השוואת אלגוריתמים נוספים, עודכנו התכונות במודלים שנבנו מנקודה זו לכדי התכונות שנבחרו על-ידי האלגוריתמים RFE והתאמות הגבוהות ביותר, שכן לפי *גרף 2* ו*גרף 1* התכונות החשובות ביותר מניבות את רמת הדיוק הגבוהה ביותר. עדכון התכונות הנבדקות לקבוצת התכונות החשובות ביותר רלוונטי כפי שתואר בפרק *מציאת תכונות חשובות*, שם מפורטים היתרונות של דרך עבודה זו. המשך עבודה זו מכאן והלאה צומצמה לשתי התכונות המפורטות בשורות השנייה והשלישית ב*טבלה 2*, עבור האלגוריתם הדוגל בשיטת האלימינציה ברקורסיה ואלגוריתם התאמות הגבוהות.

נשווה כעת בין *אלגוריתם K-NN* לבין *Linear Regression* עבור מודל מצומצם ומשופר המבוסס על הנתונים שהושגו מחישוב התכונות החשובות ביותר. הרצת אלגוריתם K-NN דורש חישוב ממוצע של ערכי ה-y הידועים לצורך הצבה בערך ה-y החזוי. לצורך השגת המספר האידיאלי של שכנים עבור הרצת האלגוריתם לעיל, הורץ האלגוריתם על מספר שכנים ההולך וגדל; זאת לצורך מציאת המספר האידיאלי של מספר הנקודות הקרובות ביותר למציאת ערך ה-y של הרשומה המבוקשת. להלן גרף המתאר את השגיאה ביחס למספר השכנים הקרובים ביותר לרשומה הנבדקת:

גרף 4: מציאת המספר האידיאלי עבור מספר השכנים לצורך הרצת אלגוריתם K-NN.



לאחר בחינת *גרף 4* בהיבטים של "שיטת המרפק" (elbow method), נשים לב כי ישנה ירידה חדה עד 4 שכנים (אימון מודל ה-K-NN עבור 4 הרשומות בעלות הערכים הקרובים ביותר) ומנקודה זו השגיאה עולה וצומחת באיטיות עבור K גבוה יותר; לכן, נבחר להריץ את אלגוריתם זה על 4 נקודות שכנות, לבנות מודל ראוי לצורך השוואה מאוחרת יותר.

אלעד דוד (206760274)
ענבר שמייה (208774026)

לפי גרף זה, בחירת 4 רשומות שכנות לצורך חישוב ממוצע עבור ערך ה-y החזוי מניבה את התוצאה המדויקת ביותר (הקרובה ביותר לערך ה-y האמיתי).

לצורך השוואת אלגוריתם K-NN לאלגוריתם נוסף, נבחר כאמור אלגוריתם *Linear Regression*. אלגוריתם זה הניב את וקטור המקדמים הבאים (או בשמו המוכר יותר- וקטור התטא):

טבלה 3: ערכי המקדמים שמודל ה-Linear Regression הניב עבור משוואת הקו הישר לתכונות החשובות ביותר.

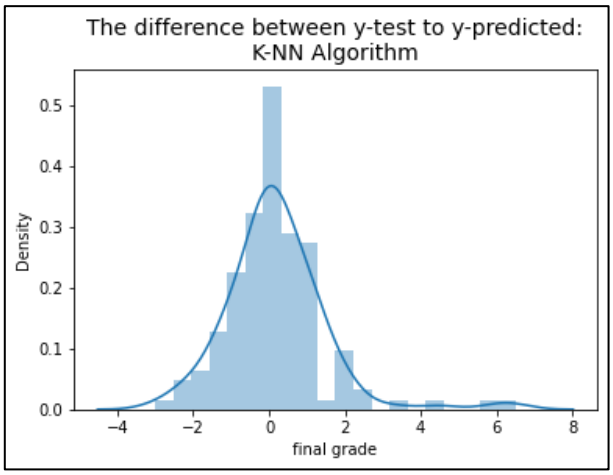
ציון בסמסטר השני	חיסורים
20.76222569	2.95777388

טבלה זו אולי נראית מצומצמת ומינימלית, אך בחירת שתי התכונות לעיל עוזרת להימנע מבעיות בזמן הריצה באמצעות צמצום מספר התכונות בדגם, דבר הנובע בשל ניסיון לייעל את ביצועי המודל. בחירת התכונות מספקת גם יתרון נוסף- פרשנות מודל; עם פחות תכונות, מודל הפלט הופך להיות פשוט וקל יותר לפרשנות.

דיון

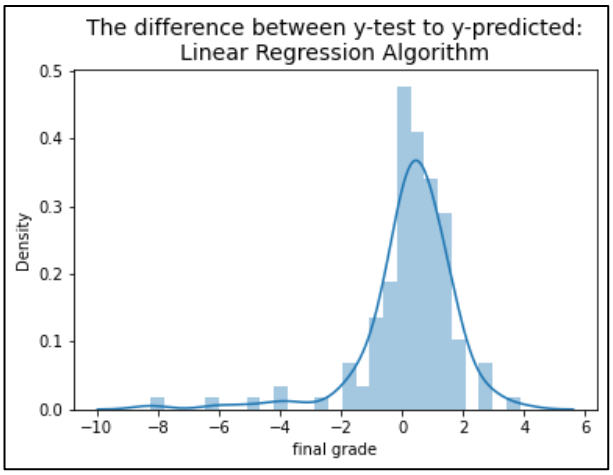
הרצת אלגוריתם K-NN עבור 4 הרשומות בעלות הערכים הדומים ביותר לרשימה הנבדקת הניבה וקטור של ערכי y חזויים. השגיאה המוערכת בשלב זה הינה ההפרש בין ערכי ה-y החזויים לערכי ה-y האמיתיים; *גרף 5* מתאר את גובה שגיאות אלה עבור כל סקאלת הציונים הסופיים:

גרף 5: הרצת אלגוריתם K-NN עם k=4. ייצוג ההפרש בין ערך ה-y האמיתי לבין ערך ה-y החזוי מופיע בצורת ברים, ונתונה הערכה מקורבת להתפלגות נורמלית בצורת קו.



הרצת אלגוריתם הרגרסיה הלינארית הניב גם הוא גרף דומה, כמתואר בגרף *6*:

גרף 6: הרצת אלגוריתם Linear Regression. ייצוג ההפרש בין ערך ה-y האמיתי לבין ערך ה-y החזוי מופיע בצורת ברים, ונתונה הערכה מקורבת להתפלגות נורמלית בצורת קו.



אלעד דוד (206760274)
ענבר שמייה (208774026)

בהשוואה בין השגיאות שהניבו שני האלגוריתמים: *אלגוריתם K-NN* ו*Linear Regression*, ניתן לראות כי סך השגיאות באלגוריתם ה-K-NN נמוך מסך השגיאות של אלגוריתם הרגרסיה הלינארית. צפייה ב*גרף 6* מלמדת כי ישנה שגיאה בגובה של 8-, והיא גבוה (בערך מוחלט) מהשגיאה המקסימלית ב*גרף 5* (שם השגיאה המקסימלית היא 6).

בהשוואה נוספת בין האלגוריתמים הנ"ל, נתבונן בערכים הבאים המפורטים ב*טבלה 4*:

טבלה 4: סיכום זמני ריצה וערכי שגיאות שונים.

	רמת הדיוק של רשומות האימון	רמת הדיוק של רשומות המבחן	ממוצע השגיאות בריבוע	זמן ריצת (בשניות)
K-NN Algorithm	0.86411	0.89812	1.83	0.580140
Linear Regression	0.80212	0.86563	2.42	0.286671

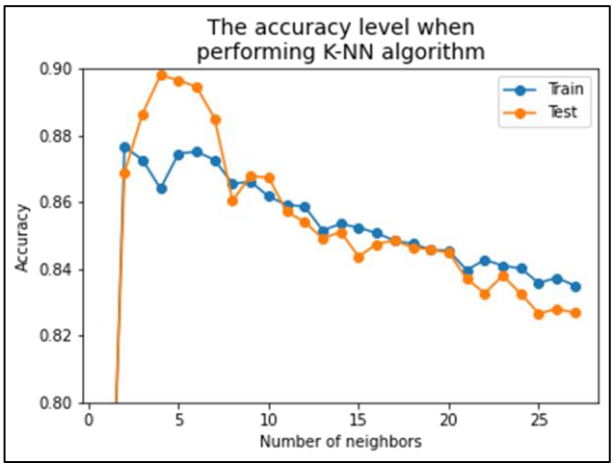
ניתן לשים-לב כי לפי *טבלה 4* רמת הדיוק בכל אחד מהפרמטרים הנבדקים הייתה גבוהה יותר באלגוריתם K-NN, לעומת אלגוריתם הרגרסיה הלינארית. עם זאת, זמן הריצה של אלגוריתם ה-K-NN היה גבוה משל אלגוריתם הרגרסיה הלינארית.

בעיית overfitting

אחת מהבעיות הגדולות ביותר במכונת למידה היא overfitting (התאמת יתר של הנתונים מקבוצת האימון לכדי מודל שיבחן את הנתונים מקבוצת המבחן). **המקרה האופטימלי עבור רמת ההתאמה, הוא כאשר השגיאה נמוכה <u>וגם</u> החיזוי של קבוצת המבחן קרוב מאוד לחיזוי של קבוצת האימון**. לפי *טבלה 4*, ניתן לראות כי בשני האלגוריתמים, רמות הדיוק בין נתוני האימון לבין נתוני המבחן קרובות יחסית ורמות הדיוק גבוהות יחסית. עם זאת, המודל של אלגוריתם ה-K-NN הוא בעל רמת overfitting נמוכה יותר משל אלגוריתם הרגרסיה הלינארית (ההפרש בין רמות הדיוק נמוך יותר).

בגרף הבא ניתן לראות כי על-אף בניית המודל באמצעות 2 תכונות בלבד, וישנו סיכון כי המודל יהיה פשוט מדי ונחווה underfitting, רמת הדיוק עודנה גבוהה יחסית. בנוסף, בגרף זה נראה כי המודל ברמת הדיוק הגבוהה ביותר עבור k=4 עבור קבוצת המבחן, כפי שנחזה בפרק *מתודולוגיה*:

גרף 7: רמת הדיוק עבור קבוצת האימון וקבוצת המבחן בהינתן k שכנים בבניית מודל K-NN.



בגרף 7 ניתן לצפות ברמות דיוק גבוהות של שני הגרפים (בין 90% ל-82% עבור קבוצת המבחן עבור כל מספרי ה-k הנבדקים), כמו כן ניתן לצפות בסמיכות של שני הגרפים, משמע אין overfitting למודל שנבנה.

רמת ההתאמה עבור אלגוריתם הרגרסיה הלינארית גבוהה גם כן (הפרש של 6% בין רמת הדיוק של קבוצת האימון לבין רמת הדיוק של קבוצת המבחן).

סיכום

עבור האלגוריתמים למציאת התכונות החשובות ביותר, ניתן לראות כי אחרי הכל ישנה חפיפה מסוימת בין התכונות שהתקבלו כחשובות ביותר. ניכר כי על אף התלות בין התכונות "ציון בסמסטר הראשון" ו-"ציון

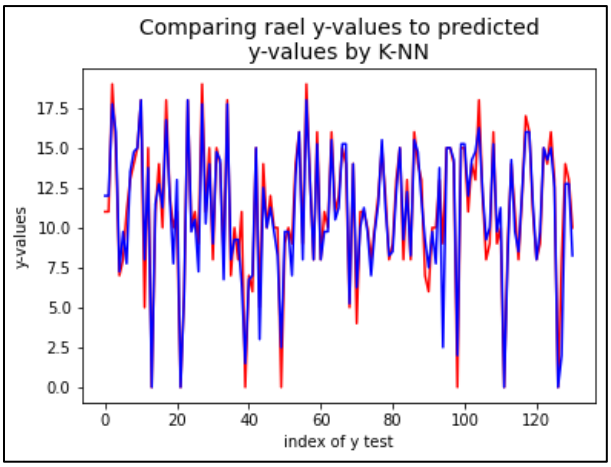
אלעד דוד (206760274)
ענבר שמייה (208774026)

בסמסטר השני" ב*טבלה 1* על ערך ה-y האמיתי, נבחרה רק תכונה אחת מתוכן (השערתנו היא שלא היה צורך באחרת אחרי שהתכונה הראשונה מביניהן הוספה).

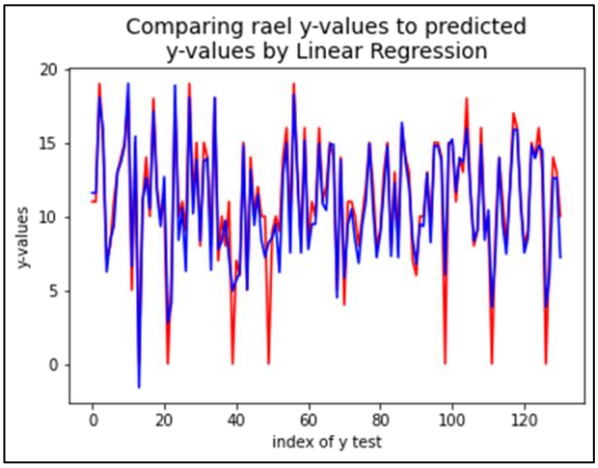
עבור הנתונים הנבחרים, אלגוריתם ה-K-NN רץ לאט יותר לעומת אלגוריתם הרגרסיה הלינארית, אך נדמה כי הוא גם "יסודי" יותר לעומתו. אלגוריתם ה-K-NN צמצם את עבודתו לכדי 4 הרשומות בעלות הערכים הקרובים ביותר לערכי הרשומה הנבדקת (לפי *גרף 4*).

לשני האלגוריתמים תופעת ה-overfitting נמצאת ברמה נמוכה מאוד, ובפרויקט זה דווקא חשודה תופעת ה-underfitting בשל התאמת 2 תכונות חשובות בלבד לבניית מודל. בהיבט של רמות ההתאמה (overfitting ו-underfitting), ייתכן כי קבוצת אימון גדולה יותר או לחלופין בחירה מרובה יותר של תכונות היו מדייקות עבורנו את האלגוריתמים ומציגות הפרש קטן אף יותר עם רמות דיוק גבוהות יותר.

גרף 8: השוואת ערכי ה-y החזויים באמצעות מודל הנבנה ע"י אלגוריתם ה-K-NN לערכי ה-y האמיתיים של קבוצת המבחן.



גרף 9: השוואת ערכי ה-y החזויים באמצעות מודל הנבנה ע"י אלגוריתם רגרסיה לינארית לערכי ה-y האמיתיים של קבוצת המבחן.



ניתן לראות בגרף ובגרף כי ישנה חפיפה גבוהה בין שנחזה לבין הערכים המקוריים, וזאת על-אף הבחירה ב-2 תכונות בלבד לצורך הסתמכות בניית האלגוריתם.

ביבליוגרפיה

https://machinelearningmastery.com/rfe-feature-selection-in-python/

https://machinelearningmastery.com/feature-selection-in-python-with-scikit-learn/

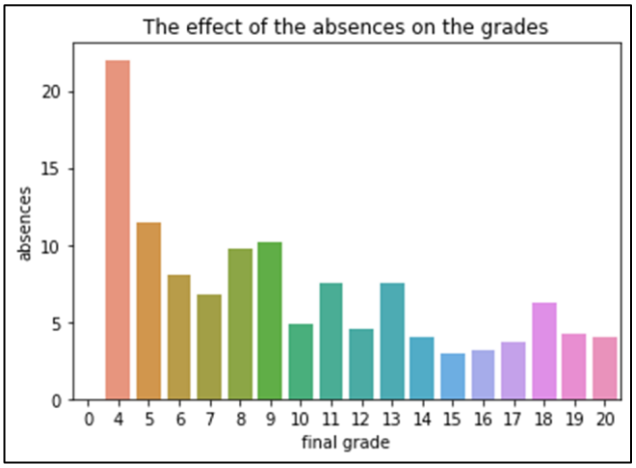
https://www.uminho.pt/EN

אלעד דוד (206760274)
ענבר שמייה (208774026)

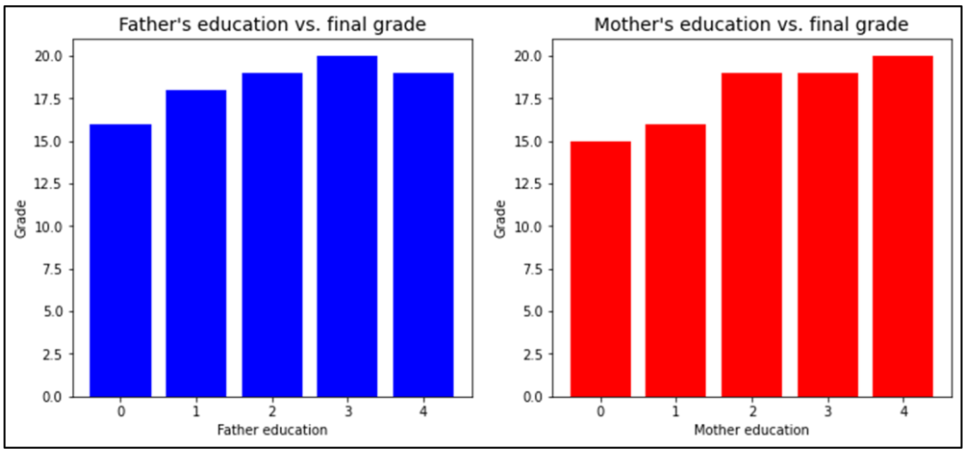
נספחים

בפרק זה תוכלו למצוא גרפים מעניינים על סטטיסטיקות שונות שחשבנו שעשויות להשפיע על הציון הסופי של סטודנטים בלימודים. על גרפים אלה לא נכתב הסבר נוסף מלבד כותרת הגרף ותוויות הצירים.

גרף 10: הציון הסופי בלימודים ביחס לכמות החיסורים של הסטודנט/ית.



גרף 11: הציון הסופי בלימודים ביחס לרמת השכלת ההורים (כל הורה בנפרד).



גרף 8: הציון הסופי בלימודים ביחס לצריכת השתייה האלכוהולית של הסטודנט/ית במהלך אמצע השבוע או במהלך סוף השבוע (בנפרד).