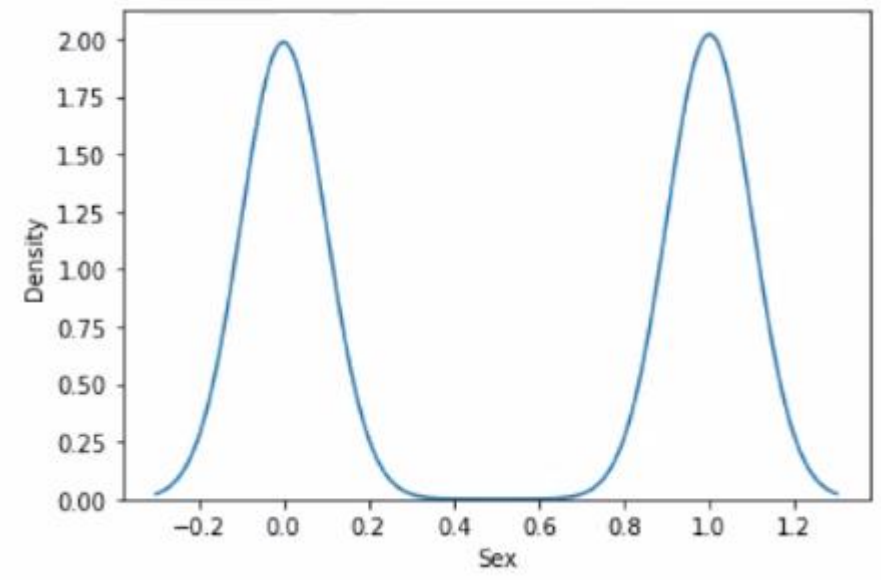
Intro To Machine Learning – Final Report HW 3

Shachar Katz 313574766

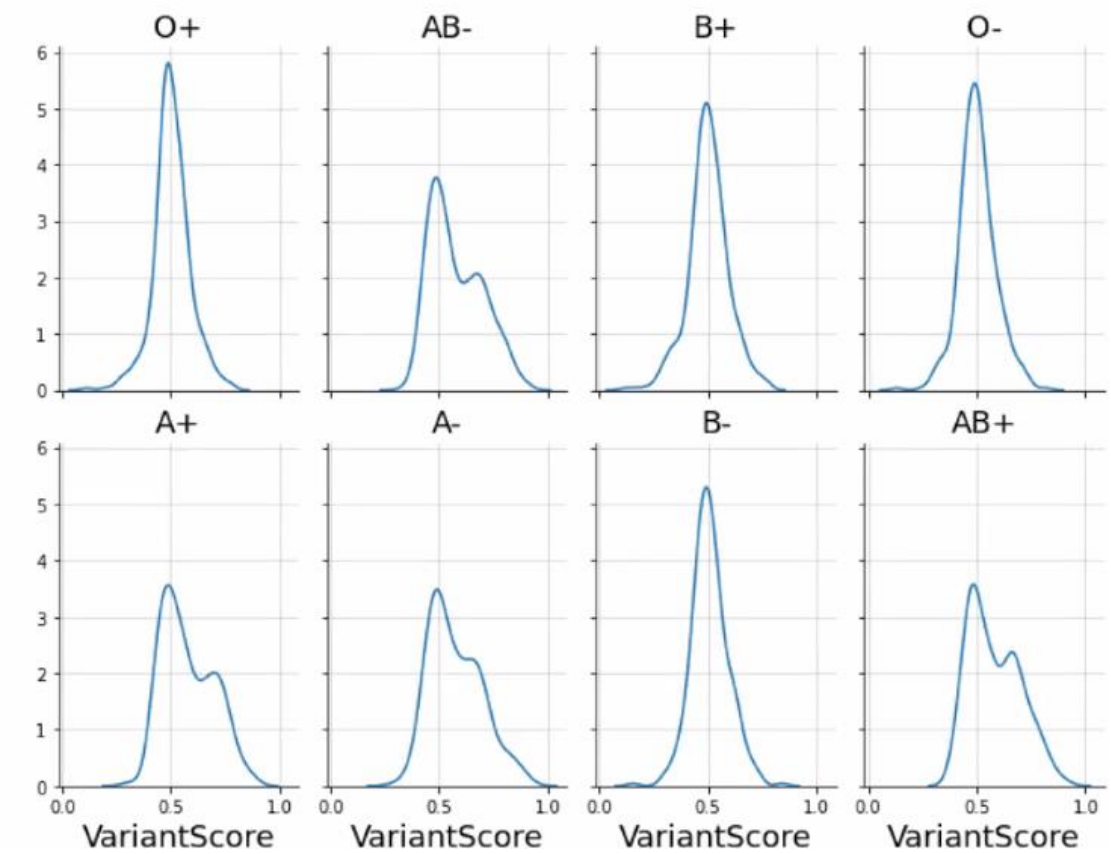
Sivan Yosebashvili 318981586

2) להלן הפלט :



כפי שמשתמע מהפלט, מספר הבנים והבנות זהה, וההתפלגות שלהם זהה.

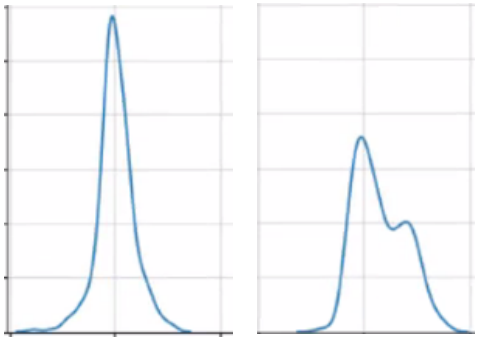
3) להלן הפלט :



ניתן לראות שקיימים שני סוגים של פלטים :
האחד דומה להתפלגות נורמלית, והשני מזכיר מאין הר עם שתי גבעות. סוגי הדם מתפלגים לפי אחת מתבניות אלו.

4) תכונת sex מכילה 2 משתנים קטגוריאליים, שמתפלגים באופן זהה, ולכן נסווג מין זכר כ-1 ומין נקבה כ-0 (באופן שרירותי, אין משמעות לערך בגלל שמדובר בשני משתנים בלבד ולכן ניתן להמיר למשתנה בינארי נומרי).

5)בעקבות תהליך הכנת הדאטה, גילינו את תופעת החלוקה ל-2 קבוצות של תכונת סוג הדם. לכן, אנו נמיר סוגי דם שמתפלגים לפי תבנית א' (מימין) ל-0 ולפי תבנית ב' (משמאל) ל-1.



נשים לב כי ישנה גישה נוספת, שמשתמשת ב .hot vector-בשיטה זו, נהפוך כל סוג דם למחרוזת בינארית באורך 8, שבה יש שבעה אפסים ואחד יחיד באינדקס המציין את סוג דם זה. גישה זו למעשה הופכת את פיצ'ר סוג דם לשמונה פיצ'רים. בגישה שאנו בחרנו לנקוט, אנו מחלקים את סוגי הדם ל-2 קבוצות, אם יש משהו שלא מוצג בפלט זה, וסוג הדם משפיע בצורה אחרת שאינה גלויה לנו כרגע על תווית הסיווג, שיטת hot vector עדיפה. אך, חסרון בולט של שיטה זו, היא הגדלת ה-sample complexity. אם היו לנו אינסוף דוגמאות, שיטה זו עדיפה, אך במצב הנתון עם כמות הדוגמאות הסופית, אנו מאמינים שהגישה שלנו תיתן תוצאה טובה יותר (כי לא נשלם את מחיר העלאת סיבוכיות המסווג).

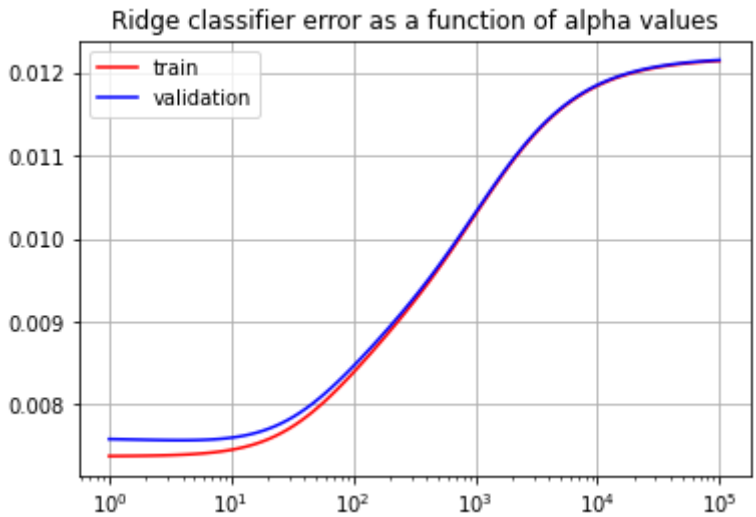
6) תהליך הכנת המידע אצלנו התחלק למספר שלבים:

* בשלב הראשון, המרנו פיצ'רים קטגוריאליים לנומריים (היו רק שניים כאלה- sex, bloodType) על פי השיקולים שהוצגו בשאלות 4+5.
* בשלב השני יצרנו לעצמנו data frame על בסיס סט האימון שלנו שכלל מידע שישמש אותנו להמשך הכנת הדאטה : ממוצע, חציון, ערך מינימום, ערך מקסימום, וערכי קצה של כל תכונה.
נציין כי השתמשנו אך ורק בסט האימון למטרה זו, כפי שלמדנו בתרגילי בית קודמים, כדי למנוע זליגה של מידע מסט המבחן לתהליך האימון (על מנת להבטיח תוצאת למידה מדויקת).
* הורדנו פיצ'רים שהם קורולטיביים אחד לשני (על פי תרגיל בית 1) על ידי הדפסת מטריצת הקורולציות והורדת פיצ'רים שהם קורולטיבים אחד לשני ביותר מ-90%.
* בשלב הבא ביצענו השלמת ערכים חסרים בסט האימון ובסט המבחן על פי ה-data frame שיצרנו בשלב 2, הבחירה בין ממוצע לחציון נעשתה לפי שיקולי התפלגות של הפיצ'רים שחקרנו בתרגיל בית 1.
* לאחר מכן, ביצענו הורדת ערכים קיצוניים (outliers) באמצעות המידע המופיע ב-data frame.
* לבסוף, ביצענו נרמול max – min על מנת שכל הערכים יהיו בין 0 ל-1 ובעלי סקאלת ערכים משותפת.
נציין שוב שכל המידע ששימש אותנו בתהליך זה ונמצא ב-data frame נאסף מסט האימון בלבד, ותהליך הכנת הדאטה בוצע הן על סט האימון והן על סט המבחן.

7) להלן תוצאות ביצועי:regressor dummy

	section	Train MSE	Validation MSE
Dummy	2	0.0121805	0.01219

8) להלן תוצאות הסיווג של Ridge כתלות בפרמטר α על סט האימון ועל סט הולידציה:




```
best score on train is :
0.007543196351274771
best score on validation is :
0.0077402046827748285
best alpha is :
4.534878508128582
```

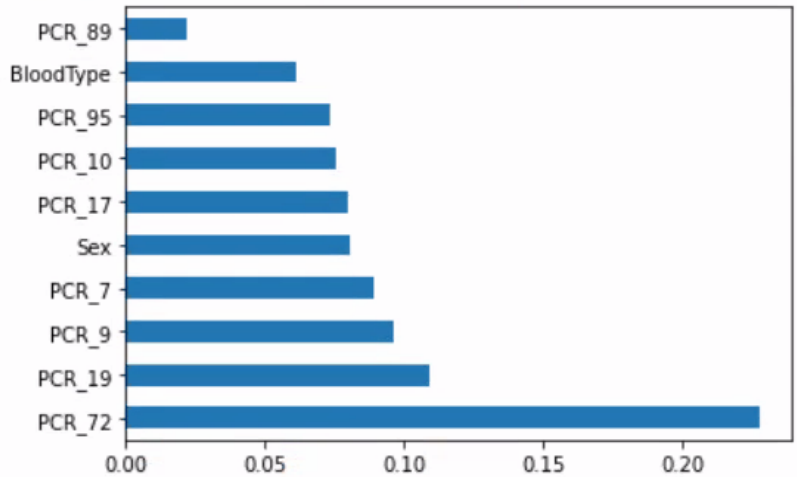
9) להלן תוצאות ביצועי: Basic Linear Regressor

	section	Train MSE	Validation MSE
Dummy	2	0.0121805	0.01219
Basic Linear	3	0.00739167	0.00756576

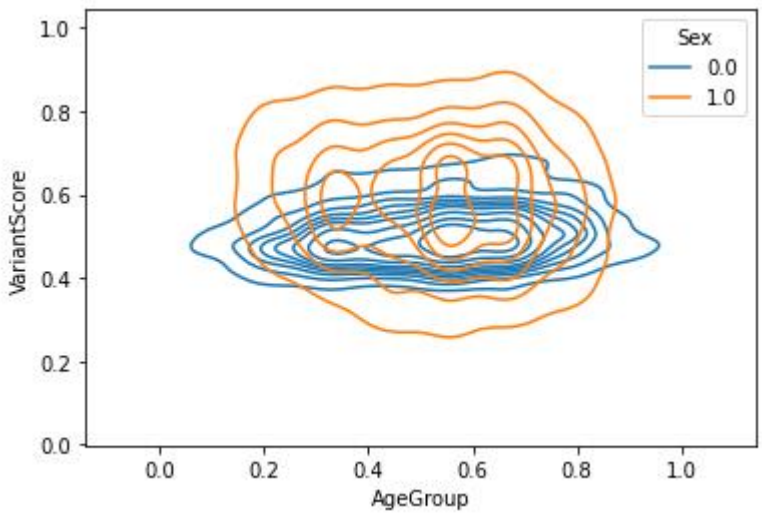
10) חמשת הפיצ'רים בעלי המקדמים הגבוהים ביותר (בערך מוחלט) מהגבוה לנמוך הינם:

$$Pcr_{72}, Pcr_{19}, Pcr_{9} Pcr_{7}, Sex$$

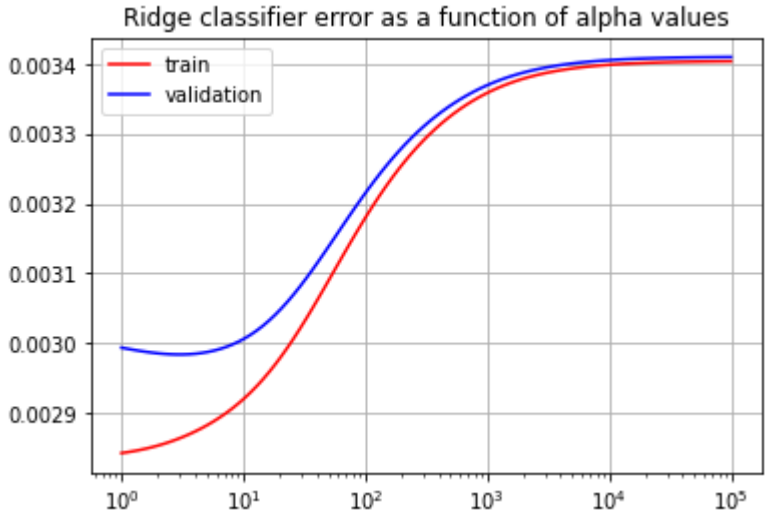
```
4.0370172585965545
w =  [-1.12581615e-02 -2.21191846e-03  6.15830469e-02 -1.77666439e-02
  1.12089201e-02 -2.02157982e-02  5.40348204e-04 -9.17061652e-03
  7.55056716e-02  7.95658711e-02 -1.09494714e-01 -5.97422825e-03
  1.55639609e-04  1.93472179e-03  8.94840756e-02 -2.27657238e-01
  9.82316108e-03 -9.98311708e-03 -4.48129732e-03 -2.23968185e-02
 -9.60426154e-02 -1.22870303e-03  7.33548653e-02  8.06308030e-02
 -1.73605912e-02  1.31288515e-03  1.65686907e-02  3.14387062e-04]
<matplotlib.axes._subplots.AxesSubplot at 0x7f8f84be9cd0>
```



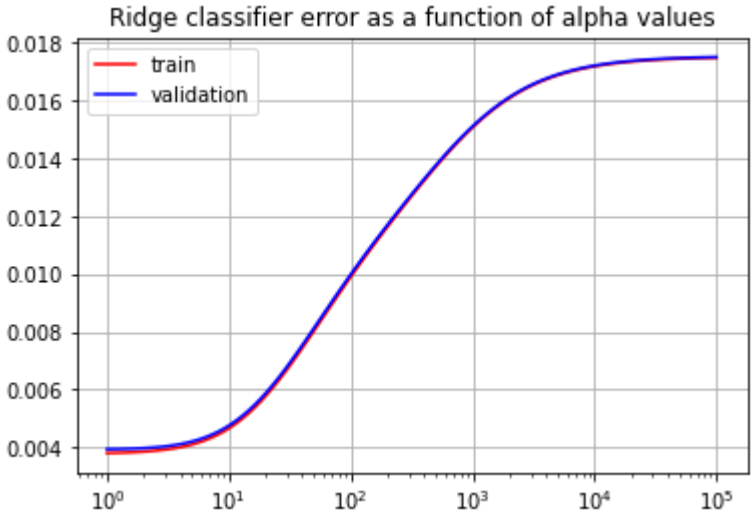
11) להלן הפלט :



12) להלן תוצאות הסיווג של Ridge כאשר נבחן על סט המידה שכולל רק דוגמאות של נשים :



להלן תוצאות הסיווג של Ridge באשר נבחן על סט המידה שכולל רק דוגמאות של גברים :



ופלט המסכם אחוזי שגיאה מינימליים ופרמטר α מיטבי :

	Train MSE	Validation MSE	Best alpha
Females	0.0028616	0.00298339	2.84804
Males	0.00379825	0.00393541	1

13) להלן הטבלה המעודכנת עם ביצועי:Multi Level Regressor

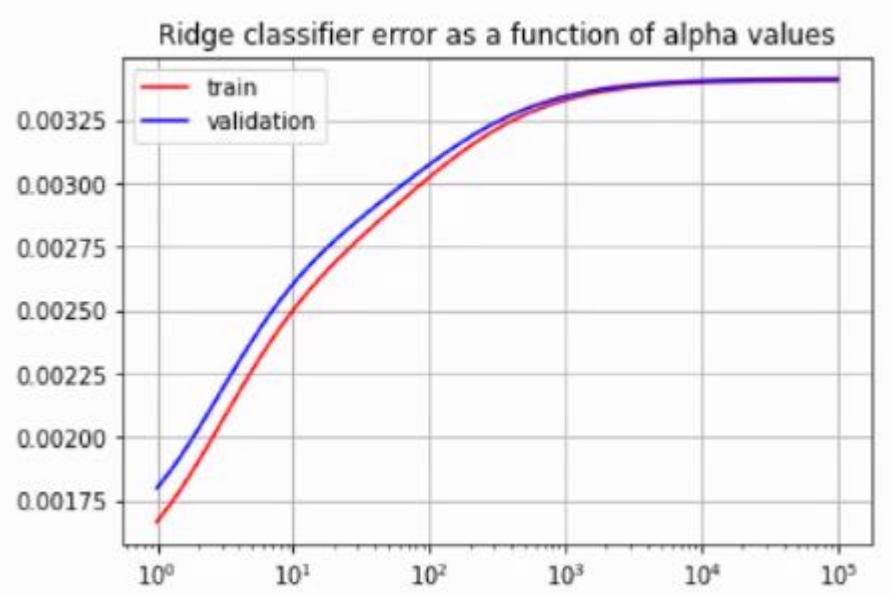
	section	Train MSE	Validation MSE
Dummy	2	0.0121805	0.01219
Basic Linear	3	0.00739167	0.00756576
Multilevel linear	4	0.00333368	0.00346586

14) ניתן להבחין בשיפור בתוצאות השגיאה, דבר שמסביר את תשובתנו לשאלה 11, המודל מצליח להתמודד טוב יותר כאשר לומד לחוד דאטה של גברים ודאטה של נשים. כלומר, קיימים קשרים בין פיצ'רים לבין תווית המטרה (כפי שהצגנו בשאלה 11) שאינם מתנהגים בצורה לינארית, ולכן, כאשר הפרדנו את הטיפול בהם, ואימנו כל אחד מהם עם מסווג לינארי, השגנו תוצאה טובה יותר משמעותית. לכן, אנו מסיקים כי בבירור תכונת sex משפיעה (יחד עם תכונות אחרות) על התפלגות הדאטה במרחב הרב מימדי בצורה לא לינארית.

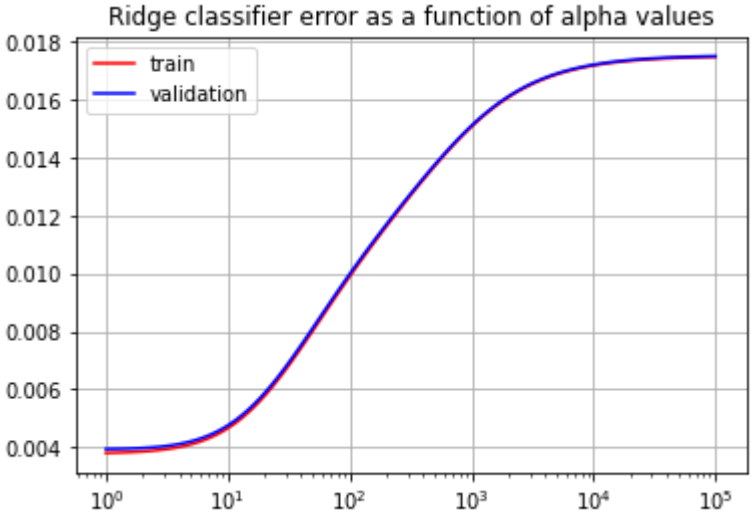
15) לפי מה שלמדנו בתרגול 9, ככל שאנו מנסים להתחקות אחרי פונקציה מדרגה גבוהה יותר, ככה תופעת ה overfitting-עולה, כי אנו יכולים להסביר כל התנהגות של הדאטה שקיבלנו כדוגמאות. כתוצאה מכך, שגיאת האימון תקטן מאוד. העלאת מספר הפיצ'רים שקולה להגדלת מרחב ההיפותזות, דבר שמעלה את ה sample-complexityועלול לגרום ל.overfitting-

16) כתוצאה מההעלאה בחזקה, אנו מצפים ששגיאת הולידציה תשתפר, **אך,** קיימת פונקציה מסוימת $f(x)$, שלה קיים קירוב פולינומיאלי כלשהו מדרגה d. ככל שננסה ללמוד פונקציה מדרגה גבוהה יותר, מצד אחד נקבל מודל אקספרסיבי יותר שיכול להסביר טוב יותר את התנהגות הדאטה הנתון, אך מנגד, החל ממקום מסוים תופעת ה Overfitting-תופסת תאוצה ותשפיע מאוד לרעה על תוצאות הולידציה כי נאבד מכלליות המודל. לכן, אם נעלה את הדרגה לדרגה הקטנה מ-d נצפה לשיפור בשגיאת האימון, וכאשר נעבור את הדרגה d כלליות המודל תאבד ושגיאת הולידציה תהיה לא טובה ותגדל. הקושי בבעיות אלו, הוא למצוא את ה"נקודה" (דרגה) המיוחדת שמבטאת את המודל בצורה אקספרסיבית מספיק מבלי לפגוע בכלליות.

17) להלן תוצאות הסיווג של Ridge באשר נבחן על סט המידה שכולל רק דוגמאות של נשים ועל דאטה שעבר: feature mapping



להלן תוצאות הסיווג של Ridge באשר נבחן על סט המידה שכולל רק דוגמאות של גברים ועל דאטה שעבר: feature mapping



	Train MSE	Validation MSE	Best alpha
Females	0.00166542	0.00179788	1
Males	0.00357682	0.0037711	1

18) להלן טבלת התוצאות :

Multilevel Model	Section	Sex	Train MSE	Valid MSE
Linear	4	M	0.00379825	0.00393541
Polynomial	5	M	0.00357682	0.0037711
Linear	4	F	0.00284185	0.00299319
Polynomial	5	F	0.00166542	0.00179788

19) ניתן לראות כי עבור גברים, feature mapping לא שיפר בצורה משמעותית את תוצאות הסיווג.
מנגד, עבור נשים, feature mapping שיפור באופן חד וברור את תוצאות השגיאה.
בדיוק כפי שראינו בהרצאות, העלאה בחזקה של הדאטה מראה תהליך של שיפור בשגיאה לצד הגדלת ה) Overfitting-ראינו זאת כתלות בחזקת הפולינום).
מגמה זו הינה הגיונית, כדי להסביר התנהגות מורכבת יש צורך בהעלאת המורכבות של אלגוריתם הסיווג, והעלאת מורכבות הסיווג גורמת כמעט באופן מיידי להעלאת ה Overfitting-לסט האימון (בגלל הקטנת הרגולריזציה והכלליות).
באופן כללי, מגמה של העלאת סיבוכיות מחלקת המסווגים מגדילה את תופעת ה-Overfittingוהמטרה שלנו היא למצוא את ה sweep spot-שמביאה תוצאות מדויקות מבלי לזרוק את כלליות המסווג ובלי להתאים לסט אימון ספציפי).

20) להלן הטבלה המעודכנת :

	section	Train MSE	Validation MSE
Dummy	2	0.0121805	0.01219
Basic Linear	3	0.00739167	0.00756576
Multilevel linear	4	0.00333368	0.00346586
Multilevel poly	5	0.00281449	0.00297481

21) להלן הטבלה בתוספת תוצאות המסווגים על סט המבחן :

	section	Train MSE	Validation MSE	Test MSE
Dummy	2	0.0121805	0.01219	0.013121
Basic Linear	3	0.00739167	0.00756576	0.007342
Multilevel linear	4	0.00333368	0.00346586	0.003166
Multilevel poly	5	0.00262943	0.00279333	0.002459

22) ניתן לראות כי המודל של multilevel poly קיבל את התוצאות הטובות ביותר על סט המבחן. באופן כללי אנו רואים מגמה של שיפור בכל התוצאות (הן על האימון, הולידציה והמבחן) ככל שאנו מתקדמים בתרגיל.
המסווג multilevel poly הגדיל את מרחב ההיפותזות (כי הכפיל את כמות הפיצ'רים), ומן הסתם ה sample complexity-גדל. תופעה שהיינו יכולים לחשוש שתקרה היא תופעת overfitting, אך התוצאות מראות שהיא אינה מתקיימת – אנו רואים **לצד** ירידה בשגיאת האימון ירידה (מתונה) בשגיאת הולידציה וכן ירידה בשגיאת המבחן. לכן, אנו יכולים להיות רגועים ולהסיק שכנראה אין התאמת יתר (יכול לקרות כי אולי לא הגענו לסף הדוגמאות הנדרש כדי שתופעת התאמת היתר תתחיל).
תופעת underfitting גם אינה מתרחשת אצלנו כי שגיאת האימון שלנו איננה עולה בשום שלב, וכן כלל שלושת השגיאות (אימון, ולידציה ומבחן) סובבות סביב אותו הערך, ולכן מהימנות.

23) כאשר הבחנו בשיפור במודל multilevel poly רצינו לשפר אותו משתי זוויות:
* אם השיפור קרה מחלוקת הדאטה לקבוצת נשים וגברים, אולי ישתפר מחלוקה נוספת ?
* כדי להקטין את מרחב ההיפותזות, שגדל פי 2 בבחירה במודל זה, נבצע בחירת פיצ'רים.

המודל שיצרנו, מחלק את הדאטה לפי מין, וכן לפי סוג דם (לפי החלוק ל-2 קבוצות שביצענו בתחילת התרגיל). כלומר, כאשר מגיעה דוגמא חדשה לסיווג, היא מסווגת על ידי מסווג שמתאים למין שלה. בתוך מסווג זה, היא פונה למסווג שיסווג אותה לפי סוג הדם שלה.

בנוסף, אימנו מסווג Ridge על כל סט האימון, ובדקנו מה ה- coef_של הפיצ'רים (בסט הפיצ'רים אחרי השימוש ב feature mapping-הריבועי). מיינו אותם לפי ערך מוחלט וכיוונו את הפרמטר n שמורה לאלגוריתם הלמידה שלנו

multi_poly_extra באיזה n פיצ'רים הוא משתמש (כאשר הם נבחרים להיות ה n-עם המקדמים הגבוהים ביותר בערך מוחלט).

כך אנו משיגים הקטנה של מרחב ההיפותזות ונמנעים מתופעת overfitting שיכולה הייתה לקרות בעקבות העלאת סיבוכיות המודל שלנו.