דו"ח – פרויקט סוף

למידה חישובית

מגישים:

תום נגר לוי 203961602

אסתי קולין 205429145

מבוא:

פרויקט זה הוא פרויקט מסכם של הקורס "למידה חישובית".

במסגרת הפרויקט ברצוננו ליישם חלק מן החומר הנלמד בכיתה, ללמוד ולהעמיק וכמו כן ללמוד דברים חדשים וליישמם.

את המידע איתו נתעסק (?) קיבלנו מחברת "אפימילק" המייצרת מערכות חליבה לרפתות בארץ ובעולם.

במסגרת העבודה קיבלנו שני קבצים המתעדים את הפרטים של הפרות בשני רפתות בקיבוצים שונים - רפת גבעת חיים ורפת סעד.

המידע שקיבלנו מתעד את כל פרטי תפוקת החלב של הפרה לאורך כל תקופת התחלובה שלה. תקופת התחלובה זו התקופה בה חולבים את הפרה, מתאריך המלטתה ועד לתקופת 305 ימים.
הפרטים של תפוקת החלב כוללים בין היתר - כמות החלב, אחוזי שומן, אחוזי חלבון וכו' (פירוט המידע המלא נמצא בנספחים.)

בעבודה זו נרצה לחזות פרטים לגבי תפוקת הפרות בסוף תקופת החליבה ביום ה-305, לאחר 54 ימי החליבה הראשונים של הפרה. הפרטים כוללים : מהו סך הכמות החלב, השומן הכולל, החלבון הכולל וכמות התאים הסומטיים לאחר 305 ימים.

שלבי הפרויקט:

החלטנו לבדוק האם הוספת תכונות נוספות (תכונות שאנחנו מוסיפים) יעזרו לשיפור המודל לכן הוספנו את התכונות הבאות:

1) WeekInMonth - מספר השבוע בתוך החודש (1 = 1-8 , 2=9-16, 3=17-24, 4=25-31)

2) WeekNumber - מספר השבוע בשנה

לאחר סינון התכונות החשובות ראינו שהתכונה WeekInMonth אינה משפיע על המודל מכיוון שה"ציון" שלה נמוך אך לעומת זאת WeekNumber קיבל "ציון" גבוהה יחסית.

לאחר קריאת הנתונים, חילקנו את הדאטא שלנו ל-3 : המידע מרפת סעד, המידע מרפת גבעת חיים והמידע משני הרפתות.

בשלב הראשון לקחנו מהדאטא רק את הדוגמאות של ה-54 ימים הראשונים בתקופת תחלובת הפרה. לאחר מכן עשינו נירמול לנתונים וחילקנו את הדאטא ל-train (70%) ו- test (30%) .

עברנו על כל אחד מסוגי הדאטא ועבורו בחנו בכל אחד מארבעת הנתונים שברצוננו לחזות את הדברים הבאים:

1. מהן התכונות החשובות - בדאטא הנתון קיבלנו כחמישים תכונות. נרצה לסנן את התכונות שלא עוזרות לצפות את הנתונים הרצויים כדי לא להעמיס על המערכת ולהתמקד בתכונות שמשפיעות בצורה המירבית. בשלב זה השתמשנו באלגוריתם החמדן שבו ווידאנו שעם התכונות הספציפיות האלה קיבלנו את "הציון" הגבוה ביותר.
2. בשלב זה חקרנו ובדקנו לגבי שיטות נוספות של Linear Models:
   - **Ordinary least squares Linear Regression** - המודל הסטנדרטי,
   - **Ridge regression** - החלטנו לבדוק את המודל הנ"ל מכיוון שהוא נותן מענה לבעיות במודל הסטנדרטי בכך שהוא מקטין את ממוצע השגיאה הכוללת, אך הבעייתיות שבו שהוא משנה מעט את האומד שלו.
   - **Lasso** - מודל שמשתמש בצמצום תכונות, הוא בנוסף לצמצום ממוצע השגיאה גם מצמצם את השונות.
   - **ElasticNet** - מודל שבעצם משלב בין ה - Ridge בכך שהוא שומר על הרגולריזציה שלו ל- Lasso שמדלל את התכונות. המודל יותר יציב בגלל שהוא משלב את שניהם.
   - **SGD - Stochastic Gradient Descent** : המודל אותו למדנו בכיתה.

על מנת לבצע את ההשוואה נשתמש במדד החציון. בצורה מדעית אופציה זו נותנת מדד שלא נותן דגש לרעשים או חריגות חד פעמיות וניתן לראות כבר בשלב הזה שחציון השגיאה המוחלטת של מודל ה- Ordinary least squares Linear Regression

3. בשלב זה יצרנו רשת נוירונים ובהמשך נשווה את המודל למודלים הליניאריים ונראה האם הרשת נוירונים עדיפה או לא.

כוח החישוב של רשתות נוירונים נובע, בראש ובראשונה, מהמבנה המבוזר והמקבילי שלהן.

החלטנו להריץ את רשת הנוירונים עם שינוי פרמטרים כך שנשנה את הפרמטר epochs (50,100,500,1500,5000) בכל הרצה ומבניהם נבחר את המודל הטוב ביותר

הרשת בנויה כך:

מספר שכבות : (5,10,20), הרחבה בשלב המסקנות

Optimizer : adam

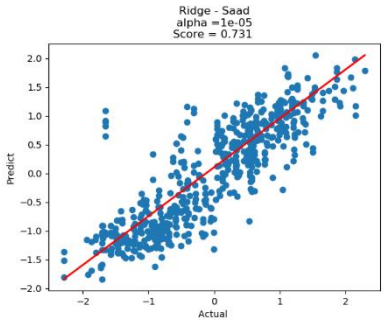
	שכבה 1	שכבות ביניים	שכבה 5
גודל	מספר התכונות חלקי 2	128	1 (תוצאה סופית)
אלגוריתם	relu	relu	linear

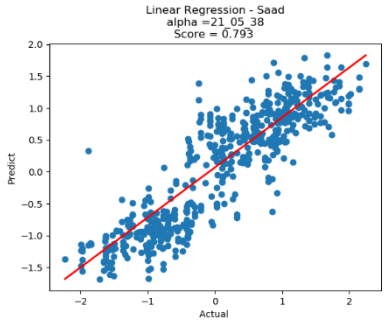
<u>תוצאות הפרויקט:</u>

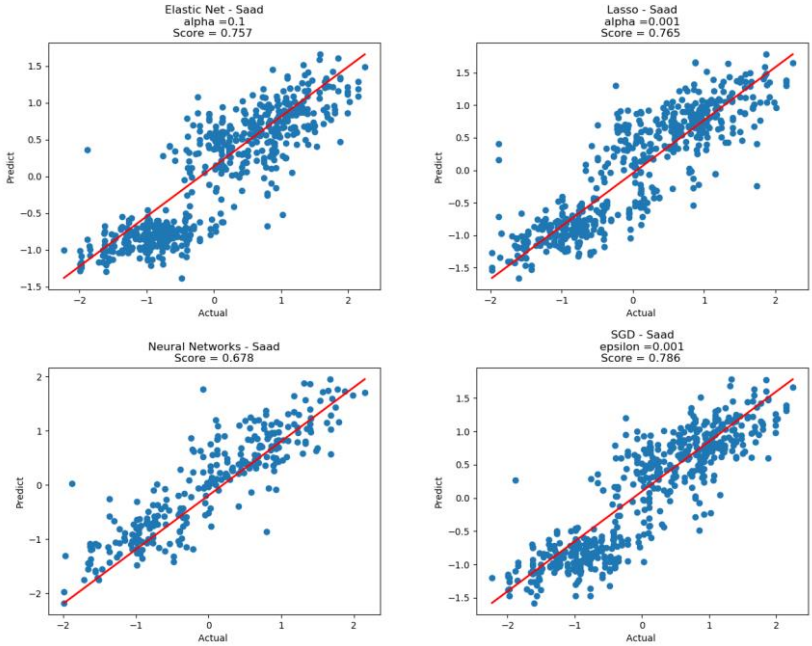
בחלק זה של הדו"ח החלטנו לצרף רק את הנתונים של הרפת בסעד. הנתונים המלאים ימצאו בנספחים.

נראה את התוצאות עבור כל y –
('sum Ecm 305', 'sum prot 305', 'sum fat 305', 'sum milk 305')

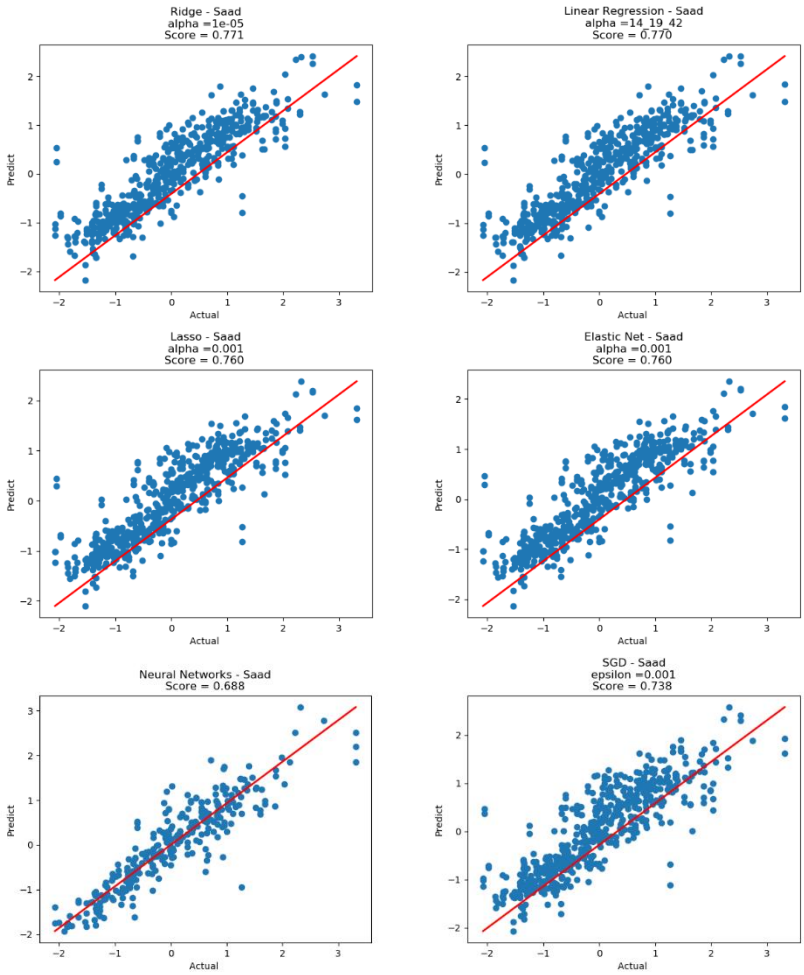
sum milk 305:



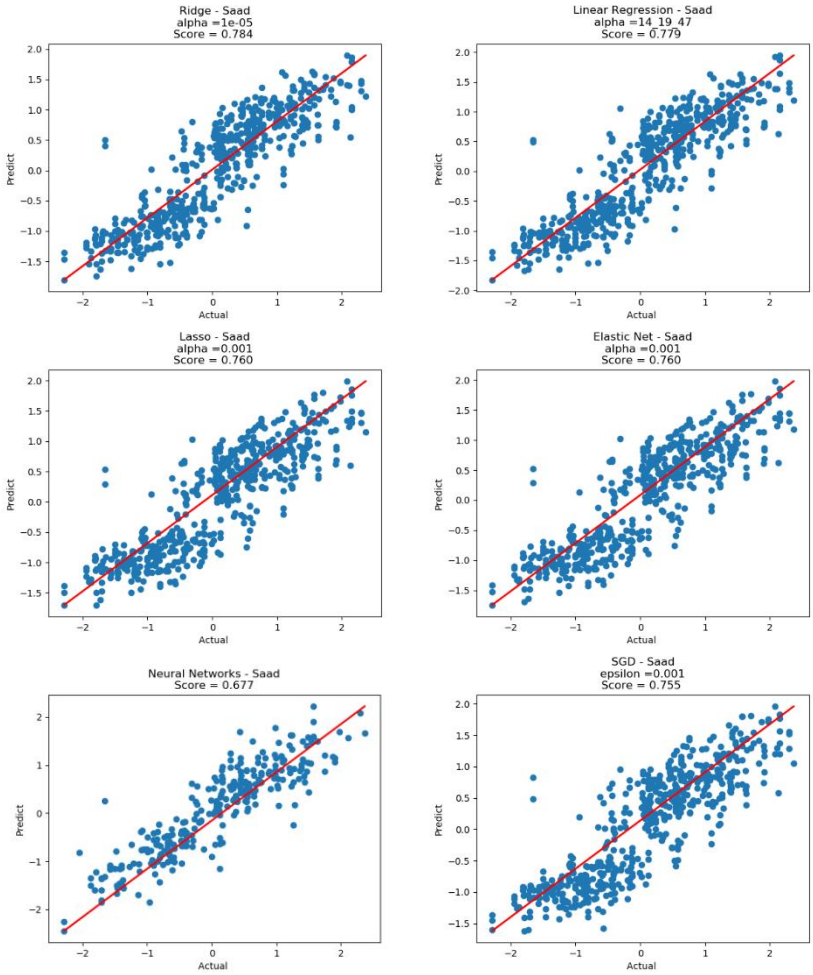



Elastic Net - Saad
alpha =0.1
Score = 0.757
Lasso - Saad
alpha =0.001
Score = 0.765
Neural Networks - Saad
Score = 0.678
SGD - Saad
epsilon =0.001
Score = 0.786
Predict
Actual

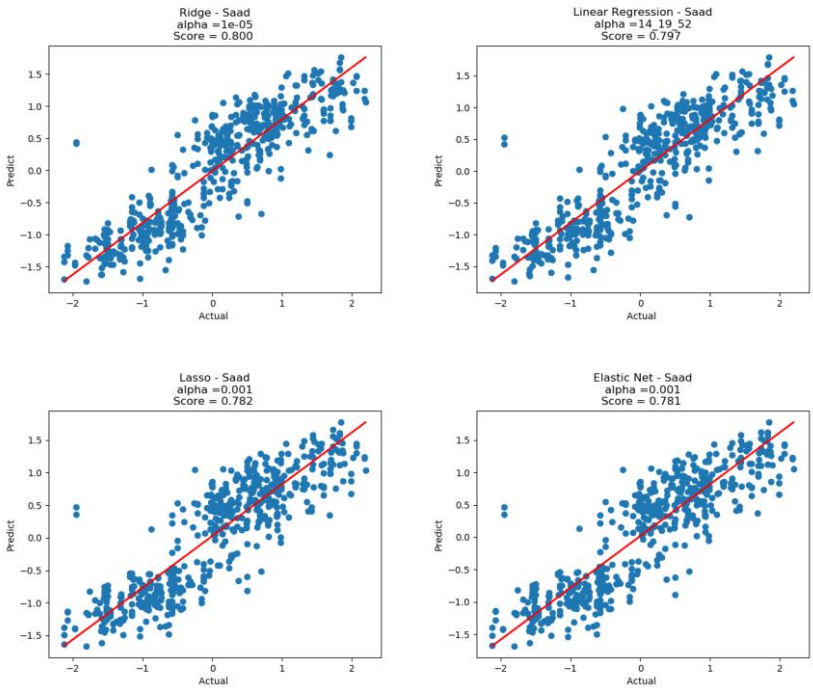
:sum fat 305

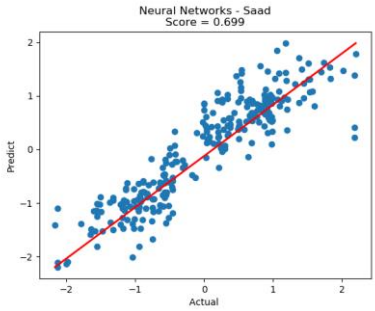

Ridge - Saad
alpha =1e-05
Score = 0.771
Linear Regression - Saad
alpha =14_19_42
Score = 0.770
Lasso - Saad
alpha =0.001
Score = 0.760
Elastic Net - Saad
alpha =0.001
Score = 0.760
Neural Networks - Saad
Score = 0.688
SGD - Saad
epsilon =0.001
Score = 0.738
Predict
Actual

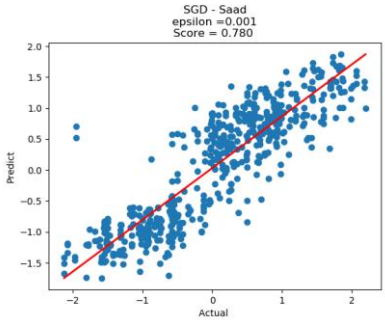
:sum prot 305



:sum Ecm 305




Neural Networks - Saad
Score = 0.699
Predict
Actual


SGD - Saad
epsilon =0.001
Score = 0.780
Predict
Actual

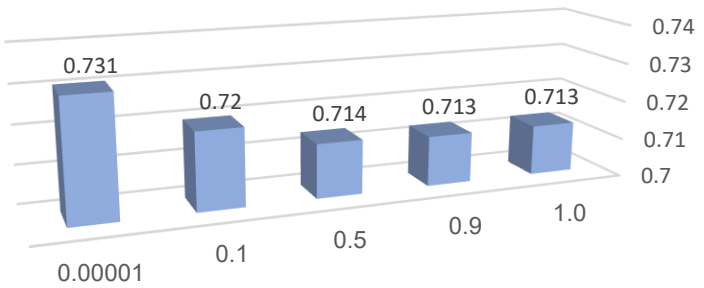
מסקנות:

ניתן לראות ברוב המקרים שמודל Ordinary least squares Linear Regression הוא המודל בעל הציון הגבוה. אך בנוסף לכך, ניתן לראות שגם אלגוריתם Ridge regression הוא גם כן בעל ציון גבוה, ובחלק מועט מן המקרים הציון המתקבל אף יותר גבוה מהמודל הסטנדרטי של Linear Regression .

כמו כן, במעבר על הפרמטרים המתקבלים בכל אחד מהאלגוריתמים. בסעיפים הבאים, ציינו מגמה שניתן לראות בכל אחד מה- y ים המתקבלים, אך בשל הרצון לא להעמיס על הדו"ח הבאנו רק את ההתייחסות ל- y= sum milk 305.

- **Ridge regression** – ערך ה-α הוא קובע את גודל השונות מהפונקציה האמיתית. ככל שנגדיל את ה-α , כך הקפיצות של פונקציית האומדן יהיו יותר קטנות ונמוכות.

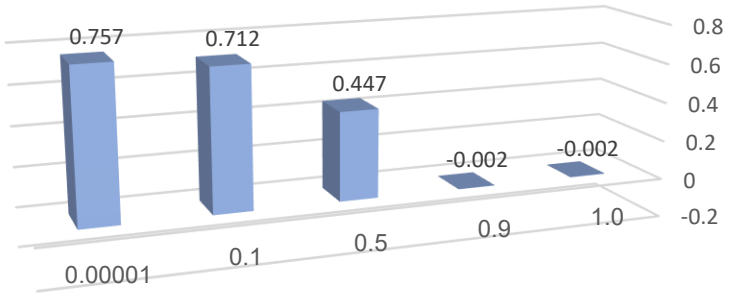
ערכים עבור α :
[1.0, 0.9 ,0.5, 0.1,0.00001]



ניתן לראות שככל שה- α קטן, הציון עבורו עולה.

- **Lasso** – כמו שציינו, האלגוריתם בעצם קובע ערך למקדמים של התכונות. חלק חשוב באלגוריתם הוא קיבעת ה- α. כאשר α=0 זה כמו רגרסיה לינארית פשוטה, כאשר לכל תכונה יש מקדם. אך ככל שנגדיל יותר את ה- α זה בעצם הסינון של התכונות שהאלגוריתם מבצע.

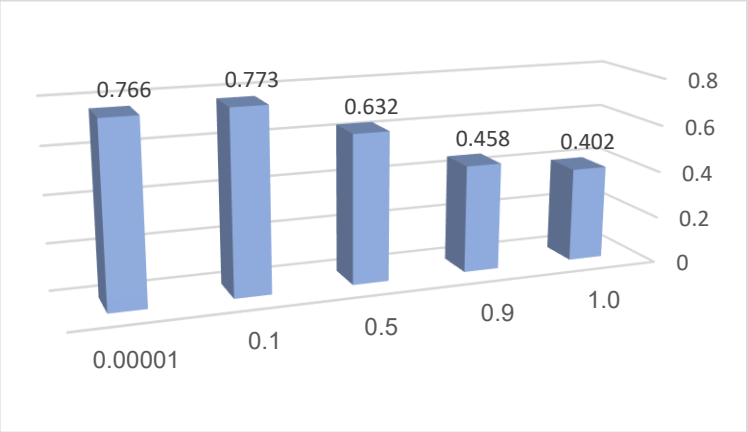
ערכים עבור α :
[1.0, 0.9 ,0.5, 0.1,0.00001]



ניתן לראות שככל שאנחנו מקטינים את α, הציון גדל. השערתנו היא שהנתונים אותם שלחנו לאלגוריתם הם נתונים שכבר עברו סינון, שכן הפעלנו את האלגוריתם של סינון התכונות החשובות ביותר. ולכן כאשר α גדלה באלגוריתם זה, האלגוריתם נאלץ לסנן תכונות שהן אכן חשובות. את הדוגמא הקיצונית לכך ניתן לראות כאשר בדקנו עבור α=0.9 ו- α=1.0 אז האלגוריתם לא הצליח לנבא דבר.

- **ElasticNet** – כאמור, אלגוריתם זה הוא משלב בין ה- Ridge לבין Lasso . באלגוריתם זה פרמטר ה- α הוא מכריע במשקלים ביניהם, כאשר α=0 מתייחס רק ל- Ridge ו- α=1 מתייחס רק ל-Lasso . בחרנו לבדוק מספר ערכים עבור α :

[0.001 ,0.1 ,0.5, 0.9 ,1.0]

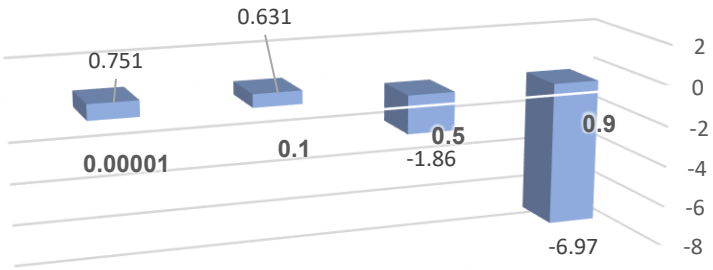


ניתן לראות שכל המקרים, עבור α=0.1 התקבל ציון יותר טוב. עבור כל בדיקה, ניתן לראות שככל שמתקרבים יותר ל- Ridge ומתרחקים מה – Lasso אז המודל יותר טוב.

- **SGD - Stochastic Gradient Descent** : באלגוריתם זה בחנו את שינוי הפרמטר אפסילון. רצינו לבחון כמה משמעותי יהיה ההבדל בין רמות האפסילון השונות.

ערכים עבור אפסילון :

[0.9 ,0.5, 0.1 ,0.001]



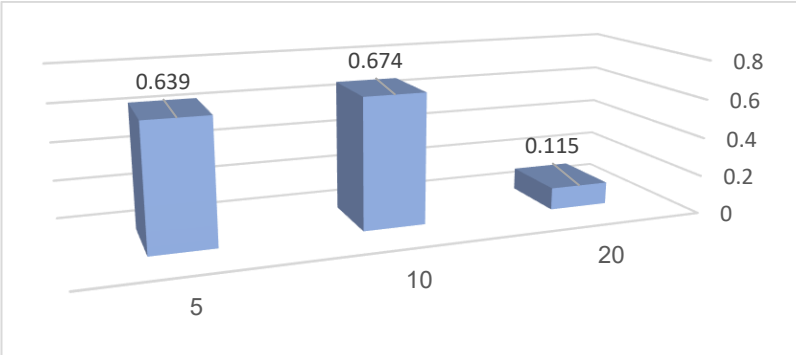
ניתן לראות שאכן ככל שהאפסילון קטן, כלומר, סך השגיאה קטן הציון המתקבל מהאלגוריתם גדל, וכאשר נתנו טווח שגיאה גדול יותר, השגיאה גדלה בהתאם.

- **רשת נוירונים** :

באלגוריתם זה רצינו לבחון שני שינויים –

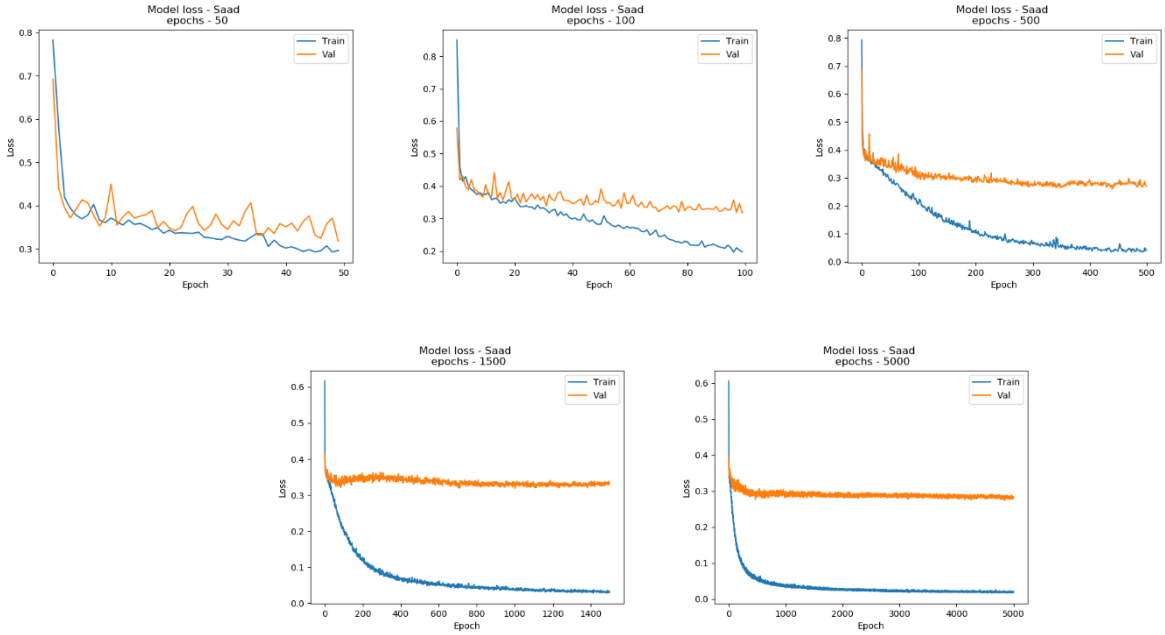
א. השינוי בין מספר שלבי הביניים (5,10,20)
ב. השינוי במספר ה- epochs. (50,100,500,1500,5000)

השערתנו הייתה שביותר שלבי ביניים האלגוריתם יעבור יותר טוב, ולהפתעתנו גילינו כי לא כך היה וכבר עצרנו בבדיקה של למעלה מ20 שלבי ביניים. ואכן ניתן לראות כי עבור 10 שלבי ביניים הציון היה המירבי.



כמו כן, היה גם פרמטר נוסף אותו ניסינו לבחון – את מספר ה- epochs.

וכמו שציפינו, ככל שהעמסנו יותר במספר epochs, כך השגיאה הלכה והתייצבה (אך לא התכנסה משמעותית). ניתן לראות שלאחר כ- 400 איטרציות כבר לא ניתן לראות שינוי משמעותי בשגיאות.



נספחים:

Data:

FarmCode	מספר חווה	DR^2	מספר ימים עד לסיום התחלובה בריבוע
DateMonth	החודש בוא נלקחה הבדיקה	TWIN	האם המליטה תאומים(0/1)
Date (DD/MM/YYYY)	התאריך בו נלקחה הבדיקה	STILL	האם הייתה לפרה מות עובר תוך ביטני(0/1)
WeekNumber	מספר השבוע בשנה	DRDIM	מספר ימים עד לסיום התחלובה * מספר ימים בתחלובה
WeekInMonth	מספר השבוע בחודש	DRDIM^2	מספר ימים עד לסיום התחלובה * מספר ימים בתחלובה בריבוע
DIM	מספר הימים בתחלובה	DRDP	מספר ימים עד לסיום התחלובה * מספר ימים בהריון
CowID	מספר פרה	DRDP^2	מספר ימים עד לסיום התחלובה * מספר ימים בהריון בריבוע
parity	מספר תחלובות של הפרה עד היום	DD	מספר ימים בייבוש
CalvingDate	תאריך המלטה	DD^2	מספר ימים בייבוש בריבוע
milk(kg)	כמות חלב בקילו	term	מספר ימי התחלובה הקודמת
accum milk (kg)	כמות חלב מצטברת בקילו	term^2	מספר ימי התחלובה הקודמת בריבוע
fat (%)	אחוז שומן	PPUD	האם הייתה לפרה דיכאון אחרי לידה(0/1)
fat(kg)	כמות שומן בקילו	FPR	יחס שומן חלבון
accum fat(kg)	כמות שומן מצטברת בקילו	LRDRm	ממוצע חלב ב10 ימים * מספר ימים עד לסיום התחלובה
protein(%)	אחוז חלבון	LR^2DRm	ממוצע חלב ב10 ימים * מספר ימים עד לסיום התחלובה בריבוע
protein(kg)	כמות חלבון בקילו	LRDRf	ממוצע שומן ב10 ימים * מספר ימים עד לסיום התחלובה
accum prot(kg)	כמות חלבון מצטברת בקילו	LR^2DRf	ממוצע שומן ב10 ימים * מספר ימים עד לסיום התחלובה בריבוע
ECM(kg)	כמות תאים סומטים בקילו	LRDRp	ממוצע חלבון ב10 ימים *מספר ימים עד לסיום התחלובה
accum ECM(kg)	כמות תאים סומטים מצטברת בקילו	LR^2DRp	ממוצע חלבון ב10 ימים *מספר ימים עד לסיום התחלובה בריבוע
cond	מוליכות החלב	LRDRecm	ממוצע תאים סומטים ב10 מים * מספר ימים עד לסיום התחלובה
DR	מספר ימים עד לסיום התחלובה	LR^2DRecm	ממוצע תאים סומטים ב10 מים * מספר ימים עד לסיום התחלובה בריבוע
CS	החודש בו ההמלטה קרתה	DP	מספר ימים בהריון
DRCS	מספר ימים עד לסיום התחלובה	sum milk 305	כמות חלב מצטברת לאחר 305 ימים
age	גיל הפרה בחודשים	sum fat 305	כמות שומן מצטברת לאחר 305 ימים
age^2	גיל הפרה בריבוע בחודשים	sum prot 305	כמות חלבון מצטברת לאחר 305 ימים
DRAI	*DRגיל של הפרה בזמן ההמלטה	sum Ecm 305	כמות תאים סומטים מצטברת לאחר 305 ימים