חיזוי בצורת בעזרת נתונים מטאורולוגיים ויבשתיים

סמינר בהנדסת תכנה – 10400  
8 במאי, 2021

נחום קלטקין - 320614654

תוכן עניינים

[רקע 1](#_Toc71279769)

[הצגת היפותזה 2](#_Toc71279770)

[תופעת הבצורת 2](#_Toc71279771)

[שיטת המחקר 2](#_Toc71279772)

[תיאור ה-Dataset 3](#_Toc71279773)

[מאגר הנתונים 3](#_Toc71279774)

[עיבוד ה-Dataset 6](#_Toc71279775)

[בדיקת קיום ערכים ריקים 6](#_Toc71279776)

[עיבוד המידע 6](#_Toc71279777)

[קורלציה וקשרים בין הפרמטרים 6](#_Toc71279778)

[איזון 9](#_Toc71279779)

[הפרדת המשתנה התלוי 9](#_Toc71279780)

[חלוקת הנתונים לאימון ולבדיקה 9](#_Toc71279781)

[אלגוריתמים 10](#_Toc71279782)

[יער אקראי – Random Forest 10](#_Toc71279783)

[תיאור 10](#_Toc71279784)

[מימוש בשיטה רגרסיבית 10](#_Toc71279785)

[תוצאות 10](#_Toc71279786)

[מימוש בשיטת הקלסיפיקציה 11](#_Toc71279787)

[תוצאות 11](#_Toc71279788)

[באייס נאיבי משלים – Complement Naïve Bayes 12](#_Toc71279789)

[תיאור 12](#_Toc71279790)

[מימוש 12](#_Toc71279791)

[תוצאות 13](#_Toc71279792)

[רגרסיה לוגיסטית – Logistic Regression 14](#_Toc71279793)

[תיאור 14](#_Toc71279794)

[מימוש 14](#_Toc71279795)

[תוצאות 14](#_Toc71279796)

[מסקנות 16](#_Toc71279797)

[ניסיון למיקוד המידע 17](#_Toc71279798)

[סיכום הסמינר 18](#_Toc71279799)

[ביבליוגרפיה 19](#_Toc71279800)

# רקע

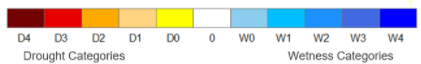
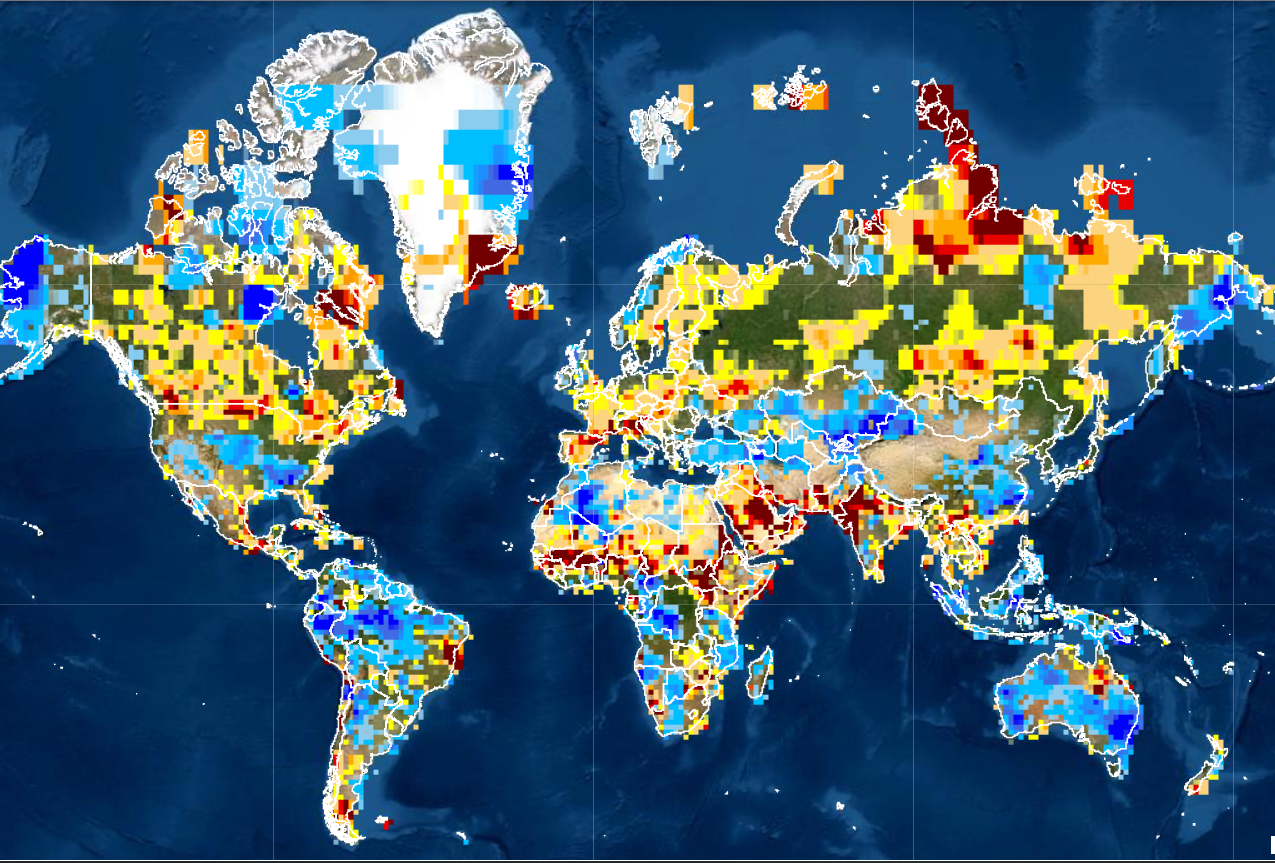
ההגדרה הפורמלית של בצורת מוגדרת על ידי מחסור במים כתוצאה ממחסור במשקעים לאורך זמן ממושך (עונה או יותר). קיימים מספר סוגים של בצורת כאשר כל אחד מהם נמדד בצורה שונה והשפעותיו שונות:

* **בצורת מטאורולוגית** - כאשר אזור מסוים חווה תקופה יבשה וארוכה, בניגוד למצופה מהממוצע הרב-שנתי לתקופה המקבילה.
* **בצורת חקלאית** - כאשר תפרוסת המשקעים במרחב ובזמן וצורתם של המשקעים אינם תואמים את דרישות הגידולים החקלאיים.
* **בצורת הידרולוגית** - כאשר רזרבות המים הזמינות ממקורות כמו אקוויפרים, אגמים, מעיינות וכדומה נמצאת מתחת לממוצע. תנאים של בצורת הידרולוגית יכולים להתפתח גם כאשר כמות המשקעים אינה נמוכה. נציין גם שמדינת ישראל נמצאת בבצורת הידרולוגית כמעט באופן תמידי.
* **בצורת אקלימית** - שנה שקיבלה כמות משקעים הקטנה מההפרש שבין כמות הגשם הממוצעת והסטייה הממוצעת לאותו מקום. כלומר, כמות המשקעים שירדה הייתה קטנה ב-30% מהממוצע הרב שנתי לאותו מקום.

לפי ארגון הבריאות העולמי כ55 מיליון אנשים בעולם מושפעים כל שנה מבצורת. בצורת הינה הסכנה הממשית הגדולה ביותר לגידולים, צאן ובקר כמעט בכל מקום בעולם. מחסור במים משפיע על כ40% מאוכלוסייה בעולם וכ700 מיליון אנשים ככל הנראה יאלצו להעתיק את מיקומם עד שנת 2030 בעקבות בצורת.

בישראל ,המדידות המטאורולוגיות בישראל החלו בשנת 1850. מאז ועד שנת 1956 הממוצע עמד על 3 שנות בצורות רצופות אחת ל-50 שנים. כאשר ב-1956 נקבע שיא חדש של 6 שנות בצורת עד שנת 1962. שיא זה נשבר בשנת 2011 לאחר 8 שנות בצורת רצופות שהחלו ב2004. החל משנות ה-2000 התרחשו שלושה אירועים של אירועי בצרות רצופים - אחד בשנים 1999-2000, אחד בשנים 2004-2011, ואחד שהחל בשנת 2013 ונמשך עד היום.   
בשנים 2013-2016 פקדה את צפון ישראל בצורת קשה. שנת 2016 הייתה שנה קשה במיוחד - ב-90 שנים האחרונות רק בשנה אחת היו פחות משקעים. הבצורת בצפון הייתה כה קשה שהביאה לתמותה של עצי חורש ועצי בר זית מיובש בגליל המערבי והתחתון ובכרמל על פי המדען הראשי של הקרן הקיימת לישראל דוקטור עמרי בונה. תופעה זו תרמה רבות לגל השרפות שפקד את ישראל בנובמבר 2016.

מפת בצורת עולמית על פי מדד GPCC (Global drought index) המבוסס על מדדים מיוחדים של אקלים (STI, SPEI,PET):

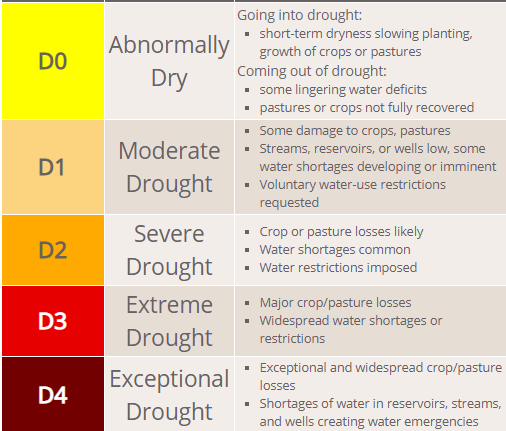


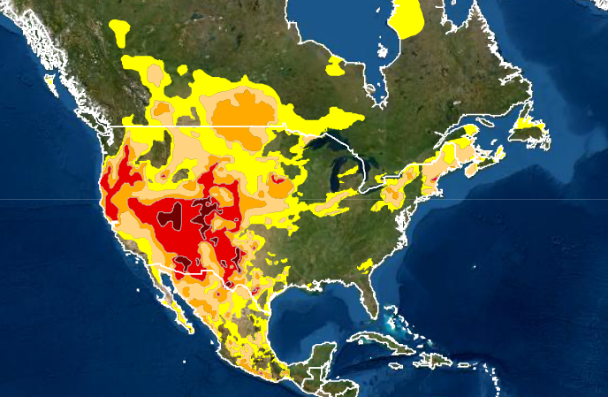
# הצגת היפותזה

## תופעת הבצורת

* בצורת הינה תופעה נורמלית כחלק מהאקלים. זוהי סכנה איטית ולכן לעתים גורמת לכך שאינה נלקחת מספיק ברצינות. ההפסדים כתוצאה מבצורת משמעותיים מאוד ועלולים לעשות נזק שלא נופל מנזק הנוצר מתופעות אקלים מהירות כמו סופות טורנדו או הוריקנים.
* בצורת פוגעת באופן ישיר בחקלאות, במקורות מים, אנרגיה, בריאות ציבורית ובבריאות חיות הטבע.
* זיהוי הגעה של בצורת בטרם עת יכול לעזור משמעותית במזעור נזקים וחיסכון כספי. ככל שנגלה יותר מוקדם כך ניתן להתכונן יותר טוב לבצורת ולמזער את הנזקים.
* דרכים מסורתיות לזיהוי הגעת בצורת הן פשוטות יחסית – השוואה לזמנים דומים בתנאים דומים משנים קודמות.
* מטרת המחקר זה היא לחקור את יעילות השימוש בפרמטרים מטאורולוגיים וגאולוגיים מרובים כדי לזהות היווצרות של בצורת בשלב מוקדם.
* מטרת פרויקט זה היא לפתח מודל שיכול לחזות בדיוק גבוה ככל שניתן את היווצרותה של בצורת מבין 17 גורמים מטאורולוגיים ו31 גורמים גאולוגיים.

## שיטת המחקר

* **ביצוע תהליכי כריית מידע מתוך בסיס נתונים משמעותי המספק מידע מטאורולוגי ויבשתי רלוונטי בנקודות גאוגרפיות הפזורות ברחבי ארצות הברית בניסיון להסיק ולחזות את התנאים המזהים להגעת בצורת.**
* **תהליכי כריית המידע יבוססו על אלגוריתמי כריית מידע.**
* **הבעיה מתאימה לאלגוריתמי רגרסיה וקלסיפיקציה וכן יבדקו שתי הדרכים על מנת למצוא פתרון אופטימלי.**
* **פרמטר ה-Score (מדד בצורת כאשר 0 – אין בצורת ל5 – בצורת קיצונית) יהיה נקודת ההשוואה.**



# תיאור ה-Dataset

## מאגר הנתונים

* קובץ ה-Dataset נלקח מאתר Kaggle.  
  <https://www.kaggle.com/cdminix/us-drought-meteorological-data>.
* המאגר התקבל משילוב של מאגרים ממרכז המחקר לנגלי של NASA ומרכז צמצום הבצורת של אוניברסיטת נברסקה, ארה"ב ([https://droughtmonitor.unl.edu/)](https://droughtmonitor.unl.edu/)%20) יחד עם המאגר ההרמוני על אדמות ושטחים של ארגון האוכל והחקלאות של האום (<http://www.fao.org/soils-portal/data-hub/soil-maps-and-databases/harmonized-world-soil-database-v12/en>).
* המאגר מכיל מידע מטאורולוגי ויבשתי על אזורים גאוגרפיים שונים ברחבי ארה"ב. אזורים אלו מאופיינים על ידי קוד FIPS. ה-Dataset מחולק לשלושה קבצי מידע מטאורולוגי וקובץ מידע גאולוגי יחיד.
* סה"כ השורות במידע המטאורולוגי עומד על 23841468 ובמידע היבשתי מספר השורות עומד על 3109.



* ניתן לראות כי בקבצים המטאורולוגיים חסר מידע רק בשורות ה"score". חוסרים אלו נובעים מכך שבצורת נמדדה פעם בשבוע ולכן קיים מידע רק בכל שורה שביעית. מכיוון שברשותנו מידע רב וגם כי אנו מנסים לנבא את העמודה הזו לא נשלים את המידע החסר אלא נמחק את שורות אלו בהמשך.
* ההצלבה בין המידע המטאורולוגי והמידע היבשתי מתבצעת לפי עמודה ה"fips" המציינת מקום גאוגרפי מסוים.

# עיבוד ה-Dataset

## בדיקת קיום ערכים ריקים

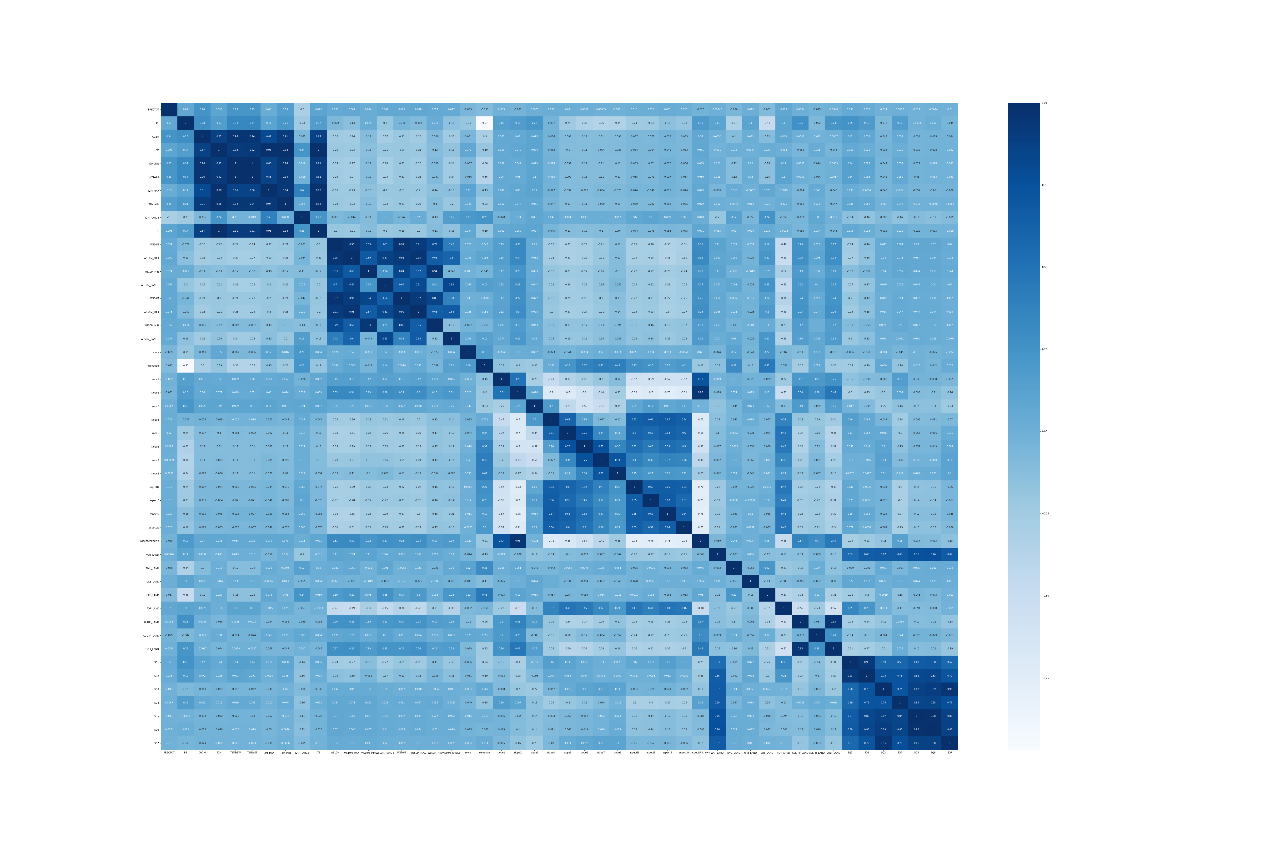
* נעלה את הקבצים ונבצע בהם פעולות בעזרת ספריית pandas.
* נבדוק כל קובץ בסיס נתונים בנפרד ונוודא:
  + תאים ריקים
  + ערכים לא הגיוניים
  + סוג הערכים
* בכל בסיס נתונים של מידע מטאורולוגי היו כ85% שורות ריקות. לא נמלא חוסרים ונמחק את כל התאים הריקים מהסיבות:
  + מדד הבצורת נמדד בעזרת מכשור ונוסחה מורכבת.
  + קיים המון מידע גם ללא ערכים ריקים אלו.
  + מדד הבצורת הינו ערך הטרגט שלנו.
* בבסיס המידע היבשתי לא היו חוסרים וערכים לא תקינים.

## עיבוד המידע

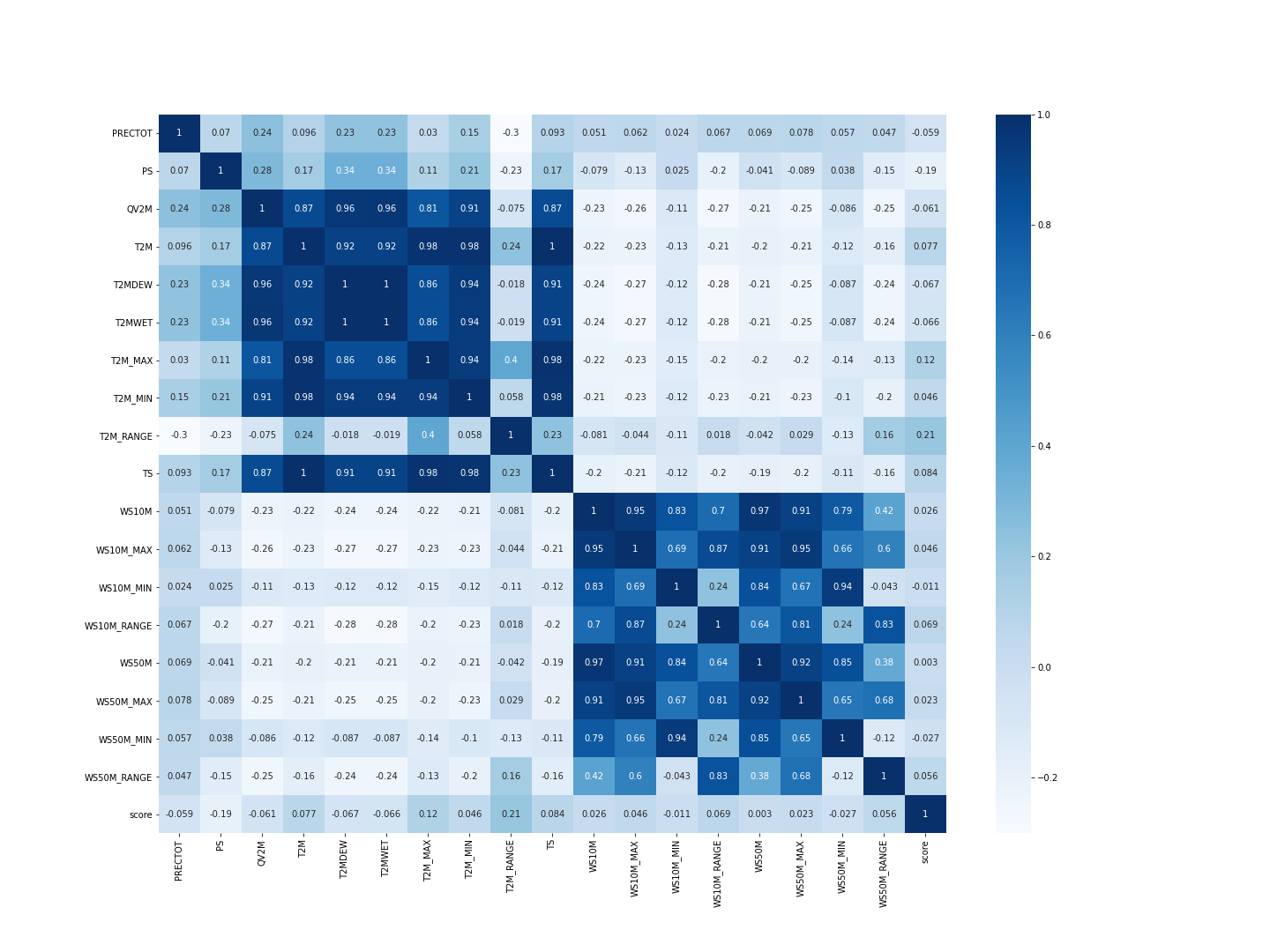
* על מנת לשמר את כמות המידע ולאחד אותו בצורה תקינה האיחוד בוצעו בכמה שלבים.
  + נאחד בנפרד כל אחת מטבלאות המידע המטאורולוגי עם המידע היבשתי לפי עמודת ‘fips’.
  + נאחד את שלושת הטבלאות המאוחדות לאחת גדולה על ידי הוספת שורות מתחת.
  + נמחק מהפריים הסופי את עמודות התאריך (date) , המיקום (fips), האורך (lon) והרוחב (lat) מכיוון שאנו מחפשים קשר בין המידע המטאורולוגי והיבשתי לבין מדד הבצורת ואין צורך באלו.
* התוצאה המתקבלת הינה פריים גדול המכיל את עמודת הטרגט שלנו אל מול כל שאר 47 הפרמטרים השונים.

## קורלציה וקשרים בין הפרמטרים

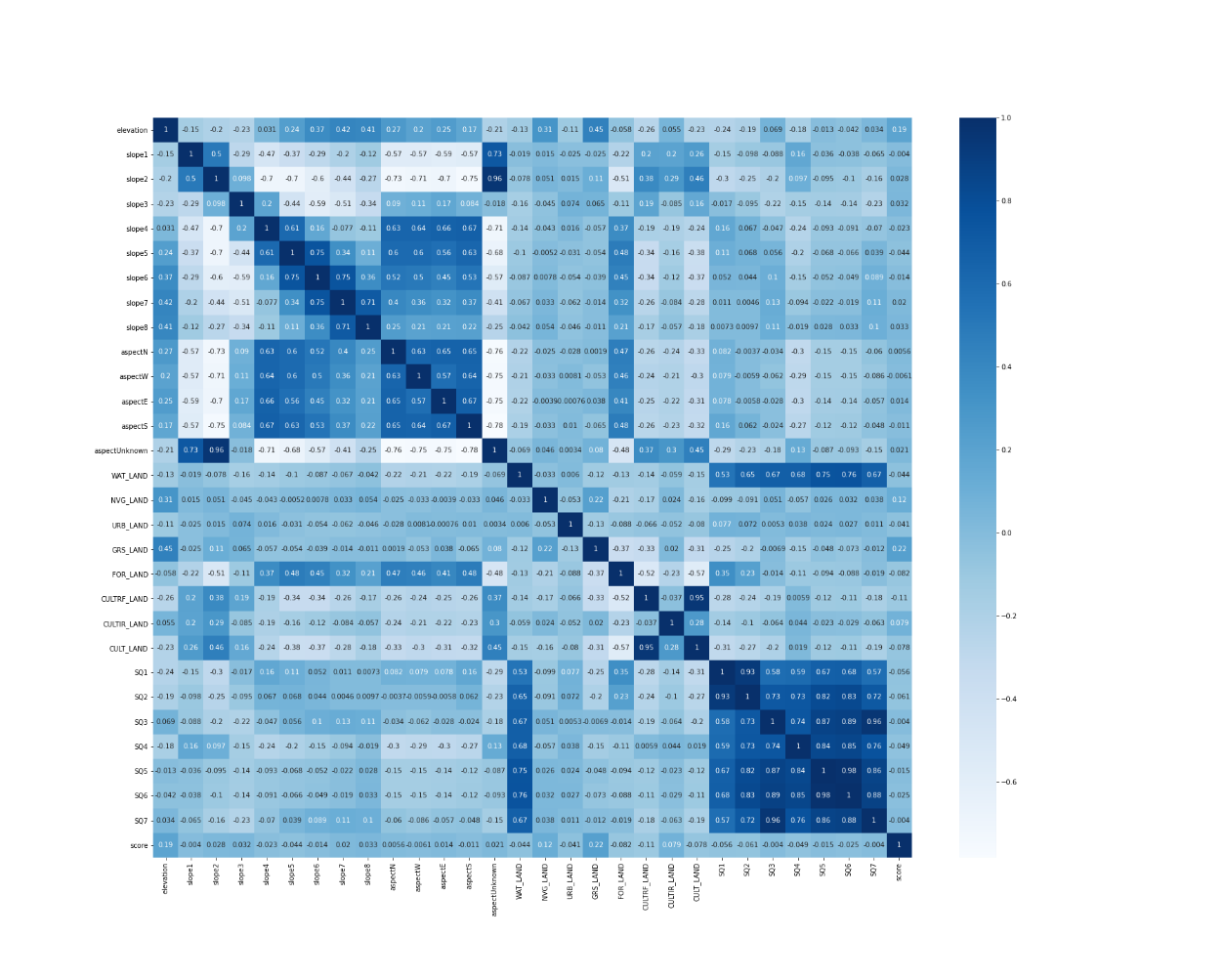
* נבצע מבחני קורלציה בעזרת מבחן פירסון על מנת לבדוק אם יש עמודות בעלות מקדמים גבוהים כך שנוכל למחוק חלק מהן.
* בוצע מבחן פירסון על הפריים הכללי. התברר כי הקורלציה מאוד נמוכה.



* + נראה שעמודות הטמפרטורה (TEMP\_RANGE) ו (T2MAX), גובה (elevation), סוג אדמה (NVG\_LAND) (GRS\_LAND) והשקיה (CULTRF\_LAND) מהווים את הקשר המשמעותי ביותר (אומנם עדיין קורלציה נמוכה).
* נבצע הפרדה ומספר מבחני פירסון נוספים:
  + בין המידע המטאורולוגי לבין מדד הבצורת



* + .בין המידע היבשתי לבין מדד הבצורת.



* + בין המשתנים שמצאנו לבין מדד ה’score’.

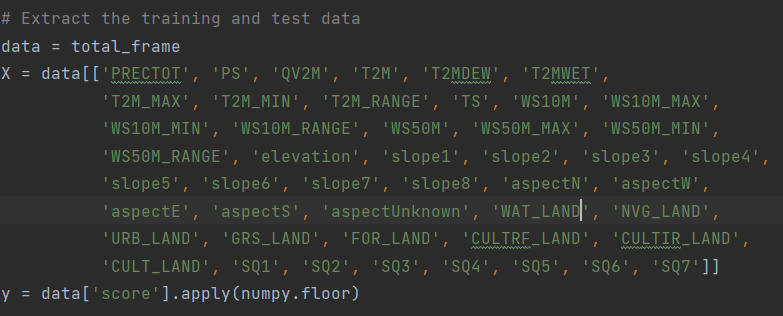


* המסקנה היא שכנראה הקשר בין הנתונים האלה יחסית נמוך ומידע זה איננו מספיק כדי לנבא בצורת וצריך מידע נוסף עם קורלציה גבוה יותר.
* בכל זאת, נבחן את הנתונים הקיימים וננסה לבדוק עד כמה ניתן לדייק בניבוי בעזרת אלגוריתמים חכמים.

## איזון

* נשים לב שהבעיה שלנו הינה בעיית קלסיפיקציה לא מאוזנת. כלומר התפלגות המידע בין ששת מצבי הבצורת שלנו איננה אחידה ועל כן האלגוריתם יהיה יותר פחות מדויק במצבים הקיצוניים יותר.   
  טבלת ההתפלגות של המידע:
* מחסור באיזון זה מוסיף לקושי של הבעיה ואף מחזק את המסקנה לאחר ביצוע קורלציה – לא ניתן להסתמך באופן מלא על יכולות החיזוי של האלגוריתמים לקשר בין בצורת לבין נתונים מטאורולוגיים ויבשתיים.

## הפרדת המשתנה התלוי

* בשלב הלמידה של האלגוריתם מתבצעת הפרדה של המשתנה על מנת להשוות לבסוף את תוצאות האלגוריתם (החיזוי) עם התוצאות הקיימות (האמיתיות).
* בשלב הבדיקות האלגוריתם משתמש בהפרדה זו על מנת לחשב אחוזי דיוק.
* קבוצת משתני ה feature:
* ערך הטרגט:

## חלוקת הנתונים לאימון ולבדיקה

* כל אלגוריתם של למידת מכונה מחולק לשלב הלמידה ולשלב המבחן.
* בשלב הלמידה נשתמש בחלק מן המידע הקיים על מנת "לאמן" את המודל שלנו על ידי יצירת הקשרים בין הפרמטרים וחיזוי של התוצאה הרצויה.
* בשלב המבחן המודל ינסה לחזות את התוצאה הרצויה והדיוק יקבע לי הפגיעה במידע האמיתי.
* מכאן, במידה ואנו מסופקים באחוזי הניבוי נוכל להשתמש במודל שלנו כדי לחזות נתונים אשר אינם קיימים.
* את הקבוצות שלנו נחלק 70% לאימון ו30% לבדיקה.
* בוצעו ניסיונות לקבוצות שונות אשר לא השפיעו על ביצועי האלגוריתם ואף פגעו ביכולותיו:
  + 25\75
  + 20\80
  + 40\60

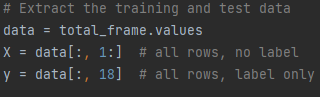
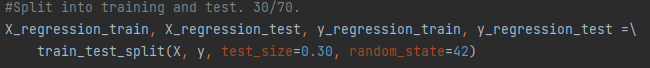
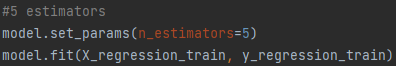
# אלגוריתמים

## יער אקראי – Random Forest

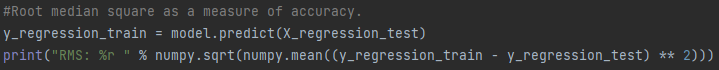
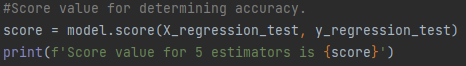
### תיאור

* יער אקראי, הינו אלגוריתם המשתייך לקבוצה של אלגוריתמים הנקראים "לומדים תחת השגחה" (Supervised).
* עצי ההחלטה אלו נוצרים לרוב על ידי דגימה מתוך המאפיינים או מתוך התצפיות. עצים אלו עובדים יחד ומכיוון שהתוצאה הסופית מושפעת מכל העצים יחד הדיוק הכולל גבוה יותר.
* אלגוריתם זה הינו כלי לקבלת החלטה. התוצר הינו גרף בצורת עץ המראה את התוצאות האפשריות.   
  על ידי הכנסת פיצ'רים ללמידה באלגוריתם נוכל "ללמד" אותו לקבל החלטה עבור שאלה מסוימת.
* זהו אחד האלגוריתמים השכיחים ביותר מעת פשטותו ויכולתו לתפקד כאלגוריתם קלסיפיקציה או רגרסיה – וזה בדיוק מה שנעשה.

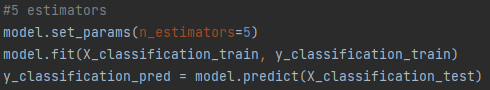
### מימוש בשיטה רגרסיבית

* בשיטה הרגרסיבית המידע מחולק מחדש כך שעמודת המטרה רציפה.
* הפרדת כל המשתנים מהטרגט (score) שלנו.
* נבצע אימון למודל.
* נאתחל את המודל.
* נאמן את האלגוריתם.

### תוצאות

* נסתכל על מדד הדיוק שלנו.   
  מדד הדיוק המתקבל הינו 0.9999 או במילים אחרות 99%.
* נסתכל על מדד הRMS (root median square) שהינו מדד השגיאה.   
  מדד הRMS הינו 1.427.
* אומנם הדיוק מעיד על 99% אך נשים לב שמדד השגיאה הריבועי שלנו עומד על 1.42.   
  אם נוציא שורש לשגיאה, נקבל שהינה 1.2~, או במילים אחרות האלגוריתם מפספס ב1.2 את המדד המתאים.  
  מאחר ומדובר בסיווג ל6 רמות שונות האלגוריתם עלול לפספס ביחידת סיווג אחת שלמה בתחזית שלו.
* מכאן נוכל להסיק שלא ניתן להסתמך על אלגוריתם יער אקראי בשיטת הרגרסיה.

### מימוש בשיטת הקלסיפיקציה

* יצירת המודל.
* אימון המודל.

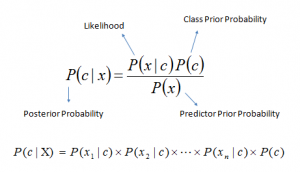
### תוצאות

* מדידת דיוק האלגוריתם.   
  דיוק האימון 0.99 או 99%. דיוק המבחן 0.71% או 71%. באופן כללי הדיוק הינו 71%.
* A picture containing graphical user interface

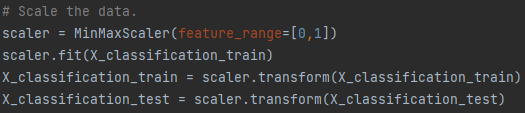
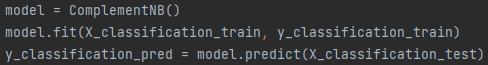
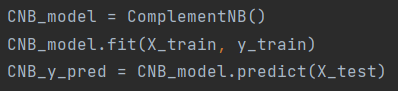
  Description automatically generatedמטריצת הבלבול.
* טבלאת קלסיפיקציה.
* קיבלנו שהדיוק עומד על 71%. מטבלת הקלסיפיקציה וממטריצת הבלבול ניתן לראות את חוסר האיזון בבסיס הנתונים. האלגוריתם מזהה בקלות את מדד 0 (recall = 0.99 ) אך יותר קשה לו לזהות את שאר המדדים.  
  בנוסף, ממדד f1 נוכל לראות שאכן הדיוק של המבחן עצמו איננו גבוה.

## באייס נאיבי משלים – Complement Naïve Bayes

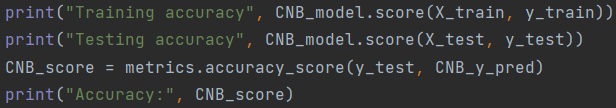
### תיאור

* אלגוריתם זה מסווג על יד שימוש במשפט Bayes תחת ההנחה שאין קשר בין הפרמטרים. (מכאן שמו "נאיבי"). המסווג הנאיבי של Bayes קל לשימוש כאשר ישנו מספר רב של מאפיינים. המסווג הנאיבי של Bayes יוצר מערך של הנחות פשטניות שסביר להניח שאינן נכונות לגמרי בפועל. עם זאת, נמצא שהמסווג הנאיבי של Bayes מועיל ביותר במצבים שונים ומגוונים. לדוגמא, המסווג הנאיבי של Bayes יעיל למדי בזיהו, דואר זבל כאשר תדירויות המילים משמשות בתור המאפיינים.
* נוסחת באייס:
* ישנם מספר סוגים של מימושים לאלגוריתם באייס הנאיבי המובדלים לפי המידע עליו הם מופעלים:
  + גאוסייני – עבור התפלגות רציפה של המשתנה הנמדד.
  + מולטינומיאלי – עבור התפלגות בדידה של המשתנה הנמדד.
  + משלים – באופן דומה למולטינומיאלי מדובר בהתפלגות בדידה של המשתנה הנמדד אך עם תוספת של חישוב סטטיסטי. משומש כדי להתמודד אל מול בעיות איזון במאגר המידע.   
    זהו הסוג בו אנו נשתמש.
  + ברנולי – עבור התפלגות ברנולי (בינארית של משתנים אשר ביניהם אין קשר).
  + קטגוריאלי – עבור התפלגויות קטגוריאליות של משתנים.

### מימוש

* נממש באמצעות אלגוריתם קלסיפיקציה באייס נאיבי במימוש משלים.
*  נרמול המידע למידע חיובי בין 0 ל 1.
* יצירת המודל ואימון האלגוריתם על המידע.

### תוצאות

* מדידת דיוק האלגוריתם.   
  דיוק האימון 0.56 או 56%. דיוק המבחן 0.56% או 56%. באופן כללי הדיוק הינו 56%.
* A picture containing table

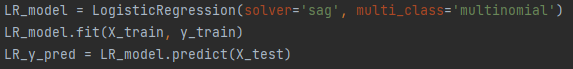
  Description automatically generatedמטריצת הבלבול.
* טבלאת קלסיפיקציה.
* קיבלנו שהדיוק עומד על 56%. מטבלאת הקלסיפיקציה וממטריצת הבלבול ניתן לראות את חוסר האיזון בבסיס הנתונים. באופן כללי ניתן לראות שהאלגוריתם באייס נאיבי במימוש משלים אינו אמין ולא מתאים לחיזוי מול הנתונים הללו.  
  אציין שאלגוריתם זה רץ הכי מהר באופן משמעותי מהיתר.

## רגרסיה לוגיסטית – Logistic Regression

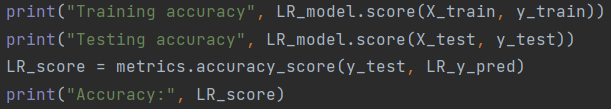
### תיאור

* רגרסיה לוגיסטית היא מודל סטטיסטי המתאר קשר אפשרי בין המשתנה אותו אנו בודקים לבין המשתנים האחרים. המודל מאפשר לאמוד את מידת ההשפעה של שינוי בערכו כל אחד מהמשתנים המסבירים על ערכו של המשתנה המוסבר. במילים אחרות, המודל מאפשר לאמוד מתאמים בין המשתנים.
* מודל הרגרסיה הלוגיסטית הוא מקרה פרטי של המודל הליניארי המוכלל.
* לרגרסיה לוגיסטית יש מספר יישומים אפשריים:
  + זיהוי קשר בין משתנים:   
    יחס הסיכויים הוא למעשה מקדם קשר בין משתנים. בעזרת רגרסיה לוגיסטית אפשר לזהות משתנים מסבירים שלהם יש קשר חזק עם המשתנה המוסבר.
  + סיווג וחיזוי: לאחר שאומדים את הפרמטרים של מודל הרגרסיה, ניתן לחשב בעזרתו את ההסתברות כי עבור פרמטר מסוים נוכל לחשב פרמטר אחר שיתאים לו.

### מימוש

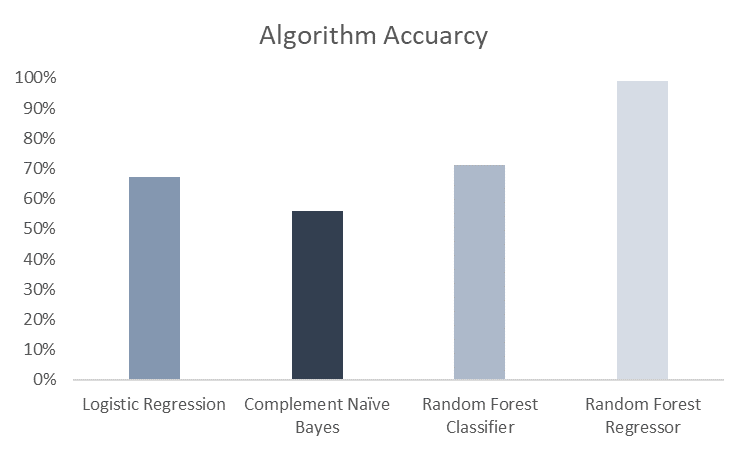
* יצירת המודל ואימון האלגוריתם על המידע. שתמש בדגל ‘sag’ על מנת להעפיל את האלגוריתם על מידע מולטינומיאלי – כלומר שיש יותר מרמת סיווג בינארי.

### תוצאות

* מדידת דיוק האלגוריתם.   
  הדיוק ההתקבל באימון הינו 0.67 או 67% ובמבחן 0.67 או 67%.  
  הדיוק הכללי עומד על 67%.
* Calendar

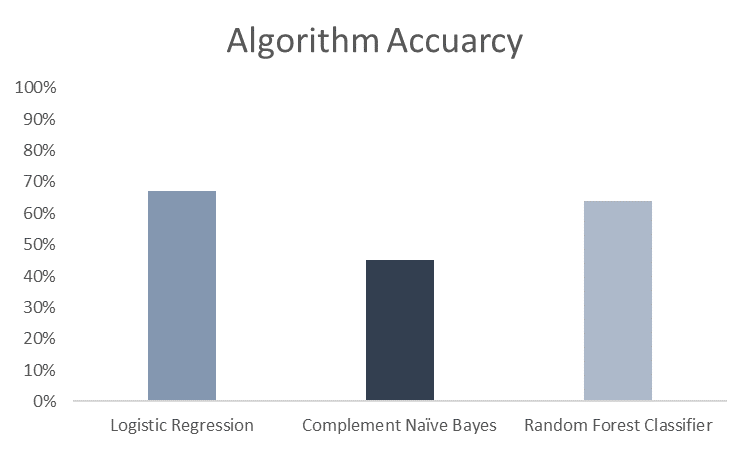
  Description automatically generatedמטריצת הבלבול.
* טבלאת קלסיפיקציה.
* קיבלנו שהדיוק עומד על 67%. מטבלאת הקלסיפיקציה וממטריצת הבלבול ניתן לראות את חוסר האיזון בבסיס הנתונים. האלגוריתם מזהה בקלות את מדד 0 (recall = 0.99 ) אך את יתר המדדים האלגוריתם אינו מזהה כלל. הדיוק מאוד נמוך.   
  לא ניתן לסמוך על תוצאות האלגוריתם הזה בחיזוי בצורת.

# מסקנות

* נסתכל על הדיוקים של האלגוריתמים:   
  כאמור, דיוק אלגוריתם יער אקראי רגרסיבי הינו אשליה מאחר והשגיאה שלו ברמת סיווג שלמה.
* בין שאר האלגוריתמים ניתן לראות שיער אקראי קלסיפיקציה בעל הסיכוי הגבוה ביותר לנחש את ערך מדד הבצורת.   
  את הדיוק של אלגוריתם זה גם כן כדאי לקחת בעירבון מוגבל. הדיוק גבוה מאוד בחיזוי מתי אין בצורת אך עבור סיווג רמות הבצורת השונות (1 עד 5) הדיוק נופל משמעותית לכ-20% בלבד.  
  תופעה דומה מתרחשת גם עבור שאר האלגוריתמים.
* בעקבות קורלציה נמוכה בין 17 פרמטרי המידע המטאורולוגי ו31 פרמטרי המידע היבשתי לבין פרמטר מדד הבצורת – לא ניתן להסתמך על אלגוריתמי למידה אלו על מנת לחזות בצורת צפויה. זוהי תוצאה הגיונית משתי סיבות:
  + הראשונה נובעת מפאת מחסור המידע ומחוסר האיזון המשמעותי בין מדדי הבצורת. יש הרבה יותר מדדים המצביעים על בצורת ברמה 0 מאשר על הבצורות החרות. כלומר המידע מוטה כלפי בצורת ברמה 0 ולכן לא ניתן להסיק לגבי בצורות אחרות.
  + השנייה נובעת מקטנותנו אל מול הטבע. לנסות לחזות תופעת טבע על ידי ניסיון קשירת מדדים הינו שאיפה גדולה ועל כן על מנת לצלוח בדבר זה יש להשקיע מאמצים כבירים ולאסוף מידע מדויק ורב ככל הניתן.

# ניסיון למיקוד המידע

* לאחר ביצוע מבחני הפירסון לבדיקת קורלציה נמצאו 6 פרמטרים בעלי קורלציה אומנם נמוכה אך כלשהי.
* פרמטרים אלו הם:
  + T2M\_RANGE – טווח טמפרטורה בגובה 2 מ'.
  + T2M\_MAX- טמפרטורה מקסימלית בגובה 2 מ'.
  + Elevation – גובה.
  + NVG\_LAND - שטחים מדבריים.
  + GRS\_LAND - שטחים עם שדות.
  + Graphical user interface, application, Teams

    Description automatically generatedCULTRF\_LAND - שטחים חקלאיים ללא השקיה.
* ננסה להריץ את האלגוריתמים שוב אך הפעם רק עם פרמטרים אלו בבדיקה ונשווה את התוצאות.  
  נתעלם מיער אקראי רגרסיבי מפאת כך שהשגיאה שלו אינה תורמת.
* נראה שצמצום הפרמטרים לא עזר ועדיין הטענה לפיה הקורלציה נמוכה מדי והיתכנות החיזוי של האלגוריתמים נמוכה מתקיימת.

# סיכום הסמינר

* הסמינר התמקד בתהליך כריית מידע וניסיון לענות על שאלה מחקרית שעולה.   
  השאלה שלי הייתה "האם ניתן לחזות מדד בצורת באמצעות נתונים מטאורולוגיים ויבשתיים?".   
  התשובה היא שלא. לא ניתן באמצעים האלו אל מול הנתונים האלו לחזות בצורת ולקבל נתון מדויק דיו עליו אפשר להסתמך ולבצע פעולות הכורכות במשאבי כסף, אנוש וזמן רבים.
* במהלך הסמינר למדתי על השלבים השונים והשיטות השונות הקיימות בעולם כריית המידע ומדעי המידע:
  + חיפוש ובחירת בסיס נתונים מלא ואיכותי.
  + ביצוע בדיקות על בסיס נתונים על מנת לאפיין אותו ולהסיק מסקנות לגביו.
  + ביצוע תיקונים, איזונים, מילוי חוסרים ופעולות נוספות על בסיס הנתונים על מנת להביאו למצב בו נוכל להפעיל עליו אלגוריתמים.
  + ביצוע מבחני התאמה ומציאת פרמטרים משמעותיים וקשרים ביניהם.
  + למידה ובחירה של אלגוריתמי למידה מבוקרת מסוגים שונים.
  + מימוש אלגוריתמים אלו באמצעות ספריות מדעי מידע.
  + פירוש התוצאות והסקת מסקנות.
* בסמינר שלי הגעתי למבוי סתום, הקורלציה בין המשתנים הייתה נמוכה והאלגוריתמים לא הניבו תוצאות משמעותיות מספיק כדי להישען עליהם לטובת הסקת מסקנות לגבי המידע.
* יתכן שהפרדת הפרמטרים והתייחסות אליהם בצורות שונות יניבו תוצאות טובות יותר לדוגמא, הפרדת מדד 0 משאר המדדים מכיוון שהינו גדול ומשמעותי מדי ומערער את איזונו של בסיס הנתונים.
* במהלך הסמינר הוכחתי לגלות שלא ניתן למדל כל מצב אמת בעולם לאלגוריתם ואכן האלגוריתמים האלו אינם קסם שיפתור את כל בעיותינו. בכל זאת, על ידי השקעת משאבים, איסוף ומיקוד מידע אני מאמין שנוכל להגיע לדיוק מאוד גבוה ואכן להסתמך על נתונים – ובעתיד, גם בשביל לחזות בצורת.

# ביבליוגרפיה

1. Predict Drought Using Weather and Soil Data, Chrsitoph Minixhofer, Kaggle  
   <https://www.kaggle.com/cdminix/us-drought-meteorological-data>
2. The United Stated Drought Monitor  
   <https://droughtmonitor.unl.edu/About/WhatistheUSDM.aspx>
3. Drought Information  
   <https://he.wikipedia.org/wiki/%D7%91%D7%A6%D7%95%D7%A8%D7%AA>  
   <http://iridl.ldeo.columbia.edu/maproom/Global/Drought/Global/CPC_GOB/Analysis.html>  
   https://www.unccd.int/actions/drought-initiative  
   <https://www.who.int/health-topics/drought#tab=tab_1>  
   <https://www.israelweather.co.il/news-weatherview.asp?id=21109>  
   <https://ecowiki.org.il/wiki/%D7%91%D7%A6%D7%95%D7%A8%D7%AA_%D7%91%D7%99%D7%A9%D7%A8%D7%90%D7%9C>  
   <https://he.wikipedia.org/wiki/%D7%91%D7%A6%D7%95%D7%A8%D7%AA>  
   https://experience.arcgis.com/experience/5dceec104a384df094e65af12a274959/
4. Random Forest Resources  
   <https://towardsdatascience.com/understanding-random-forest-58381e0602d2>  
   <https://status.co.il/%D7%90%D7%9C%D7%92%D7%95%D7%A8%D7%99%D7%AA%D7%9D-%D7%99%D7%A2%D7%A8-%D7%90%D7%A7%D7%A8%D7%90%D7%99-random-forest-%D7%95%D7%A9%D7%99%D7%98%D7%95%D7%AA-%D7%9E%D7%95%D7%A8%D7%9B%D7%91%D7%95%D7%AA-ense/>  
   <https://scikit-learn.org/stable/modules/ensemble.html>  
   https://dataaspirant.com/random-forest-algorithm-machine-learing/
5. Naibe Bayes Resources  
   <https://towardsdatascience.com/all-about-naive-bayes-8e13cef044cf>  
   https://status.co.il/%D7%94%D7%9E%D7%A1%D7%95%D7%95%D7%92-%D7%94%D7%A0%D7%90%D7%99%D7%91%D7%99-%D7%A9%D7%9C-%D7%91%D7%99%D7%99%D7%A1-naive-bayes-classifier/  
   <https://scikit-learn.org/stable/modules/naive_bayes.html#complement-naive-bayes>
6. Logistic Regression Resources  
   https://towardsdatascience.com/introduction-to-logistic-regression-66248243c148  
   https://status.co.il/%d7%9e%d7%95%d7%93%d7%9c%d7%99%d7%9d-%d7%9c%d7%99%d7%a0%d7%90%d7%a8%d7%99%d7%99%d7%9d-%d7%91%d7%91%d7%a2%d7%99%d7%95%d7%aa-%d7%a1%d7%99%d7%95%d7%95%d7%92-%d7%a8%d7%92%d7%a1%d7%99%d7%94-%d7%9c%d7%95/  
   <https://www.hamichlol.org.il/%D7%A8%D7%92%D7%A8%D7%A1%D7%99%D7%94_%D7%9C%D7%95%D7%92%D7%99%D7%A1%D7%98%D7%99%D7%AA>  
   https://scikit-learn.org/stable/modules/generated/sklearn.linear\_model.LogisticRegression.html#sklearn.linear\_model.LogisticRegression
7. Data Science Resources  
   https://machinelearningmastery.com/what-is-imbalanced-classification/