**אוניברסיטת בן גוריון בנגב**

**הפקולטה למדעי ההנדסה**

**המחלקה להנדסת תעשייה וניהול**

חיזוי ריכוזי נוטריינטים בצמחים באמצעות למידת מכונה וטכניקות ספקטרוסקופיות

מאת: דביר רחבי, לידור ארז וגיא מזרחי

מרצה: פרופ' בעז לרנר

תאריך הגשה: 1.1.2025

# תקציר

פרויקט זה עוסק בפיתוח מודל כמומטרי לחיזוי ריכוזי נוטריינטים חיוניים - חנקן, סוכר ועמילן - בצמחים, תוך שימוש בטכניקות מתקדמות של למידת מכונה וניתוח ספקטרלי. המודל יסתמך על מאגר נתונים רחב היקף, המכיל מדידות ספקטרליות (1557 אורכי גל) וריכוזי נוטריינטים שנאספו מעלים של הדרים.

במסגרת הפרויקט, נתמודד עם אתגר חיזוי בו זמני של שלושה משתני מטרה (Multi-Output Regression) ואתגר צמצום ממדי הנתונים (Dimensionality Reduction) הנובע מריבוי אורכי הגל. לשם כך, יבחנו אלגוריתמים שונים של למידת מכונה, ביניהם Random Forest, XGBoost ו PLS-R , וכן שיטות להקטנת הממד, כגון בחירת תכונות (Feature selection).

המודל הצפוי להתפתח במסגרת המחקר יהווה כלי יישומי משמעותי לחקלאות מדייקת, ויאפשר ניטור רמות נוטריינטים בצמחים בזמן אמת. היכולת לחזות את צרכי הדישון של גידולים שונים בדיוק רב תתרום לייעול תהליכי הדישון, תצמצם את ההשפעות הסביבתיות השליליות של דישון מופרז, ותתמוך בהגדלת יבולים חקלאיים.

**\*בסוף העבודה להוסיף תוצאות ומסקנות\***

**מילות מפתח**: חקלאות מדייקת, ניהול חנקן, ספקטרוסקופיה, למידת מכונה, Multi-Output Regression, Dimensionality Reduction, Random Forest, XGBoost, PLS-R, Feature selection.

# תוכן עניינים

[תקציר i](#_Toc184075427)

[תוכן עניינים ii](#_Toc184075428)

[רשימת סימנים וקיצורים iii](#_Toc184075429)

[1.מבוא והבנת התחום...................................................................................................................................1](#_Toc184075430)

[2.הבנת הבעיה............................................................................................................................................2](#_Toc184075431)

[3.הכנת הנתונים..........................................................................................................................................3](#_Toc184075432)

[4.מידול.....................................................................................................................................................4](#_Toc184075433)

[5.הערכה....................................................................................................................................................5](#_Toc184075434)

[6.סיכום, דיון ומסקנות..................................................................................................................................6](#_Toc184075435)

[מקורות......................................................................................................................................................7](#_Toc184075436)

# רשימת סימנים וקיצורים

NIR Near-Infrared

NIRS Near-Infrared Spectroscopy

PCA Principal Component Analysis

PLS-R Partial Least Squares Regression

RF Random Forest

RMSE Root Mean Squared Error

VIS-NIR-SWIR Visible-to-Shortwave Infrared

XGBoost Extreme Gradient Boosting

# מבוא והבנת התחום

חקלאות מדייקת (Precision Agriculture) מתבססת על איסוף וניתוח נתונים ברזולוציה גבוהה כדי לייעל את תהליכי הגידול ולהתאים את הטיפול בצמחים לצרכים הספציפיים שלהם (Mohamad, 2016). גישה זו חיונית כיום יותר מתמיד לאור הצורך להגביר את ייצור המזון העולמי תוך צמצום ההשפעות הסביבתיות השליליות של החקלאות (Wolfert et al., 2017). ניהול יעיל של חומרי דישון, ובפרט חנקן (Nitrogen Management) הינו אחד האתגרים המרכזיים בתחום החקלאות המדייקת. חנקן הינו יסוד חיוני לצמיחה והתפתחות תקינה של צמחים, המשפיע על תהליכים מרכזיים כמו פוטוסינתזה, סינתזת חלבונים, וביולוגיה של השורשים (Lawlor, 2001). לכן, קיים צורך בפיתוח שיטות אמינות ונגישות לניטור רמות נוטריינטים בצמחים בזמן אמת, כדי לאפשר לחקלאים לקבל החלטות מושכלות לגבי דישון ולמנוע בזבוז משאבים ונזקים סביבתיים. יחד עם זאת, דישון מופרז עלול להוביל למגוון השפעות שליליות (Vitousek et al., 2009).

דישון יתר גורר עימו שורה של השפעות שליליות, ביניהן זיהום מי תהום ופליטת גזי חממה (Erisman et al., 2013). בפרט, פליטת תחמוצת חנקן (N2O) מאדמות חקלאיות מהווה מקור משמעותי לגז חממה זה, אשר חזק פי 300 מפחמן דו-חמצני ביכולתו לפגוע בשכבת האוזון (Davidson & Kanter, 2014). עודף חנקן בקרקע עלול להוביל גם לצמיחה וגטטיבית מוגזמת על חשבון התפתחות פירות, להגביר את רגישות הצמחים למחלות ומזיקים, ולפגוע בטעם ובאיכות התוצרת (Albornoz, 2016). לכן, ניהול מושכל של דישון חנקני מחייב מעקב רציף אחר רמות הנוטריינטים בצמחים, כדי להתאים את כמויות הדשן לצרכים האקטואליים של הגידול ולמנוע בזבוז משאבים ופגיעה בסביבה.

שיטות ההערכה המסורתיות לניטור רמות נוטריינטים בצמחים, המבוססות על דגימת עלים ובדיקות מעבדה, הינן תהליכים יקרים, גוזלים זמן, ומוטים לדגימה נקודתית בזמן. כתוצאה מכך, הן אינן מספקות תמונה דינמית ומדויקת של מצב ההזנה בצמח ולעיתים קרובות אינן יעילות בזיהוי מצבי עודף (Araújo et al., 2023). ספקטרוסקופיה (Spectroscopy), ובפרט ספקטרוסקופיית (Near-Infrared Spectroscopy) NIR ובפרט ספקטרוסקופיית VIS-NIR-SWIR (Visible-to-Shortwave Infrared Spectroscopy) (400-2500 ננומטר) מציעות חלופה מבטיחה לשיטות הקונבנציונליות, המאפשרת ניטור לא פולשני של הצמח בזמן אמת (Osborne et al., 1993) המבוססות על דגימת עלים ובדיקות מעבדה. שיטות אלו מבוססות על הקרנת אור על הצמח ומדידת האור המוחזר/מועבר, תוך ניתוח "החתימה" הספקטרלית הייחודית לכל תרכובת. היתרון המשמעותי של טכנולוגיות אלו הוא ביכולתן לספק מידע רב בזמן קצר ובעלות נמוכה יחסית, ובכך לאפשר קבלת החלטות מהירה ומושכלת יותר לגבי ניהול הדישון. יחד עם זאת, הנתונים הספקטרליים הם מורכבים ומצריכים ניתוח מתקדם כדי לחלץ מהם את המידע הרלוונטי.

\*לחשוב על לנפח\* כדי לתרגם את הנתונים הספקטרליים המורכבים למידע מעשי הנגיש לחקלאי, נעשה שימוש בכלי ניתוח מתקדמים מעולם למידת מכונה (Machine Learning). הכלים הללו מאפשרים לפתח מודלים כמומטרים (Chemometric Models) המסוגלים לחזות את ריכוזי הנוטריינטים ברקמות הצמח. אלגוריתמים כגון Random Forest (Chen & Guestrin, 2016), XGBoost (Breiman, 2001), ו- R-PLS מסוגלים ללמוד את הקשר המורכב בין "החתימה" הספקטרלית לבין ריכוזי הנוטריינטים בצמח. האלגוריתמים נמצאו כיעילים במיוחד במשימות סיווג ורגרסיה בתחום החישה מרחוק (Sensing Remote) וחקלאות מדייקת.

# הבנת הבעיה

מטרת מחקר זה היא לרתום את היתרונות של הספקטרוסקופיה ולמידת המכונה לשם פיתוח מודל כימומטרי אשר יאפשר לחזות באופן מדויק את ריכוזי חנקן, סוכר ועמילן בצמחים. המודל יסתמך על מאגר נתונים גדול הכולל מדידות ספקטרליות (1557 אורכי גל) ומדידות מעבדה של ריכוזי הנוטריינטים בצמחים שונים. היכולת לחזות באופן מדויק שלושה משתני מטרה בו זמנית חשובה במיוחד, שכן קיים קשר הדוק בין ריכוז החנקן לבין תהליכי ייצור ואגירת האנרגיה בצמח. חוסר איזון בין המרכיבים הללו עלול להוביל לפגיעה בצמיחה, לירידה באיכות הפרי ולבזבוז משאבים.

אחד האתגרים המרכזיים בפיתוח המודל הוא הצורך בהתמודדות עם ״קללת המימד״ (Curse of Dimensionality). הנתונים הספקטרליים מתאפיינים בממדיות גבוהה (מספר גדול של אורכי גל המהווים תכונות), אשר עלולה להוביל למספר קשיים בפיתוח המודל. ראשית, ממדיות גבוהה מגדילה את מורכבות המודל ואת הזמן הדרוש לאימון (Training) שלו. שנית, כמות גדולה של תכונות עלולה להוביל להתאמת יתר (Overfitting), מצב בו המודל לומד את נתוני האימון בדיוק רב מדי כולל את הרעש, ולכן מתקשה להכליל (Generalize) לנתונים חדשים. ממדיות גבוהה מקשה על הבנת חשיבותם היחסית של המשתנים המסבירים (Feature Importance).

לשם התמודדות עם אתגר זה, נעשה במחקר זה שימוש בטכניקות להקטנת ממדיות הנתונים (Dimensionality Reduction). שיטות אלו נועדו לצמצם את מספר התכונות המשמשות את המודל, תוך שמירה על כמות המידע המרבית האצורה בנתונים המקוריים. שיטה נפוצה להקטנת ממדיות היא בחירת תכונות (Feature Selection). שיטה זו מאפשרת לצמצם את מורכבות המודל, לשפר את יכולת ההכללה שלו לנתונים חדשים, ולקצר את זמן האימון - נראה לי שלא צריך את החלק הזה.

המודל הכימומטרי שיפותח במחקר זה צפוי להוות כלי עזר משמעותי לניהול יעיל ויותר של דישון בחקלאות מדייקת, ולתרום לתחומים חשובים כמו קיימות סביבתית וביטחון תזונתי.

אולי להוסיף פה על בדיקת המשתנים התלויים – להגיד שאנחנו בבעיית MultiOutput Regression ולהצדיק למה השתמשנו ב Multi Output עם איזה גרף בנספחים.

# הכנת הנתונים

שלב הכנת הנתונים הינו קריטי להצלחת פרויקט למידת מכונה בכלל, ובפרט בתחום הכימומטריה והחקלאות המדייקת. מטרת שלב זה היא להפוך את הנתונים הגולמיים, במקרה שלנו נתונים ספקטרליים ומדידות מעבדה של ריכוזי נוטריינטים, לפורמט המתאים לאימון והערכת מודל כימומטרי יעיל ומדויק. תהליך הכנת הנתונים במחקר זה יכלול מספר שלבים מרכזיים:

1. ניקוי נתונים (Data Cleaning):

**טיפול בערכים חסרים:**

אני חושב שצריך לנסח משהו בסגנון:

בפרויקט שלנו היו ככה וככה ערכים חסרים... ככה התמודדנו איתם... (חציון / ממוצע)...

במאגרי נתונים גדולים נפוץ למצוא ערכים חסרים, בין אם בגלל שגיאות מדידה, בעיות בהעברת המידע, או גורמים אחרים. במקרה של ערכים חסרים, נוכל לבחור מתוך מגוון שיטות טיפול, בהתאם לכמות הערכים החסרים ולסוג הנתונים. אפשרויות נפוצות הן: השמטת רשומות (Rows) או עמודות (Columns) עם ערכים חסרים, מילוי הערכים החסרים בממוצע (Mean) או בחציון (Median) של העמודה לטעמי זה מיותר.. צריך להסביר מה עשינו ואיך הכנו את הנתונים ופחות לספר למה צריך לטפל ואיך... .

**טיפול בערכים חריגים:**   
בזה לא נגענו בכלל.. אולי צריך? אולי לא? לי הגיוני שלא חייב להכניס ..

ערכים חריגים הם ערכים החורגים באופן משמעותי משאר הערכים במאגר הנתונים. ערכים אלו עלולים להיות תוצאה של שגיאות מדידה או גורמים אחרים, והם עלולים להטות את המודל ולפגוע בדיוקו. קיימות שיטות שונות לזיהוי וטיפול בערכים חריגים, כגון: שימוש בכלל האצבע הסטטיסטי (למשל, ערכים הנמצאים מעל או מתחת לשלוש סטיות תקן מהממוצע), ניתוח גרפי (למשל, באמצעות דיאגרמת קופסה - Boxplot), או שימוש באלגוריתמים לזיהוי אנומליות (Anomaly Detection).

2. טרנספורמציה והנדסת תכונות (Feature Engineering and Transformation):

טרנספורמציות מתמטיות: לעיתים קרובות נבצע טרנספורמציות מתמטיות על הנתונים כדי לשפר את התאמתם למודל הכימומטרי. לדוגמה, נוכל לבצע טרנספורמציה לוגריתמית (Logarithmic Transformation) כדי להתמודד עם התפלגות נתונים א-סימטרית (Skewed Distribution) או טרנספורמציית שורש ריבועי (Square Root Transformation) כדי להקטין את השפעתם של ערכים גדולים.

הנדסת תכונות: שלב זה כולל יצירת תכונות (משתנים) חדשות מתוך התכונות הקיימות במאגר הנתונים. לדוגמה, נוכל לחשב יחסים בין שני אורכי גל שונים (Spectral Indices), לבצע פעולות מתמטיות על מספר אורכי גל (למשל, ממוצע, סטיית תקן), או ליצור משתנים דמה (Dummy Variables) המבוססים על ערכים קטגוריאליים. הנדסת תכונות טובה יכולה לשפר באופן משמעותי את ביצועי המודל על ידי הדגשת הקשרים הנסתרים בנתונים.

לנסח את החלק הזה בצורה אחרת לגמרי ופשוט להסביר שגיא הפך את אורכי הגל מרציפים לבדידים כדי שנוכל למדל את זה.

3. בחירה והקטנת ממדים (Feature Selection and Dimensionality Reduction):

כאן אפשר להכניס את זה שעשינו PLS כדי להוריד את מימד הנתונים.. לגבי Feature Selection לא בדיוק עשינו ואני חושב שלא צריך לעשות..

בחירת תכונות (Feature Selection): במאגרי נתונים ספקטרליים, לעתים קרובות נדרשת בחירת תכונות כדי לצמצם את מספר אורכי הגל המשמשים את המודל. שיטות נפוצות לשם כך הן: בחירה בהתבסס על מתאם (Correlation) עם משתנה המטרה, בחירה בהתבסס על חשיבות המשתנה (Feature Importance) המתקבלת ממודלים מבוססי עצים (Tree-Based Models), ובחירה היברידית המשלבת מספר שיטות.

הקטנת ממד (Dimensionality Reduction): שיטות אלו נועדו לצמצם את ממד הנתונים תוך שמירה על כמות המידע המרבית. שיטה נפוצה להקטנת ממד היא ניתוח מרכיבים עיקריים (PCA - Principal Component Analysis), המאפשרת לייצג את הנתונים בממד נמוך יותר באמצעות צירופים לינאריים חדשים של המשתנים המקוריים.

4. נורמליזציה וסטנדרטיזציה (Normalization and Standardization):  
  
כאן אפשר לפרט שזה בוצע נטו לפני ה PLSR ולא לפני ה XGBOOST ו – RANDOM FOREST כי אותם זה לא מעניין אם זה מנורמל או לא..

נורמליזציה (Normalization): תהליך זה מביא את כל המשתנים לטווח ערכים משותף, בדרך כלל בין 0 ל-1. דבר זה חשוב במיוחד כאשר המשתנים במאגר הנתונים נמדדים ביחידות שונות.

סטנדרטיזציה (Standardization): תהליך זה מביא את כל המשתנים לאותו סדר גודל על ידי הפיכתם לבעלי ממוצע 0 וסטיית תקן 1. דבר זה חשוב במיוחד עבור אלגוריתמים רגישים להבדלים בסדרי גודל בין המשתנים.

שלב הכנת הנתונים הינו תהליך איטרטיבי הדורש ניתוח מעמיק של הנתונים והתאמה לבעיה הספציפית שברצוננו לפתור. בחירה מושכלת של שיטות הכנת הנתונים היא קריטית כדי לבנות מודל כימומטרי מדויק, אמין ורלוונטי לצורך חיזוי יעיל של רמות נוטריינטים בצמחים.

# מידול

בשלב המידול בחרנו לעשות שימוש באלגוריתמים הבאים: Partial Least Squares Regression, XGBoost, Random Forest בשל שימושם הרב בפתרון בעיות מסוג זה כפי שראינו בחלק של סקירת הספרות. הדרך שבה בחרנו לאמן כל מודל היא כדלקמן:

תחילה, חיפשנו את סט הקונפיגורציה לכל מודל אשר ימזער את שורש השגיאה הריבועית הממוצעת (RMSE) שכן אנו פותרים בעיית רגרסיה. כלומר, כל מודל אומן על סט האימון כמה פעמים כך שבכל פעם הוא עם סט היפר פרמטרים שונה ולאחר האימון נבדקו ביצועיו על סט הולידציה וחושבה השגיאה הריבועית הממוצעת שלו. מכיוון שאנו פותרים בעיית רגרסיה שבה אין פלט אחד, אלא שלושה, נעשה שימוש ב RMSE הממוצע על פני שלושת המשתנים התלויים שלנו. לאחר מציאת ההיפר פרמטרים אשר ממזערים את השגיאה הריבועית הממוצעת אומן המודל הסופי על סט האימון בעזרת CV-10. לבסוף, נבדקו ביצועי המודלים השונים על סט הבדיקה. חשוב לציין כי סט האימון, הבדיקה והוולידציה חולקו עוד לפני שלב המידול על מנת שיהיו אותו הדבר ונוכל להשתמש בהם בכדי להשוות בין ביצועי המודלים השונים – אולי לפרט על זה בחלק של הכנת הנתונים?. כעת, נפרט על סט הקונפיגורציות של כל מודל אשר נבדק במהלך האופטימיזציה:

**צריך להרחיב על המודלים**

1. PLSR – היפר פרמטר בודד (n\_components) אשר מייצג את מספר הקומפוננטים שייצגו את הנתונים לאחר שלב הורדת המימד. נבדקו 50 ערכים שונים למספר הקומפוננטים.
2. XGBoost –
3. Random Forest –

את ביצועיי המודלים Random Forest ו – XGBoost אימנו ובחנו גם על המשתנים שהוציא לנו מודל ה PLSR הטוב ביותר (הקונפיגורציה שממזערת את הRMSE).

# הערכה

להכניס את התוצאות של המודלים בשלב הבדיקה - Multi PLSR, Multi XGB, Multi RF גרפים בנספחים + לדווח עליהם.

להציג אולי את התוצאות של ה CV10?

להציג את התוצאות של ההיפר פרמטר טיונינג?

# סיכום, דיון ומסקנות

זה די פשוט ברגע שיש לנו הכל... לדבר קצת על התוצאות מה יצא לנו... איזה מודל היה הכי טוב מבחינת ביצועים.. מה זה יכול לתת התוצאות האלה ומה אנחנו חושבים שצריך להיות המשך למחקר... (אפשר לדחוף את זה שאולי צריך לנסות למדל את זה כסדרת זמן בעזרת LSTM שזה בגדול מה שגיא עושה בתזה שלו).

# מקורות

Albornoz, F. (2016). Crop responses to nitrogen overfertilization: A review. In *Scientia Horticulturae* (Vol. 205, pp. 79–83). Elsevier B.V. <https://doi.org/10.1016/j.scienta.2016.04.026>

Araújo, S. O., Peres, R. S., Ramalho, J. C., Lidon, F., & Barata, J. (2023). Machine Learning Applications in Agriculture: Current Trends, Challenges, and Future Perspectives. In *Agronomy* (Vol. 13, Issue 12). Multidisciplinary Digital Publishing Institute (MDPI). <https://doi.org/10.3390/agronomy13122976>

Breiman, L. (2001). *Random Forests* (Vol. 45).

Chen, T., & Guestrin, C. (2016). XGBoost: A scalable tree boosting system. *Proceedings of the ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, *13-17-August-2016*, 785–794. <https://doi.org/10.1145/2939672.2939785>

Davidson, E. A., & Kanter, D. (2014). Inventories and scenarios of nitrous oxide emissions. *Environmental Research Letters*, *9*(10). <https://doi.org/10.1088/1748-9326/9/10/105012>

Erisman, J. W., Galloway, J. N., Seitzinger, S., Bleeker, A., Dise, N. B., Roxana Petrescu, A. M., Leach, A. M., & de Vries, W. (2013). Consequences of human modification of the global nitrogen cycle. *Philosophical Transactions of the Royal Society B: Biological Sciences*, *368*(1621). <https://doi.org/10.1098/rstb.2013.0116>

Lawlor, D. W. (2001). *Carbon and nitrogen assimilation in relation to yield: mechanisms are the key to understanding production systems*. <https://academic.oup.com/jxb/article/53/370/773/2908378>

Mohamad, B. (2016). *Variable rate application of fertilizer in rice precision farming*. <https://www.researchgate.net/publication/332060576>

Osborne, B. G., Fearn, T., & Hindle, P. H. (1993). *Practical NIR Spectroscopy with Applications in Food and Beverage Analysis. Longman Scientific and Technical, Harlow.*

Vitousek, P. M., Naylor, R., Crews, T., David, M. B., Drinkwater, L. E., Holland, E., Johnes, P. J., Katzenberger, J., Martinelli, L. A., Matson, P. A., Nziguheba, G., Ojima, D., Palm, C. A., Robertson, G. P., Sanchez, P. A., Townsend, A. R., & Zhang, F. S. (2009). Nutrient imbalances in agricultural development. In *Science* (Vol. 324, Issue 5934, pp. 1519–1520). <https://doi.org/10.1126/science.1170261>

Wolfert, S., Ge, L., Verdouw, C., & Bogaardt, M. J. (2017). Big Data in Smart Farming – A review. In *Agricultural Systems* (Vol. 153, pp. 69–80). Elsevier Ltd. <https://doi.org/10.1016/j.agsy.2017.01.023>