אוניברסיטת בן גוריון בנגב

הפקולטה למדעי ההנדסה

המחלקה להנדסת תעשייה וניהול

חיזוי ריכוזי נוטריינטים בצמחים באמצעות למידת מכונה וטכניקות ספקטרוסקופיות

מאת:

דביר רחבי 207206624

לידור ארז 318661444

גיא מזרחי 314975442

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, התמודדנו 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. התוצאות הצביעו על כך ש XGBoost השיג את הדיוק הטוב ביותר עם RMSE של 0.046 על סט הבדיקה ובנוסף שיטת PLS להורדת המימד שיפרה את ביצועי המודלים האחרים.

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.

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
תוכן עניינים ii
רשימת סימנים וקיצורים iii
1.מבוא והבנת התחום..... 1
2.הבנת הבעיה..... 2
3.הכנת הנתונים..... 2
4.מידול..... 3
5.הערכה..... 6
6.סיכום, דיון ומסקנות..... 8
מקורות..... 10

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
MOR	Multi Output Regressor

1. 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 NIR (Near-Infrared Spectroscopy) 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.

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- PLS -R מסוגלים ללמוד את הקשר המורכב בין "החתימה" הספקטרלית לבין ריכוזי הנוטריינטים בצמח. האלגוריתמים נמצאו כיעילים במיוחד במשימות סיווג ורגרסיה בתחום החישה מרחוק (Remote Sensing) וחקלאות מדייקת.

2. הבנת הבעיה

מטרת מחקר זה היא לרתום את היתרונות של הספקטרוסקופיה ולמידת המכונה לשם פיתוח מודל כמומטרי 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 בטכניקה להקטנת ממדיות הנתונים (Dimensionality Reduction). אשר נועדה לצמצם את מספר המשתנים המשמשים את המודל בתהליך האימון, תוך שמירה על כמות המידע המרבית האצורה בנתונים המקוריים. בנוסף לאתגר זה, קיים אתגר נוסף שבו אנו מנסים לחזות שלושה משתנים תלויים במקביל מקרה שבו יש להבין האם אנו נדרשים לבנות מודל נפרד לכל משתנה (כזה אשר ילמד לחזות רק משתנה אחד) או מודל אחד אשר מנבא את שלושתם. לצורך התמודדות עם אתגר נבצע ניתוח קשרים בין המשתנים התלויים בכדי להבין האם קיים קשר כלשהו אשר יכול להצדיק את השימוש במודל מסוג MOR.

לסיכום, המודל הכמומטרי אשר יפותח במחקר זה צפוי להוות כלי עזר משמעותי לניהול יעיל ויותר של דישון מדויק יותר בחקלאות, ולתרום לתחומים חשובים כמו קיימות סביבתית וביטחון תזונתי.

3. הכנת הנתונים

שלב הכנת הנתונים מהווה את הבסיס להצלחת כל פרויקט בתחום למידת המכונה, ובמיוחד בתחומים הדורשים עיבוד מדויק של נתונים מדעיים, כגון כמומטריה וחקלאות מדייקת. תהליך זה הוא קריטי משום שהוא מאפשר לקחת את הנתונים הגולמיים, אשר לעיתים קרובות לא ישימים ישירות לאימון מערכות לומדות, ולהפוך אותם לפורמט אשר יאפשר למערכת הלומדת ללמוד בצורה מיטבית ולהפיק תובנות מהימנות ככל הניתן. במקרה הנוכחי, מדובר בנתונים ספקטרליים מדויקים ומדידות מעבדה של ריכוזי נוטריינטים, אשר יש לעבד ולארגן אותם על מנת להפיק מהם מידע משמעותי ומועיל. לצורך הכנתם של הנתונים בחנו את כמות הערכים החסרים, טיפלנו בערכים חריגים, ואף ביצענו הורדת מימד וסטנדרטיזציה במידת הצורך. להלן, פירוט מפורט על כל מה שעשינו במהלך שלב זה:

1. ניקוי הנתונים (Data Cleaning):

במאגרי נתונים גדולים נפוץ להיתקל בערכים חסרים, הנובעים משגיאות מדידה, בעיות באיסוף הנתונים, או שיבושים בתהליך העברת המידע. ערכים חסרים עלולים לפגוע בביצועי המודלים ולגרום להטיה בתוצאות הניתוח, ולכן חשוב לטפל בהם בצורה שיטתית. בפרויקט שלנו, זיהינו מספר ערכים חסרים במספר משתנים, ובמקום למלא אותם באמצעות חישובים כמו ממוצע או חציון, החלטנו למחוק את הרשומות שבהן נמצאו ערכים חסרים. החלטה זו התקבלה מכיוון שמספר הערכים החסרים היה נמוך יחסית ולא השפיע מהותית על מאגר הנתונים הכולל. כך, יכולנו לשמור על פשטות הניתוח ולהימנע מהכנסת הנחות אשר עלולות להטות את תוצאות המודלים.

2. טיפל בערכים חריגים (Outliers):

ערכים חריגים הם נתונים שחורגים משמעותית מהתפלגות הערכים הכללית, ולעיתים נובעים משגיאות מדידה, תקלות באיסוף הנתונים, או אירועים יוצאי דופן. ערכים אלו עלולים לפגוע בדיוק התחזיות ולגרום להטיה בתוצאות הניתוח, במיוחד במקרים שבהם החישובים תלויים בהתפלגות הנתונים. במהלך תהליך ניתוח הנתונים, זיהינו ערכים חריגים, ובפרט ערכים שליליים במשתנים שאמורים להיות חיוביים בלבד. כדי לטפל בבעיה מבלי להסיר את הערכים הללו ולפגוע בשלמות מאגר

הנתונים, החלטנו להחליף את הערכים החריגים בחציון של אותו משתנה. גישה זו שמרה על עקביות הנתונים והפחיתה את השפעתם של הערכים החריגים, תוך שמירה על מבנה הנתונים המקורי ככל האפשר.

3. הקטנת מימד הנתונים (Dimensionality Reduction):
בשל הקורלציה הגבוהה בין המשתנים הבלתי תלויים, ובשל היחס הלא מאוזן בין מספר המשתנים למספר התצפיות, זיהינו את הצורך לבצע הורדת ממד לנתונים כדי לשפר את ביצועי המודלים. לצורך הורדת המימד, בחרנו להשתמש בשיטת PLS שהינה שיטה יעילה להורדת ממד המשלבת שמירה על הקשר בין המשתנים הבלתי תלויים למשתנים התלויים. שיטה זו נבדלת מטכניקות אחרות בכך שהיא ממקסמת את השונות המשותפת בין המשתנים התלויים לבלתי תלויים ולא מתמקדת רק בשונות הפנימית של המשתנים הבלתי תלויים, כפי שנעשה ב-PCA. במסגרת התהליך, PLS יצרה משתנים חדשים שהינם שילובים ליניאריים של המשתנים המקוריים, תוך שמירה על מידע חיוני שמסייע לניבוי המשתנים התלויים. לאחר הורדת הממד, נבנו סטים חדשים של נתונים (אימון, בדיקה ווליד‌ציה), שכללו את המשתנים החדשים שיצרה השיטה. תהליך זה חשוב לפני שלב המידול מכיוון שיכול להפחית את הסיכון ל-Overfitting אשר נובע משימוש ביותר מדי משתנים באופן יחסי למספר התצפיות.

4. פיצול הנתונים לסטים של אימון, ולידציה ובדיקה:
לצורך הכנת הנתונים לקראת שלב המידול ולאחר כל שלבי ההכנה המקדמים כגון, מחיקת ערכים חסרים והפיכת ערכים שליליים לממוצע המשתנה פיצלנו את הנתונים שלנו לשלושה סטים אשר שימשו את שלושת המודלים בהם השתמשנו לצורך אימון, ולידציה ובדיקה ובכך דאגנו לשמור על עקביות במהלך כל שלב. כתוצאה מכך שלושת המודלים התאמנו ונבדקו על אותם הנתונים.

5. נורמליזציה וסטנדרטיזציה (Standardization) :
לפני השימוש בשיטת PLS להורדת ממד, ביצענו תהליך של נורמליזציה וסטנדרטיזציה על הנתונים, שכן שיטה זו רגישה מאוד להיקפים (Scales) ולערכים הקיצוניים של המשתנים הבלתי תלויים. תהליך זה הבטיח שכל המשתנים יהיו מיוצגים באותה סקאלה, מה שמנע מצב שבו משתנים בעלי ערכים גדולים יותר משפיעים בצורה לא פרופורציונלית על יצירת המשתנים החדשים של PLS. עם זאת, עבור אלגוריתמים אחרים כמו XGBoost ו-Random Forest, לא נדרשנו לבצע נורמליזציה או סטנדרטיזציה, מכיוון שאלגוריתמים אלו אינם רגישים לטווח הערכים של הנתונים. לכן, בוצעה סטנדרטיזציה אשר זה הוא תהליך המביא את כל המשתנים לסדר גודל דומה (לדבר באותה שפה) על יד הפיכתם לבעלי תוחלת השווה לאפס ושונות השווה לאחד רק לפני השימוש בPLS בכדי שלא לפגוע בביצועי השיטה.

4. מידול

בשלב המידול בחרנו לעשות שימוש בכמה אלגוריתמי רגרסיה אשר נמצאו כיעילים לפתרון בעיות דומות, כפי שצוין בחלק של סקירת הספרות. האלגוריתמים שבחרנו הינם:RF,PLS-R ו- XGBoost. הבחירה באלגוריתמים אלו נבעה מהיתרונות הטמונים בכל אחד מהם PLS - מצטיין בהורדת ממד ושימור קשרים בין משתנים XGBoost ,מבדל את עצמו ביכולת להתמודד עם נתונים לא ליניאריים וייצוג קשרים מורכבים בכך שמוריד את הטיה על ידי שימוש בעצי החלטה כך שכל אחד מהם מאומן על הטעויות של אלו שאומנו לפניו. ו - Random Forest אשר מצוין בהפחתת הסיכון ל Overfitting בכך שמוריד את השונות הכוללת של עצי החלטה שונים. השלב הראשון בתהליך המידול כלל חיפוש אחר סט הקונפיגורציה (היפר פרמטרים) של כל מודל, אשר ימזער את שורש השגיאה הריבועית הממוצעת (RMSE). מאחר ואנו עוסקים בבעיה שבה ישנם שלושה משתנים תלויים, חישבנו את ה RMSE-הממוצע על פני שלושתם. כל מודל אומן על סט האימון במספר מחזורי אימון, כאשר בכל פעם שונו ההיפר פרמטרים ונבדקו ביצועיו על סט הווליד‌ציה בכל מחזור. לבסוף, נבחר סט הקונפיגורציה אשר הוביל לממוצע ה RMSE הקטן ביותר על סט הווליד‌ציה ועל פני שלושת המשתנים התלויים שלנו. לאחר תהליך החיפוש אחר סט הקונפיגורציה הטוב ביותר, כל מודל אומן פעם נוספת על סט האימון באמצעות חיבור של טכניקת Cross-Validation בעשרה קיפולים (CV-10) בכדי לאמוד את ביצועי המודלים על נתונים שמעולם לא ראו. בשלב האחרון, ביצענו הערכה של ביצועי המודלים על סט הבדיקה, כדי להעריך את יכולת המודל להתמודד עם נתונים חדשים שלא נראו באף אחד מהשלבים המצוינים לעיל ובכדי לבדוק את יכולת ההכללה שלו. לאחר הניבוי חושב מדד ה RMSE בכדי לאמוד את ביצועי המודלים על סט הבדיקה. את המודלים XGBoost ו – RF אימנו על סט הנתונים הרגיל

וגם על סט הנתונים שיצרנו בעזרת PLS בכדי להקטין את זמן האימון של XGBoost ו – RF ולבדוק האם שיטה זו מאפשרת לשפר את ביצועי XGBoost ו- RF באופן כללי.

PLSR – Partial Least Squares Regression

PLS הינה שיטה להורדת מימד אשר משתמשים בה במקרים בהם קיים הבדל גדול בין מספר התצפיות למספר העמודות ו/או כאשר קיימת קורלציה גבוהה בין המשתנים הבלתי תלויים ונדרשת שמירה על התלות בין X ל Y. מטרתה המרכזית של שיטת ה PLS היא להוריד את המימד של הנתונים תוך שמירה על כמה שיותר מידע שימושי לצורך ניבוי המשתנים התלויים (Y). כלומר, הייחודיות של PLS-R היא בכך שהוא ממקסם את השונות המשותפת בין X ל- Y (Cov(X,Y)) ושומר על הקשר ביניהם. לעומת שיטות אחרות כמו PCA אשר ממקסמות את השונות בתוך המשתנים הבלתי תלויים (X) מבלי להתייחס למשתנים התלויים. במסגרת השימוש במודל זה נעשה שימוש בהיפר-פרמטר היחיד שלו השולט על **מספר הקומפוננטים (n_components)** פרמטר זה מייצג את מספר ממדי המידע החדשים שייווצרו לאחר שלב הורדת הממד. במהלך שלב בחירה הקונפיגורציה הטובה ביותר נבדקו 50 ערכים שונים עבור מספר הקומפוננטים, תוך שימוש בסט אימון ובסט הווליד‌ציה. המטרה הייתה למצוא את מספר הקומפוננטים אשר שומר על אחוז גבוה של שונות הנתונים לאחר שלב הורדת הממד על ידי שימוש בגרף מרפק (Elbow Plot). תהליך זה מאפשר למודל לא רק להוריד ממד אלא גם לשמור על ביצועים אופטימליים בניבוי ואף למנוע התאמת יתר. לבסוף, נבדקו ביצועי המודל על סט הבדיקה לצורך השוואה עם שאר המודלים.

(eXtreme Gradient Boosting) – XGBoost

אלגוריתם XGBoost משלב מספר רב של לומדים חלשים (עצי החלטה) ליצירת מודל חיזוי חזק תוך מזעור של פונקציית ההפסד הנבחרת (במקרה שלנו RMSE).

המודל מחושב על פי המשוואה הבאה:

$$\hat{y}_{\mathrm{i}} = \sum_{k=1}^{K} f_k(x_i) \quad , \quad f_k \in \mathcal{F}$$

כך ש – $\hat{y}_{\mathrm{i}}$: התוצאה החזויה עבור דגימה i, x_i: מערך התכונות עבור דגימה i ,

f_k: עץ החלטה מתוך K עצים, $\mathcal{F}$: מרחב העצים האפשריים

באופן כללי, המודל ממזער את פונקציית ההפסד הבאה:

$$\mathrm{L}(\emptyset) = \sum_{i=1}^{n} \ell(y_i, \hat{y}_i) + \sum_{k=1}^{K} \Omega(f_k)$$

פונקציית ההפסד בנויה מהאלמנטים הבאים -

$\ell(y_i, \hat{y}_i)$: פונקציית ההפסד, לדוגמא -שורש ממוצעי הריבועים הפ גחותים: RMSE

$\Omega(f_k)$: פונקציית רגולריזציה למניעת התאמת יתר. $\Omega(f_k) = \gamma \mathrm{T} + \frac{1}{2}\lambda\|\omega\|^2$

מיפוי היפר-פרמטרים שנבדקו במהלך שלב האופטימיזציה:

- <u>*Learning Rate* (η)</u> – קצב הלמידה מגדיר עד כמה "מתקדמים" בכל איטרציה – ערכים קטנים מבטיחים למידה איטית יותר ויציבה (*Exploitation*) ככל שיגדל יכול להימנע מאופטימום לוקאלי (*Exploration*) הערכים שנבחנו – [0.2 ,0.1 ,0.01].
- <u>*Max depth*</u> – מגדיר את העומק המרבי של כל עץ ובכך משפיע על מספר העלים בעץ (T). עומק רב יותר מגדיל את מורכבות המודל, עשוי ללכוד פרטים נוספים, אך מסתכן בהתאמת יתר. הערכים שנבחנו - [3, 5, 7].
- <u>*N Estimators* (K)</u> – מגדיר את מספר העצים במודל, ניתן גם להגיד כמציין את מספר סיבובי ההגברה. סיבובים נוספים משפרים את יכולת המודל, אך מגדילים את זמן האימון. ערכים שנבחנו - [50, 100, 200].
- <u>*Subsample*</u> – מייצג את החלק היחסי של הדגימות שבהן משתמשים לאימון כל העץ, ערכים נמוכים מונעים התאמת יתר באמצעות הוספת אקראיות. הערכים שנבחנו – [0.6, 0.8, 1]
- <u>*Colsample bytree*</u> – מייצג את החלק היחסי של משתנה מסביר x_i לבניית כל עץ. בחירת תכונות אקראית משפרת את מגוון העצים. הערכים שנבחנו – [0.6, 0.8, 1]
- <u>*Gamma* (γ)</u> – פרמטר מפונקציית רגולריזציה $\Omega(f_k)$, מגדיר את ההפחתה המינימלית בהפסד הנדרש לפיצול. ערכים גבוהים יותר מעודדים מודלים פשוטים יותר על ידי הגבלת הפיצולים. ערכים שנבדקו [0, 0.1, 0.2]

- *Reg lambda* – מונח הקשור לרגרסיית L2 אשר מעניש משקלים גדולים. עוזר למניעת התאמת יתר. ערכים שנבחנו [2 ,1]

מתוך רשת ההיפר פרמטרים שנבנתה, 100 תצורות אקראיות נדגמו לצורך הערכה. כל תצורה הוערכה על פי ביצועי RMSEעל סט הווליד‌ציה ולאחר בחירתה של הקונפיגורציה המיטבית נמשך תהליך האימון כמפורט בהתחלה.

(Random Forest) – RF

המודל בנוי ממספר רב של עצי החלטה, כאשר כל עץ מאומן על תת-קבוצה אקראית של נתוני האימון ותכונותיהם. גישה זו מאפשרת למודל לשלב את התחזיות מכל העצים כדי לשפר את הדיוק ולצמצם את הסיכון להערכת יתר של הנתונים בעזרת האפקט של האקראיות. במקרה של סיווג, המודל מחליט על התוצאה לפי התוצאה של רוב העצים, ובמקרה של רגרסיה (כמו במקרה שלנו) ממוצע התחזיות מכל העצים מהווה את הפלט הסופי. יתרונו המרכזי טמון ביכולת להתמודד עם נתונים בעלי ממדים רבים ועם משתנים שאינם ליניאריים, תוך שמירה על עמידות מפני רעשים ונתונים חסרים. מודל זה לעומת מודלים אחרים כמו XGBoost שהזכרנו אינו משתמש בפונקציית הפסד אחת כללית לטובת המודל, אלא הוא מתבסס על קריטריון פיצול ברמת העץ, כגון GINI בכדי להחליט על הפיצולים האופטימליים במהלך בניית כל עץ בנפרד ללא קשר לשאר העצים האחרים.

מיפוי פרמטרים שהשתמשנו:

- *N Estimators* - מספר העצים ביער האקראי. ככל שמספר העצים גדול יותר, הביצועים של המודל עשויים להשתפר, אך גם זמן האימון יגדל. עצים רבים יותר עוזרים להפחית את השונות (variance) אך עשויים להגדיל את זמן החישוב. הערכים שנבחנו – [50,100,200]
- *Max Depth* - מגדיר את העומק המקסימלי של כל עץ ביער. עומק רב יותר מאפשר למודל ללמוד פרטים נוספים מהנתונים אך עלול להוביל להתאמת יתר .(overfitting) עומק קטן יותר מונע התאמת יתר אך עשוי לפספס דפוסים חשובים. הערכים שנבחנו – [10,20,None]
- *Min Samples Split* - מגדיר את המספר המינימלי של דוגמאות הנדרשות כדי לפצל צומת בעץ. ערך נמוך מדי עשוי להוביל לעצים מסובכים מאוד ,(overfitting) בעוד שערך גבוה מדי עלול לפספס מידע חשוב. הערכים שנבחנו – [5,2]
- *Min Samples Leaf* - מספר המינימלי של דוגמאות הדרושות כדי ליצור עלה (leaf) בעץ. פרמטר זה עוזר לשלוט בגודל העלים ובעומק העץ הכולל. ערכים גבוהים יותר יכולים למנוע התאמת יתר על ידי יצירת עלים גדולים יותר עם יותר דוגמאות. הערכים שנבחנו – [2,1]

מתוך סט הפרמטרים הללו נבדקו 36 קומבינציות. כל קומבינציה נבדקה לפי ביצועי ה RMSE על סט הווליד‌ציה. הקומבינציה בעלת התוצאות הטובות ביותר נבחרה לצורך האימון הסופי של המודל.

5. הערכה

PLSR

כפי שצוין קודם לכן, נבחרו 50 ערכים שונים למספר הקומפוננטים ונבדק אחוז השונות המוסברת המצטברת לאחר שלב הורדת המימד. על פי ניתוח גרף 1.1, נראה כי השונות המוסברת במשתנים N_Value ו – SC_Value הגיעה לנקודת רוויה אך במשתנה ST_Value נראה כי ניתן להמשיך להעלות את מספר הקומפוננטים שכן אחוז השונות המוסברת ממשיך להעלות. לכן בכדי להימנע מהתאמת יתר של האלגוריתם נבחר המספר 15. לאחר מכן, אומן מודל PLSR יחד עם מספר הקומפוננטים הטוב ביותר (15) בעזרת CV10 בכדי להעריך את יכולות המודל על נתונים שלא ראה. להלן, ביצועי המודל על כל משתנה תלוי במהלך CV10 כפי שניתן לראות בגרף 1.2:

- RMSE – N Value היה נמוך לאורך האימון (מתחת ל 0.1) מלבד עלייה ב Fold 6 ו Fold 7 ששם נצפתה קפיצה בRMSE (0.3 ו – 0.5 ב Fold 6 ו – Fold 7 בהתאמה).
- RMSE – SC Value בין 0.1 ל 0.2.
- RMSE– ST Value בין 0.3 לבין 0.6. הגבוה ביותר בזמן האימון.
- RMSE ממוצע – בין 0.15 ל0.4.

XGBoost

XGBoost הראה יכולת התמודדת עם הנתונים הכמומטרים ואף שילובו עם המשתנים שנוצרו כתוצאה מהשימוש ב-PLS המחיש כיצד הפחתת ממדיות לצד אלגוריתמי חיזוי מתקדמים יכול לשפר את ביצועיהם. כפי שצוין קודם לכן אומנו שתי מודלי XGBoost כאשר הראשון אומן על הנתונים המקוריים והשני אומן על הנתונים של ה PLSR. לכל אחד מן המודלים נבחר סט קונפיגורציה מתאים אשר ממזער את ה RMSE. להלן התוצאות:

– XGBoost

- קצב למידה (Learning rate): 0.2
- עומק העץ (Max Depth): 5
- כמות עצים (N Estimators): 100
- אחוז התצפיות שידגמו לכל עץ (subsample): 0.6
- אחוז המשתנים שידגמו לכל עץ (colsample bytree): 0.8
- פרמטר רגולריזציה (gamma): 0.2
- פרמטר רגולריזציה (lambda): 2

– XGBoost המשלב PLSR

- קצב למידה (Learning rate): 0.1
- עומק העץ (Max Depth): 3
- כמות עצים (N Estimators): 200
- אחוז התצפיות שידגמו לכל עץ (subsample): 0.6
- אחוז המשתנים שידגמו לכל עץ (colsample bytree): 1
- פרמטר רגולריזציה (gamma): 0.1
- פרמטר רגולריזציה (lambda): 2

לצורך קבלת תמונת מצב להשוואה בין שני המודלים אמדנו את ביצועיהם על סט הווליד‌יציה. כפי שצפינו, התוצאות של ה-RMSE בעבור כל אחד ממשתני המטרה היו טובות יותר לאחר הורדת המימד של הנתונים (ראה גרף 2.1) לכן בחרנו להמשיך איתו.

RMSE XGBoost:

- N_Value: 0.31
- SC_Value: 4.34
- ST_Value: 20.46
- AVG_RMSE: 8.37

RMSE XGBoost המשלב PLSR:

- N_Value: 0.16
- SC_Value: 3.49
- ST_Value: 14.95
- AVG_RMSE: 6.2

עוד הצדקה להעדפת XGBoost המשלב PLSR היא שכאשר בחנו את גרף השאריות על סט הולידציה של שני המודלים צפינו הבדל בהתקבצות השאריות סביב אפס כפי שניתן לראות בגרפים 2.2, 2.3, 2.4.

לאחר מכן, אומן המודל הטוב יותר (XGBoost משולב עם PLSR) בעזרת CV10. להלן, ביצועי המודל על כל אחד מהמשתנים התלויים כפי שניתן לראות בגרף 2.5:

- RMSE – N Value בין 0.2 לבין 0.25.
- RMSE – SC Value נע בין 0.15 לבין 0.4.
- RMSE– ST Value בין 0.25 לבין 0.3
- RMSE ממוצע – בין 0.2 0.3.

Random Forest

הקדמה: את מודל הRF אימנו לאחר מודל הXGBoost ולאור השיפור בתוצאות XGBoost עם השימוש בנתוני PLS, החלטנו להשתמש רק בנתונים אלו לאימון מודל RF, כך שבמודל זה השתמשנו במשתנים לאחר הורדת המימד.
סט הקונפיגורציה שנבחר הוא:

- מספר עצים (N Estimators) : 200
- עומק מקסימלי(Max Depth) : None (בלי הגבלה)
- מספר מינימלי לפיצול צומת (Min Samples Split) : 5
- מספר מינימלי לעלה (Min Samples Leaf) : 1

לאחר מכן, אומן מודל הRF יחד עם סט הקונפיגורציה הנ"ל בעזרת CV10. להלן, ביצועי המודל על כל משתנה תלוי במהלך CV10 כפי שניתן לראות בגרף 3:

- RMSE – N Value הערכים היו נמוכים לאורך האימון.
- RMSE – SC Value הערכים היו יציבים לאורך האימונים ונעו בין 2 ל-4.
- RMSE – ST Value הערכים היו גבוהים לאורך האימון ונעו בין 14 ל-16.
- RMSE ממוצע – נע בין 5.8 ל-6.

לבסוף בוצעה השוואה בין ביצועי המודלים על סט הבדיקה. להלן התוצאות כפי שמתואר בגרף 4:

מודל PLSR:

- N Value – 1.186

- 1.663 – SC Value
- 0.468 – ST Value
- 1.106 – RMSE ממוצע

מודל XGBoost המשלב PLSR:

- 0.039 – N Value
- 0.037 – SC Value
- 0.064 – ST Value
- 0.046 – RMSE ממוצע

מודל Random Forest המשלב PLSR:

- 0.044 – N Value
- 0.044 – SC Value
- 0.070 – ST Value
- 0.052 – RMSE ממוצע

6. סיכום, דיון ומסקנות

בפרויקט זה פותחו מודלים כמומטרים מתקדמים לחיזוי ריכוזי נוטריינטים בצמחים על בסיס נתונים ספקטרליים, תוך התמקדות בשימוש בשיטות מתקדמות של למידת מכונה. המודלים בהם נעשה שימוש במחקר הראו ביצועים שונים על פי מדד ה RMSE. מודל ה XGBoost (בשילוב עם הורדת המימד של PLS) הראה את הביצועים הטובים ביותר על סט הבדיקה ביחס לשני המודלים האחרים. המודל הצליח לשלב בין הורדת מימד לבין שמירה על קשר חזק בין המשתנים התלויים והבלתי תלויים מה שהמחיש את ההתאמה של המודל לנתונים שנעשה בהם שימוש במהלך המחקר.

מסקנות

1. המחקר הדגים את הפוטנציאל הגבוה של מודלים כמומטרים, ובמיוחד מודל ה XGBoost, ככלים אמינים ויעילים לחיזוי ריכוזי נוטריינטים בצמחים. המודל הוכיח כי שיטות כמו Boosting יכולות לשפר משמעותית את היכולת להכליל על נתונים חדשים.
2. **שילוב שיטות**: התוצאות ממחישות את היתרון בשילוב טכניקות שונות – כמו הורדת מימד (PLS) ושיטות חיזוי מתקדמות (XGBoost) ליצירת פתרונות מדויקים ויעילים יותר שכן הורדת המימד שיפרה משמעותית את ביצועי האלגוריתם.
3. **פוטנציאל יישומי:** הכלים שפותחו יכולים להוות בסיס לטכנולוגיות מתקדמות לניהול דישון חכם ומדויק, מה שיכול לתרום לייעול תהליכי החקלאות ולהפחתת השפעות סביבתיות שליליות.

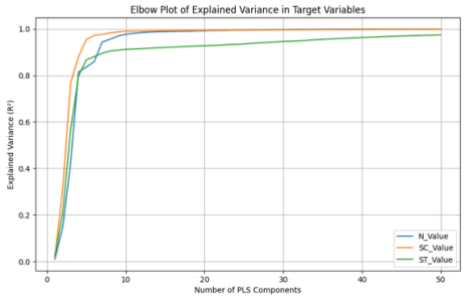
כיווני מחקר עתידיים

1. **חיזוי דינמי באמצעות סדרות זמן:** ניתן לבחון מודלים מבוססי LSTM לניתוח דינמי של ריכוזי נוטריינטים לאורך זמן, דבר שיכול להוות כלי עזר מתקדם לניהול חקלאי מותאם.
2. **הרחבת מאגר הנתונים:** הגדלת מגוון הדגימות והתצפיות, כולל סוגי צמחים שונים, יכולה לשפר את המודלים ולהפוך אותם לרלוונטיים במגוון רחב יותר של סביבות חקלאיות.

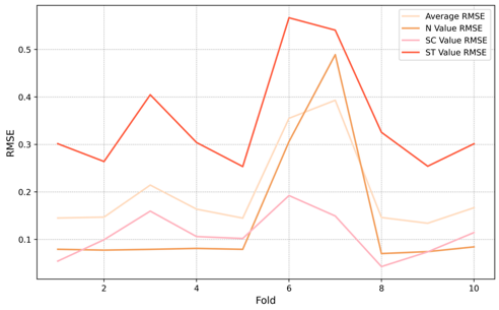
נספחים

קישור לקוד ב GitHub - https://github.com/GuyMizrahi1/Nitrogen-Status-By-Spectroscopy/tree/main

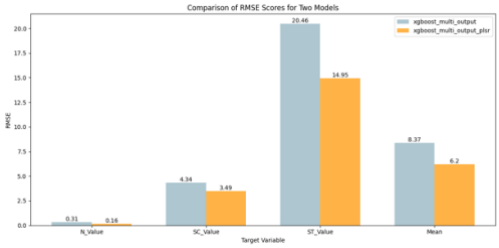
גרף 1.1



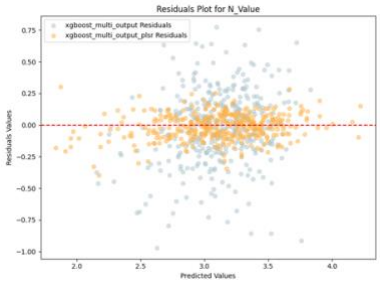
גרף 1.2



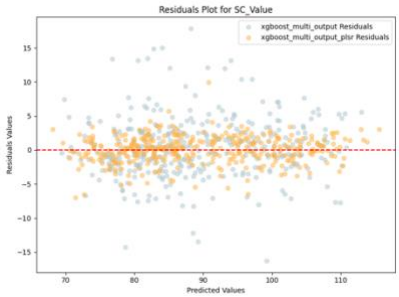
גרף 2.1



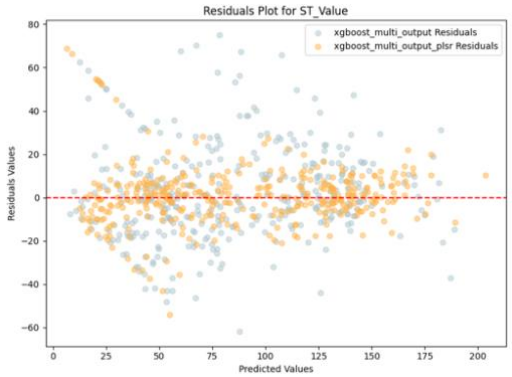
גרף 2.2



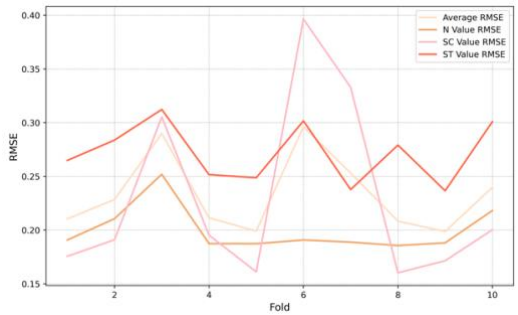
גרף 2.3



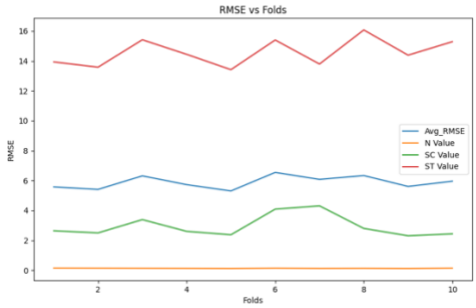
גרף 2.4



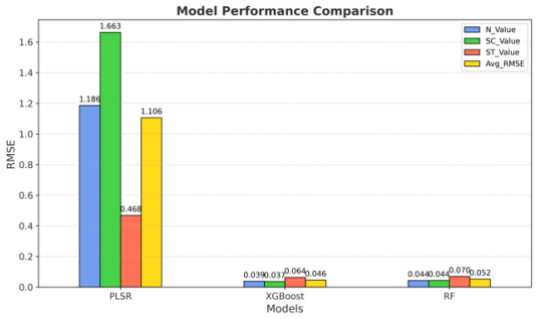
גרף 2.5



גרף 3



גרף 4



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