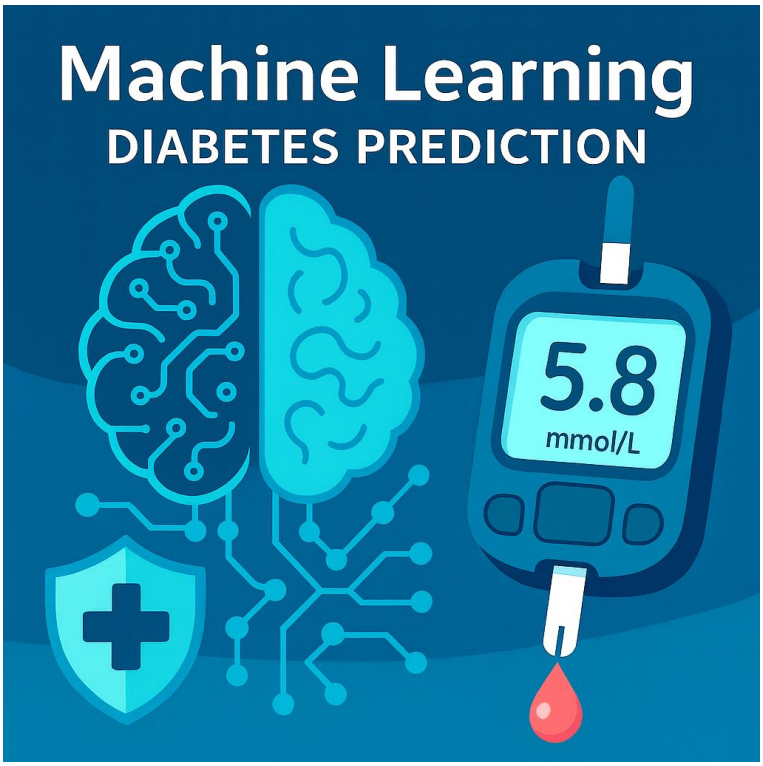
המחלקה להנדסת תעשייה וניהול

דוח פרויקט כריית ידע ולמידת מכונה



שם המרצה: ד"ר רוני הורביץ

מגישים:

נועה עזרי	207875436
אלון אחיטוב	208968354
ניב מאירוביץ'	315224519

קישור לקובץ הנתונים הגולמי

קישור למחברת

מבוא

מחלת הסוכרת היא אחת המחלות הכרוניות הנפוצות ביותר בעולם, עם השפעות בריאותיות רבות. לפי ארגון הבריאות העולמי, מאות מיליוני אנשים חיים כיום עם סוכרת, כאשר מספר החולים נמצא בעלייה מתמדת. אבחון מוקדם של סוכרת חשוב במיוחד, משום שהוא מאפשר להתחיל טיפול בשלב מוקדם, למנוע סיבוכים רפואיים חמורים ולשפר את איכות חייהם של החולים. עם זאת, תהליך האבחון הרפואי מבוסס על בדיקות רפואיות ולעיתים אינו מזהה חולים פוטנציאליים בשלב מוקדם מספיק.

בפרויקט זה מטרתנו הייתה לבחון האם ניתן להשתמש בטכניקות של כריית ידע ולמידת מכונה כדי לחזות האם אדם עלול לחלות בסוכרת, בהתבסס על מאפיינים אישיים, רפואיים ודמוגרפיים.

הבעיה מוגדרת כבעיית סיווג בינארי:

Outcome = 0 ← האדם אינו צפוי לחלות בסוכרת.

Outcome = 1 ← האדם צפוי לחלות בסוכרת.

לשם כך נעשה שימוש במאגר נתונים ייעודי בתחום הסוכרת, הכולל מאות אלפי רשומות של נבדקים. כל רשומה מייצגת אדם אחד ומתארת מגוון רחב של משתנים רפואיים, פיזיולוגיים ודמוגרפיים. בין המשתנים ניתן למצוא:

- מדדים רפואיים כגון רמות גלוקוז בדם, מדד מסת גוף (BMI), לחץ דם והיסטוריה משפחתית של סוכרת.
- מדדים דמוגרפיים כמו גיל, מין, רמת השכלה ורמת הכנסה.
- מדדים התנהגותיים הכוללים מידע על עישון, פעילות גופנית וגורמי סיכון נוספים.

להלן רשימת העמודות הקיימות במאגר הנתונים יחד עם המשמעות של כל אחת מהן:

שם עמודה	תיאור עמודה
Diabetes_binary	סכרתי/בריא
HighBP	לחץ דם גבוה
HighChol	כולסטרול גבוה
CholCheck	בדיקת כולסטרול
BMI	מדד מסת הגוף
Smoker	מעשן/לא מעשן
Stroke	שבץ/ לא שבץ
HeartDiseaseorAttack	מחלת או התקף לב / אין
PhysActivity	פעילות גופנית
Fruits	אכילת פירות
Veggies	אכילת ירקות
HvyAlcoholConsump	צריכת אלכוהול גבוהה
AnyHealthcare	ביטוח רפואי
NoDocbcCost	יכולת כלכלית לרופא
GenHlth	בריאות כללית
MentHlth	בריאות נפשית
PhysHlth	בריאות פיזית
DiffWalk	קושי לעליית מדרגות
Sex	מין
Age	קטגוריית גיל
Education	קטגורית חינוך
Income	קטגוריית הכנסה

המחלקה להנדסת תעשייה וניהול

שלב ראשון: ניתוח מקדים של הנתונים (EDA)

טרם בניית המודלים בוצע ניתוח אקספלורטיבי של הנתונים, שמטרתו הייתה להבין את מבנה הנתונים, לאתר בעיות אפשריות ולהפיק תובנות ראשוניות לגבי המשתנים.

מבנה הנתונים

מאגר הנתונים כלל ב-253,680 רשומות ו-22 עמודות. כל המשתנים נשמרו כערכי float. לא נמצאו ערכים חסרים.

כבר בשלב זה התברר כי הנתונים אינם מאוזנים: כ־86% מהנבדקים אינם חולי סוכרת, בעוד שרק כ-14% מאובחנים כחולים.

ניתוח BMI

לצורך בחינת תרומתו של מדד מסת הגוף למחלת הסוכרת, נבדקה ההתפלגות הכללית של ערכי ה-BMI בכלל המדגם, חושבו שיעורי חולי הסוכרת בכל רמת BMI, ונעשה שימוש בנירמול הנתונים כך שהשוואה תתבצע על סמך אחוזים ולא על פי מספר מקרים מוחלט.

בעקבות ניתוח זה נמצא כי ככל שה-BMI גבוה יותר, כך גדלה ההסתברות לסוכרת, מה שמצביע על חשיבותו של משתנה זה כחיזוי מרכזי.

ניתוח נתונים דמוגרפיים

הקשר בין משתנים דמוגרפיים לבין סוכרת נבדק על ידי חישוב שיעורי המחלה בכל קבוצה של משתנה: הכנסה, גיל והשכלה.

המסקנות מהניתוח היו חד-משמעיות:

הכנסה גבוהה - הסתברות קטנה יותר לחלות בסוכרת.

קבוצות גיל מבוגרות יותר - הסתברות גדולה יותר לחלות בסוכרת.

רמת השכלה גבוהה - הסתברות קטנה יותר לחלות בסוכרת.

ניתוח נתוני בריאות

כדי לבחון את הקשר בין מצב בריאותי כללי לבין מחלת הסוכרת, נבדקו שלושה משתנים עיקריים: בריאות נפשית, בריאות פיזית ודירוג בריאות כללי. בדומה לניתוח ה- BMI, תחילה נותחו ההתפלגויות של כל אחד מהמשתנים, ולאחר מכן חושבו שיעורי החולים בסוכרת בכל רמת דירוג. בנוסף, בוצעה השוואה בין קבוצות נפרדות כדי לזהות פערים מובהקים.

מהבדיקות עלה כי בכל שלושת המשתנים קיים קשר עקבי וברור: ככל שרמת הבריאות ירודה יותר (נפשית, פיזית או כללית)- כך עולה ההסתברות לחלות בסוכרת.

ניתוח משתנים נוספים

בנוסף, נבחנו משתנים רפואיים והתנהגותיים נוספים. נמצא קשר חזק בין סוכרת לבין משתנים כגון: לחץ דם גבוה, רמות כולסטרול גבוהות, בדיקת כולסטרול, עישון, מחלות לב, צריכת אלכוהול גבוהה, חוסר בפעילות גופנית, קושי בעליית מדרגות ושבץ.

לעומת זאת, משתנים כמו אכילת פירות וירקות, יכולת כלכלית לרופא, ביטוח רפואי או מגדר נמצאו כבעלי תרומה נמוכה יותר.

המחלקה להנדסת תעשייה וניהול

שלב שני: עיבוד הנתונים (Data Preparation)

בשלב זה הוכנו הנתונים לצורך בניית המודלים החיזויים. מצבם ההתחלתי של הנתונים היה שכל המשתנים התקבלו כ-Float, ללא הבחנה בין משתנים בינאריים, קטגוריאליים או נומריים רציפים. מצב זה יצר קושי, משום שהוא לא שיקף את טבעו האמיתי של כל משתנה ועלול היה לפגוע ביכולת המודלים ללמוד באופן נכון.

לכן בוצעו מספר פעולות מרכזיות להכנת הנתונים:

- הגדרת טיפוסי נתונים מתאימים: משתנים שהתקבלו כ-Float הותאמו לטיפוס הנכון: בינארי (0/1), קטגוריאלי או נומרי רציף.
- One Hot Encoding / Dummies: עבור משתנים קטגוריאליים בוצע קידוד לערכים בינאריים, כך שכל קטגוריה קיבלה עמודת 0/1 נפרדת.
- Normalization: עבור משתנים רציפים בוצע נרמול, כך שכולם יהיו באותו טווח מספרי.
- סינון משתנים: הוסרו עמודות שהציגו קשר חלש במיוחד לעמודת המטרה או תרומה שולית לחיזוי, ונשמרו רק המשתנים המרכזיים.

הטבלה הבאה מסכמת את תהליך עיבוד הנתונים, ומציגה עבור כל עמודה את הטיפוס המקורי שבו התקבלה (Float), הטיפוס שהותאם לה לאחר עיבוד, והאם נשמרה במודל הסופי או הוסרה.

שם עמודה	נשאר / לא נשאר	Dtype מקורי	Dtype רצוי
Diabetes_binary	✓	Float	Binary
HighBP	✓	Float	Binary
HighChol	✓	Float	Binary
CholCheck	✓	Float	Binary
BMI	✓	Float	Normalized
Smoker	✓	Float	Binary
Stroke	✓	Float	Binary
HeartDiseaseorAttack	✓	Float	Binary
PhysActivity	✓	Float	Binary
Fruits	✗	Float	-
Veggies	✗	Float	-
HvyAlcoholConsump	✓	Float	Binary
AnyHealthcare	✗	Float	-
NoDocbcCost	✗	Float	-
GenHlth	✓	Float	Categorical
MentHlth	✓	Float	Normalized
PhysHlth	✓	Float	Normalized
DiffWalk	✓	Float	Binary
Sex	✗	Float	-
Age	✓	Float	Categorical
Education	✓	Float	Categorical
Income	✓	Float	Categorical

המחלקה להנדסת תעשייה וניהול

שלב שלישי: בנייה ואימון המודלים

כבסיס הבנייה ואימון המודלים הוגדר כי המדד המרכזי להערכת ההצלחה יהיה Recall, שכן החמצת חולה סוכרת בפועל (False Negative) עלולה להוביל להשלכות חמורות, בעוד שטעות מסוג False Positive מביאה בעיקר לעלויות נוספות של בדיקות אך אינה מסכנת חיים.
לכן, לכל אורך תהליך בניית המודלים, השאיפה המרכזית הייתה למקסם Recall, גם במחיר ירידה מסוימת ב-Precision או ב-Accuracy. בהתאם לכך, תהליך הפיתוח התבצע כשלבי אב-טיפוס מתקדם: תחילה נבנה מודל בסיסי, ובהמשך נוספו שיפורים בארכיטקטורה, איזון הנתונים ופרמטרי האימון. בכל מודל הושוו תוצאות המודל לתוצאות המודלים הקודמים שנבנו, מתוך מטרה לזהות את השילוב האופטימלי שמספק את התוצאות הטובות ביותר ביחס למדד היעד. כך גובש בהדרגה המודל "המנצח", שנבחר לא על בסיס ניסיון חד פעמי, אלא מתוך תהליך שיטתי של ניסוי והשוואה.

Pre Undersampling Model | Diabetes Classifier1

הצעד הראשון היה לבנות מודל בסיסי על הנתונים המקוריים, גם אם הוא לא מאוזן. הבחירה להתחיל דווקא בנתונים הלא מאוזנים נבעה מהחשיבות לראות כיצד המודל מתמודד עם הנתונים כפי שהם במציאות. בעולם האמיתי נתונים רפואיים כמעט תמיד סובלים מחוסר איזון משמעותי בין אוכלוסיות (מספר החולים קטן בהרבה ממספר הבריאים), ולכן רצינו לבחון תחילה מהי נקודת המוצא של המודל בתנאים אלו. בנוסף, מודל כזה מהווה נקודת ייחוס שממנה ניתן להעריך את התרומה של שיטות איזון שונות ושל שיפורים בהמשך.

המודל הכיל רשת עצבית פשוטה עם שכבות Fully Connected, פונקציית הפעלה ReLU ו-Dropout. בזמן האימון, המודל הציג ירידה עקבית בפונקציית ההפסד, אך בפועל בניתוח התוצאות נובאו כמעט את כל המקרים כלא-חולים. אמנם התקבל Accuracy גבוה (כ86%), אך Recall נמוך מאוד. תוצאה זו נובעת ישירות מחוסר האיזון המשמעותי בנתונים: המודל למעשה למד לנצל את חלוקת המקרים הלא שוויונית, ולכן הצליח להשיג דיוק גבוה יחסית, אך ה-Recall שהוא המדד המרכזי מבחינתנו היה נמוך מאוד. כלומר, המודל אמנם היה מדויק במקרים רבים, אך החמיץ שיעור ניכר של חולים אמיתיים.

הממצאים הללו חיזקו את הצורך להתמודד באופן ישיר עם בעיית חוסר האיזון. לאחר שנבחנו מספר אפשרויות, נבחר ליישם שיטת Under-Sampling, המפחיתה באופן יזום את מספר הדוגמאות מהמחלקה הגדולה (בריאים) כדי להתקרב ליחס מאוזן יותר בין חולים ללא-חולים, בנתונים המשמשים לאימון המודלים בלבד. גישה זו נבחרה מאחר שהיא פשוטה ליישום ומאפשרת למודל להיחשף באופן מאוזן יותר לשתי הקבוצות, ובכך להעלות את הסיכוי לזהות חולי סוכרת בפועל.

Undersampling Modeling (75:25 Strategy) | Model 1 | Diabetes Classifier1

לשם כך אוזנו הנתונים ליחס של כ־75:25, נוצר מופע חדש של אותו מודל והוא אומן מחדש. כתוצאה מכך, נצפתה ירידה מתונה בדיוק הכולל (לכ-80%), אך במקביל נרשמה עלייה ניכרת ב-Recall, כך שהמודל הצליח לסווג את המקרים באופן מאוזן יותר.
בנוסף, לצורך שיפור נוסף של ה-Recall, הופחת סף ההחלטה ל־0.4, מה שהגדיל עוד יותר את שיעור החולים שזוהו בהצלחה.

Undersampling Modeling (75:25 Strategy) | Model 2 | Diabetes Classifier2

בשלב הבא נבחן האם השימוש ב-Dropout אכן תורם לשיפור ביצועי המודל או שמא ניתן לוותר עליו. לשם כך נבנה מודל דומה לארכיטקטורה המקורית, אך ללא שכבות Dropout, אשר אומן על אותם הנתונים שעברו איזון. לאחר ההשוואה נמצא כי המודל הראשון (עם Dropout) הציג ביצועים טובים יותר, במיוחד במדד ה-Recall.
בעקבות ממצא זה הוחלט כי בכל המודלים הבאים Dropout יישאר כרכיב קבוע בארכיטקטורה.

Undersampling Modeling (75:25 Strategy) | Model 3 | Generic Model

לאחר קביעת מבנה בסיסי שכלל Dropout כחלק אינטגרלי, הוגדר מודל גנרי גמיש יותר, אשר אפשר שליטה על פרמטרים שונים בארכיטקטורה ובאימון. מבנה המודל כלל שכבות Fully Connected, בשילוב Batch Normalization ו־Dropout, על מנת לשפר יציבות ולמנוע למידה עודפת. מתוך מודל זה נבנו שמונה מופעים שונים, אשר נבדלו זה מזה בשלושה פרמטרים עיקריים:

- פונקציית הפסד: BCE או MSE
- אופטימייזר: Adam או SGD
- Batch size: 32 או 64

כל שילוב אומן על גבי 15 Epochs בלבד, כדי לצמצם את זמן הריצה ולאפשר השוואה מהירה. עבור כל מופע חושבו המדדים המרכזיים והתוצאות הושוו ביניהן. מן ההשוואה עלה כי השילוב האופטימלי לבעיה שלנו היה BCE + Adam + Batch size = 32, אשר סיפק את האיזון הטוב ביותר בין הדיוק הכולל לשיפור ב-Recall.

Undersampling Modeling (75:25 Strategy) | Model 4 | Chosen Model

בהתבסס על המסקנות מהמודל הגנרי, נבנה מודל נוסף עם השילוב האופטימלי שנבחר (BCE + Adam + Batch size = 32). הפעם, על מנת להפיק את המרב מארכיטקטורה זו, הוגדל מספר ה-Epochs ובמקביל הונמך סף ההחלטה ל-0.35, במטרה לשפר עוד יותר את מדד ה-Recall.

תוצאות האימון הראו שכאשר מספר ה-Epochs גדל וסף ההחלטה הונמך, שיעור החולים שזוהו בהצלחה (Recall) עלה, אך חלה ירידה בדיוק הכולל. עם זאת, בהתאם למטרתנו המרכזית בפרויקט, הירידה בדיוק הכולל נחשבה סבירה ביחס לשיפור המשמעותי ביכולת המודל לזהות חולי סוכרת.

Undersampling Modeling (50:50 Strategy) | Model 5 | Chosen Model50

בשלב הבא נבדקה השפעת איזון נוסף של הדאטה על ביצועי המודל. המודל הנבחר אומן על נתונים המאוזנים ביחס 50:50 בין חולים ללא-חולים.

התוצאות הצביעו על עלייה משמעותית ב-Recall, אולם במקביל נרשמה ירידה ניכרת ב-Accuracy, בעיקר משום שכמות הרשומות הצטמצמה באופן חד, והמודל נטה לנבא שיעור גבוה יחסית של חולים. עם זאת, המודל עמד במטרה המרכזית של הפרויקט ואכן השיג Recall גבוה.

Undersampling Modeling (60:40 Strategy) | Model 6 | Chosen Model60

כדי לבחון האם איזון פחות אגרסיבי תוך שמירה על מספר רב יותר של רשומות ישפר את התוצאות, בוצע איזון של הנתונים ביחס 60:40.

כאשר סף החלטה היה 0.35 התקבל Recall גבוה מאוד, אך במחיר ירידה חדה ב-Accuracy, שכן המודל ניבא אחוז גבוה מדי מהנבדקים כחולי סוכרת.

כאשר הועלה סף ההחלטה חזרה ל־0.5, התקבלה תוצאה מאוזנת יותר; גם Recall גבוה וגם דיוק משופר. ניסוי זה הראה כי איזון מתון של הנתונים, בשילוב התאמת סף ההחלטה, יכול לספק תוצאות טובות יותר מהאיזון הקיצוני- ובנקודה זו היינו שבעי רצון מהתוצאות שהושגו.

בסיום שלבי האימון וההשוואה, שלושת המודלים שנמצאו כמתאימים ביותר לצורכי הפרויקט היו:

- Model 4 (Chosen Model)
- Model 5 (Chosen Model50)
- Model 6 (Chosen Model60) עם סף החלטה של 0.5

מודלים אלו הציגו את האיזון הטוב ביותר בין Recall גבוה לדיוק, ובהם התמקד השלב הבא- הפעלת המודל הכלכלי, שנועד להעריך את המשמעות המעשית של ביצועי המודל במונחי עלות ותועלת.

המחלקה להנדסת תעשייה וניהול

שלב רביעי: המודל הכלכלי

לאחר בחירת שלושת המודלים המובילים, נבחנו ביצועיהם לא רק במונחים סטטיסטיים (Recall, Precision, Accuracy), אלא גם בהיבט יישומי-כלכלי. לשם כך פותח מודל כלכלי שמטרתו להעריך את המשמעות המעשית של טעויות הסיווג, באמצעות כימות של העלויות והתועלות הנלוות לכל תוצאה אפשרית (True Positive, False Positive, False Negative, True Negative).

בשלב הראשון הוגדרו כלל רכיבי העלות והתועלת המרכיבים את המודל הכלכלי. רכיבים אלו שימשו כבסיס לחישוב המשמעות המעשית של כל אחד מסוגי הסיווג (TP, FP, FN, TN):

- עלות הרצת המודל לכל נבדק (C_screen) = 1$.
- עלות בדיקת אימות/שלילה לכל מטופל שנחזה כחיובי (C_confirm) = 30$.
- עלות טיפול מוקדם למקרים שאומתו כחיוביים (C_treat) = 200$.
- תועלת (חיסכון) הנובעת מזיהוי מוקדם של חולה סוכרת, כלומר הוצאות עתידיות שנמנעות (B_TP) = 6000$.
- עלות שלילי כוזב (FN), המבטאת את המחיר הגבוה של פספוס חולה אמיתי (C_FN) = 12000$.
- עלות עודפת לחיובי כוזב (FP), מעבר לעלות הבדיקה המאשרת (C_FP_extra) = 20$.

הפרמטרים הכמותיים הוגדרו בהתאם למידע שנאסף ממקורות באינטרנט על עלויות בדיקות רפואיות וטיפולים, והותאמו לשם ביצוע סימולציה.

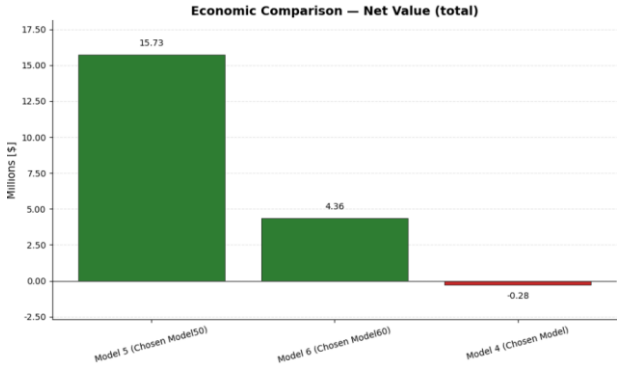
לאחר הגדרת רכיבי העלות והתועלת, יושם המודל הכלכלי על שלושת המודלים שנבחרו:
Model 4 (Chosen Model), Model 5 (Chosen Model50) ו-Model 6 (Chosen Model60)

בחלק מההערכה של כל אחד מהמודלים, חושבו ערכי TP/FP/FN/TN מתוך ה-Confusion Matrix. על בסיסם חושבו:

- עלות כוללת: שילוב של עלות הרצת המודל, בדיקות אישור, טיפולים, פספוסים ועלויות חיוביים כוזבים.
- תועלת כוללת: החיסכון מגילוי מוקדם (TP).
- ערך נטו: סך התועלות פחות סך העלויות.

תוצאות אלו הוצגו במספר גרפים עבור כל מודל, ולבסוף הוצג גרף מסכם אשר משווה את הערך הכלכלי הכולל בין שלושת המודלים.

מהשוואת הערך הכלכלי הכולל (Net Value) בין שלושת המודלים, התקבלה תוצאה ברורה:
Model 5 (Chosen Model50): הציג את התוצאה הטובה ביותר, עם ערך כלכלי חיובי גבוה במיוחד של כ-15.73 מיליון דולר.
תוצאה זו מעידה כי אף על פי שסטטיסטית המודל הציג ירידה בדיוק, השילוב של Recall גבוה מאוד יחד עם מניעת עלויות עתידיות הניב חיסכון כלכלי משמעותי.
Model 6 (Chosen Model60): הציג גם הוא ערך כלכלי חיובי של כ-4.36 מיליון דולר.
מודל זה תרם לחיסכון, אם כי בהיקף קטן בהרבה לעומת מודל 5.
Model 4 (Chosen Model): הציג ערך כלכלי שלילי של כ-0.28 מיליון דולר במינוס, מה שמעיד כי למרות ביצועים סטטיסטיים טובים יחסית, תרומתו בהיבט הכלכלי אינה משתלמת.



לסיכום, המודל עם איזון 50:50 (Model 5) הוכיח את עצמו כבחירה המשתלמת ביותר בהיבט הכלכלי, והציג את החיסכון הגבוה ביותר מבין כל החלופות שנבדקו. ממצא זה מדגיש את החשיבות של שילוב בין בחינה סטטיסטית לבחינה כלכלית-יישומית: מודל שאינו מצטיין בכל המדדים הסטטיסטיים עשוי להיות בכל זאת הפתרון המועדף כאשר בוחנים את המשמעות המעשית של יישומו בעולם האמיתי.