המכללה האקדמית עמק יזרעאל
ע"ש מקס שטרן
החוג למערכות מידע

The Max Stern Academic
College of Emek Yezreel
Dept. of Information Systems

מסמך בחירת נושא

במסמך זה עליכם להגדיר מהו הנושא שבחרתם לפרויקט הגמר.

1. שמות הסטודנטים

נועם פטיש - 212205082
ניקול בלישר - 328654546

2. נושא הפרויקט

חיזוי מניות – בפרויקט אנו ננתח נתוני מניות במטרה לחזות את התנהגותן בעתיד. הרעיון המרכזי הוא להשתמש בנתוני המניות ובמדדים פיננסיים נוספים כדי לנסות לזהות דפוסים ומגמות שיאפשרו להעריך האם מחיר המניה יעלה או ירד.

חתימת הסטודנט: **נועם פטיש**

חתימת הסטודנט: **ניקול בלישר**

חתימת המנחה: _____

תאריך 10/11/2025

דו"ח התקדמות להגשה לשלב Choosing Subject+Business Understanding

שמות הסטודנטים : ניקול בלישר נועם פטיש

מס' ת"ז : 328654546 212205082

שם הפרויקט : חיזוי מניות

שם המנחה : ג'ניה גוטפריד

אנא פרטו את שלבי ביצוע הפרויקט עד כה מהדוח הקודם.
במידה ואין התקדמות עד כה יש לפרט את הסיבות.

*** שלבי ביצוע מהדוח האחרון עד לדוח זה: בדקנו את המקורות מידע שלנו, חידדנו את מטרת הפרויקט שלנו. בדיקת מניות שאותם אנו נרצה להתמקד ולבדוק, בדקנו באיזה כלים ומודלים נרצה להשתמש.**

חתימת הסטודנט : נועם פטיש

חתימת הסטודנט :ניקול בלישר

חתימת המנחה _____

תאריך :28/11/25

הגדרת הבעיה של הפרויקט:

גולשים רבים המנסים להשקיע בחברות שונות רוצים להבין את גורם העליה והירידה שלה החברות בהן רוצות להשקיע.

הבעיה אותה אנו רוצים לפתור- האם ניתן לחזות מגמת עליה או ירידה של מניות גדולות באמצעות שילוב נתונים מאקרו (ריבית הפד, אינפלציה, אמון הצרכנים נתוני, VIX של וול סטריט...), נתוני שוק ונתוני מניה.

מטרת הפרויקט:

פיתוח מודל חיזוי למניות במידה אם יעלו או ירדו, בשילוב נתונים ממספר מקורות. ניתוח השפעה של משתנה על התנהגות מניה, וזיהוי פיצ'רים משמעותיים המשפיעים על המגמה בצורה ניכרת. ולבסוף נציג תוצאות באמצעות גרפים.

התועלת הצפויה מהפרויקט:

ניתן לקבל תחזית למספר גורמים כגון משקיעים כלי המספק להם תחזית מניה, לסטודנטים הלומדים מקצועות פיננסים, לאנליסטים תובנות על הקשרים בין הנתונים לבין ביצועי המניות, וכמו כן תרומה טכנולוגית איסוף דאטה ניתוח בניית מודלים וכדומה.

זמינות נתונים:

מקורות הנתונים שנאסוף בפרויקט זמינים בפלטפרמות שונות כגון :

YAHOO FINANCE-

FRED API-

FINNHUB /NEWSAPI-

ובהמשך נחפש בעוד מקורות במידת הצורך.

בחינת פתרונות קיימים לבעיה:

חיזוי מניות הוא תחום שבו לא קיים פתרון מוחלט וודאי ולכן לא קיימת הבטחה לחיזוי מושלם. הפרויקט שלנו מציע מודל שמנסה לשפר את איכות החיזוי באמצעות שילוב מספר מקורות מידע תוך הבנה שייתכן ומדובר בחוסר וודאות.

טכניקות וכלים לפרוייקט:

תחילה נאסוף נתונים מהמקורות באמצעות API, ספריות בפיתון וכדומה. כמו כן נעשה ניתוח באמצעות כלים ויזואליים כגון גרפים וקשרים בין הפיצ'רים לנתונים. נעשה שימוש במודלים כגון רגרסיה לוגיסטית,KNN, ועוד מודלים שנראה לנכון לנתח באמצעותם בהמשך הפרויקט. נציג את הנתונים באמצעות POER BI, DASHBORADS וכדומה.

הערך המוסף שלנו:

בניה מערכת חיזוי המסייעת למשתמשים לא לבחור מניות בצורה אקראית, אלא לקבל המלצות שמתאימות לאופי, להרגלים ולרמת הסיכון האישית שלהם. כך תהליך ההשקעה יהפוך ברור ונגיש יותר, גם למי שלא מגיע מעולם הפיננסיים. הערך המוסף הוא יצירת מסלול השקעה חכם שמכוון אנשים לבחור נכון לעצמם ולבנות יציבות לאורך זמן.

דו"ח התקדמות להגשה לשלב Data Understanding

שמות הסטודנטים : נועם פטיש, ניקול בלישר

מס' ת"ז : 212205082, 328654546

שם הפרויקט : חיזוי מניות

שם המנחה : ג'ניה גוטפריד

<u>שלבי התקדמות הפרויקט:</u>

בשלב זה, אספנו את מקורות המידע בחרנו את המניות הרלוונטיות שעליהם נרצה להתמקד בפרויקט, הרחבנו את המדדים שעליהם נבדוק את המודל שלנו שיתן לנו בהמשך את החיזוי הכי אמין. התחברנו לAPI של אותם המקורות וחילצנו מהם נתונים בעזרת קוד בפייתון והימרנו את הנתונים לטבלאות בCSV ועשינו ייבוא לPOWER BI .

הצגנו אותם לתרשים ERD המחולקים לטבלאות שבכל טבלה קיימים מאפיינים ונתונים על מנת שיקדמו אותנו בהמשך לתוצאה הרצויה, ועשינו בהתאם 3 ויזואליזציות בהתאם.

חתימת הסטודנט : נועם פטיש

חתימת הסטודנט : ניקול בלישר

חתימת המנחה _____

תאריך : 06/01/2026

דוח מספר 2 - Data Understanding :

מקורות המידע:

1. Yahoo Finance - מספק נתוני שוק (Stock Prices)
2. FRED - המספק נתוני מאקרו כלכלה.

הנתונים שלנו כוללים טווחים של התאריכים בין 2018-2025.

בדיקת ראשונית של הנתונים:

<u>סוגי הנתונים :</u>

- נתונים מספריים רציפים - מחירי פתיחה, סגירה תשואות יומיות, אינפלציה, ריבית וכו'.
-נתונים מספריים בדידים - נפח מסחר המייצג ספירה של העסקאות שבוצעו.
- נתונים קטגוריאליים - מניה.
- נתונים זמן - תאריכים.

<u>היבטים מתוך המידע :</u>

- ניתן להבחין במחירי מניה שנמדדים כל יום לאורך תקופה ארוכה.
- ישנם משתני מאקרו המשפיעים על השוק.
- ישנו קשר בין נפח המסחר לתנודות חדות של המגמה.

<u>כמות דגימות :</u>

בכל טבלה יש אלפי רשומות ומספר מאפיינים המספיקים לנו לניתוח סטטיסטי לצורך הסקת מסקנות ולניתוח חיזוי אמין.

טווח הנתונים משתרע על כ-8 שנים עם נתונים יום יומיים.

<u>מיזוג מקורות ואתגרים :</u>

הפרויקט דורש מיזוג של נתוני מניות, לכן הם נלקחו ממקורות שונים שמרנו בטבלאות שונות ונפרדות וחיברנו באמצעות מפתח משותף שהוא המאפיין התאריך.

האתגר המרכזי הוא חוסר אחידות בזמינות נתונים (נתוני מאקרו חסרים).

<u>טיפול במידע חסר :</u>

-ימים ללא מסחר הוסרו מנתוני המניות
-ערכים חסרי של מדדי מאקרו הושלמו באמצעות הערך האחרון
-רשומות ראשונות ללא ערך תשואה הוסרו

תיאור הנתונים:

<u>כמות:</u>

dim_symbol - 3 שורות ומאפיין 1 (המניה).

fact_macro_daily - 2,086 שורות ו-7 מאפיינים.

fact_prices_daily - 6,024 שורות ו-8 מאפיינים.

dim_date- 2922 שורות ו-6 מאפיינים.

<u>סוגי ערכים:</u>

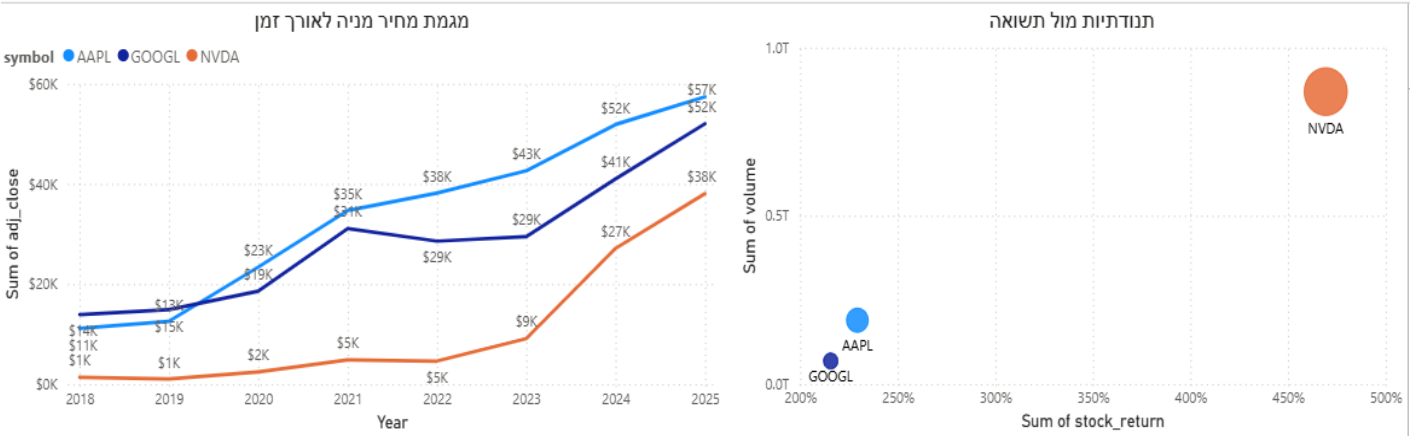
בטבלאות שלנו יש נתונים קטגוריאליים, תאריכים, מספרים שלמים, מספרים רציפים.

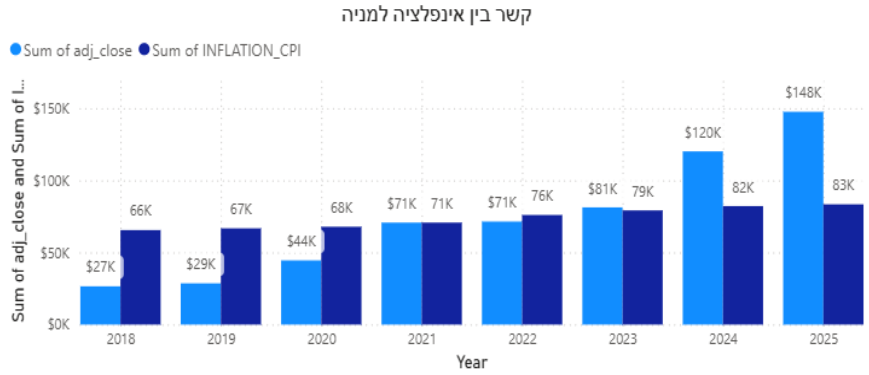
<u>קידוד:</u>

לא קיימים ייצוגים שונים לאותו ערך, שמות המשתנים ויחידות מידה אוחדו לצורך אחידות וניתוח תקין.

חקירת הנתונים:

לצורך חקירת הנתונים עשינו שימוש ב Power BI ליצירת ויזואליזציות. נראה כי לפי החקירה הראשונית נראה זיהוי הבדלים ברמת תנודתיות בין המניות והדגישה היא על NVDIA במניה בעלת תשואה גבוהה יחסית.





איכות הנתונים:

מידע חסר - מצאנו מידע חסר בימי מסחר חסרים, ולכן טיפלנו באמצעות הסרה או השלמה בהתאם לקישור.

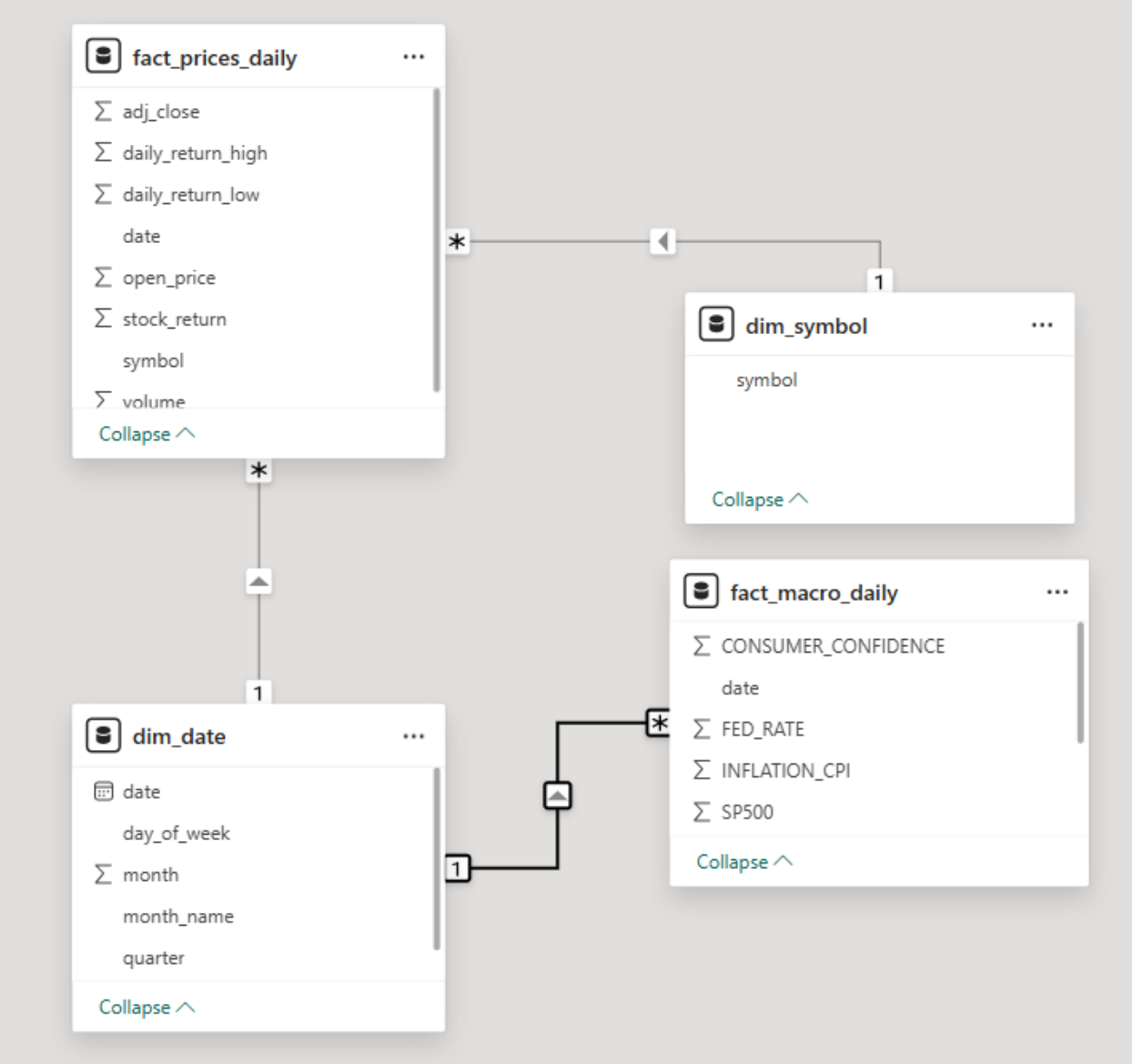
שגיאות - לא היה לנו בנתונים, היות והנתונים נאספו ממאגרים רשמיים.

בעיות מדידה – לצורך אחידות את מדדי המאקרו בוצעו התאמה לתדירות יומית.

חוסר אחידות – אוחדו יחידות מידה אחוזים מחירים והוגדרו בפורמט תצוגה אחיד POWER BI באמצעות NEW MEASURE.

מטא־דאטה שגוי - לא נמצא אי התאמה, נבדקה התאמה מלאה בין משמעות השדות לבין התוכן בפועל.

הכנת תרשים ERD ראשוני:



3. Data Preparation

בחירת נתונים

הרשומות והמאפיינים הרלוונטיים שלנו הם:

- נתוני מחירי המניות: AAPL,GOOGL,NVDA

-נתוני המאקרו: fed rate, inflation cpi, consumer confidence, sp500, VIX,US10Y

-משתנה המטרה: Stock return

הסיבה להשארת הנתונים

בשלב זה בחרנו להשאיר את המאפיינים כי ראינו בהם השפעה על התשואות היומיות. נתוני המאקרו משקפים את מצב הכלכלה ומשפיעים באופן ישיר על שוק ההון.

ביצוע ניקוי נתונים

-נבדקו ערכי NULL בטבלאות שייבאנו לSQL באמצעות שאילתות LEFT JOIN , ובוצעו בדיקות COUNT. נמצא כי אין עמודות NULL,ואין תאריכים חסרים.

-בעת ייבוא CSV חלק מהעמודות הוגדרו כ VARCHAR.

-בוצעו בדיקות כפילויות.

-בוצעו אחידות של תאריכים, נבנה טבלת Dim_Date המאחדת את התאריכים מהטבלאות באמצעות UNION. לצורך יצירת מבנה נתונים והגדרת קשרי Foreign Key .

טיפול בנתונים

-נבדקו outliners בתשואות במסגרת הEDA וראינו כי ישנם כמה תצפיות חריגות אך לא הוסרו, זאת מאחר והם מייצגים אירועי שוק אמיתיים ולא כשגיאות נתונים.

-נתוני המאקרו יושרו לתדירות יומית באמצעות פונקציית תדירות ASFRENQ המבצעת רצף של תאריכים גם במידה ואין יום/ערך, ובאמצעות forward fill ממלאת לפי יום קודם. כל הנתונים נשמרו בטבלה fact_macro_daily והחיבור לנתוני המניות נעשה על בסיס השדה date.

בכך הומרו נתוני המאקרו לתדירות יומית התואמת לנתוני המניות, כך שכל יום מסחר קיבל ערך מאקרו רלוונטי.

Feature Engineering

-בחרנו ב-Stock Return כמשתנה מטרה משום שהוא מתאר את השינוי היחסי במחיר המניה מיום ליום, ולא את המחיר עצמו, כך ניתן להשוות בין מניות שונות גם אם רמות המחיר שלהן שונות.

בנוסף, תשואות יומיות מאפשרות לזהות תגובות מהירות לאירועי שוק ולשלב בצורה טבעית נתוני מאקרו הנמדדים על בסיס יומי.

-החישוב של Stock return בוצע על בסיס מחירי הסגירה היומיים המתואמים שנשלפו ממקור הנתונים תוך השוואה למחיר יום המסחר הקודם עבור כל מניה.

-לא נוצרו משתני Lag במסד נתונים, רק ביצענו הזזה של יום אחד לאחור על מנת שנוכל לחשב חישוב התשואה.

- בשלב זה לא נוצרו מאפיינים מסוג Moving Averages או Rolling Features.

מאפיינים אלו מחושבים על פני חלון זמן מתגלגל (מספר ימים אחורה לכל תאריך), ואינם תלויים בטווח התאריכים הכולל של מסד הנתונים.

נרמול/סטנדרטיזציה

לא בוצע נרמול למשתנה המטרה כי הוא מחושב כיחס ולכן כבר מנורמל באופן כפי שלקחנו מהמקורות.

אינטגרציה של נתונים

בוצעה אינטגרציה בין: fact_prices_daily ו - fact_macro_daily

באמצעות JOIN על השדה DATE .

עיצוב ופורמט נתונים

הנתונים הוכנסו למסד הנתונים לפי השדות הבאים:

Dimensions:

- dim_date (PK: date)
- dim_stock (PK: symbol)

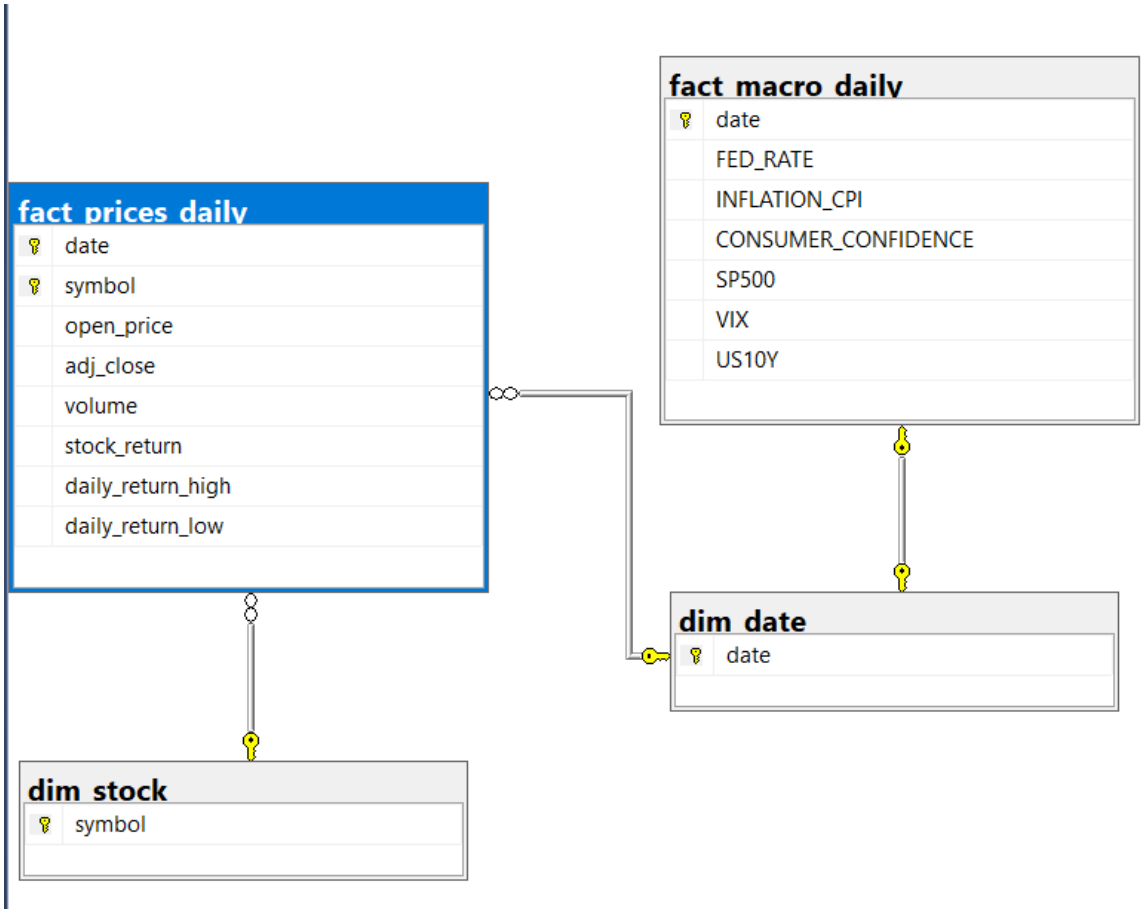
Fact Tables:

- fact_prices_daily (PK: symbol, date)
- fact_macro_daily (PK: date)

Foreign Keys:

- fact_prices_daily -> dim_date
- fact_prices_daily -> dim_stock
- fact_macro_daily -> dim_date

צירוף צילום של הERD הסופי:



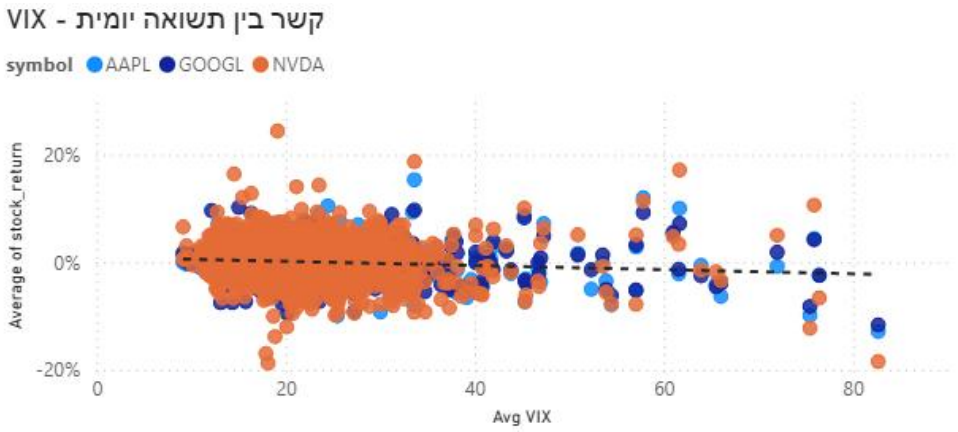
ניתוח חוקר EDA

<u>סטטיסטיקה תיאורית- משתנה מטרה STOCK RETURN:</u>

0.24	0.0015	-0.19	0.02
Max Stock retur...	Avg Stock return	Min Stock return	Std Stock return

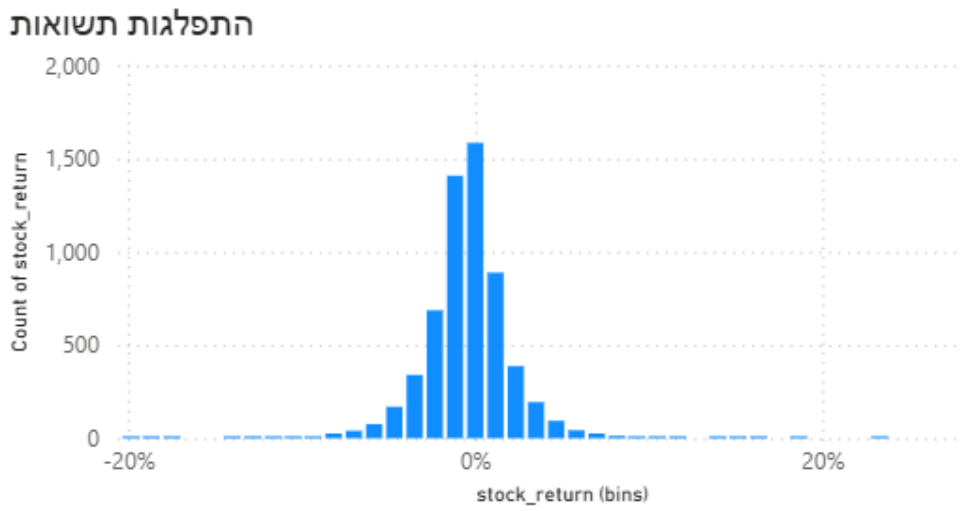
בוצעו: ממוצע תשואה יומית, סטיית תקן, מינימום יומי, מקסימום יומי.

<u>סטטיסטיקה תיאורית למאפיינים חשובים- נתוני מאקרו:</u>



ראינו כי מדד הVIX בעל הקשר החזק ביותר- קשר שלילי מתון למדד התשואה היומית.

<u>היסטוגרמה להצגת התפלגות תשואות:</u>



ריכוז התשואות סביב ה0 אשר מצביע כי רוב ימי הסחר מאופיינים בשינויים יחסית קטנים, לא נראה הטיה מובהקת לכיוון חיובי או שלילי.

Corr Return Fed Rate	Corr Return SP500	Corr Return US10Y	Corr Return VIX
-0.02	0.01	-0.03	-0.13

symbol
- [x] AAPL
- [ ] GOOGL
- [ ] NVDA

ביצענו קורלציה ו מהניתוח עולה כי הקשר בין התשואה היומית למשתני המאקרו שנבדקו חלש באופן כללי. מדד ה־VIX בולט כמשתנה היחסי המשמעותי ביותר, בעוד ששאר המשתנים מראים קשר זניח לתשואה היומית.

Corr Return Fed Rate	Corr Return SP500	Corr Return US10Y	Corr Return VIX
0.01	0.04	0.00	-0.12

symbol
- [ ] AAPL
- [x] GOOGL
- [ ] NVDA

גם במקרה של GOOGL ניתן לראות כי מדד ה־VIX הוא המשתנה היחיד שמראה קשר כלשהו לתשואה היומית, בעוד שהשפעת יתר המשתנים כמעט ואינה ניכרת.

Corr Return Fed Rate	Corr Return SP500	Corr Return US10Y	Corr Return VIX
0.02	0.02	0.00	-0.10

symbol
- [ ] AAPL
- [ ] GOOGL
- [x] NVDA

גם עבור מניית NVDA מתקבל דפוס דומה, שבו מדד ה־VIX מציג את הקורלציה הגבוהה ביותר לתשואה היומית, בעוד שיתר משתני המאקרו מציגים קשר חלש מאוד או זניח.
ניתוח הקורלציה עבור AAPL, GOOGL ו-NVDA מראה דפוס עקבי שבו מדד ה־VIX הוא המשתנה המאקרו בעל הקשר החזק ביותר לתשואה היומית, אם כי ברמה חלשה.
לעומת זאת, משתני ריבית, אג״ח ומדד S&P500 מציגים קשר זניח, דבר המחזק את המסקנה כי ברמה היומית השפעתם מוגבלת.

את שני המדדים Inflation CPI ו־Consumer Confidence לא הכנסו לניתוח הקורלציה כי מדובר במדדים חודשיים המשתנים בקצב איטי, ולכן אינם מתאימים לניתוח קשרים ברמה של תשואה יומית, שבה נדרשת תגובה מהירה לאירועי שוק.

דו"ח התקדמות להגשה לשלב Data Preparation

שמות הסטודנטים ניקול בלישר נועם פטיש

מס' ת"ז 328654546 212205082

שם הפרויקט : חיזוי מניות

שם המנחה ג'ניס גוטפריד

לקחנו את הנתונים והעברו למסד נתונים של SQL SERVER, ואז לאחר מכן עשינו בדיקות באמצעות שאילתות במידה ויש NULL וכפילויות. ואז חיברנו את הטבלאות באמצעות מפתחות והגדרנו קשרים. ראינו לנכון להשאיר את הנתונים כיוון שהם משקפים ומשפיעים על השוק וזה מה שאנו מנסות לחזות באמצעות המודל. עשינו ניתוח EDA המשקף את הניתוחים הסטטיסטיים את המדדים, ותרשימים אשר פירטנו והסברנו כיצד המדדים משפיעים ואילו הכי הרבה על השוק.

חתימת המנחה _____

תאריך 25/02/2026

4. Modelling

מטרת שלב זה:

לבחור את המודלים בהם תשתמשו, להכיר אותם, לתכנן את הניסוי, לבנות פייפליין מסודר בקוד להרצת המודלים תוך קריאות אל מסד הנתונים של הפרויקט, בנוסף ביצוע הרצות ראשוניות של המודל/מודלים לקבלת תוצאות.

בדוח זה עליכם לפרט על מה שביצעתם (מה שלא רלוונטי ניתן לא לכלול):

- **בחירת שיטות מודל רלוונטיות -**
- Random Forest: מודל המבוסס על "יער" של עצי החלטה הפועלים במקביל לחיזוי תוצאה משותפת. בחרנו בו בשל יכולתו להתמודד עם נתונים פיננסיים תנודתיים מבלי לבצע "התאמת יתר" (Overfitting) ובזכות היציבות שלו בזיהוי קשרים מורכבים בין משתני מאקרו למחיר המניה.

- XGBoost: אלגוריתם מתקדם בשיטת ה-Boosting, הלומד מטעויות קודמות ומשפר את עצמו בכל איטרציה. בחרנו בו כיוון שהוא נחשב למודל החזק ביותר כיום לחיזוי נתונים טבלאיים, המאפשר להגיע לדיוק גבוה מאוד תוך ניצול יעיל של כוח עיבוד.

- Logistic Regression: מודל סטטיסטי קלאסי המשמש כנקודת ייחוס (Baseline) לבדיקת קשרים ליניאריים פשוטים. השתמשנו בו כדי להוכיח כי השוק מתנהג בצורה מורכבת שאינה ניתנת לחיזוי באמצעות קו ישר, מה שחיזק את הצורך במודלים המורכבים יותר שהרצנו.

- **תכנון ניסוי – (Test Design) -**

בשלב הראשון, חילקנו את הנתונים לפי שיטת החלוקה Train/Validation/Test.

בשלב השני, הגדרנו קריטריונים להצלחת המודל לפי המדדים הבאים המתבססים על מטריצת הבלבול:

- **Accuracy (דיוק כללי):** זהו המדד הראשוני ליציבות המודל. יעד ההצלחה הוגדר כ-60%, והמודל הנבחר (Random Forest) עקף אותו משמעותית עם 86% דיוק.

- **Precision (דיוק בחיזוי):** זהו המדד החשוב ביותר למשקיע כדי למנוע כניסה לעסקאות מפסידות (False Positive). המודל השיג 0.90 (90%) בקטגוריה זו, מה שמעיד על רמת אמינות גבוהה מאוד של סיגנל הקנייה.

- **Recall (זיהוי הזדמנויות):** המודל השיג 84% מה שמראה שהוא יודע "לתפוס" את רוב הזדמנויות הרווח בשוק ולא להישאר בחוץ בימים ירוקים (יום בו מחיר המניה נסגר במחיר גבוהה יותר ממחיר הפתיחה שלו או ממחיר הסגירה של יום הקודם).
- **F1-Score (המדד המאזן):** מדד זה (שעמד על 0.87) מוכיח שהמודל מאוזן - הוא לא רק "מנחש" עליות כל הזמן, אלא מבצע הפרדה איכותית בין ימי רווח לימי הפסד.

חלוקת נתונים - השתמשנו בשיטת Train/Validation/Test ולשם כך בחרנו לחלק את הנתונים לפי סדר כרונולוגי ולא באקראי מהסיבה שבשוק ההון לא ניתן לערבב זמנים, המודל חייב ללמוד מהעבר כדי לחזות את העתיד.

- לקחנו את 70% התאריכים הראשונים ביותר (4,155 שורות) כדי שהמודל יתאמן ויבין איך השוק מתנהג.
- המשכנו עם 15% נוספים (891 שורות) כקבוצת בדיקת ביניים, שדרכה אנחנו מוודאים שהמודל לא סתם מנחש ומבצעים התאמות.
- בסוף, שמרנו את ה-15% הכי עדכניים (891 שורות) למבחן הסופי והאמיתי - לראות איך המודל מתמודד עם נתונים חדשים לגמרי שהוא מעולם לא ראה.

החלוקה הזו היא קריטית כי היא מונעת זליגת מידע מהעתיד ומדמה מצב של מסחר אמיתי בזמן אמת.

- במידת הצורך יש להפעיל אלגוריתם **לבחירת מאפיינים (Feature selection)** – הפעלנו אלגוריתם בקוד להצגת ה5 מאפיינים שהכי השפיעו על התוצאות של המודל. הוא מדרג את המשתנים לפי תרומתם לדיוק הפיצולים בעצי ההחלטה, ככל שמשתנה מסוים עזר להפריד טוב יותר בין ימי עליות לימי ירידות, כך דירוגו במודל עלה.

<u>5 המאפיינים המובילים במודל Random Forest:</u>

```
daily_return_high    0.340273
daily_return_low     0.316228
VIX                  0.057080
volume               0.053226
SP500                0.046468
dtype: float64
```

<u>5 המאפיינים המובילים במודל Random Forest:</u>

```
daily_return_low       0.467775
daily_return_high      0.196680
adj_close              0.047603
CONSUMER_CONFIDENCE    0.041343
SP500                  0.039188
dtype: float32
<Figure size 800x600 with 0 Axes>
```

תיעוד הנחות והגדרות -

- הנחות לגבי נתונים:

1. הנחנו שהעבר משפיע על העתיד ולכן כשחילקנו את הנתונים ל-Train ו-Test, השתמשנו ב-shuffle=False. הנחנו שסדר הימים הוא קריטי.
2. הנחנו ששורות עם ערכים חסרים הן "רעש" שעלול להטעות את המודל ולכן הסרנו אותן.
3. הנחנו שהדבר הכי חשוב למשקיע הוא העלייה/ירידה של המניה ולא המחיר המדויק באגורות, ולכן הפכנו את הבעיה לחיזוי של 0 או 1.

- התאמות המודל:

1. המרנו את סוגי הנתונים שהגיעו מטבלאות הSQL כ"טקסט" או אובייקט. ביצענו התאמה והמרנו אותם למספרים כי המודל שלנו יודע לבצע חישובים רק על מספרים.
2. טיפלנו בערכים חסרים על ידי ניקוי הטבלאות משורות ריקות כדי למנוע שגיאות אימון.
3. יצרנו עמודה חדשה של 0 ו1 לצורך דיוק המודל והחלטה עסקית.
4. סידרנו את כל ה-DataFrame לפי תאריך לפני החלוקה, כדי להבטיח שהמודל לומד על תקופה מוקדמת ובודק את עצמו על תקופה מאוחרת יותר.

- **הציגו תוצאות ראשוניות של הרצת מודל**

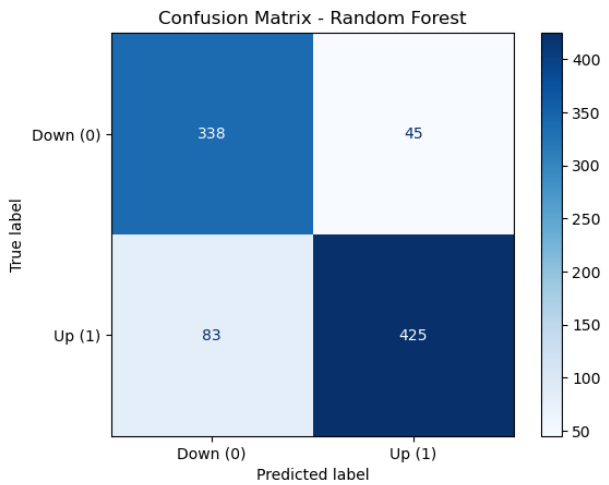
Random Forest

```
Classification Report - Random Forest:
              precision    recall  f1-score   support

           0       0.80      0.88      0.84       383
           1       0.90      0.84      0.87       508

    accuracy                           0.86       891
   macro avg       0.85      0.86      0.85       891
weighted avg       0.86      0.86      0.86       891
```

<u>Confusion Matrix</u>



הפרמטרים שהשתמשנו במודל:

```
--- Features Importance: Random Forest ---
daily_return_high       0.340273
daily_return_low        0.316228
VIX                     0.057080
volume                  0.053226
SP500                   0.046468
open_price              0.044214
adj_close               0.043690
US10Y                   0.041024
CONSUMER_CONFIDENCE     0.022106
INFLATION_CPI           0.019378
FED_RATE                0.016313
dtype: float64
```

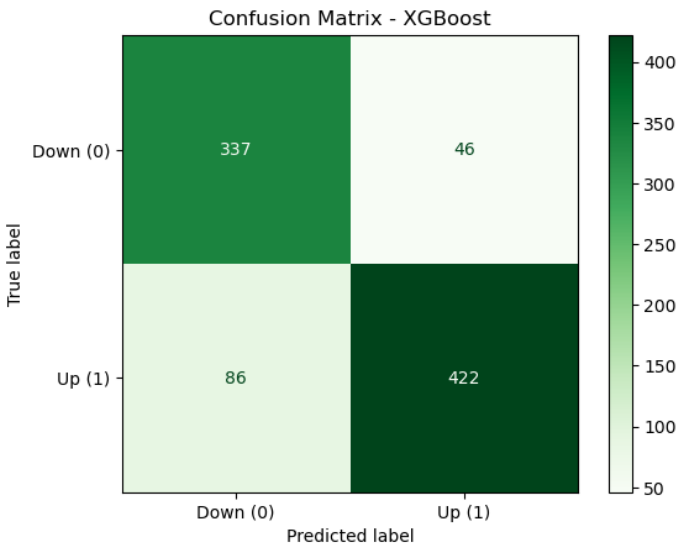
XGBoost

```
Classification Report - XGBoost:
              precision    recall  f1-score   support

           0       0.80      0.88      0.84       383
           1       0.90      0.83      0.86       508

    accuracy                           0.85       891
   macro avg       0.85      0.86      0.85       891
weighted avg       0.86      0.85      0.85       891
```

Confusion Matrix



הפרמטרים שהשתמשנו במודל:

```
--- Features Importance: XGBoost ---
daily_return_low        0.467775
daily_return_high       0.196680
adj_close               0.047603
CONSUMER_CONFIDENCE     0.041343
SP500                   0.039188
VIX                     0.036697
FED_RATE                0.036138
open_price              0.034511
INFLATION_CPI           0.034249
US10Y                   0.033923
volume                  0.031893
dtype: float32
```

מודל ה Random Forest - נבחר כמודל המועדף מאחר שהשיג דיוק כללי מעט גבוה יותר (כ-86%) בהשוואה למודל XGBoost שהגיע לכ-85%. בנוסף, ערכי ה Precision, Recall-ו F1-Score-במודל זה מצביעים על איזון טוב בין זיהוי נכון של ימי עלייה לבין הימנעות מסיווגים שגויים. ערך F1-Score הגבוה יחסית מעיד כי המודל מצליח לשלב בצורה יעילה בין דיוק החיזוי לבין כיסוי נכון של המקרים החיוביים. רמת ה Precision-הגבוהה מצביעה על כך שרוב התחזיות החיוביות אכן תואמות את התוצאה בפועל, דבר שמפחית את הסיכון לקבל החלטות שגויות. לכן, בשל השילוב בין דיוק מעט גבוה יותר ויציבות במדדי הביצוע, נבחר מודל Random Forest להמשך העבודה והניתוח בפרויקט.

האקדמית עמק יזרעאל
ע"ש מקס שטרן

דו"ח התקדמות להגשה לשלב Modelling

שמות הסטודנטים : נועם פטיש וניקול בלישר

מס' ת"ז : 212205082, 328654546

שם הפרויקט : חיזוי מניות

שם המנחה : ג'ניה גוטפריד

בשלב זה חיברנו את המסד נתונים שנמצא בSQL לקוד בפייתון להרצת המודל.
שלפנו נתונים, ניקינו ערכים חסרים, המרנו טקסטים ואובייקטים למספרים לצורך חישובים.
לאחר מכן, ביצענו חלוקת נתונים לפי שיטת Train/Validation/Test , בחרנו כמה מודלים לצורך השוואה ובדיקה של המודל.
ביצענו הרצה ראשונית ומצאנו כי באלגוריתם של Random Forest כי רמת הדיוק יותר גבוהה ולכן אנו רואות לנכון לבחור בשיטה זו בהמשך הפרויקט שלנו.

חתימת הסטודנט : נועם פטיש

חתימת הסטודנט : ניקול בלישר

חתימת המנחה _____

תאריך 13/04/2026

השוואת מודלים על בסיס קריטריונים בחרנו את DAILIY RETURN LOW-HIGH :

במהלך שלב ההערכה זוהה כי המשתנים daily_return_high ו־daily_return_low עשויים ליצור מצב של Data Leakage, מאחר שהם מבוססים על נתוני מסחר מאותו יום של התחזית.
החשד לכך עלה לאחר ניתוח Feature Importance, שבו שני משתנים אלו קיבלו את ערכי החשיבות הגבוהים ביותר במודל בהשוואה לשאר המשתנים.
דבר שהראה כי המודל נשען עליהם בצורה משמעותית לצורך ביצוע החיזוי.

```
# 5. המשתנים שהכי השפיעו
importances = pd.Series(model_rf.feature_importances_, index=X_train.columns).sort_values(ascending=False)
print("\nהמשתנים שהכי השפיעו על החיזוי:")
print(importances.head(5))
```

```
המשתנים שהכי השפיעו על החיזוי:
daily_return_high    0.340273
daily_return_low     0.316228
VIX                  0.057080
volume               0.053226
SP500                0.046468
```

על מנת לבדוק האם המודלים אכן מסתמכים על מידע שאינו זמין בזמן אמת למשקיע, בוצע ניסוי נוסף שבו המשתנים daily_return_high ו־daily_return_low הוסרו ממערך הנתונים.

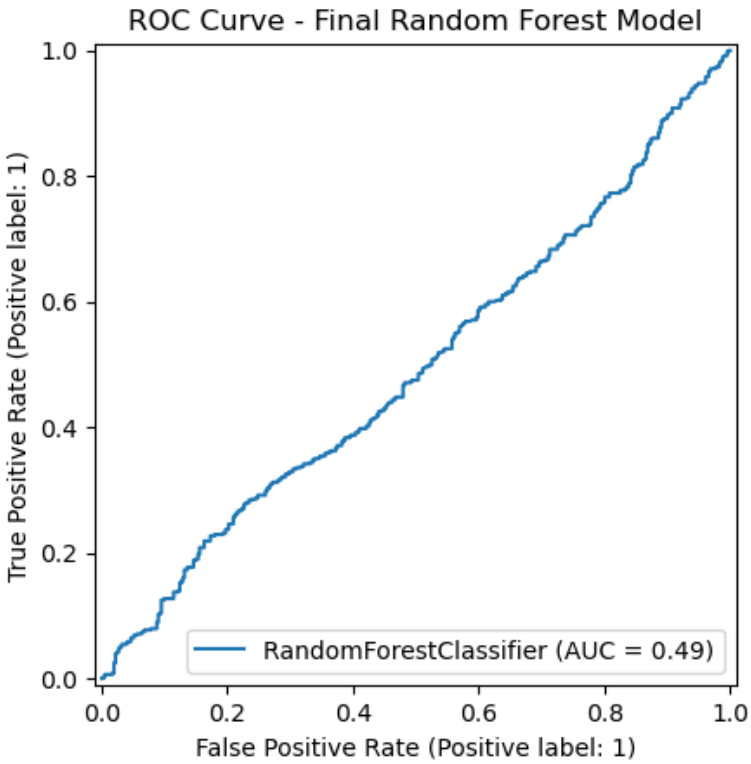
```
X = df.drop(columns=[
    'stock_return',
    'date',
    'symbol',
    'daily_return_high',
    'daily_return_low'
], errors='ignore')
```

לאחר הסרתם נצפתה ירידה משמעותית בביצועי המודלים :

- Random Forest ירד מ־86% Accuracy ל־52%
- XGBoost ירד מ־85% Accuracy ל־55%

ממצא זה מחזק את ההשערה כי המודלים הסתמכו במידה רבה על מידע הקשור ישירות לתוצאת יום המסחר עצמו, ולכן תוצאות הניסוי השני נחשבות לאמינות וריאליסטיות יותר לצורך חיזוי אמיתי של שוק ההון.

דירוג מודים לפי מדדים כמותיים:



בוצע חישוב ROC-AUC לצורך הערכה נוספת של יכולת הסיווג של המודל. תוצאת ה־AUC שהתקבלה הייתה 0.49, ערך הקרוב ל־0.5, דבר המעיד כי יכולת ההבחנה של המודל בין עלייה לירידה במניה נמוכה יחסית ודומה לביצוע אקראי.

גם גרף ה־ROC הציג עקומה הקרובה לאלכסון, דבר המחזק את המסקנה כי חיזוי תנועות יומיות בשוק ההון הוא משימה מורכבת במיוחד לאחר הסרת משתנים שיצרו Data Leakage.

עם זאת, תוצאות אלו נחשבות לאמינות יותר, מאחר שהמודל נבחן על נתונים שלא נראו במהלך האימון וללא שימוש במידע שעלול לדלוף מתוצאת יום המסחר עצמו.

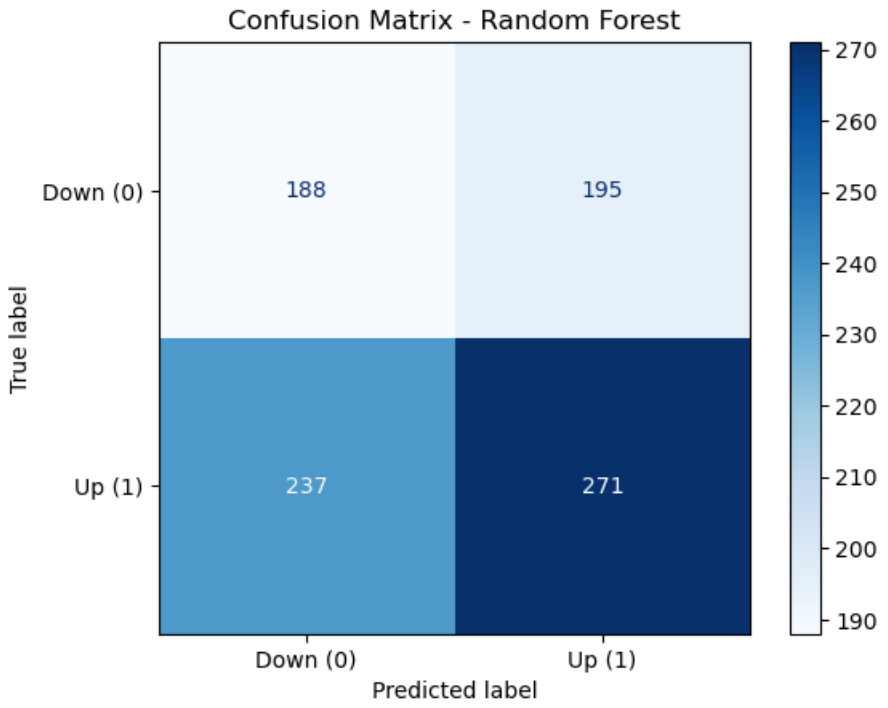
השוואת המודלים ובחירת מדדי ההערכה:

Model	Accuracy	Precision	Recall	F1
RF מקור	0.86	0.90	0.84	0.87
XGB מקור	0.85	0.90	0.83	0.86
RF ללא leakage	0.52	0.58	0.53	0.56
XGB ללא leakage	0.55	0.59	0.66	0.62

מהטבלה ניתן לראות כי לאחר הסרת משתני ה־ Data Leakage חלה ירידה משמעותית בביצועי שני המודלים. עם זאת, תוצאות אלו משקפות בצורה אמינה יותר את יכולת החיזוי האמיתית של המודלים, מאחר שהן אינן מסתמכות על מידע שאינו זמין בזמן אמת.

בפרויקט זה בחרנו להתמקד במדדי Precision ו־ F1-score. הסיבה לכך היא שבחיזוי שוק ההון חשוב לא רק לדעת כמה תחזיות היו נכונות באופן כללי, אלא גם עד כמה ניתן לסמוך על תחזיות העלייה שהמודל מספק למשקיע.
מדד Precision בוחן את אחוז תחזיות העלייה שהיו נכונות בפועל, ולכן הוא חשוב במיוחד לצמצום מצבים שבהם המודל ממליץ על השקעה שבסופו של דבר אינה משתלמת. בנוסף, מדד F1-score משלב בין Precision לבין Recall ומספק הערכה מאוזנת יותר של ביצועי המודל. מסיבה זו התבססנו בעיקר על מדדים אלו בעת הערכת המודלים ובחירת המודל הסופי.

:CONFUSION MATRIX



בנוסף למדדי הביצועים, נבנתה גם Confusion Matrix לצורך ניתוח מפורט יותר של תחזיות המודל.
מהמטריצה ניתן לראות כי המודל הצליח לזהות בצורה טובה יותר מצבים של עלייה במניה (Up) בהשוואה למצבים של ירידה (Down) .
המודל חזה נכון 271 מקרים של עלייה במניה ו־188 מקרים של ירידה. עם זאת, ניתן לראות כי קיימים גם מספר לא קטן של סיווגים שגויים, דבר המעיד על הקושי בזיהוי מדויק של תנועות יומיות בשוק ההון.
ממצאים אלו תואמים למדדי ה־ Accuracyוה־ AUC שהתקבלו, ומחזקים את המסקנה כי לאחר הסרת משתנים שיצרו Data Leakage, ביצועי המודל הפכו לריאליסטיים יותר אך גם מאתגרים יותר מבחינת יכולת החיזוי.

Fine Tuning וכוונון פרמטרים:

```
n_estimators=200
max_depth=10
min_samples_split=5
```

בשלב ה־Fine Tuning בוצעו מספר ניסויים למודל Random Forest באמצעות שינוי פרמטרים מרכזיים, כגון מספר העצים במודל (n_estimators) ועומק העצים (max_depth). בתחילה בוצע כוונון בסיסי באמצעות קוד ידני פשוט, שבו נוסו מספר ערכים בודדים לפרמטרים המרכזיים. בניסוי זה התקבל שיפור קטן בלבד, כאשר ביצועי המודל השתפרו מ־52% ל־53% Accuracy.

Fine Tuning מורחב:

```
models_to_test = {
    "RF_Default": RandomForestClassifier(n_estimators=100, random_state=42),

    "RF_200_depth10": RandomForestClassifier(
        n_estimators=200,
        max_depth=10,
        random_state=42
    ),

    "RF_300_depth15": RandomForestClassifier(
        n_estimators=300,
        max_depth=15,
        random_state=42
    )
}

for name, model in models_to_test.items():

    model.fit(X_train, y_train_class)

    y_pred = model.predict(X_val)

    print("\n", name)

    print(f"Accuracy: {accuracy_score(y_val_class, y_pred):.2f}")

    print(classification_report(y_val_class, y_pred))
```

לאחר מכן הוחלט לבצע תהליך Fine Tuning רחב ומסודר יותר, שבו הורחב הקוד ונבדקו מספר שילובים שונים של פרמטרים עבור מודל Random Forest. מטרת התהליך הייתה לבדוק כיצד שינוי הפרמטרים משפיע על יכולת החיזוי של המודל.

במסגרת הניסויים נמצא כי המודל בעל n_estimators=300 ו־max_depth=15 השיג את התוצאה הטובה ביותר עם Accuracy של 55%, זה תוצאה טובה יותר מ53%.
למרות השיפור שהתקבל לאחר כוונון הפרמטרים, מדובר בשיפור מתון יחסית, דבר שממחיש את המורכבות של חיזוי מגמות יומיות בשוק ההון לאחר הסרת משתנים שעלולים לגרום ל־Data Leakage.

ביצוע הרצה על נתונים חדשים - TEST:

```
y_test_class = (y_test > 0).astype(int)

final_model = RandomForestClassifier(
    n_estimators=300,
    max_depth=15,
    random_state=42
)

final_model.fit(X_train, y_train_class)

y_pred_test = final_model.predict(X_test)
print()
print("Final Test Results-RANDOM FOREST")

print(f"Accuracy: {accuracy_score(y_test_class, y_pred_test):.2f}")

print(classification_report(y_test_class, y_pred_test))
```

תוצאות הרצה ראשונית על ה-TEST:

```
Final Test Results-RANDOM FOREST
Accuracy: 0.50
              precision    recall  f1-score   support

           0       0.45      0.43      0.44       411
           1       0.53      0.55      0.54       480

    accuracy                           0.50       891
   macro avg       0.49      0.49      0.49       891
weighted avg       0.50      0.50      0.50       891
```

לאחר בחירת המודל הסופי וביצוע תהליך Fine Tuning, בוצעה הרצה סופית על קבוצת ה־Test, אשר הופרדה מראש ממערך הנתונים ולא שימשה במהלך האימון או תהליך הכוונון. החלוקה בוצעה כך ש־70% מהנתונים שימשו לאימון המודל, 15% ל־Validation לצורך השוואת מודלים וכוונון פרמטרים, ו־15% נוספים נשמרו כקבוצת Test לבדיקה הסופית בלבד.

מטרת שלב זה הייתה לבדוק את יכולת ההכללה של המודל על נתונים שלא נראו במהלך תהליך הלמידה. תוצאות ההרצה הסופית הראו כי מודל Random Forest השיג Accuracy של 50% על קבוצת ה־Test, לצד ערכי Precision ו־F1-score מאוזנים יחסית בין מחלקות העלייה והירידה.

ביצוע השוואה עם ה-Validation:

```
y_train_class = (y_train > 0).astype(int)
y_val_class = (y_val > 0).astype(int)

# 1. הגדרת המודל
model_rf = RandomForestClassifier(n_estimators=100, random_state=42)

# 2. אימון המודל (שימי לב שאנחנו משתמשים ב-y_train_class החדש)
model_rf.fit(X_train, y_train_class)

# 3. ביצוע תחזית על נתוני ה-Validation
y_pred_val = model_rf.predict(X_val)

# 4. הדפסת התוצאות (משווים מול y_val_class)
print(" תוצאות הרצה ראשונית (Validation Set)")
print(f"Accuracy (דיוק כללי): {accuracy_score(y_val_class, y_pred_val):.2f}")

print("\nClassification Report - Random Forest:")
print(classification_report(y_val_class, y_pred_val))
```

תוצאות הרצה ראשונית על ה-Validation:

```
תוצאות הרצה ראשונית (Validation Set)
Accuracy (דיוק כללי): 0.52

Classification Report - Random Forest:
              precision    recall  f1-score   support

           0       0.44      0.49      0.47       383
           1       0.58      0.53      0.56       508

    accuracy                           0.52       891
   macro avg       0.51      0.51      0.51       891
weighted avg       0.52      0.52      0.52       891
```

בהשוואה לתוצאות שהתקבלו על קבוצת ה־Validation, נצפתה ירידה מסוימת בביצועי המודל. ממצא זה מצביע על כך שחיזוי תנועות יומיות בשוק ההון הוא משימה מורכבת מאוד, במיוחד לאחר הסרת משתנים שעלולים לגרום ל־Data Leakage. למרות הירידה בביצועים, תוצאות אלו נחשבות לאמינות יותר מאחר שהמודל נבחן על נתונים חדשים לחלוטין מבחינת תהליך הלמידה.

דו"ח התקדמות להגשה לשלב Evaluation

שמות הסטודנטים ניקול בלישר נועם פטיש

מס' ת"ז 328654546 212205082

שם הפרויקט חיזוי מניות

שם המנחה ג'ניה גוטפריד

בשלב זה ביצענו תהליך Evaluation למודלי החיזוי בפרויקט באמצעות חלוקת הנתונים ל־ Train, Validation ו־Test. הורצו מודלי Random Forest ו־XGBoost ונמדדו ביצועיהם באמצעות Accuracy, Precision, F1-score ו־ROC-AUC. במהלך העבודה זוהתה בעיית Data Leakage במשתנים daily_return_high ו־daily_return_low, ולכן בוצע ניסוי נוסף ללא משתנים אלו לקבלת תוצאות אמינות יותר. בנוסף בוצע Fine Tuning למודל Random Forest ונבחר המודל בעל הביצועים הטובים ביותר. לבסוף בוצעה הרצה סופית על קבוצת Test והמודל נשמר בקובץ PKL לצורך שימוש עתידי בשלב ה־Deployment.

חתימת הסטודנט ניקול בלישר

חתימת הסטודנט נועם פטיש

חתימת המנחה _____

תאריך _____09/06/2026_____