דוח הכנת הנתונים

השקעות בסיכון נמוך זה אצלנו-NexTrade

מגישים – אריאל קריב ומשי בר

תקציר

דוח זה מסכם את תהליך עיבוד הנתונים שנעשה לקראת שלב המידול בפרויקט. הנתונים נאספו, נוקו, ועובדו על מנת להתאים לדרישות המודלים שנעשה בהם שימוש. **LSTM** לניתוח סדרות זמן ו **BERT**-לניתוח חדשות כלכליות. כמו כן, חושבו אינדיקטורים פיננסיים ונעשה ניתוח חקר נתונים (EDA) לזיהוי מגמות וקשרים בין המשתנים.

תוכן עניינים

1. בחירת מקורות הנתונים

לפרויקט זה השתמשנו בשני מקורות עיקריים:

1. נתוני חדשות כלכליות- יובאו באמצעות API פרימיום של Alpha Vantage.

2. נתוני מדד S&P 500- התקבלו מהאתר Yahoo Finance (YFinance) הכוללים מחירי מניות יומיים.

הצדקה לבחירת הנתונים

- מאגר החדשות הכלכליות מספק ניתוח סנטימנט של אירועים כלכליים, המסייע בהבנת השפעות השוק.

- נתוני מדד S&P 500 מכילים נתוני מחירים היסטוריים ואינדיקטורים טכניים (MACD, RSI), אשר עוזרים בזיהוי מגמות השוק.

- מחירי המדד מושפעים מגורמים כלכליים שונים כגון החלטות ריבית, נתוני תעסוקה, אינפלציה, תוצר מקומי גולמי ומדדים פיננסיים נוספים.

- בשלב זה, שני מקורות הנתונים אינם צריכים להתמזג לטבלה אחת, שכן כל אחד מהם ייכנס למודל שמתאים לטיפוס הנתונים שלו. כפי שצוין בדוח הקודם:

מודל (Long Short-Term Memory) LSTM - יישום לניתוח נתוני S&P 500. מודל זה מתאים במיוחד לניתוח סדרות זמן, ולכן ישמש לחיזוי מגמות במדד S&P 500 בהתבסס על אינדיקטורים טכניים כמו MACD ו-RSI.

מודל (Bidirectional Encoder Representations from Transformers) BERT - יישום לניתוח חדשות כלכליות. זהו מודל מתקדם לניתוח שפה טבעית, המאפשר להבין את ההקשר של הסנטימנט הכלכלי וכיצד הוא עשוי להשפיע על השוק.

שילוב הנתונים בהמשך הפרויקט יתבצע בצורה עקבית כך שהאימון הראשוני של המודלים יתבצע בנפרד:

-מודל LSTM ינותח על סדרות הזמן של נתוני המדד בלבד (סדרת זמן של 30 יום).

-מודל BERT ינותח על חדשות כלכליות בלבד.

-לאחר קבלת התחזיות מכל מודל, ניתן לבצע "איחוד" של התוצאות בהתאם לתאריכים.

נוכל לחבר בין שני סוגי הנתונים על-ידי יצירת Pipeline חכם שמאפשר שילוב של הקלטים משני המודלים לתוך מודל רב-ערוצי (Multi-Input Neural Network).

Pipeline- (צינור נתונים) הוא תהליך עבודה אוטומטי שבו הנתונים עוברים דרך שלבים שונים באופן מסודר ומובנה. במקרה שלנו, ה-Pipeline יבטיח שכל הנתונים יוזנו בצורה נכונה לכל מודל, שהתוצרים שלהם יישמרו בצורה מתואמת, ושבסופו של דבר ניתן יהיה לשלב אותם יחד באופן עקבי ומדויק.

-מודל רב-ערוצי (Multi-Input Neural Network) הוא רשת נוירונים שיכולה לקבל שני סוגי קלט שונים במקביל ולנתח אותם יחד. במקרה שלנו :

ערוץ אחד יקבל נתוני סדרות זמן של (LSTM) S&P 500.

ערוץ שני יקבל נתוני טקסט מהחדשות הכלכליות (BERT).

השכבות של כל אחד מהמודלים יעובדו בנפרד, אך לבסוף יתמזגו לשכבה משותפת, שתאפשר לנו לקבל תחזית משוקללת המשלבת גם מידע חדשותי וגם נתוני שוק בפועל.

חשוב לציין, אין צורך לאחד מראש את קבצי הנתונים לטבלה אחת, שכן כל מקור נתונים יאומן בנפרד על פי המודל המתאים לו, ורק לאחר מכן נבחן כיצד לחבר בין התחזיות השונות לטובת קבלת תובנות חזקות יותר.

בחירת שורות

- טבלת נתוני חדשות כלכליות :

בטבלה זו ריכזנו רק חדשות כלכליות העוסקות במדדים מרכזיים למשל כמו- CPI, החלטות ריבית, נתוני תעסוקה וכו'. לכן נבחרו רק כותרות ידועות ורלוונטיות כפי שצוין בדוח הקודם, זאת מתוך הבנה שלכותרות הראשיות יש השפעה משמעותית על סנטימנט השוק. הכותרות שנבחרו-

- "Core CPI"
- "Average Hourly Earnings"
- "S&P Global Services PMI"
- "ISM Manufacturing PMI"
- "CPI"
- "GDP"
- "Fed Interest Rate Decision"
- "Core PCE Price Index"
- "ISM Manufacturing Prices"
- "Unemployment Rate"
- "ECB Interest Rate Decision"
- "CB Consumer Confidence"
- "JOLTS Job Openings"
- "Crude Oil Inventories"

כותרות אלו נבחרו באופן מדויק מכיוון שהן התיאור לחדשות הכלכליות הקבועות והרלוונטיות ממקורות מסחר שמסווג כל חדשה לפי החשיבות שלה על השוק (לדוגמא אתר investing).

עקב מגבלות זמן ואיכות המידע, הוגבלו הנתונים לשלוש שנים בלבד ולא לחמש שנים כפי שתוכנן בתחילת הכנת הפרויקט. טבלה זו מכילה 36,628 שורות.

שורות כפולות הוסרו כדי למנוע השפעה על ניתוח המודל.

-טבלת נתוני מדד S&P 500 :

בטבלה זו ריכזנו את כל הנתונים היומיים מהבורסה של SPY (S&P 500) לתקופה של 7 שנים (2018-2025), תוך הכללת כל יום מסחר מלבד סופי שבוע וחגים שבהם לא התקיים מסחר.

למרות שהמיקוד המרכזי בפרויקט הוא ממרץ 2020 עד 2025, הוחלט להוריד נתונים מ-2018 כדי לאפשר ניתוח מגמות לטווח רחב יותר וגם לשפר את טווח החישוב המדויק של המדדים אשר מבוססי חישוב ממוצעים על סמך נתוני ימי מסחר הקודמים להם. טבלה זו מכילה 1,575 שורות.

בחירת עמודות- טבלת החדשות הכלכליות

עמודות שנבחרו מטבלת מהחדשות הכלכליות :

- Date- תאריך הפרסום של הכתבה (משמש כמפתח לקישור לטבלת נתוני מדד s&p500).

- Title- כותרת הכתבה הכלכלית, מספקת מידע ישיר על הנושא המדובר.

- Authors- שדה זה הומר לערכים בינאריים (0 אם אין מחבר, 1 אם יש מחבר), כדי לבדוק האם יש השפעה על הטון החדשותי לקיום שם מחבר. אם כן נמצא השפעה או מגמה מסוימת, נבדוק אם קיימת השפעה לזהות המחבר החדשה.

- Summary- סיכום הכתבה אשר מספק תיאור נוסף לתוכן החדשותי.

- Source- מקור הכתבה, מסייע להבין את ההטיה האפשרית במידע.

- Category_within_source- קטגוריה של הכתבה בתוך המקור (כגון "מקרו כלכלה", "שוק ההון").

- Topics- נושא הכתבה (CPI, החלטות ריבית, אבטלה וכו').

- Overall Sentiment Score- ציון רמת הסנטימנט של הכתבה.

- Overall Sentiment Label- קטגוריזציה מילולית של רמת הסנטימנט (Bullish, Bearish וכו').

- Ticker Sentiment- רמת הסנטימנט הספציפי כלפי מניות או מדדים מסוימים.

עמודות שהוסרו- טבלת החדשות הכלכליות

- URL- אינו נמצא כתורם למודל.

- -Source_domain- מידע זהה לעמודת `Source`.

- Banner_image- תמונת הכתבה אינה רלוונטית למידול.

עמודות שנבחרו ממדד S&P 500:

- Date- תאריך יום המסחר.

- Close- מחיר הסגירה של המדד.

- High- המחיר הגבוה ביותר שנרשם באותו יום מסחר.

- Low- המחיר הנמוך ביותר שנרשם באותו יום מסחר.

- Open- מחיר הפתיחה של המדד באותו יום.

- Volume- כמות המניות שנמסרו במהלך יום המסחר.

- MACD- אינדיקטור ממוצעים נעים (משמש לזיהוי מומנטום בשוק).

- RSI- מדד חוזק יחסי (לזיהוי מצב של קנייה יתרה/מכירה יתרה).

לא בוצעו הסרות של עמודות בטבלת נתוני המדד אשר מכילות מידע נכון לחלק זה, בכדי לבדוק את חשיבות המדדים והשפעתם במודל עצמו. לעומת זאת, שינויים קוסמטיים ועיצוביים בוצעו ישירות בקובץ האקסל של נתוני המדד ללא צורך בשימוש קוד וללא השפעה על אמינות ואיכות הנתונים. השינויים שביצענו-

1. מחיקת תוכן תא 1A שהכילה כותרת לשורת שמות העמודה "Price".

2. העברת עמודת "DATE" מ1C ל1A יחד עם שורת כותרות תוכל העמודות בטבלה.

3. מחיקת תוכן שורה 2 מ-A עד -F שבתאים אלו היה פשוט רשום "SPY ".

4. צמצום הרווח שנוצר משורה 4 שבו הנתונים של העמודות כבר היו רשומים לשורה מספר 2 במצוד לשורת הכותרות עמודה.

2. ניקוי הנתונים

ניקוי הנתונים הוא שלב חיוני שבו מתוקנים, מוסרים או מתעדכנים נתונים כדי להבטיח שהם תקינים, מסודרים וללא בעיות. שלב זה כלל תיקון מבנה תאריכים, הסרת עמודות לא רלוונטיות, טיפול בכפילויות, עדכון ערכים חסרים, המרת פורמטים ושמירת נתונים נקיים להמשך המידול. חשוב לציין, כל השינויים והתהליכים שיצוינו בחלק זה עד שיצוין אחרת בוצעו על נתוני החדשות הכלכליות בלבד.

תיקון מבנה התאריכים

ביצענו המרה של time_published בפורמט YYYYmmddTHHMMSS לפורמט DD/MM/YYYY כדי ליצור עקביות עם נתוני S&P 500 בכל שורות טבלת החדשות הכלכליות.

בנוסף, ביצענו שינוי בשם השדה time_published → Date בכל שורות טבלת החדשות הכלכליות כדי להבטיח התאמה במיזוג עתידי במידת הצורך עם טבלת נתוני המדד.

אחידות בפורמט התאריכים תאפשר שילוב נתונים חלק בניתוחים עתידיים שנבצע במידת הצורך.

הסרת עמודות לא רלוונטיות

- banner_image – תמונת הכתבה (לא רלוונטי למידול).
- url – קישור חיצוני למקור (לא משפיע על הניתוח).
- source_domain – חופף לשדה source, ולכן הוסר.

הסרת מידע שאינו תורם למודלים ולא משמש לצורך חיזוי מסייע לדיוק מציאת החיזויים המדויקים ללא הפרעות.

הסרת כפילויות

- בכדי למנוע חזרה של נתונים שעלולים לגרום להטיה של המודל ביצענו ווידא של מחיקת שורות להן יש כפילויות.
- סך השורות הכפולות שאותרו ונמחקו : 476,382.
  - הסיבה לכמות המחיקות/כפילויות היא שנמצא לאחר הורדת הנתונים שהורדנו את הנתונים בטעות 4 פעמים לאותה בטבלה, כך נוצרו הרבה כפילויות שהסרנו.

טיפול בערכים חסרים

- category_within_source – כ- 19,781 ערכים חסרים הוחלפו ב-"Nan".
- Topics – כ- 2,151 שורות בהן לא צוין נושא, הומרו לערך "No Topic".
- ticker_sentiment – כ- 5,730 שורות בהם חסר מידע על מניה מסוימת, הומר לערך "No Ticker".

כך ניתן יהיה להתייחס למקרים שבהם המידע חסר בלי להשפיע על מודל החיזוי.

המרת עמודת Authors לערכים בינאריים

המרה של עמודת authors לערכים 0/1:

1 – אם קיים שם מחבר.

0 – אם השדה ריק או מכיל [].

כך נוכל לבדוק אם יש השפעה של קיום שם מחבר על הטון החדשותי.

אם יימצא קשר משמעותי, נוכל לבדוק גם את ההשפעה של זהות המחבר בהמשך.

המרת קטגוריות סנטימנט לכתב אחיד-אותיות קטנות

המרה זו בוצע על מנת למנוע התפלגות לא אחידה של קטגוריות הסנטימנט. מכך המשתנים יותר מדויקים עבור המודל ונמענו שגיאות טקסט מיותרות. נמצא כי בתהליך ההמרה- 36,625 שורות הומרו והותאמו לפורמט במוקש להצלחת הפרויקט.

שמירת הנתונים לאחר כל השינויים

הנתונים הנקיים נשמרו בקובץ CSV חדש לצורך המשך העבודה. זהו קובץ הנתונים הסופי שנשמש בו לשלב המידול.

לסיכום, ניקוי הנתונים בוצע במטרה להבטיח איכות, עקביות והתאמה למידול. תהליך זה כלל מספר שלבים חיוניים שנועדו להסיר רעשים, לשפר את מבנה הנתונים ולשמור על עקביות בין מקורות שונים. במהלך התהליך זוהו מספר בעיות ורעשים בנתונים, כולל חוסר אחידות בתאריכים, כפילויות בין קבצים שונים, מידע חסר ועמודות לא רלוונטיות. גישות מוצלחות ששימשו להסרת רעשים כללו סינון נתונים, התאמת פורמטים, המרת ערכים חסרים ויצירת אחידות בתכונות קריטיות.

כמוסבר למעלה חלק מהשדות המקוריים לא נכללו בקובץ הסופי ,מאחר שהם לא סיפקו ערך משמעותי למידול או יצרו חוסר עקביות. הסרת עמודות אלו נועדה למנוע עומס נתונים מיותר ולהתמקד במידע החשוב באמת לניתוח ולתחזיות.

3. דוח בניית נתונים חדשים

בשלב זה, בוצעה בנייה של נתונים חדשים כדי להעשיר את מערך הנתונים ולשפר את דיוק המודלים החיזויים. הנתונים החדשים נגזרו מהנתונים הקיימים, תוך ביצוע התאמות ושיפור איכותם כך שיהיו מתאימים למודלים מבוססי למידת מכונה.

יש לציין, לפני שהגענו לבניית הדוח הנוכחי, יצרנו את העמודות הללו באמצעות נתוני המסחר מ- Yahoo Finance (YFinance), כך שנתוני RSI ו-MACD כבר חושבו ונוספו לנתונים שלנו. במידת הצורך, אם נזדקק למדדים נוספים, לא נצטרך לחשבם באופן ידני, מאחר שרכשנו מנוי פרימיום באתר Alpha Vantage, שמאפשר ייבוא של מדדים כלכליים נוספים שיכולים להיות רלוונטיים להמשך הפרויקט.

יצירת תכונות חדשות מנתוני המדד S&P 500

<u>חישוב RSI (Relative Strength Index)</u>

RSI הוא אינדיקטור טכני פופולרי שמודד חוזק יחסי של מחירים, על-ידי השוואת ימי עליות מול ימי ירידות בטווח זמן מוגדר.

כיצד הוא מחושב ?

- RSI מבוסס על 14 הימים האחרונים ומחשב את היחס בין הרווחים הממוצעים לבין ההפסדים הממוצעים.

- לאחר מכן, הערך מנורמל לטווח של 0-100.

משמעות ערכי ה-RSI :

- RSI > 70 → השוק קנוי מדי (Overbought) → עשויה להיות ירידה קרובה.

- RSI < 30 → השוק נמצא במכירת יתר (Oversold) → עשויה להיות עלייה קרובה.

<u>חישוב MACD (Moving Average Convergence Divergence)</u>

MACD הוא אינדיקטור טכני חשוב המסייע לזהות שינויים במגמות השוק.

כיצד הוא מחושב?

- חישוב EMA (Exponential Moving Average) :

  - EMA קצר = 12 ימים → מגיב מהר לשינויים בשוק.

  - EMA ארוך= 26 ימים → מזהה מגמות יציבות יותר.

- MACD מחושב כהפרש בין שני הממוצעים :

  - MACD חיובי → מעיד על מגמה חיובית (Bullish).

  - MACD שלילי → מעיד על מגמה שלילית (Bearish).

התאמת הנתונים לדרישות המודל

<u>נרמול נתונים מספריים</u>

ביצענו נרמול ל-RSI ול-MACD לטווח `[0,1]` כדי למנוע מצב שבו משתנה אחד ישפיע יותר מדי על המודל.

- נרמול מבטיח שכל התכונות יישקלו באותו סדר גודל, ללא קשר לסקאלה המקורית שלהן.

4. דוח שילוב נתונים

שילוב הנתונים בהמשך הפרויקט יתבצע בצורה עקבית, כאשר כל מודל יאומן בנפרד בהתאם לסוג הנתונים שלו. מודל LSTM ינותח על סדרות הזמן של מדד S&P 500, תוך שימוש בחלונות זמן של 30 יום, ואילו מודל BERT ינותח על חדשות כלכליות בלבד. לאחר קבלת התחזיות מכל אחד מהמודלים, ניתן יהיה לבצע איחוד של התוצאות בהתאם לתאריכים, כך שנוכל לזהות קשרים בין הסנטימנט החדשותי לבין ביצועי המדד.

כדי לאחד את המידע באופן חכם, נשתמש ב-Pipeline אוטומטי, שיבטיח שכל הנתונים יוזנו כראוי לכל מודל, שהתוצרים שלהם יישמרו בצורה מתואמת, ושבסופו של דבר ניתן יהיה למזג אותם יחד באופן עקבי ומדויק. מודל רב-ערוצי (Multi-Input Neural Network) ישמש לשלב זה, כאשר ערוץ אחד יקבל נתוני סדרות זמן של (LSTM) S&P 500, וערוץ שני יקבל נתוני טקסט מהחדשות הכלכליות (BERT). השכבות של כל אחד מהמודלים יעובדו בנפרד, אך לבסוף יתמזגו לשכבה משותפת שתאפשר יצירת תחזית משוקללת, המשלבת מידע חדשותי עם נתוני השוק בפועל.

חשוב לציין כי אין צורך לאחד מראש את קובצי הנתונים לטבלה אחת, שכן כל מקור נתונים יאומן בנפרד על פי המודל המתאים לו. רק לאחר שלב האימון נבחן כיצד ניתן לשלב בין התחזיות, על מנת לייצר תובנות מדויקות יותר ולשפר את יכולת החיזוי של המערכת.

5. דוח עיצוב הנתונים

לאחר תהליך הניקוי והעיבוד, ביצענו מספר התאמות קריטיות בפורמט הנתונים כדי להכין אותם לשלב המידול. כל ההתאמות שנעשו הן בהתאם לדרישות המודלים שבהם נשתמש.

1. יצירת משתני זמן (Time Features) עבור נתוני S&P 500

ביצענו המרת עמודת Date לפורמט תקני (DD/MM/YYYY) כדי להבטיח אחידות.

יצירת עמודות זמן חדשות מתוך התאריך, כולל:

- Year – השנה בה נרשם הנתון.
- Month – החודש בו נרשם הנתון.
- Day – היום בחודש בו נרשם הנתון.
- Day_of_Week – היום בשבוע.

מטרת ההתאמה היא לזהות דפוסים עונתיים ומגמות מחזוריות במדד S&P 500.

2. התאמת נתוני S&P 500 למודל LSTM (חלונות זמן של 30 יום)

ביצענו יצירת חלונות זמן של 30 יום כך שלכל שורה יש רצף של 30 תצפיות קודמות. שימוש בעמודות Close, MACD, ו-RSI כקלט למודל.

הגדרת עמודת המטרה (Target) כערך Close של היום הבא, כך שהמודל ילמד לחזות מחירים עתידיים.

מטרת ההתאמה היא להתאים את הנתונים לפורמט שבו מודל LSTM יכול ללמוד מסדרות זמן.

3. עיבוד טקסטים והכנת נתוני החדשות הכלכליות למודל BERT (NLP)

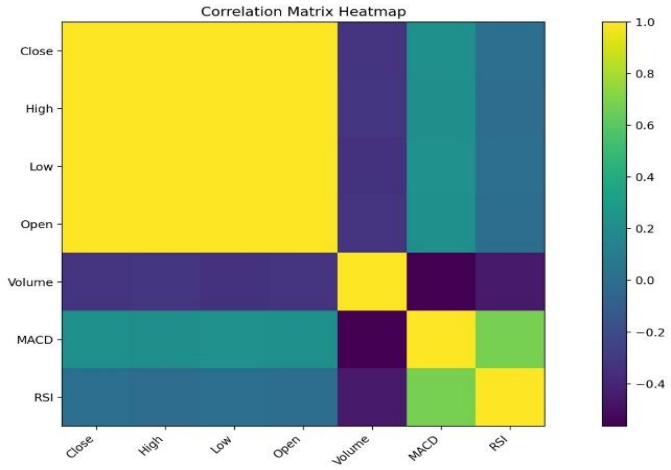
ביצענו ניקוי הטקסט בעמודת summary, כולל:

- המרת כל הטקסט לאותיות קטנות.
- הסרת מספרים ותווים מיוחדים (כדי להשאיר טקסט "נקי").
- הסרת מילים חסרות משמעות (Stop Words) כדי לשמור רק על מילים משמעותיות.
- יצירת עמודה חדשה בשם Cleaned_Text שמכילה את הגרסה המעובדת של הטקסט.

ובכך הכנו את הנתונים לפורמט מתאים למודל BERT, כך שהוא יוכל לנתח את משמעות החדשות והשפעתן על השוק.

6. ניתוח חקר הנתונים – קורלציות בין משתנים

התרשים מציג את מטריצת הקורלציות בין מספר משתני מפתח במדד S&P 500 :



- מחירי מסחר : מחיר סגירה (Close), מחיר גבוה יומי (High), מחיר נמוך יומי (Low), מחיר פתיחה (Open).
- נפח מסחר : Volume – כמות המניות שנקנו או נמכרו ביום המסחר.
- אינדיקטורים טכניים : MACD ו-RSI.

ציר הצבעים מייצג את עוצמת הקורלציה

- קורלציה שלילית חזקה (בסביבות `-1`) מוצגת בגוונים כהים (סגול/כחול כהה).
- קורלציה חיובית חזקה (`+1`) מוצגת בגוונים צהובים.
- ערכים קרובים ל-0 מייצגים קורלציה נמוכה או חוסר תלות בין המשתנים.

קורלציות בין מחירי המניה (Close, High, Low, Open)

תוצאה : קורלציה כמעט מושלמת בין כל מחירי המסחר (פתיחה, סגירה, גבוה ונמוך). ניתן לראות ריבועים צהובים בוהקים, המעידים על **קורלציה קרובה ל-1**.

תוצאה זו צפויה בשוק ההון, כיוון שכל ארבעת משתני המחיר מבטאים את תנודות השוק באותו יום או בימים סמוכים.

משמעות תפעולית

בשל הקשר החזק בין משתני המחיר, שימוש בכולם יחד במודל עלול ליצור כפילויות מידע (Multicollinearity). לרוב נהוג לבחור משתנה אחד או שניים בלבד כדי לייצג את מגמות המחירים (למשל, Close ו-Open).

קורלציה בין נפח המסחר (Volume) למחירי המניה

תוצאה: קורלציה שלילית או נמוכה בין Volume למחירי המניה. ניתן לראות גוונים סגולים בשורת `Volume`, המעידים על מתאם שלילי קל או חוסר קשר ישיר עם המחירים. כלומר, כשהמחיר עולה או יורד, נפח המסחר לא בהכרח משתנה באופן ליניארי.

משמעות תפעולית

נפח מסחר לא בהכרח נע יחד עם מחיר המניה, ולעיתים אף מציג קשר שלילי חלש. הסיבה לכך היא שמשקיעים לא תמיד קונים יותר מניות כשהמחיר עולה, ולעיתים דווקא בירידות נרשמות רכישות גדולות. ייתכן גם שמחזורי מסחר גבוהים משקפים אי-ודאות בשוק ולא בהכרח תנועה מובהקת במחיר.

קורלציה בין אינדיקטורים טכניים (MACD, RSI) למחירים

MACD לעומת מחירי המניה

ניתן לראות גוונים ירוקים-טורקיזיים, המצביעים על מתאם חיובי בינוני (0.3-0.6) בין MACD למחירים. תוצאה זו צפויה, כיוון ש-MACD מחושב על בסיס ממוצעים נעים של מחירי המניה. ככל שהמחיר עולה לאורך זמן, גם ה-MACD צפוי לעלות בהתאם.

RSI לעומת MACD והמחירים

ה-RSI נמצא בקורלציה בינונית (ירוק/צהבהב) עם MACD, כיוון ששניהם מודדים מומנטום שוק.

מול מחירי המניה עצמם, ה-RSI מציג קורלציה נמוכה עד בינונית (גוונים טורקיזיים), כיוון שהוא מחושב לפי עוצמת שינויי המחיר ולא לפי מחיר מוחלט.

משמעות תפעולית

שילוב של MACD ו-RSI במודל עשוי לשפר תחזיות לזיהוי מגמות שוק. כיוון שהם לא מראים קורלציה של 1 עם המחיר, הם עשויים להוסיף ערך חיזוי משמעותי מעבר לנתוני המחיר עצמם.

מסקנות לשלב זה

- מחירי הסגירה, הפתיחה, הגבוה והנמוך מתואמים חזק מאוד זה עם זה, כצפוי.
- נפח המסחר (Volume) אינו מתקדם באותו כיוון כמו המחיר, ולעיתים אף יש ביניהם מתאם שלילי קל.
- MACD מציג מתאם חיובי בינוני עם מחירי המניה, כיוון שהוא מבוסס על ממוצעים נעים.
- RSI מציג קשר בינוני עם MACD וקורלציה נמוכה עד בינונית עם מחירי המניה.
- בעת בניית מודלים, מומלץ לבחור בקפידה אילו משתנים להכניס – שילוב אינדיקטורים טכניים לצד מחירי המניה יכול לספק תחזיות טובות יותר.

ניתוח הקורלציות מספק תובנות קריטיות על היחסים בין המשתנים בשוק ההון. ממצאים אלו יסייעו לנו לזהות אילו תכונות הכי שימושיות למודלים החיזויים, ולקבוע כיצד לשלב את הנתונים בצורה היעילה ביותר.