המחלקה למתמטיקה ולמדעי המחשב

פרויקט מתקדם במדעי המחשב 22997

מערכת היברידית לסיווג מגמות במכשירים פיננסיים - FXMiner

שם המנחה: ד״ר הרמן מיה

שם הסטודנט: קריימר אנדריי, 314589490

נובמבר 2017

תוכן עניינים

1 תקציר....3
2 מבוא....4
3 מתודולוגיה....6
3.1 הגדרת דרישות המערכת....6
3.2 ניתוח ועיצוב המערכת....8
3.3 נתונים....13
3.4 בניית מודלים....19
3.5 בחינת איכות הסיווג....20
4 כריית מידע....21
4.1 Ensemble Classifier....21
4.2 Logistic Regression....22
4.3 Random Forest....22
4.4 Stochastic Gradient Descent....23
4.5 AdaBoost....23
4.6 Bagging....24
4.7 Best Parents....24
5 מימוש....25
5.1 ארכיטקטורת המערכת....25
5.2 מודולים עיקריים במערכת....26
5.3 כלי פיתוח....27
5.4 ממשק משתמש....28
5.5 בדיקות....30
6 מקרה בוחן....30
6.1 שלבי ביצוע מקרה הבוחן....31
6.2 מטריקות איכות....31
6.3 ניתוח תוצאות....32
7 סיכום והצעות להמשך מחקר....37
8 נספחים....38
8.1 פרסומים....38
8.2 הוראות התקנה....38
8.3 הוראות הרצה....39
9 רשימת מקורות....39

1 תקציר

כיום קיימים מספר רב של מוצרים להמלצות השקעה ומסחר. רוב הפתרונות הקיימים מבוססים על תבניות ידניות של סוחרים ואינן מבוססות על כריית מידע או ניתוח מתוחכם של הנתונים ההיסטוריים.

במסגרת הפרויקט פותחה מערכת סיווג היברידית למגמות במכשירים פיננסיים (מניות, תעודות סל ומטבע חוץ). המערכת מהווה כלי עזר לקבלת החלטות מסחר ומשתמשת בנתונים היסטוריים (ציבורי) בכדי לכרות מידע (מגמות עליה/ירידה) ולהקל על הסוחר הביתי.

ייחודיות המערכת באה לידי ביטוי בשלושה ממדים. מבחינת עיבודי האצווה מתבצע ניתוח מעמיק של הנתונים ההיסטוריים. כמות הנתונים שהמערכת מעבדת כרגע עומדת על למעלה משמונת אלפים מכשירים פיננסיים שנסחרים בבורסות השונות בארה״ב. מבחינת זמן אמת מתבצע ניתוח של סחירות, נתוני היצע, ביקוש ושלל מטה מאפיינים בעת החיזוי. מבחינת למידת המכונה מתבצע שימוש בסיווג אנסמבלי של כמה מסווגים ממשפחות שונות כולל אלגוריתם חדשני שפותח לקבל תחזית מדויקת ויציבה יותר. שלושת הממדים הללו והאינטגרציה ביניהם במערכת זו מהווה את לב ליבה של ייחודיות המערכת שפותחה בהשוואה למוצרים הקיימים כיום בשוק.

הקלט של המערכת הוא מידע היסטורי של מחירי המכשירים הפיננסיים בתוספת אינדיקטורים טכניים (מתנדים שמשמשים את רוב הסוחרים בעת קבלת החלטות מסחר כגון ממוצעים נעים, נפח המסחר וכו׳) ואינדיקטורים מהונדסים (Feature Engineering, כלומר יצירה של מאפיינים שמבוססים על מאפייני היסוד כגון אגרגציות של מספר טווחי זמן אחרונים או נקודות מינימום ומקסימום עבור מספר טווחי זמן אחרונים וכו׳).

המערכת מעבדת את הנתונים תוך שימוש בכריית מידע. סיווג המגמות מתבצע על ידי מיצוע של מספר מודלים מסווגים כגון Logistic Regression, Random Forest, SGD, Boosting -ו Bagging. בנוסף למכלול האלגוריתמים הקלאסיים מתחום למידת המכונה, משולב במערכת אלגוריתם חדשני לבניה של רשת בייסיאנית אופטימלית – Best Parents. כל זה מבוצע על ידי תהליך סיווג אנסמבלי (Ensemble Learning). סיווג אנסמבלי נחוץ להגברת הביצועים של המסווגים השונים וקבלת דעה מדויקת יותר בעת החיזוי של התנודות.

הפלט של המערכת הוא איתותי מסחר עבור מכשירים פיננסיים במבנה של כיוון התנודה, טווח התנודה וההסתברות לתנודה. באמצעות מידע זה ניתן להשקיע בצורה נוחה ויעילה.

2 מבוא

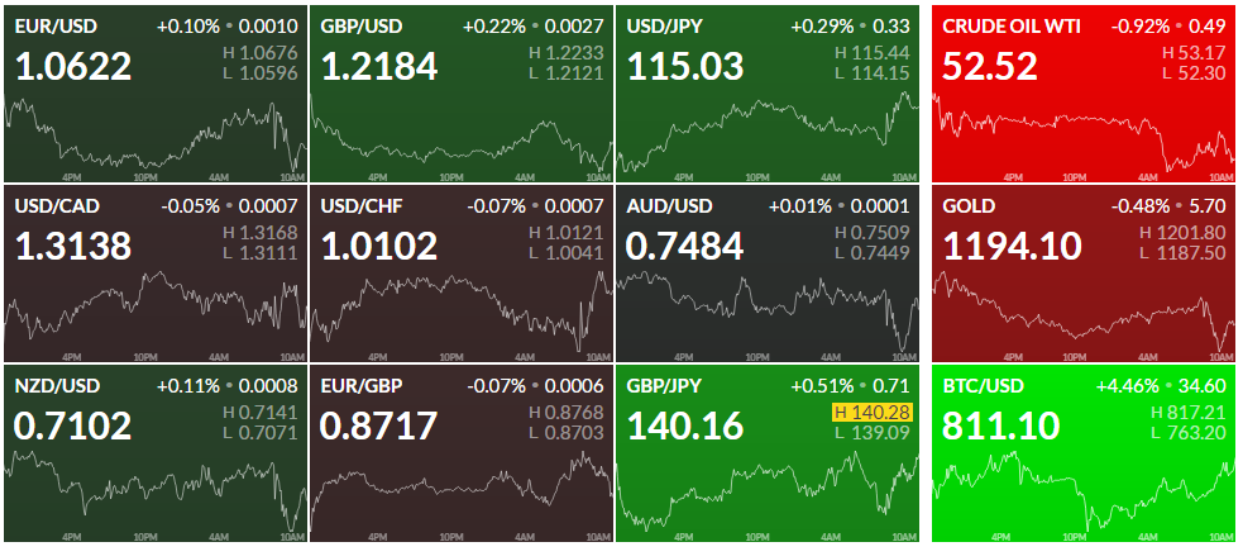
שווקים מאופיינים בנפחי פעילות גדולים. כמויות הנתונים והמכשירים הפיננסיים שנסחרים גבוהות ביותר. לדוגמה, בארה״ב ישנם למעלה משמונת אלפים מכשירים פיננסיים, וסה״כ בעולם ישנם למעלה מ-26 אלף מכשירים פיננסים שונים (מניות, אג״ח, תעודות סל, מדדים, צמדי מט״ח ועוד) [15]. באיור 1 מודגם שינוי יומי במניות ארה״ב לפי קטגוריות (מקבץ מניות ממדד ה-S&P500), באיור 2 מתוארים מספר צמדי מט״ח והשינוי היומי [21].

בעת מסחר נסתמך על ידע אישי, אתרים פיננסיים והצעתם של מומחים בתחום הפיננסי. נכון להיום, אין זה אפשרי לנתח ידנית את כלל הנתונים הקיימים: מחירים, דוחות היסטוריים, חדשות ותחזיות של אנליסטים באופן ידני. בנוסף לכך קיימות קורלציות בין מכשירים שונים העלולות להגדיל משמעותית את החשיפה והסיכון. ניתן לראות בעיה זו כבעיית אופטימיזציה כאשר פונקציית המטרה היא להגיע למקסימום רווחים תוך מינימום סיכון וחשיפה. מכאן, שנכון להיום בעת החלטות מסחר מקובל להסתייע בעקרונות וטכניקות מתחום מדעי המחשב [1, 2, 3].

איור 1: מניות שונות במדד S&P500 [21]

המערכת שפותחה בפרויקט מסייעת באחת המשימות הקשות ביותר בימינו – השקעה במכשירים פיננסיים. היכולת לקנות או למכור מכשירים פיננסיים בזמן הנכון ובכמות הנכונה

מהווה בעיה שרוב הסוחרים הביתיים בימינו לא יודעים להתמודד איתה. המערכת שפותחה מתמודדת עם הבעיה תוך שימוש באלגוריתמים של כריית מידע ולמידת מכונה.



איור 2 : צמדי מט"ח שונים [21]

קיימות בתחום מספר עבודות המיישמות מספר גישות לפתרון הבעיה כגון חיזוי שערי המכשירים הפיננסיים, סיווג מגמות ושימוש במספר טווחי זמן לנתונים ההיסטוריים [13, 17, 19]. החסרונות העיקריים בעבודות אלו הם התמקדות בשוק אחד בלבד. בניגוד לעבודות הקיימות בספרות, בפרויקט נבחן טווח חיפוש רחב הרבה יותר והמאפשר איתור שך מודלים טובים בהרבה לצורכי השקעה. הבדל מהותי נוסף הוא ביכולת ההפשטה ועיבוד הנתונים שכן המערכת שלנו מאפשרת לנתח כל מכשיר פיננסי ואינה מוגבלת לשווקים מסוימים.

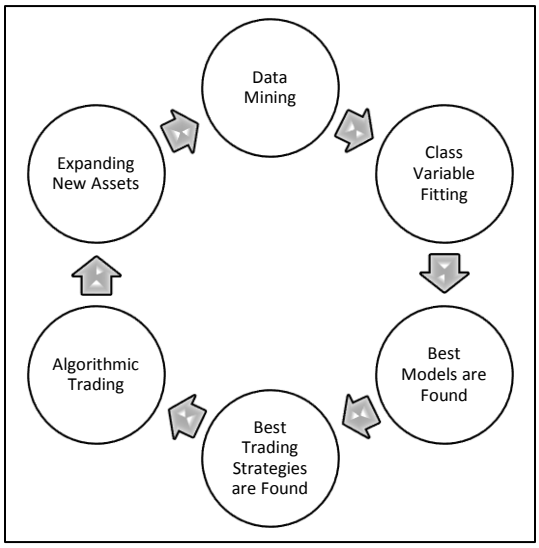
בעוד עבודות אחרות בתחום מתארות גישות לפתרון הבעיה, רובן לא מציגות וידוא קונקרטי לפתרון שמוצג. כלומר הבחינה מתבצעת על הנייר תוך שימוש במטריקות וחסמים סטנדרטיים בלמידת מכונה, ללא וידוא אמיתי של האיתותים או השיטה במסחר מדומה או חי. ננקוט בגישה שונה של מסחר מדומה על מנת לבחון את איכות האיתותים והמסחר. כמו כן מתבצעת השוואת ביצועים מול המדד הרלוונטי בשוק [2, 4, 7].

בפרויקט זה פותחה מערכת היברידית לסיווג מגמות במכשירים פיננסיים שונים. המערכת משלבת תהליכי אצווה ותהליכי זמן אמת. במערכת שולבו מספר אלגוריתמי סיווג שונים כגון : Logistic Regression, Random Forest, Bagging, SGD וגם פותח אלגוריתם חדש לבניה של רשתות בייסיאניות אופטימליות – Best Parents. כלל המסווגים ששולבו במערכת הוערכו באמצעות מטריקות מתורת המידע וגם מסחר חי בשוק [8, 10, 14, 15].

מדובר במערכת היברידית לסיווג מגמות במכשירים פיננסיים שונים. מדובר במערכת המשלבת תהליכי אצווה וזמן אמת. תוצאות המערכת הושמו עם סימולציה של מסחר (Paper Trading).

3 מתודולוגיה

תקיפת הבעיה מבוצעת באמצעות תהליך מחזורי של חיפוש היוריסטי, איתור מודלים טובים, מסחר באמצעות מודלים אלו ובחינה של מכשירים פיננסיים ו/או נתונים חדשים כמתואר באיור 3. כך ניתן ליצור מודלים חזקים למסחר והשקעות מדי יום.



איור 3 : מחזור החיים של המערכת

3.1 הגדרת דרישות המערכת

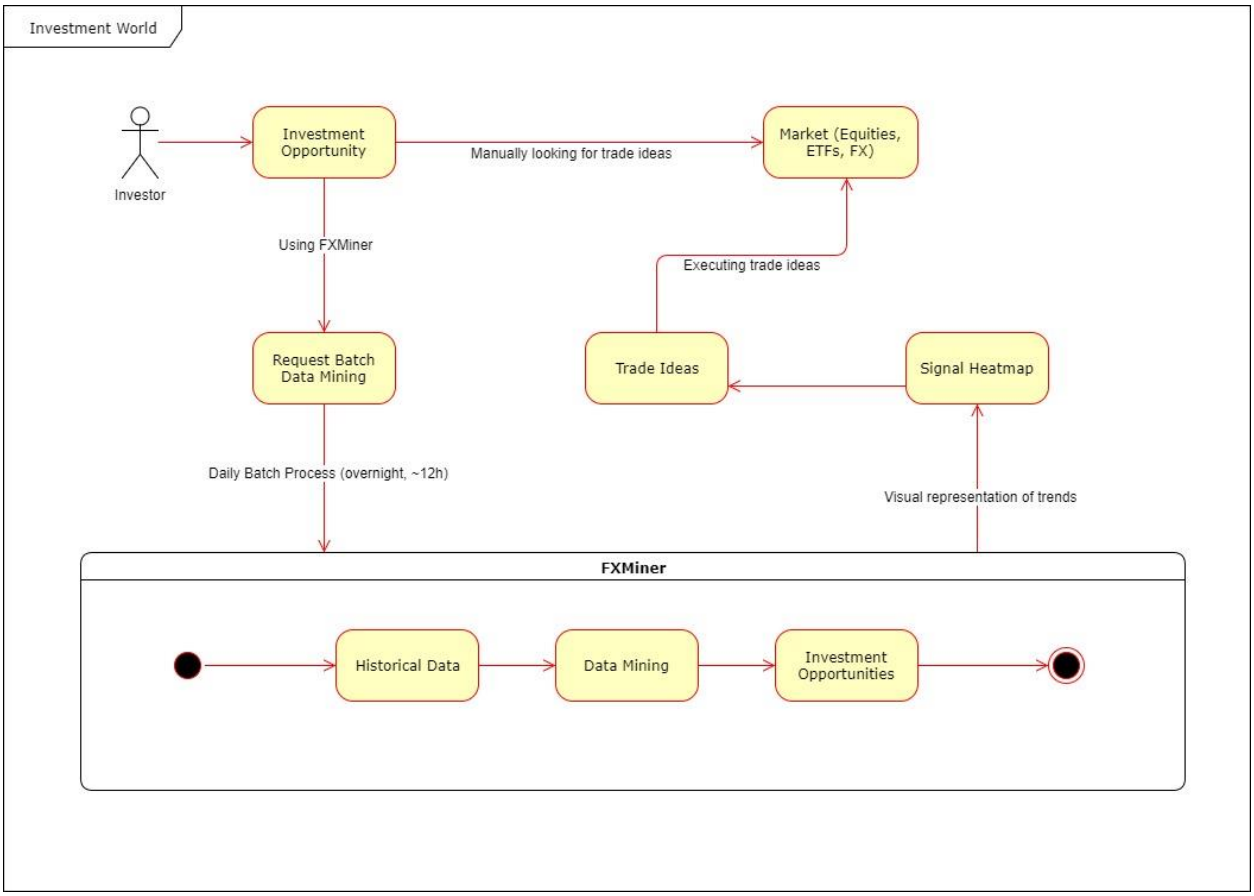
דרישות פונקציונליות: 1. עיבוד מקדים שיכלול ניקוי, הטבה והעשרה של הנתונים (מטה-מאפיינים). 2. בניית מודלים מנבאים/מסווגים לתנודות במכשירים הפיננסיים. 3. הצגת החלטות העבר בעניין כניסה לפוזיציות ע"י המודל המסווג. סוחר יוכל לקבל חוות דעת מדעית בזמן אמת ולבחון ויזואלית את ביצועי העבר. 4. הפקת סטטיסטיקות על ביצועי האיתותים והשוואתם אל מול המדד בשוק הרלוונטי. 5. אחזור נתוני עבר מברוקרים ומספר מקורות מידע.

דרישות לא פונקציונליות: המערכת מייצרת איתותי מסחר באמצעות שליפה, העשרה וכריית מידע היסטורי. באמצעות שיטות אופטימיזציה (מינימיזציה ומקסימיזציה של מטריקות המודלים). המערכת מאפשרת בחינה ויזואלית של התנודות והאיתותים על ידי

ממשק המשתמש. כמו כן המערכת מאפשרת שליפה של דו״חות וביצועים היסטוריים על מנת לשקף את איכות הביצועים.

הנחות יסוד: א. ידע בסיסי במסחר והשקעות. על המשתמש להיות בקיא ברמה בסיסית בשווקים פיננסיים ולהבין את הסיכונים והסיכויים. ביצוע המסחר בפועל הוא לא חלק מהמערכת. ב. המערכת מסתמכת על תקינות הנתונים הציבוריים המסופקים ע״י Yahoo Finance ו-MetaTrader 5. ג. המערכת אינה יכולה לשלוט על טיב הנתונים.

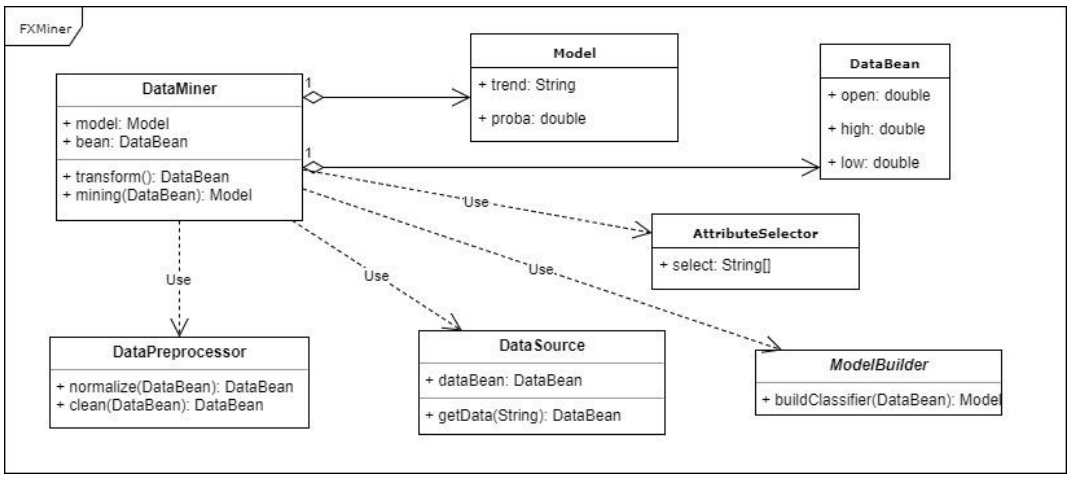
באיור 3.1 מתואר שילוב מערכת FXMiner בעולם ההשקעות הכולל בורסות שונות וסוחרים. בעקרון, קיימת אפשרות לסרוק את השווקים בצורה ידנית או להתבסס על מערכת ה-FXMiner לקבלת איתותי מסחר. המשקיע מקבל איתותי מסחר מדי יום לאחר תהליך אצווה לילי של כ-12 שעות ריצה. בסוף התהליך מתקבלת דיאגרמת heat map אשר מציגה מספר מכשירים פיננסיים והתנודות הצפויות.



איור 3.1: דיאגרמת תרחישים לשילוב מערכת ה-FXMiner בעולם ההשקעות

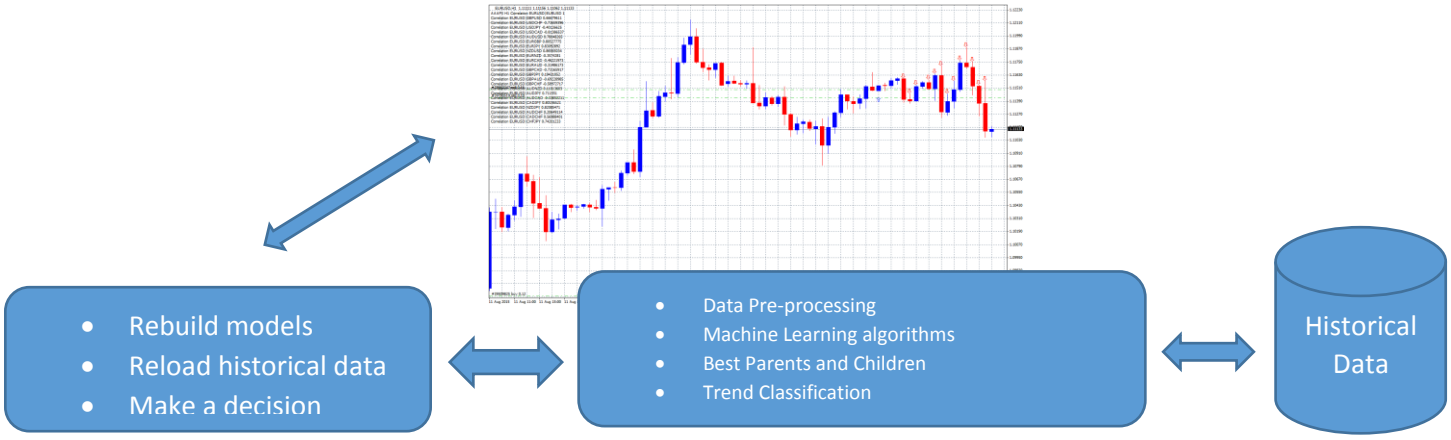
3.2 ניתוח ועיצוב המערכת

FXMiner הינה מערכת שרת-לקוח המורכבת מממשק משתמש וצד שרת כך שתעבורה וניתוח המידע מתבצע בשני מסלולים. ממשק המשתמש מאפשר לענות בצורה ויזואלית ונוחה על השאלה האם כדאי לקנות או למכור מכשירים פיננסיים מסוימים. צד השרת אחראי על איסוף, עיבוד הנתונים ובניית המודלים המנבאים. להלן תרשים סכמתי.



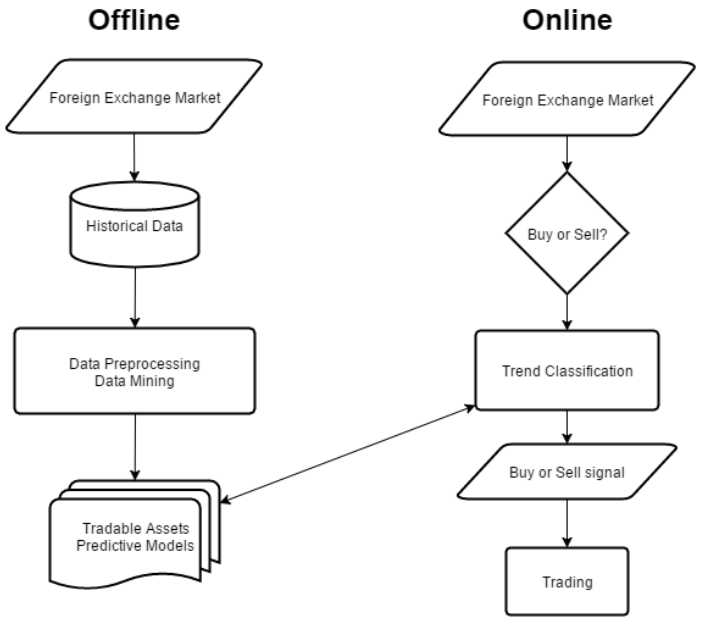
איור 3.2 : דיאגרמת מחלקות במערכת ה-FXMiner

באיור 3.2 מתוארת דיאגרמת מחלקות מרכזיות של המערכת. כל מחלקה אחראית על תהליך מרכזי במערכת. רכיב ה-DataMiner אחראי על אגרגציה של כלל התהליכים: הורדת הנתונים, עיבוד מקדים וכריית מידע. המחלקה DataPreprocessor אחראית על עיבוד מקדים של הנתונים. המחלקה DataSource אחראית על הורדת הנתונים ההיסטוריים. המחלקה ModelBuilder אחראית על בניית המודלים. המחלקה AttributeSelector אחראית על בחירת המאפיינים.



איור 3.3 : מבנה מערכת ה-FXMiner

באיור 3.3 מתואר מבנה כללי של מערכת FXMiner, הבעיה וכיצד המערכת פותרת את הבעיה. בשלב הראשון הסוחר מבקש לקבל תחזית לעליה או ירידה של המכשיר הפיננסי שבחר. בשלב השני המערכת (מבעוד מועד) אוספת נתונים היסטוריים ובאמצעות תהליך של כריית מידע יוצרת תחזית לעליה או ירידה של המכשיר הפיננסי, טווח התחזית וההסתברות לתחזית. המידע הזה מאפשר לסוחר להמשיך בביצוע עסקאות במידה וזה תואם את דעתו ו/או לאתר מכשיר פיננסי אחר בצורה מחזורית.

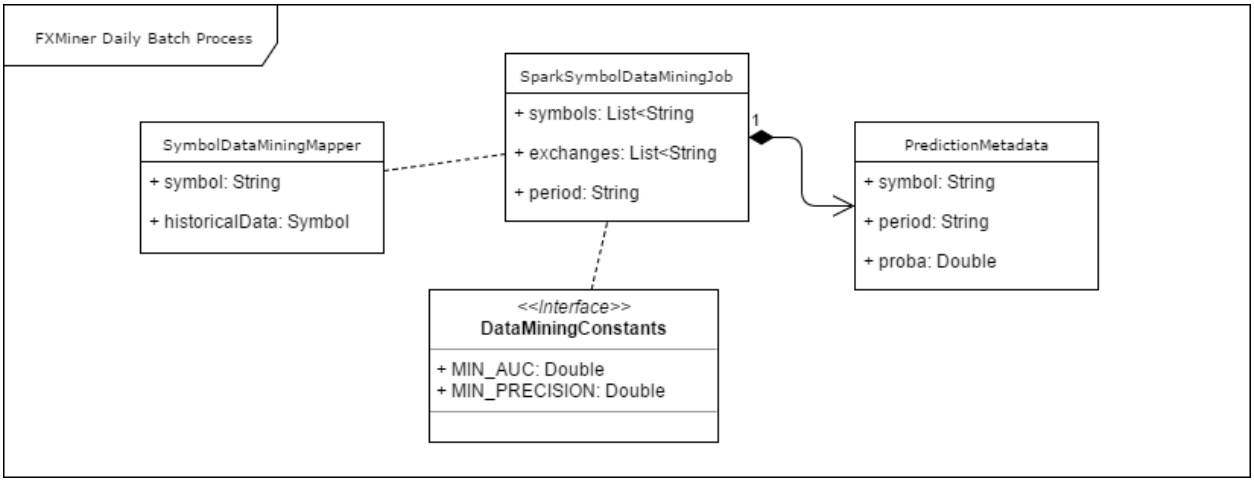


איור 3.4 : תהליכי אצווה וזמן אמת

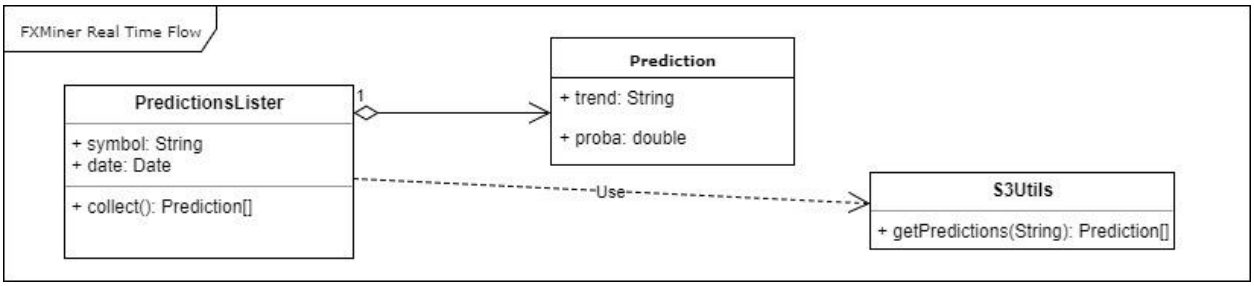
כמו גם באיור 3.5, באיור 3.4 מודגמים כלל התהליכים במערכת. תהליכים אלו בחלקם תהליכי אצווה ובחלקם תהליכים בזמן אמת. תהליכי האצווה אחראים על איסוף המידע, ניקוי ועיבוד מקדים ובניית מודלים מנבאים. תהליכי האצווה אחראים על שליפת הנתונים, ניקוי, השבחה, טרנספורמציה ושמירה בדיסק. לאחר מכן יתבצע תהליך של כריית מידע שכולל בחירה של מאפיינים, חלוקה למספר קבוצות אימון ומבחן, בחינה של מספר מטריקות ועמידה בדרישות סף (ראו סעיף "למידת מכונה"). כל תהליכי האצווה מתבצעים באמצעות שירותי הענן של אמזון (Amazon Web Services) ומאפשרים הגדלה אלסטית של כמות הנתונים המעובדים בצורה בלתי תלויה (Elastic Map-Reduce).

תהליכי זמן אמת אחראים על שליפה של איתותים קיימים למכשירים פיננסיים שונים כמודגם באיור 3.6. הנתונים בפרויקט נשמרים כקבצים שטוחים בדיסק. קיימת אפשרות

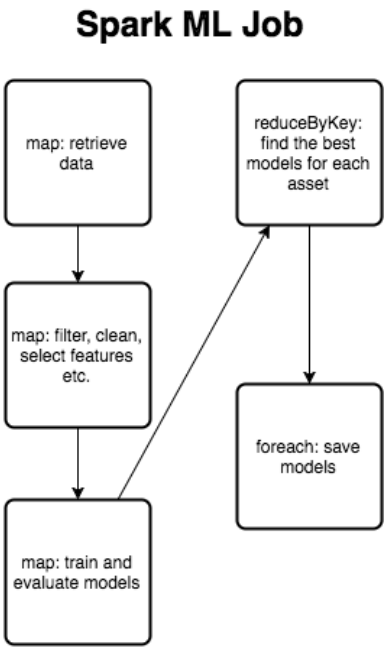
לחבר אמצעי אחסון נוספים כגון קבצים שטוחים בענן או בסיס נתונים סטנדרטי. כמו כן תהליכי זמן האמת אחראים על הצגת המידע ההיסטורי והמכשיר הפיננסי המתאים, והצגה של איתותי מסחר על הגרף. מאחר ומדובר בגרף יומי, סיווג התנודה יתבצע בשלב הפתיחה של המסחר (לאחר שכל המידע הדרוש עבור הסיווג קיים). ישנה אפשרות לחבר את איתותי המסחר למנוע ביצוע עסקאות.



איור 3.5 : היררכיית המחלקות של תהליך האצווה היומי



איור 3.6 : היררכיית המחלקות של תהליכי זמן אמת

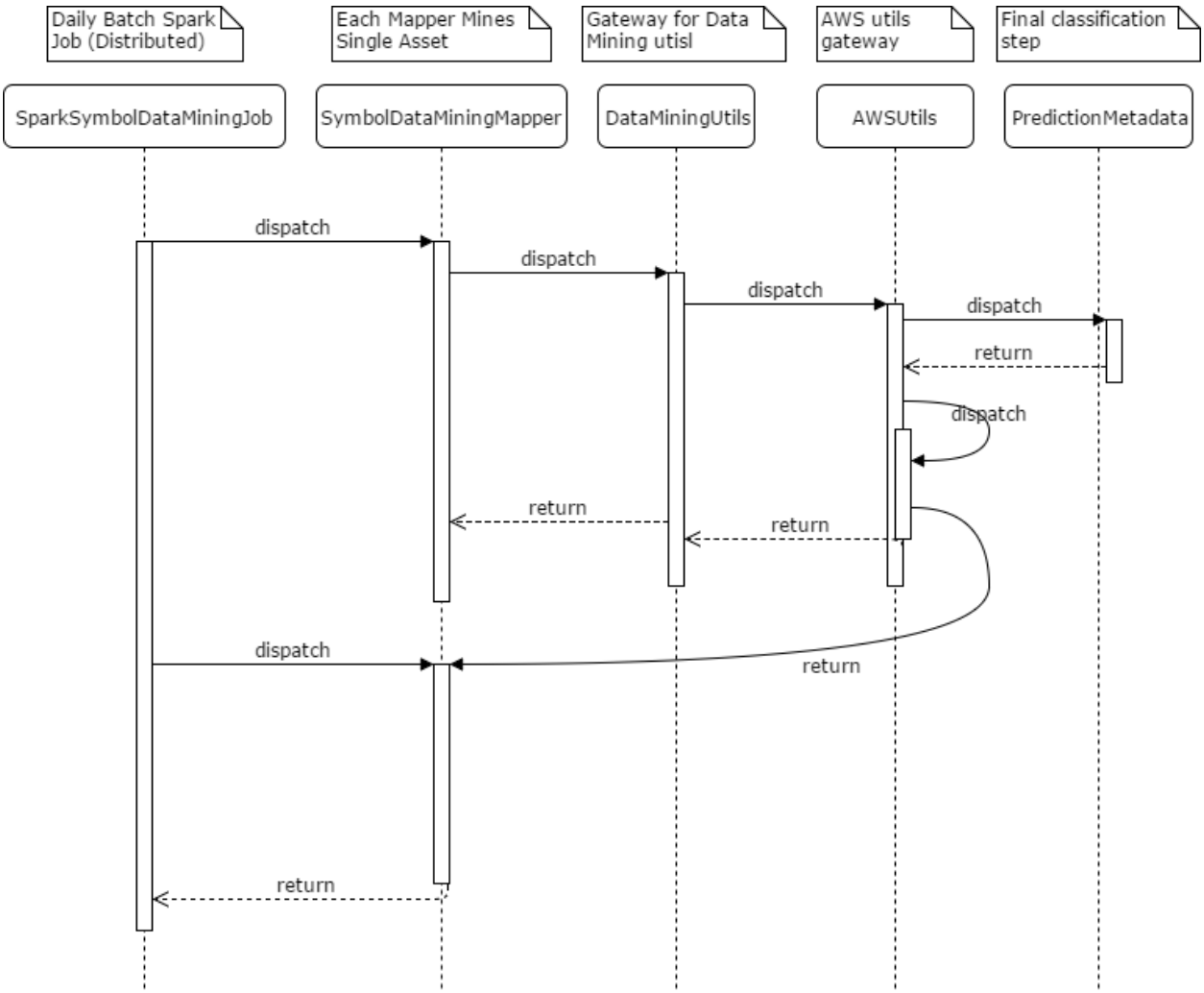


איור 3.7 : תיאור סכמיטי של ה-Spark Job

איור 3.7 מתאר סכמטית את תהליך האצווה היומי ומבנה ה-Spark Job [23], כלומר המבנה הלוגי של התכנית שרצה מדי יום. להלן קוד המקור של SparkSymbolDataMiningJob, האחראי על הרצת תהליך יומי וחלוקת העבודה על פני מספר שרתים.

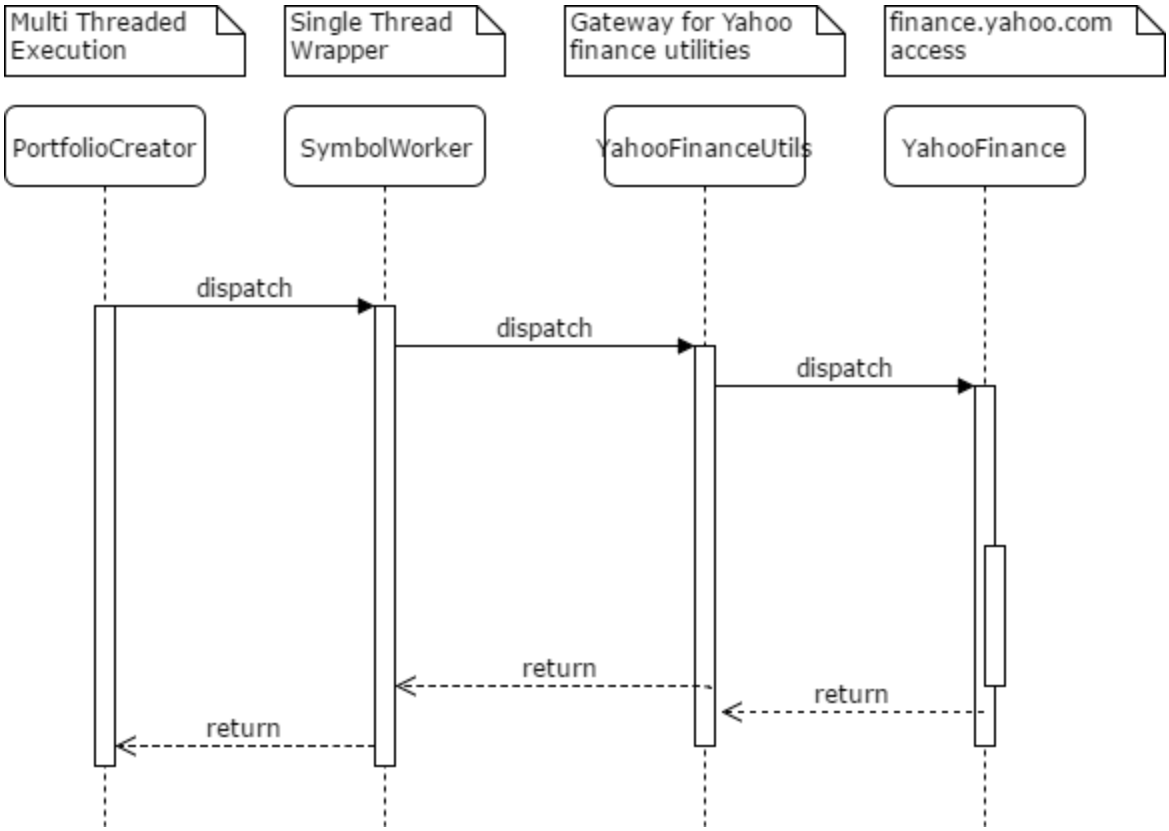
```
                                                    // configure Spark
                javaSparkContext = SparkUtils.createJavaSparkContext();

                                                  // get params & create symbol list
                                                   HashSet<String> fullSymbolList =
    SymbolUtils.createSymbolListFromNasdaq(Arrays.asList(args));

                                 // exclude symbols that already have predictions
                HashSet<String> symbols = reduceSymbolList(fullSymbolList);

                                                       // create RDD of symbols
                                                  JavaRDD<String> symbolRdd =
    javaSparkContext.parallelize(new ArrayList<String>(symbols));

                                                  // find best class for each symbol
                                JavaRDD<SymbolDataMiningBean> dataMiningBeanRdd =
                                symbolRdd.map(new SymbolDataMiningMapper());

                                                    // save models and classes files
    dataMiningBeanRdd = dataMiningBeanRdd.map(new SaveModelsMapper());

                                                                          // run
                    logger.info("rdd count: " + dataMiningBeanRdd.count());
```

באיור 3.8 מתוארת דיאגרמת רצף של תהליך האצווה היומי : איסוף, טיוב, כריית המידע וסיווג המגמה ליום המסחר הבא.



איור 3.8 : דיאגרמת רצף עבור תהליך האצווה היומי

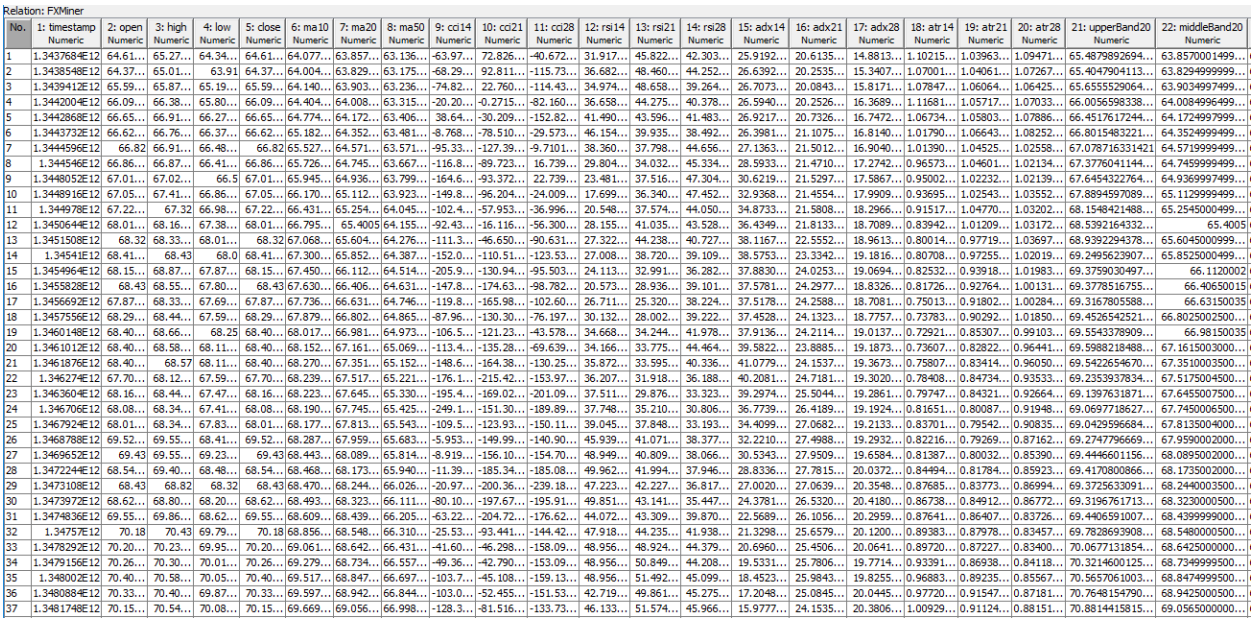
באיור 3.9 מודגמת דיאגרמת רצף ליצירה של Portfolio, כלומר רשימת המכשירים הפיננסיים שאותם נרצה לחקור ועליהם נריץ את תהליך האצווה. התהליך מתבסס על מניות סחירות לפי האתר של NASDAQ.



איור 3.9 : דיאגרמת רצף ליצירת ה-Portfolio

3.3 נתונים

נתוני הפרויקט הם המאפיינים הבסיסיים של מכשיר פיננסי סחיר כדוגמת מחיר פתיחה, גבוה, נמוך, סגירה ונפח והמסחר (כמות המניות שנרכשו נמכרו באותו יום). מקורם של הנתונים במאגרים היסטוריים של ברוקרים. הנתונים זמינים בכל מערכת מסחר מסוג MetaTrader 5 עבור מט״ח והאתר Yahoo Finance עבור מניות ותעודות סל. הנתונים ייאספו באופן אוטומטי על ידי סקריפטים לייבוא נתונים וממשקי פיתוח (API). נתונים אלו יטוייבו בעזרת מספר אלגוריתמים של עיבוד מקדים כדוגמת דיסקרטיזציה מונחית וסילוק מאפיינים אונריים. איור 3.10 מתאר מדגם של נתוני האימון. העמודות מכילות נתונים גולמיים שהתקבלו מהברוקר. העמודות המעושרות של מאפיינים יוצרו באופן דינאמי על ידי חישובים של אינדיקאטורים טכניים [10,15].



Relation: FXMiner

No.	1: timestamp Numeric	2: open Numeric	3: high Numeric	4: low Numeric	5: close Numeric	6: ma10 Numeric	7: ma20 Numeric	8: ma50 Numeric	9: cci14 Numeric	10: cci21 Numeric	11: cci28 Numeric	12: rsi14 Numeric	13: rsi21 Numeric	14: rsi28 Numeric	15: adx14 Numeric	16: adx21 Numeric	17: adx28 Numeric	18: atr14 Numeric	19: atr21 Numeric	20: atr28 Numeric	21: upperBand20 Numeric	22: middleBand20 Numeric
1	1.3437684E12	64.61...	65.27...	64.34...	64.61...	64.077...	63.857...	63.136...	-63.97...	72.826...	-40.672...	31.917...	45.822...	42.303...	25.9192...	20.6135...	14.8813...	1.10215...	1.03963...	1.09471...	65.4879892694...	63.8570001499...
2	1.3438548E12	64.37...	65.01...	63.91	64.37...	64.004...	63.829...	63.175...	-68.29...	92.811...	-115.73...	36.682...	48.460...	44.252...	26.6392...	20.2535...	15.3407...	1.07001...	1.04061...	1.07267...	65.4047904113...	63.8294999999...
3	1.3439412E12	65.59...	65.87...	65.19...	65.59...	64.140...	63.903...	63.236...	-74.82...	22.760...	-114.43...	34.974...	48.658...	39.264...	26.7073...	20.0843...	15.8171...	1.07847...	1.06064...	1.06425...	65.6555529064...	63.9034997499...
4	1.3442004E12	66.09...	66.38...	65.80...	66.09...	64.404...	64.008...	63.315...	-20.20...	-0.2715...	-82.160...	36.658...	44.275...	40.378...	26.5940...	20.2526...	16.3689...	1.11681...	1.05717...	1.07033...	66.0056598338...	64.0084996499...
5	1.3442868E12	66.65...	66.91...	66.27...	66.65...	64.774...	64.172...	63.406...	38.64...	-30.209...	-152.82...	41.490...	43.596...	41.483...	26.9217...	20.7326...	16.7472...	1.06734...	1.05803...	1.07886...	66.4517617244...	64.1724997999...
6	1.3443732E12	66.62...	66.76...	66.37...	66.62...	65.182...	64.352...	63.481...	-8.768...	-78.510...	-29.573...	46.154...	39.935...	38.492...	26.3981...	21.1075...	16.8140...	1.01790...	1.06643...	1.08252...	66.8015483221...	64.3529999499...
7	1.3444596E12	66.82	66.91...	66.48...	66.82	65.527...	64.571...	63.571...	-95.33...	-127.39...	-9.7101...	38.360...	37.798...	44.656...	27.1363...	21.5012...	16.9040...	1.01390...	1.04525...	1.02558...	67.078716331421	64.5719999499...
8	1.344546E12	66.86...	66.87...	66.41...	66.86...	65.726...	64.745...	63.667...	-116.8...	-89.723...	16.739...	29.804...	34.032...	45.334...	28.5933...	21.4710...	17.2742...	0.96573...	1.04601...	1.02134...	67.3776041144...	64.7459999499...
9	1.3448052E12	67.01...	67.02...	66.5	67.01...	65.945...	64.936...	63.799...	-164.6...	-93.372...	22.739...	23.481...	37.516...	47.304...	30.6219...	21.5297...	17.5867...	0.95002...	1.02232...	1.02139...	67.6454322764...	64.9369997499...
10	1.3448916E12	67.05...	67.41...	66.86...	67.05...	66.170...	65.112...	63.923...	-149.8...	-96.204...	-24.009...	17.699...	36.340...	47.452...	32.9368...	21.4554...	17.9909...	0.93695...	1.02543...	1.03552...	67.8894597089...	65.1129999499...
11	1.344978E12	67.22...	67.32	66.98...	67.22...	66.431...	65.254...	64.045...	-102.4...	-57.953...	-36.996...	20.548...	37.574...	44.050...	34.8733...	21.5808...	18.2966...	0.91517...	1.04770...	1.03202...	68.1548421488...	65.2545000499...
12	1.3450644E12	68.01...	68.16...	67.38...	68.01...	66.795...	65.4005	64.155...	-92.43...	-16.116...	-56.300...	28.155...	41.035...	43.528...	36.4349...	21.8133...	18.7089...	0.83942...	1.01209...	1.03172...	68.5392164332...	65.4005
13	1.3451508E12	68.32	68.33...	68.01...	68.32	67.068...	65.604...	64.276...	-111.3...	-46.650...	-90.631...	27.322...	44.238...	40.727...	38.1167...	22.5552...	18.9613...	0.80014...	0.97719...	1.03697...	68.9392294378...	65.6045000999...
14	1.34541E12	68.41...	68.43	68.0	68.41...	67.300...	65.852...	64.387...	-152.0...	-110.51...	-123.53...	27.008...	38.720...	39.109...	38.5753...	23.3342...	19.1816...	0.80708...	0.97255...	1.02019...	69.2495623907...	65.8525000499...
15	1.3454964E12	68.15...	68.87...	67.87...	68.15...	67.450...	66.112...	64.514...	-205.9...	-130.94...	-95.503...	24.113...	32.991...	36.282...	37.8830...	24.0253...	19.0694...	0.82532...	0.93918...	1.01983...	69.3759030497...	66.1120002
16	1.3455828E12	68.43	68.55...	67.80...	68.43	67.630...	66.406...	64.631...	-147.8...	-174.63...	-98.782...	20.573...	28.936...	39.101...	37.5781...	24.2977...	18.8326...	0.81726...	0.92764...	1.00131...	69.3778516755...	66.40650015
17	1.3456692E12	67.87...	68.33...	67.69...	67.87...	67.736...	66.631...	64.746...	-119.8...	-165.98...	-102.60...	26.711...	25.320...	38.224...	37.5178...	24.2588...	18.7081...	0.75013...	0.91802...	1.00284...	69.3167805588...	66.63150035
18	1.3457556E12	68.29...	68.44...	67.59...	68.29...	67.879...	66.802...	64.865...	-87.96...	-130.30...	-76.197...	30.132...	28.002...	39.222...	37.4528...	24.1323...	18.7757...	0.73783...	0.90292...	1.01850...	69.4526542521...	66.8025002500...
19	1.3460148E12	68.40...	68.66...	68.25	68.40...	68.017...	66.981...	64.973...	-106.5...	-121.23...	-43.578...	34.668...	34.244...	41.978...	37.9136...	24.2114...	19.0137...	0.72921...	0.85307...	0.99103...	69.5543378909...	66.98150035
20	1.3461012E12	68.40...	68.58...	68.11...	68.40...	68.152...	67.161...	65.069...	-113.4...	-135.28...	-69.639...	34.166...	33.775...	44.464...	39.5822...	23.8885...	19.1873...	0.73607...	0.82822...	0.96441...	69.5988218488...	67.1615003000...
21	1.3461876E12	68.40...	68.57	68.11...	68.40...	68.270...	67.351...	65.152...	-148.6...	-164.38...	-130.25...	35.872...	33.595...	40.336...	41.0779...	24.1537...	19.3673...	0.75807...	0.83414...	0.96050...	69.5422654670...	67.3510003500...
22	1.346274E12	67.70...	68.12...	67.59...	67.70...	68.239...	67.517...	65.221...	-176.1...	-215.42...	-153.97...	36.207...	31.918...	36.188...	40.2081...	24.7181...	19.3020...	0.78408...	0.84734...	0.93533...	69.2353937834...	67.5175004500...
23	1.3463604E12	68.16...	68.44...	67.47...	68.16...	68.223...	67.645...	65.330...	-195.4...	-169.02...	-201.09...	37.511...	29.876...	33.323...	39.2974...	25.5044...	19.2861...	0.79747...	0.84321...	0.92664...	69.1397631871...	67.6455007500...
24	1.346706E12	68.08...	68.34...	67.41...	68.08...	68.190...	67.745...	65.425...	-249.1...	-151.30...	-189.89...	37.748...	35.210...	30.806...	36.7739...	26.4189...	19.1924...	0.81651...	0.80087...	0.91948...	69.0697718627...	67.7450006500...
25	1.3467924E12	68.01...	68.34...	67.83...	68.01...	68.177...	67.813...	65.543...	-109.5...	-123.93...	-150.11...	39.045...	37.848...	33.193...	34.4099...	27.0682...	19.2133...	0.83701...	0.79542...	0.90835...	69.0429596684...	67.8135004000...
26	1.3468788E12	69.52...	69.55...	68.41...	69.52...	68.287...	67.959...	65.683...	-5.953...	-149.99...	-140.90...	45.939...	41.071...	38.377...	32.2210...	27.4988...	19.2932...	0.82216...	0.79269...	0.87162...	69.2747796669...	67.9590002000...
27	1.3469652E12	69.43	69.55...	69.23...	69.43	68.443...	68.089...	65.814...	-8.919...	-156.10...	-154.70...	48.949...	40.809...	38.066...	30.5343...	27.9509...	19.6584...	0.81387...	0.80032...	0.85390...	69.4446601156...	68.0895002000...
28	1.3472244E12	68.54...	69.40...	68.48...	68.54...	68.468...	68.173...	65.940...	-11.39...	-185.34...	-185.08...	49.962...	41.994...	37.946...	28.8336...	27.7815...	20.0372...	0.84494...	0.81784...	0.85923...	69.4170800866...	68.1735002000...
29	1.3473108E12	68.43	68.82	68.32	68.43	68.470...	68.244...	66.026...	-20.97...	-200.36...	-239.18...	47.223...	42.227...	36.817...	27.0020...	27.0639...	20.3548...	0.87685...	0.83773...	0.86994...	69.3725633091...	68.2440003500...
30	1.3473972E12	68.62...	68.80...	68.20...	68.62...	68.493...	68.323...	66.111...	-80.10...	-197.67...	-195.91...	49.851...	43.141...	35.447...	24.3781...	26.5320...	20.4180...	0.86738...	0.84912...	0.86772...	69.3196761713...	68.3230000500...
31	1.3474836E12	69.55...	69.86...	68.62...	69.55...	68.609...	68.439...	66.205...	-63.22...	-204.72...	-176.62...	44.072...	43.309...	39.870...	22.5689...	26.1056...	20.2959...	0.87641...	0.86407...	0.83726...	69.4406591007...	68.4399999000...
32	1.34757E12	70.18	70.43	69.79...	70.18	68.856...	68.548...	66.310...	-25.53...	-93.441...	-144.42...	47.918...	44.235...	41.938...	21.3298...	25.6579...	20.1200...	0.89383...	0.87978...	0.83457...	69.7828693908...	68.5480000500...
33	1.3478292E12	70.20...	70.23...	69.95...	70.20...	69.061...	68.642...	66.431...	-41.60...	-46.298...	-158.09...	48.956...	48.924...	44.379...	20.6960...	25.4506...	20.0641...	0.89720...	0.87227...	0.83400...	70.0677131854...	68.6425000000...
34	1.3479156E12	70.26...	70.30...	70.01...	70.26...	69.279...	68.734...	66.557...	-49.36...	-42.790...	-153.09...	48.956...	50.849...	44.208...	19.5331...	25.7806...	19.7714...	0.93391...	0.86938...	0.84118...	70.3214600125...	68.7349999500...
35	1.348002E12	70.40...	70.58...	70.05...	70.40...	69.517...	68.847...	66.697...	-103.7...	-45.108...	-159.13...	48.956...	51.492...	45.099...	18.4523...	25.9843...	19.8255...	0.96883...	0.89235...	0.85567...	70.5657061003...	68.8474999500...
36	1.3480884E12	70.33...	70.40...	69.87...	70.33...	69.597...	68.942...	66.844...	-103.0...	-52.455...	-151.53...	42.719...	49.861...	45.275...	17.2048...	25.0845...	20.0445...	0.97720...	0.91547...	0.87181...	70.7648154790...	68.9425000500...
37	1.3481748E12	70.15...	70.54...	70.08...	70.15...	69.669...	69.056...	66.998...	-128.3...	-81.516...	-133.73...	46.133...	51.574...	45.966...	15.9777...	24.1535...	20.3806...	1.00929...	0.91124...	0.88151...	70.8814415815...	69.0565000000...

איור 3.10 : מדגם של נתוני האימון

הטיפול בנתונים נחלק לשלב האימון (offline) ולשלב הסיווג בזמן אמת (real-time). בשלב האימון הנתונים נשלפים על ידי מספר מקורות מידע שונים :

- Broker Historical Data– נתונים היסטוריים עבור שוק המט"ח.
- Yahoo Finance API – שליפה של נתונים על מניות.
- Yahoo Finance – Quandl API לעיתים אינו זמין ולא יציב, לכן עבור מקרים מסוימים נגבה את שליפת הנתונים על ידי ספרייה נוספת.

הקבצים נשמרים כקבצים שטוחים בדיסק מאחר ומדובר בקבצים קטנים יחסית (קבצים של עד 1G). לאחר מכן הנתונים עוברים מספר שלבי ניקוי : הורדה של רשומות מלאות חלקית או בעלי ערכים מחוץ לתחום (אינסוף, NAN וכדומה), השבחה והעשרה במטה מאפיינים. בשלב הניקוי נוריד מאפיינים כגון זמן או מזהה ייחודי. בשלב ההשבחה נוסיף מאפיינים חדשים על בסיס הנתונים הקיימים תוך אגרגציה של המאפיינים כגון ממוצעים נעים ורגרסיה ליניארית עבור מספר מסוים של טווחים [10]. מידע זה נשלח למערכת בניית מודלים שנעזרת במספר שיטות לבחירה של מאפיינים ובניה של מודל מסווג המבוססות על מדדים השאולים מתורת המידע כגון Info-Gain, Gain Ratio ואנטרופיה. המודלים נשמרים בענן לשימוש בזמן אמת לאחר מכן.

להלן מבנה חלקי של הנתונים בפורמט ARFF (Attribute Relation File Format), זהו פורמט נתונים מיוחד של WEKA. בניגוד לקבצי CSV רגילים, קבצי ARFF מכילים תיאור של טיפוס הנתונים, כך שקריאת הנתונים מתבצעת באופן חד משמעי.

@relation FXMiner – שם הרלציה

@attribute timestamp numeric - מזהה ייחודי, זמן

%OHLC - Open, High, Low, Close - מחירי פתיחה, גבוה, נמוך וסגירה

@attribute open numeric

@attribute high numeric

@attribute low numeric

@attribute close numeric

%MA – ממוצעים נעים

@attribute ma10 numeric

@attribute ma20 numeric

@attribute ma50 numeric

%CCI – מתנדים שונים

@attribute cci14 numeric

@attribute cci21 numeric

@attribute cci28 numeric

%RSI

@attribute rsi14 numeric

@attribute rsi21 numeric

@attribute rsi28 numeric

%ADX

@attribute adx14 numeric

@attribute adx21 numeric

@attribute adx28 numeric

%ATR

@attribute atr14 numeric

@attribute atr21 numeric

@attribute atr28 numeric

%Bands

@attribute upperBand20 numeric

@attribute middleBand20 numeric

@attribute lowerBand20 numeric

@attribute upperBand50 numeric

@attribute middleBand50 numeric

@attribute lowerBand50 numeric

%MACD

@attribute MACD1226 numeric

@attribute MACDSignal1226 numeric

@attribute MACDHist1226 numeric

%Stochastic

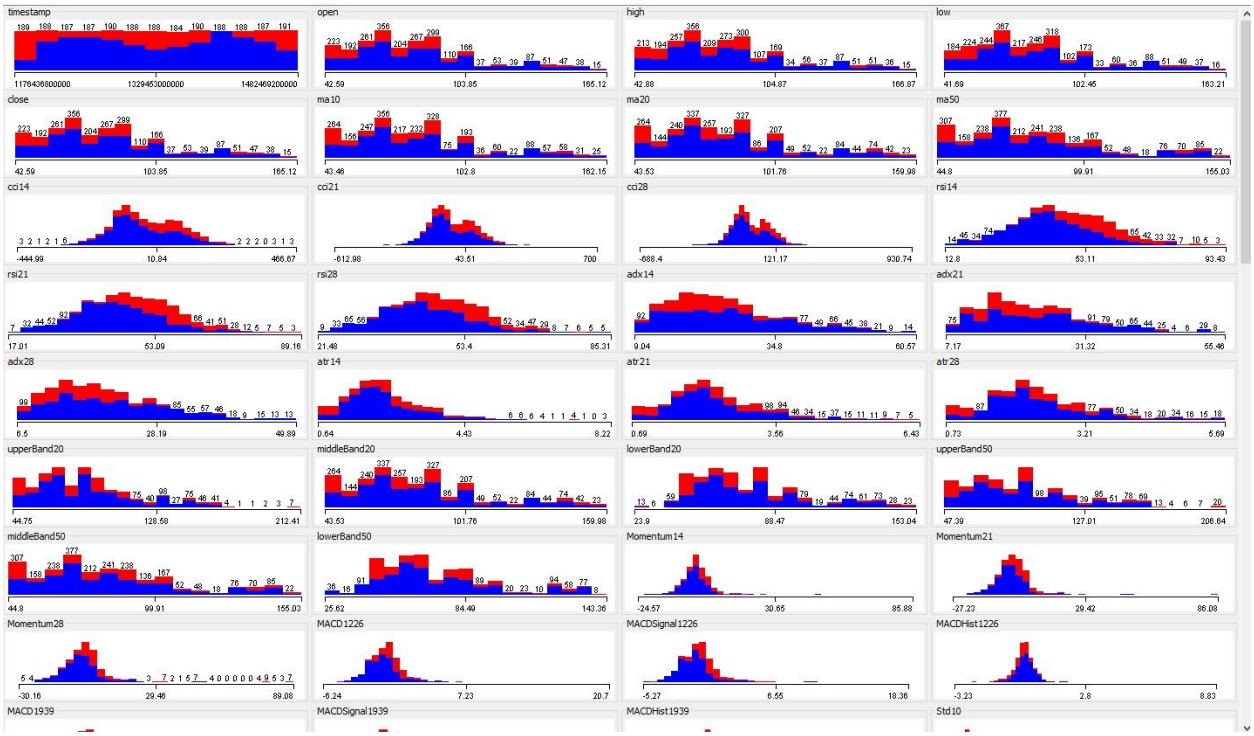
@attribute StochasticSlowK335 numeric

@attribute StochasticSlowD335 numeric

%Trend – משתנה מטרה בינארי (עליה או ירידה של המחיר)

@attribute classifiedTrend {Up, Down}

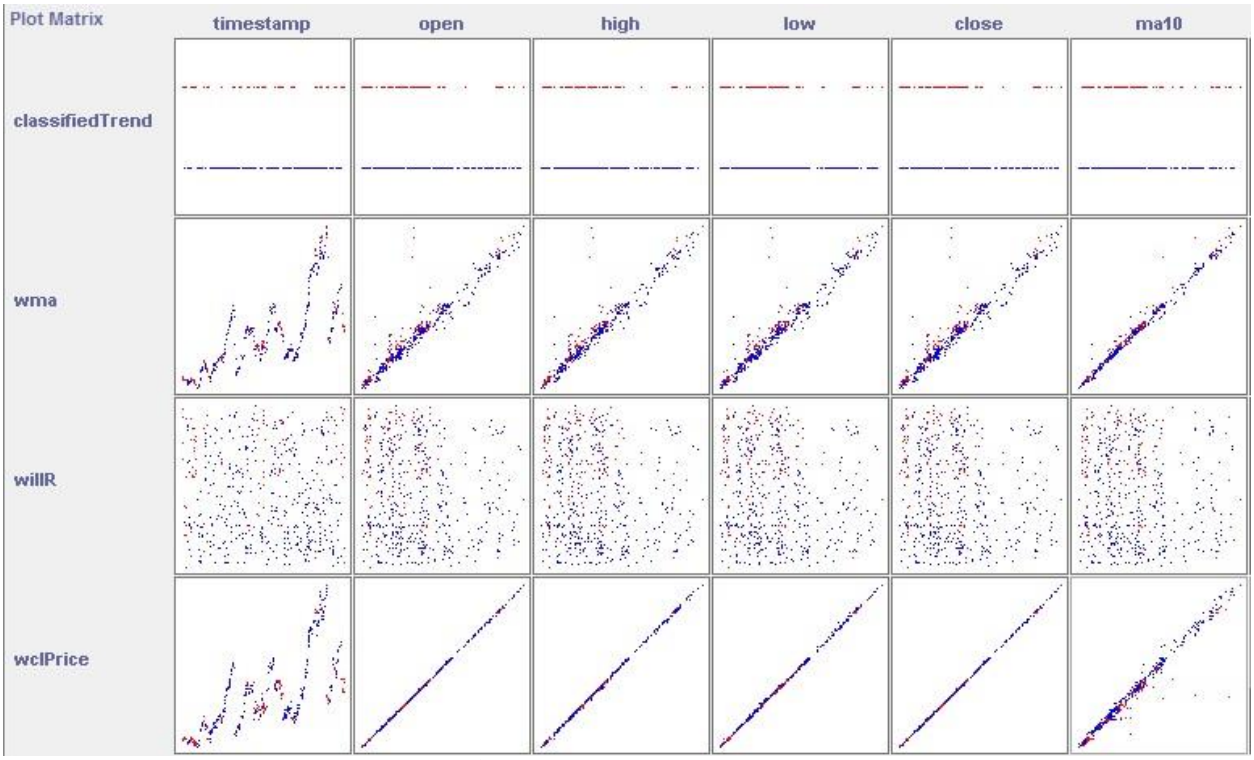
נשתמש בנתונים היסטוריים עבור כלל המכשירים הפיננסיים. לרוב המכשירים הפיננסיים ישנם בד"כ 10 שנים של נתונים היסטוריים. על בסיס ארבעת וקטורי היסוד (פתיחה, גבוה, נמוך וסגירה) נבנה קרוב ל-100 מאפיינים המבוססים על אינדיקטורים טכניים. באיור 3.11 מתוארות היסטוגרמות עבור מספר מאפיינים בכדי לקבל רושם ראשוני על ההתפלגויות של הנתונים.



איור 3.11 : היסטוגרמה של מאפיינים בקבוצת האימון

בפרויקט הוספו מטה-מאפיינים תוך שימוש בניתוח טכני. הגישה הטכנית למסחר מספקת מגוון רחב של אינדיקטורים טכניים. רוב האינדיקטורים מבוססים על נוסחאות ואגרגציות של מספר רב של תצפיות עבר [15]. לרוב האינדיקטורים קיימים כללים ברורים להיותם איתותי קניה או מכירה. כלל האינדיקטורים הקיימים בספרייה TA-LIB ישולבו בפרויקט, ויוספו מטה מאפיינים על ידי דיסקרטיזציה מונחית וידנית.

באיור 3.12 מתוארים מספר מאפיינים והקורלציות ביניהם. ניתן לראות כי קיימות קורלציות לא זניחות בין מספר מאפיינים. נתייחס לבעיה זו בעת בחירת המאפיינים.



איור 3.12 : צמדי מאפיינים והקורלציות ביניהם

רוב המאפיינים נוצרים על ידי Feature Engineering כלומר בניה מלאכותית של מאפיינים. כתוצאה מכך ניתן לקבל מאפיינים בעלי קורלציה גבוהה ביניהם העשויים להוסיף רעש למודל. כדי להימנע ממצבים אלו, נבצע בדיקה מקדימה של איכות המאפיינים ובניה של מודלים ליניאריים שונים בין תת קבוצות של מאפיינים.

3.4 בניית מודלים

נעזר במספר מסווגים ממשפחות שונות (ראה סעיף למידת מכונה) והכללתם על ידי הצבעה תיצור מודל חזק מבחינת יכולת ההכללה שלו על נתונים חדשים. בדומה לשימוש במסווג יחיד לפתרון בעיות בלמידת מכונה, כאן יבוצעו מספר סיווגים עבור כל מסווג שבנינו ושיערוך התחזית על פי ההסתברויות שהתקבלו מכל מסווג. דהיינו, סיווג אנסמבלי [17].

נבחן מספר רב של משתני מטרה – היכולת לחזות תנודות בטווחים שנעים בין שבוע לשלושה חודשים קדימה. מודלים המספקים את המטריקות הטובות ביותר, הם המודלים שלבסוף יבחרו. הבחירה של המודלים מתבצעת לפי מטריקות של AUC, Precision ו-Recall [5]. לכל אחד מהערכים נקבע סף מינימלי. מודלים העוברים את הסף, נחשבים למודלים טובים. בשלב החיזוי נשתמש במודלים שנבנו בשלב האימון על מנת לסווג תנודות במכשירים

הפיננסיים. המודלים שנבנו מאפשרים להכליל את בעיית הסיווג למספר ימים ושבועות קדימה.

3.5 בחינת איכות הסיווג

נעזר במספר כלים וטכניקות בתחום למידת המכונה בכדי למצוא מודלים מנבאים חזקים. כל זאת בכדי לאפשר יכולת הכללה של המודלים וקבלת החלטות מסחר מדויקות ביותר. חשוב לציין שמבחינתנו מודלים שלא עומדים בסף המטריקות, לא נכללים במערכת. כלומר מבחינתנו עדיף לוותר על מודל בינוני ולהישאר רק המודלים הטובים ביותר.

נסתייע בשלוש מטריקות לבחינת איכות המודלים: Precision, Recall ו-AUC. מטריקות אלה מאפשרות בחינה של מודלים מסווגים בינריים (במקרה שלנו מדובר במשתנה מטרה בינרי: עליה או ירידה). מודלים שאינם עומדים בסף של 0.8 בשלושת המטריקות יפסלו. חשוב לציין שסף של 0.5 שקול להטלת מטבע ואילו סף של 0.9 נחשב לסף גבוה. בחרנו ב-0.8 על מנת להקל במידה על סינון המודלים שכן מדובר בבעיה קשה. חיזקנו את הבחירה במספר רב של הרצות ואופטימיזציה של הסף האופטימלי. לבסוף הגענו לסף של 0.8 (ראו סעיף מטריקות איכות).

כל שלבי האימון והבדיקה מתבצעים על נתונים מפוצלים לפי שכבות: k-fold cross validation. כלומר הנתונים חולקו למספר שכבות ב"ת הממוינות לפי זמן כך שהאימון מתבצע על נתוני העבר ובחינת המודל מתבצעת על נתוני העתיד. זו נקודה מהותית בתחום למידת המכונה בסדרות עיתיות. יש לציין שאמנם הבעיה שאיתה אנו מתמודדים היא לא ניבוי סדרות עיתיות, החלוקה של הנתונים בזמנים ממוינים היא קריטית לבחינה הוגנת של המודלים.

שיטה נוספת לבחינת איכות הסיווג תהיה מערכת של back-testing, כלומר הרצת הסיווג על נתוני העבר ובחינה של תוצאות במסחר מדומה. בנוסף לכך נבחן את הסיווג על ידי מסחר בזמן אמת בחשבון אמתי. ניתוח התוצאות יתבסס על מטריקות להערכת המסווגים וגם מטריקות להערכת טיב המסחר והתשואה כגון ROI – Return On Investment, או PF – Profit Factor.

4 כריית מידע

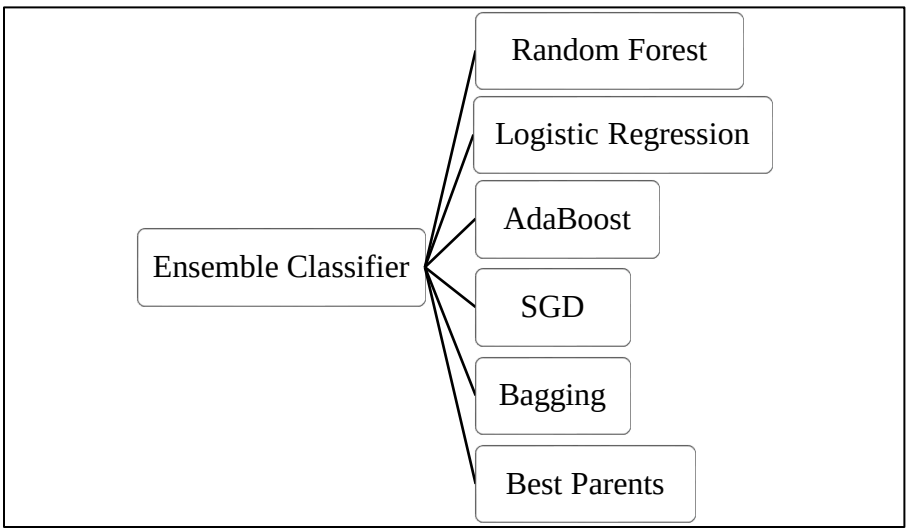
בעבודה זו נשתמש במספר רב של אלגוריתמי סיווג ומטה-מודל לשילוב של מספר החלטות סיווג והצבעה. הטכניקה מאוד פופולרית בתעשייה וידועה כדרך בטוחה להגברת יכולת הסיווג וההכללה. סיווג אנסמבלי מבוסס על מספר שלבים עיקריים. בשלב הראשון מאמנים מודלים מסווגים לחוד על קבוצת אימון אחידה. בשלב השני מבצעים תחזיות על התצפיות באמצעות כל אחד מהמסווגים. בשלב השלישי מבצעים מיצוע של הסתברויות התחזיות וכך מקבלים הצבעה של מספר מסווגים.

נעזר בספריית WEKA אשר כוללת מימוש לכל האלגוריתמים ומאפשרת עבודה עם כמויות נתונים בינוניות. בנוסף למכלול האלגוריתמים הקיים בספרייה, אני נעזרים באלגוריתם שפיתחנו תחת הספרייה הזו [1,5].

מאחר ומדובר בבעיית סיווג בינארי נעזר במטריקות סטנדרטיות לסינון מודלים חלשים. AUC, Precision ו-Recall, מאפשרים לבחון את עוצמת המסווג ויכולת ההכללה שלו. בנוסף למטריקות, נבצע חלוקה שכבתית לקבוצות אימון ובדיקה לפי סדר כרונולוגי.

4.1 Ensemble Classifier

Ensemble Classification או מטה מסווג היא שיטה להגברת יכולת הסיווג וההכללה של מודלים מסווגים חלשים על ידי בניית מודל מסווג מעל שכבה של מודלים פשוטים. בשלב הראשון בונים מודלים מסווגים על אותם הנתונים עבור כל מודל בנפרד. בשלב השני נבצע הצבעה בין מספר מודלים לקבלת דעה על פי רוב. באפשרותנו לבנות מודל מסווג נוסף על



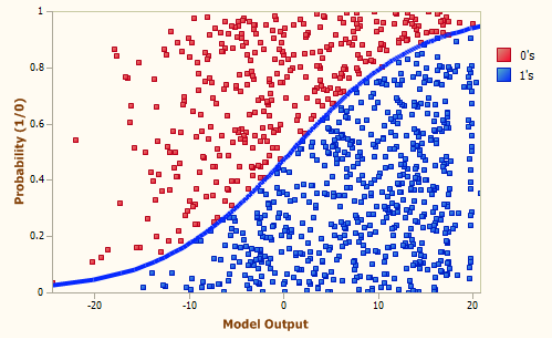
תוצאות הסיווג של המודלים מהשלב הראשון.

איור 4: שילוב של מספר מסווגים על ידי מטה-מסווג

בחרנו במספר מסווגים אשר שונים אחד מהשני באופן פתרון בעיית הלמידה. איור 4 מתאר את המטה-מסווג ואת המסווגים מהם הוא מורכב. המשקלים בעת השערוך זהים בין כלל המסווגים.

4.2 Logistic Regression

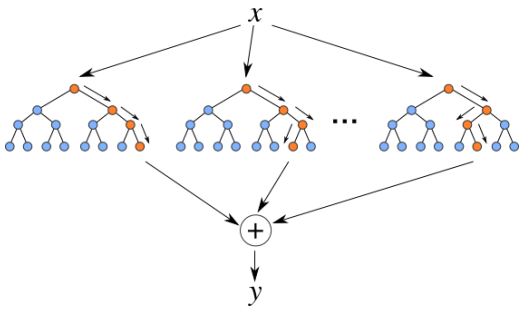
רגרסיה לוגיסטית הוא אלגוריתם סיווג מהיר אשר נעזר במינימיזציה של פונקציית המטרה (שגיאת האימון). לרוב האלגוריתם מבצע נורמליזציה של הנתונים לטווח אחיד ופועל טוב יותר כשאר הנתונים נפרשים לייצוג בינארי (One Hot Encoding). האלגוריתם מבוסס על מודל ליניארי לפילוג שתי קבוצות אימון לפי שתי מחלקות (משתנה מחלקה בינארי בלבד). האלגוריתם נעזר בפונקציה לוגיסטית לחישוב ההסתברויות. האלגוריתם פופולרי בתעשייה בגלל קלות היישום והשימוש בזמן אמת, כמו כן בעל זמן אימון קצר יחסית [5]. באיור 4.1 מתואר מסד נתונים בעל שתי מחלקות ויצירה של פונקציית הפרדה בין שתי הקבוצות.



איור 4.1: סיווג נתונים ע"י LR

4.3 Random Forest

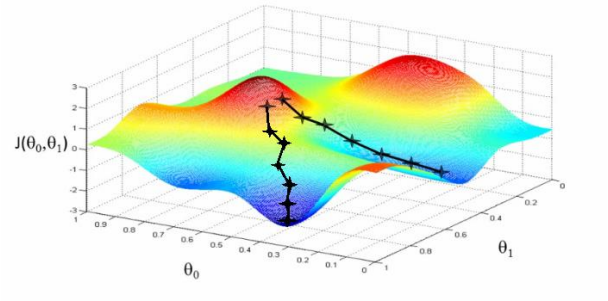
Random Forest הוא מטה-מסווג אשר מורכב ממספר עצי סיווג אקראיים. כל עץ סיווג נבנה תוך שימוש בתת רשימה של מאפיינים שנבחרו באקראי מתוך כלל המאפיינים. הבחירה האקרית מונעת מצבי התאמת יתר ומאפשרת בניית מודלים שמכלילים יותר טוב על נתוני זמן אמת. איור 4.2 מציג יער אקראי שנתקבל משלושה עצי סיווג אקראיים. בשלב החיזוי התוצאה מתקבל על ידי מיצוע של התחזיות עבור כל עץ סיווג אקראי [5].



איור 4.2 : סיווג הנתונים על ידי עצי סיווג אקראיים

4.4 Stochastic Gradient Descent

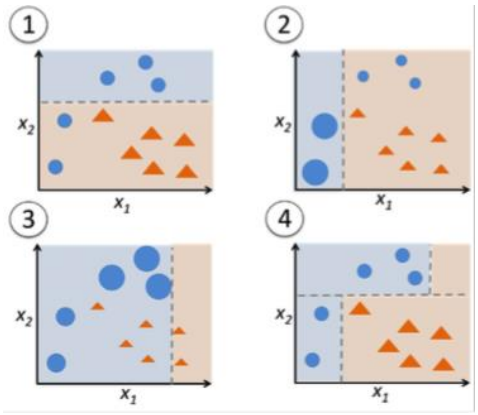
SGD (Stochastic Gradient Descent) הוא אלגוריתם סיווג אקראי. האלגוריתם מתבסס על דיפרנציאלים שליליים בכדי לאתר נקודות מינימום מקומי. בנוסף לכך האלגוריתם נעזר באקראיות בכדי להימנע מכניסה לאזורי מינימום מקומי. בדומה לרגרסיה ליניארית האלגוריתם פותר בעיית מינימיזציה בצורה של שגיאת האימות. האלגוריתם מאפשר פלישה לכלל מרחב החיפוש בכדי למצוא את המינימום המוחלט [5]. איור 4.3 מתאר את תהליך האיתור של מינימום מקומי על ידי השיפוע הגבוה ביותר בכיוון המטרה.



איור 4.3 : אופטימיזציה סטוכסטית

4.5 AdaBoost

בדומה ליער אקראי (Random Forest), שיטה זו משתמשת בעצי סיווג בכדי לקבל מודל ראשוני. לאחר מכן מתחיל תהליך איטרטיבי של למידה מטעויות. בשלב הראשון בונים עצי סיווג. בשלב השני כל התצפיות השגויות מהשלב הראשון מקבלות משקל מוגבר בכדי ללמוד טוב יותר מהטעיות. התהליך חוזר על עצמו עד אשר מגיעים לטווח טעות רצוי [5]. איור 4.4 מתאר למידה הדרגתית מטעויות קודמות הקטנה איטרטיבית של הבעיה.



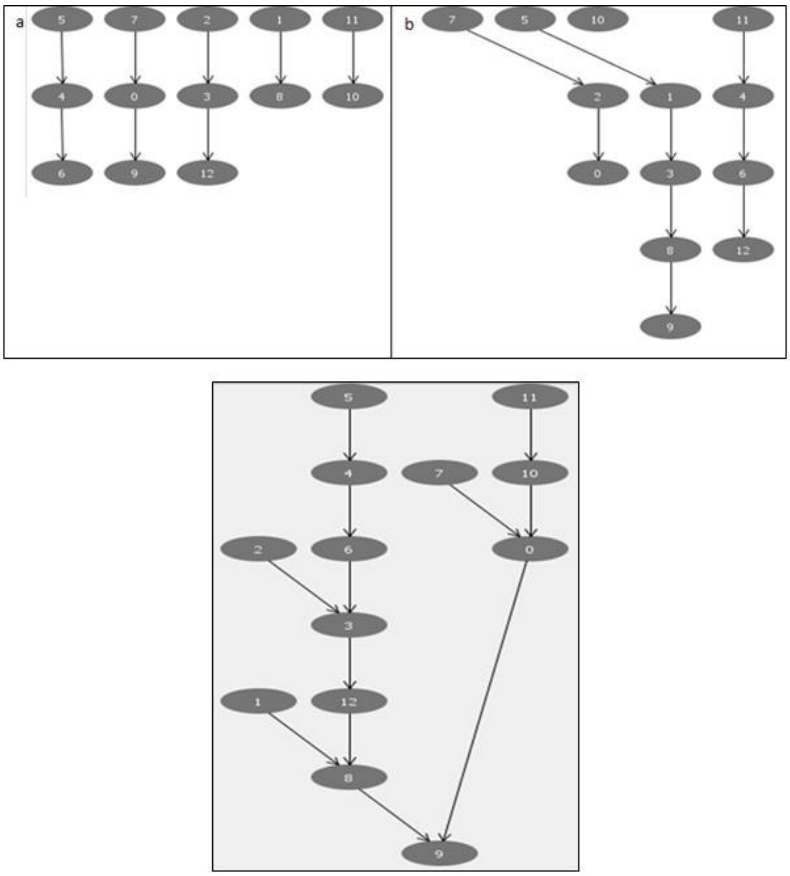
איור 4.4 : הקטנת קבוצת האימון בצורה איטרטיבית

4.6 Bagging

אלגוריתם זה בדומה לאלגוריתם Random Forest בד"כ מתבסס על עצי סיווג. האלגוריתם משתמש בהרכבה מתוחכמת של קבוצת האימון : דגימה נשנית עם חזרות של תצפיות בקבוצת האימון שמונעת התאמת יתר. השיטה מאוד שימושית במקרים בהם קיים חוסר איזון בין המחלקות של משתנה המטרה. כך מאזנים את משתנה המטרה על ידי בחירה עם חזרות.

4.7 Best Parents

אלגוריתם לבניה של רשת בייסיאנית אופטימלית – שילוב של אימון מהיר ודיוק גבוה [5, 1]. האלגוריתם מבוסס על חיפוש חמדני של צמדי מאפיינים מיטביים לפי אנטרופיה (ראו נספח לתיאור מפורט של האלגוריתם). איור 4.5 מתאר את שלבי הבניה של הרשת הבייסיאנית באמצעות האלגוריתם. בשלב הראשון מחפשים יחסים חזקים בין צמדי מאפיינים, בשלב השני ממזגים את הצמדים על פי דירוג היחסים תוך שימוש באנטרופיה מונחית.



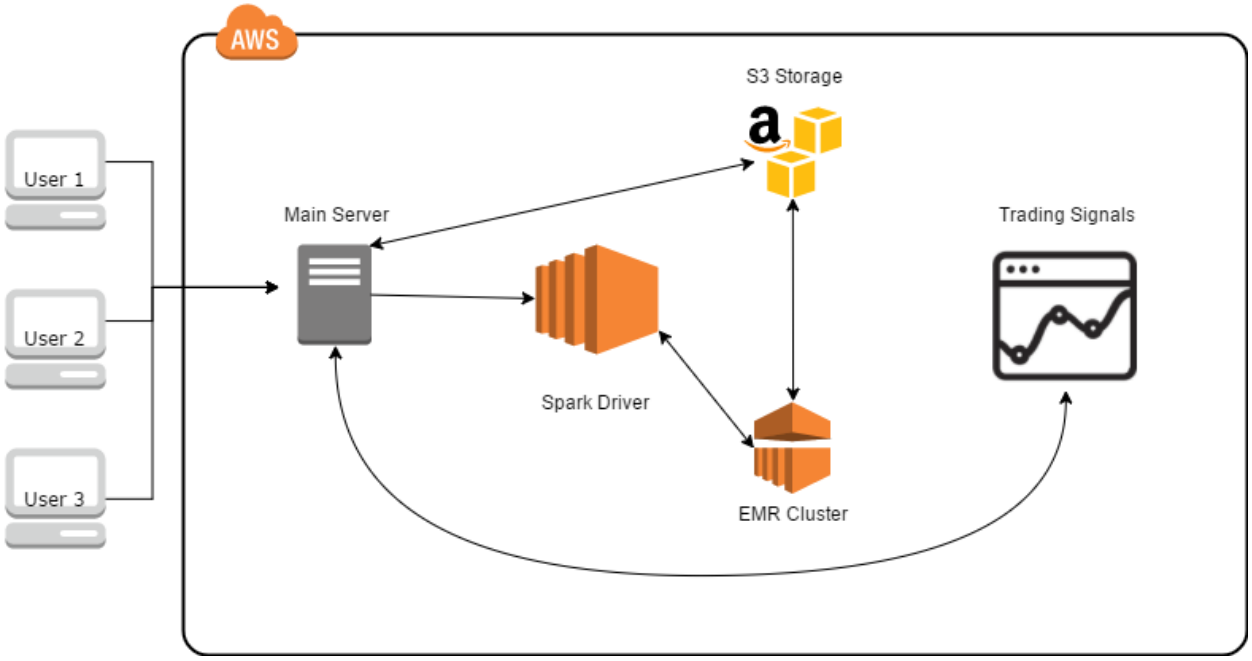
איור 4.5 : בנייה של רשת בייסיאנית לפי אלגוריתם Best Parents

5 מימוש

המערכת הינה מערכת שרת לקוח מבוסס שרותי ענן וטכנולוגיות נתוני עתק. נשתמש בעיבוד מקבילי ושירותי מחשוב ענן בכדי למקבל את התהליכים ולעבד כמויות גדולות של נתונים. בצורה כזו מפרידים לגמרי את גודל הנתונים מתהליך כריית המידע.

5.1 ארכיטקטורת המערכת

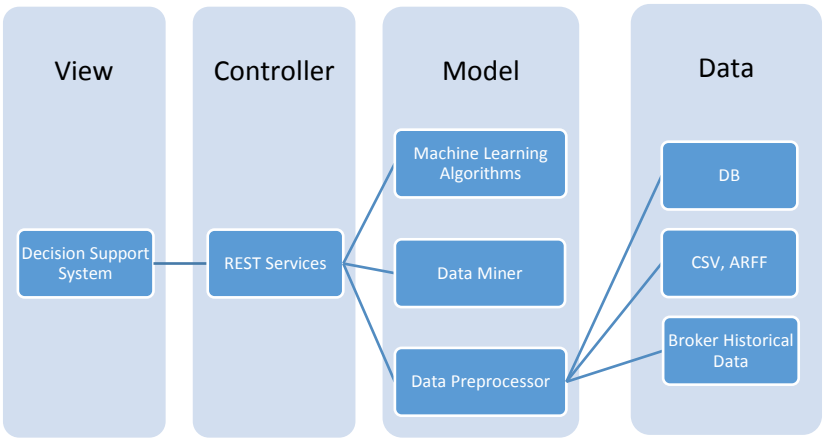
המערכת פותחה בגישת פיתוח Web. המבנה הלוגי הוא תבנית העיצוב MVC (Model-View-Controller). הגישה למידע מתבצעת על ידי תבנית העיצוב DAO (Data Access Object) ו-DTO (Data Transfer Object). תבניות עיצוב אלו הן הבסיס לפיתוח Web וגישה מופשטת לתעבורת לוגיקה ומידע. איור 5 מתאר את מבנה המערכת והשרתים הרלוונטיים.



איור 5 : ארכיטקטורת המערכת

5.2 מודולים עיקריים במערכת

המערכת מכילה מספר מודולים עיקריים. איור 5.1 מתאר את החלוקה הלוגית למודלים לפי תבנית העיצוב Model View Controller – MVC.



איור 5.1 : MVC במערכת FXMiner

View

- Decision Support System – ממשק משתמש גרפי מבוסס טכנולוגיות WEB.

Controller

- REST Services – מכלול שירותי REST לניתוח הנתונים: טיוב הנתונים, ביצוע תחזיות וסיווג מגמות כתוצאה מכריית מידע.

Model

- Data Preprocessor – עיבוד מקדים הכולל תהליכי טיוב וטרנספורמציה של הנתונים.
- אינטגרטור בין אלגוריתמי עיבוד מקדים ותהליכי כריית מידע שונים.
- Machine Learning Algorithms – אלגוריתמים לכריית מידע מסוג: סיווג, חיזוי, קיבוץ וחוקי הקשר.
- Data Miner – מרכיב מרכזי שמשתמש במכלול האלגוריתמים הקיים ומבצע הצבעה פנימית לצורך מתן אינטרפרטציה מדעית לשוק. זוהי הליבה של המערכת וזה הרכיב שמספק בסופו של דבר את המלצות המסחר.

Data

- CSV, ARFF – קבצי טקסט שטוחים.
- Broker Historical Data– נתונים היסטוריים עבור שוק המט"ח.
- Yahoo Finance API – שליפה של נתונים על מניות.

5.3 כלי פיתוח

הפרויקט נבנה תוך שימוש בטכנולוגיות WEB ו-Cloud Computing. להלן רשימה של הכלים המרכזיים בפרויקט.

- Java 8 – גישה פונקציונלית, מיפוי הפחתה.
- Amazon Web Services – שירותי מחשוב ענן של אמזון. מאפשרים שמירה נוחה של נתונים ועיבוד מקבילי של נתוני עתק תוך שימוש בעיבוד מקבילי (Cluster Computing).

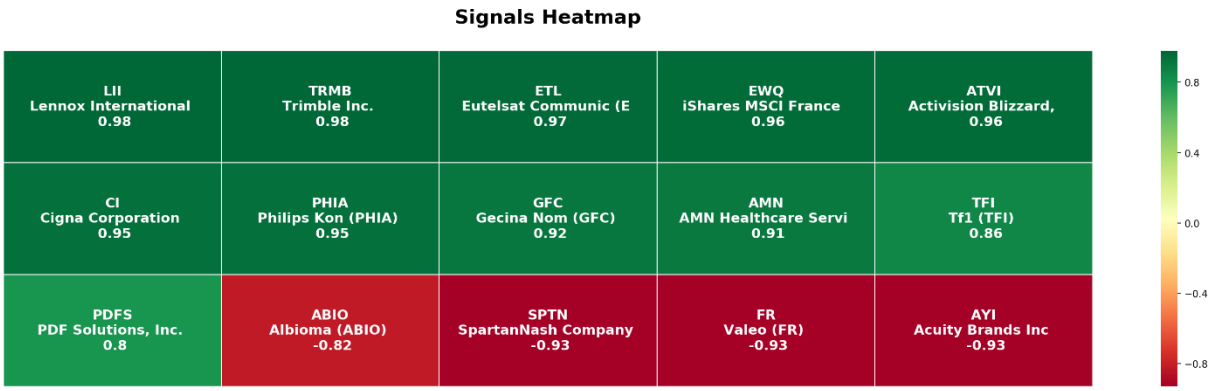
- Apache Spark – מערכת ליצירה של אפליקציות מקביליות לעבודה עם נתוני עתק. פרויקט קוד פתוח שתפס תאוצה בשנים האחרונות.
- WEKA – ספריית קוד פתוח לכריית מידע ולמידת מכונה.
- TA-LIB – ספרייה לחישוב אינדיקטורים טכניים.
- MetaTrader 5 – פלטפורמת המסחר במט"ח לסוחר הביתי.
- Yahoo Finance API – ממשק לשליפת נתוני עבר לנתונים היסטוריים.
- Quandl – ממשק נוסף לשליפת נתוני עבר לנתונים היסטוריים.
- Apache Tomcat 8 – שרת קוד פתוח. מאפשר הרצה של אפליקציות דינאמיות.
- JUnit – כלי לביצוע בדיקות יחידה.

5.4 ממשק משתמש

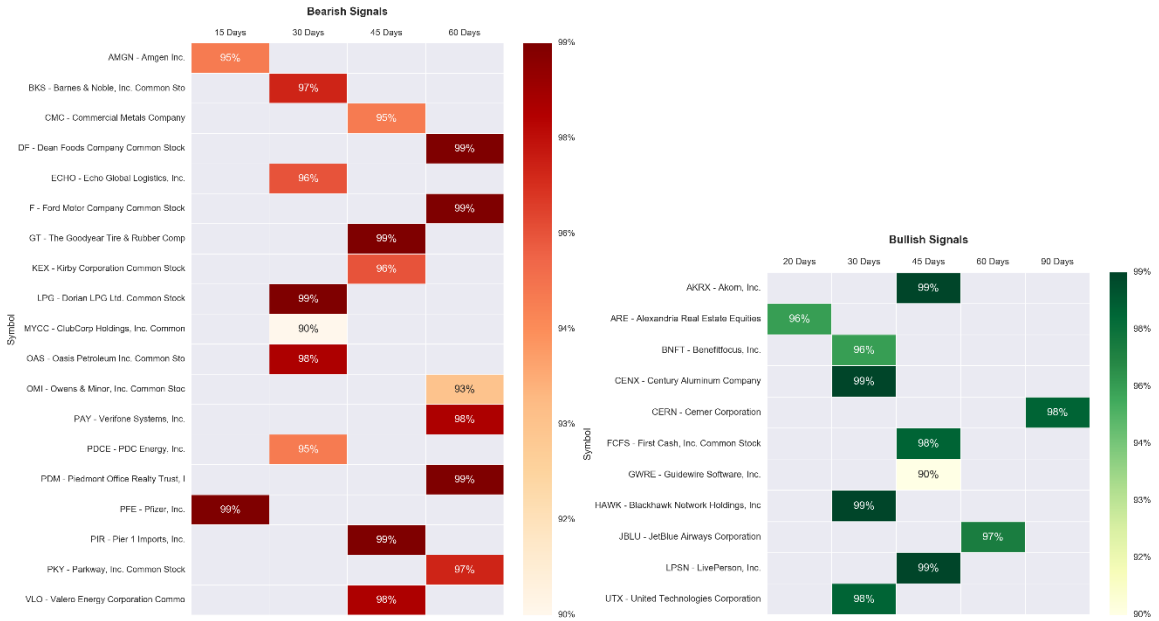
ממשק המשתמש מציג את המכשיר הפיננסי והאיתותים שנוצרו עד כה. כמו כן הממשק מאפשר חיפוש של מכשירים פיננסים וסינון הנתונים לתקופות של חודש, שלושה חודשים או שנה. איור 5.2 מתאר את ממשק המשתמש של מערכת ה-FXMiner. הממשק מאפשר לחפש מכשירים פיננסיים ולקבל תצוגה של איתותים היסטוריים. ממשק המשתמש מבוסס על טכנולוגיות WEB, כגון HTML, CSS ו-JSP.

איור 5.2 : מניה ומספר איתותים לקניה/מכירה

איור 5.3 מציג רשימה של איתותי קניה ומכירה למניות תוך שימוש בדיאגרמת Heat map, המציגה את הנתונים בצורה של טמפרטורות לפי עוצמת האות. עוצמת האיתות נקבעת לפי ההסתברות לתחזית. בדוגמה זו האיתותים הם בין 97% לעליה של מניות כגון ETL לעומת 93% לירידה של מניות כגון FR.



איור 5.3 : איתותי קניה ומכירה של מניות



איור 5.4 : איתותי קניה ומכירה של מניות לפי טווחי החיזוי

איור 5.4 מציג רשימה של איתותי קניה ומכירה למניות תוך שימוש בדיאגרמת Heat map (מאפשרת תיאור ויזואלי בצורה של חום וקור לערכים נומריים) נוספת כך שהפעם החלוקה

היא לפי המניה, טווח החיזוי וההסתברות לתחזית. טווח החיזוי הוא הטווח שהאיתות תקף לגביו. בדוגמה זו מניית ARE צפויה להעלות ב-20 הימים הקרובים ואילו מניית PAY צפויה לרדת ב-60 הימים הקרובים. שני האיתותים הם בעלי הסתברות של למעלה מ-0.9.

5.5 בדיקות

להלן מספר בדיקות יחידה (Unit Test):

- בדיקות נתונים – בדיקות גישה לממשקי המשתמש, כולל ניקוי הנתונים.
- בדיקות להוספה של נתונים חדשים – בונים מספר רב של אינדיקטורים טכניים על בסיס הנתונים הגולמיים. התהליך פגיע לשגיאות וערכים חריגים. נעזר בבדיקות לשלב הזה.
- בדיקות של חישוב התנודות – חישוב התנודות לפי השיטה שהצגנו הוא ייחודי ולכן דורש בדיקות מעמיקות לערכי קצה ונתונים חסרים.
- בדיקות תקינות ממשקי פיתוח התכנה לאיסוף נתונים היסטוריים – בחינה של אפשרות לאסוף נתונים היסטוריים עבור מספר אקראי של מניות (שירותי איסוף הנתונים כגון Yahoo Finance ידועים בכך שאינם זמינים לעיתים קרובות).

להלן רשימה של בדיקות האינטגרציה הכללית במערכת שמבוצעות מדי שבוע:

- בדיקה של תקינות המכשירים הפיננסיים – בדיקה של כלל המכשירים הפיננסיים שעבורם בונים מודלים. הבדיקה כוללת וידוא התאמה בין המכשיר לבין תקינות הנתונים ועמידה בדרישות סף מבחינת נתוני ביקוש והיצע. הבדיקה מתבצעת אחת לשבוע.
- בדיקת אינטגרציה רחבה – מורכבת ממספר בדיקות ומשמשת לשינויים רוחביים בגרסה.

6 מקרה בוחן

בפרק זה נדגים מקרה בוחן. מטרות מקרה הבוחן הן: 1. הצגת יכולת סיווג וביצועים דומים לאלו שבספרות [2, 4, 7]. 2. הצגת הבדל מהותי בשיטת ההשוואה של המודלים אל מול עבודות קיימות – הן בסימולציית המסחר והן בהצבת רף גבוה יותר לסינון המודלים. בסדרת ניסויים שערכנו במשך מעל שלושה חודשים נציג תוצאות מובהקות ומסחר מורווח עבור למעלה מ-1000 עסקאות.

6.1 שלבי ביצוע מקרה הבוחן

מקרה הבוחן מחולק למספר שלבים עיקריים. בשלב הראשון נבנו מודלים אחת לשבוע תוך שימוש בתהליך האצווה היומי. בשלב השני בוצע מסחר באמצעות המודלים שנבנו. בשלב השלישי בוצע מעקב אחר תוצאות המסחר, ניתוח התוצאות ובדיקת מובהקות לרווחיות.

בניגוד לעבודות הקיימות בתחום בחרנו להראות תוצאות של מסחר לעומת סטטיסטיקה של איתותים. המטרה של הניסוי הייתה לסמלץ מסחר אמיתי תוך שימוש באיתותים שנוצרים נכון לאותו רגע ולא להציג סטטיסטיקה אפוסטריורית.

השתמשנו בפלטפורמת המסחר MetaTrader 5 ונתונים היסטוריים של שוק המט"ח בכדי לסחור במשך מספר חודשים על חשבון מדומה. המודלים נבו על ידי אופטימיזציה של טווחי החיזוי. להלן שתי דוגמאות לאופטימיזציות שונות, תוך חיפוש מקסימום מקומי.

6.2 מטריקות איכות

הוגדרו מספר מטריקות לבחינה של הביצועים ואימון המודלים. להלן המטריקות והשיטות לחישוב [5].

$$TP - True\ Positive$$

$$TN - True\ Negative$$

$$FP - \ False\ Positive$$

$$FN - False\ Negative$$

$$Precision = \frac{TP}{TP + FP}$$

$$Recall = \frac{TP}{TP + FN}$$

$$AUC = \int_{\infty}^{-\infty} TP(x)\big(-FP'(x)\big)dx$$

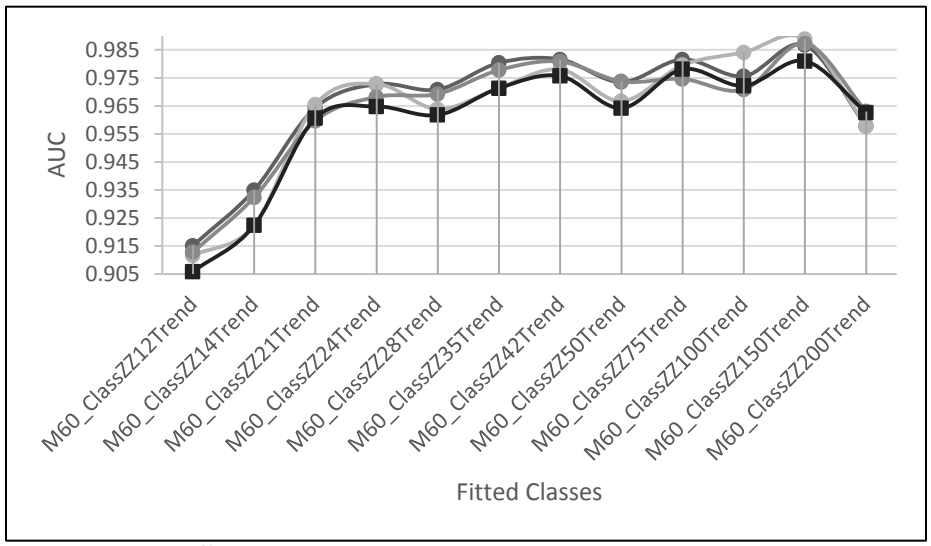
לסינון המודלים נשתמש בסף שנע בין 0.8 ל-0.9 עבור AUC, Precision ו- Recall.
המספרים מבוססים על בחינה של עשרות מודלים ואופטימיזציה של מקסימום מקומי. AUC היא מטריקה בלעדית למשתני מטרה בינריים ואילו Precision ו-Recall משמשים גם משתני מחלקה מולטינומיים. מטריקות אלו מאפשרות לאתר מודלים טובים על ידי הצבת סף גבוה. ערכים של AUC=0.5 נחשבים למודל בעל יכולת חיזוי הדומה להטלת מטבע ואילו AUC=1 הוא מודל בעל טעות חיזוי של 0 (כלומר המודל האופטימלי, מה שאינו ריאלי בד"כ).

6.3 ניתוח תוצאות

ניתוח התוצאות כולל הן שלבי סינון המודלים והן ניתוח לבחינת איכות המודל המורכב מהניתוח של איכות המודלים שהתקבלו אפריורית (לפני הרצת הסימולציה או הנפקת איתותי המסחר) ומניתוח איכות המודלים לאחר הנפקת האיתותים והמסחר לפיהם.

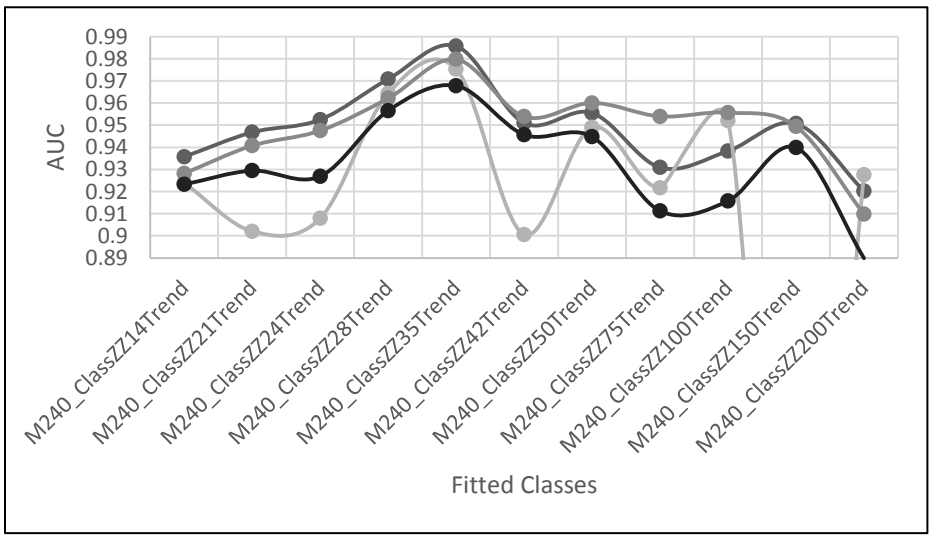
סינון מודלים המייצרים איתותים מתבצע על ידי בעיית חיפוש ומספר היוריסטיקות. בשלב הראשון נבחן מספר רב של מחזורים אפשריים לטווח החיזוי כגון 5, 10, 15 ימים וכך הלאה קפיצות של 5 עד 300, כאשר 300 יום מציינים תחזית או מודל שיודע לתת תחזית ל-300 ימי מסחר קדימה. נציב סף מינימלי של 0.9 לשלושת המטריקות AUC, Precision ו-Recall בשילוב עם מספר חלוקות לקבוצות אימון וניסוי.

באיור 6 ניתן לראות כי טווח החיזוי עבור 150 ימים קדימה מקבל מקסימום מקומי ולכן יכנס כמודל מועדף בשלב הסיווג בזמן אמת. ניתן לראות כי מדובר בבעיית חיפוש. מרחב החיפוש האידיאלי יהיה כאמור לבדוק את כלל הטווחים הקיימים בין יום לשלושה חודשים. מכיוון שמרחב החיפוש גדול, נעזר בהיוריסטיקה של הפחתה במספר הטווחים לפי טווחים של שבועות.



איור 6 : אופטימיזציה של טווח החיזוי בצמד EURUSD

באיור 6.1 מתוארת דוגמה נוספת לאופטימיזציה של טווח החיזוי. במקרה הזה המקסימום המקומי עומד על 25 ימי חיזוי קדימה ולכן נשתמש במודל זה לסיווג התנודות בזמן אמת. חשוב לציין כי בהקלה זו עלולים לפספס את המקסימום המקומי ומצד שני אין באפשרותנו לעבור על כל הטווחים האפשריים.

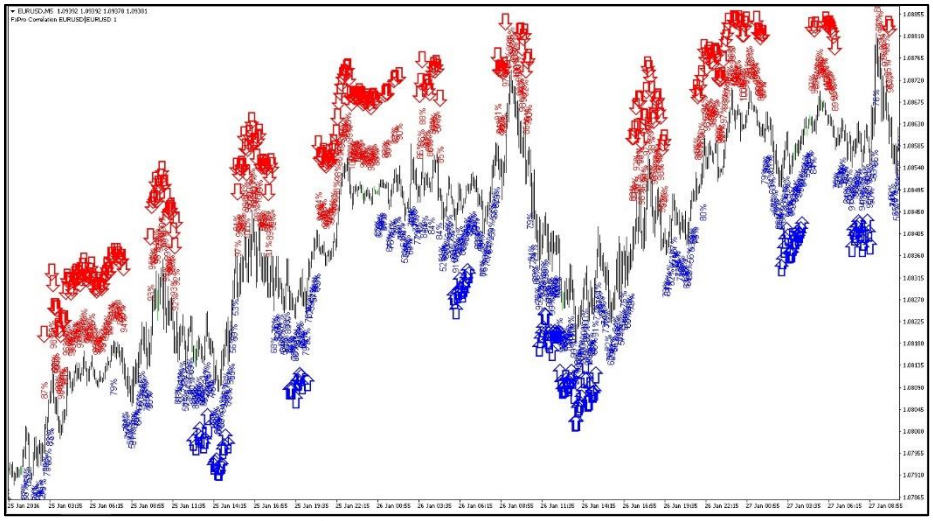


איור 6.1 : אופטימיזציה של טווח החיזוי בצמד EURUSD

אחת המטריקות המרכזיות שמבדילות עבודה זה לבין העבודות הקיימות היא הסימולציה של המערכת במסחר מדומה. בניגוד לעבודות הקיימות נתבסס על מסחר מדומה לעומת סימולציה של מסחר היסטורי. מסחר מדומה מהווה קירוב טוב יותר לביצועי המודל. איור 6.2 מציג את ביצועי המסחר של המערכת כעבור מספר חודשים. ניתן לראות כי המערכת מורווחת עם קו רגרסיה יציב. נתבסס על חוק המספרים הגדולים, כלומר היכולות לבצע מספר רב של עסקאות.

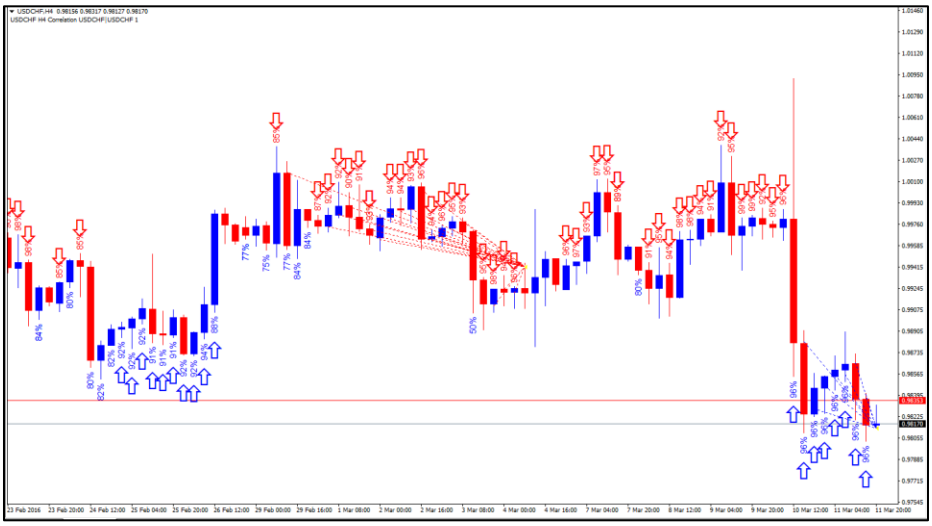
איור 6.2 : ביצועי המסחר המדומה

איור 6.3 מציג דוגמה לסיווג בזמן אמת (בשלב המסחר). החיצים האדומים הם איתותי מכירה ואילו החיצים הכחולים הם איתותי קניה. נקבע סף הסתברות מינימלי לאיתות, במקרה הזה הסף נקבע ל-0.75. קביעת הסף מתבצעת על ידי בחינה של ממוצע ההסתברויות של התחזיות.



איור 6.3 : סיווג בזמן אמת של תנודות

איור 6.4 מציג דוגמה נוספת לסיווג בזמן אמת עם ההסתברויות לעליה או ירידה של המכשיר הפיננסי. החצים האדומים הם איתותי מכירה והחיצים הכחולים הם איתותי קניה. המחיר של צמד המט"ח (במקרה זה USDCHF, דולר ארה"ב לפרנק שווייצרי) מתואר על ידי סדרה עיתית של מחירי פתיחה, גבוה, נמוך וסגירה (נרות יפניים) [15].



איור 6.4 : סיווג תנודות בזמן אמת

בנוסף למסחר המדומה, המערכת מאפשרת הנפקת סטטיסטיקות על ביצועי האיתותים שנוצרו בעבר. להלן מספר דוגמאות לאיתותים והביצועים המתאימים :

איתותי קניה:

MOS bullish signal (86%) for the next 20 days. Since 2016-11-25 we've made: 3.96 Open: 27.45 Close: 31.41 Change: 14.43%

AIZ bullish signal (92%) for the next 85 days. Since 2016-11-25 we've made: 8.64 Open: 86.19 Close: 94.83 Change: 10.02%

ARRY bullish signal (88%) for the next 45 days. Since 2016-12-23 we've made: 1.50 Open: 8.61 Close: 10.11 Change: 17.42%

GM bullish signal (93%) for the next 15 days. Since 2016-11-25 we've made: 3.65 Open: 33.86 Close: 37.51 Change: 10.78%

WDC bullish signal (92%) for the next 10 days. Since 2016-11-25 we've made: 9.71 Open: 61.04 Close: 70.75 Change: 15.91%

HES bullish signal (89%) for the next 10 days. Since 2016-11-25 we've made: 5.65 Open: 53.20 Close: 58.85 Change: 10.62%

SNI bullish signal (87%) for the next 15 days. Since 2016-12-05 we've made: 6.93 Open: 68.47 Close: 75.40 Change: 10.12%

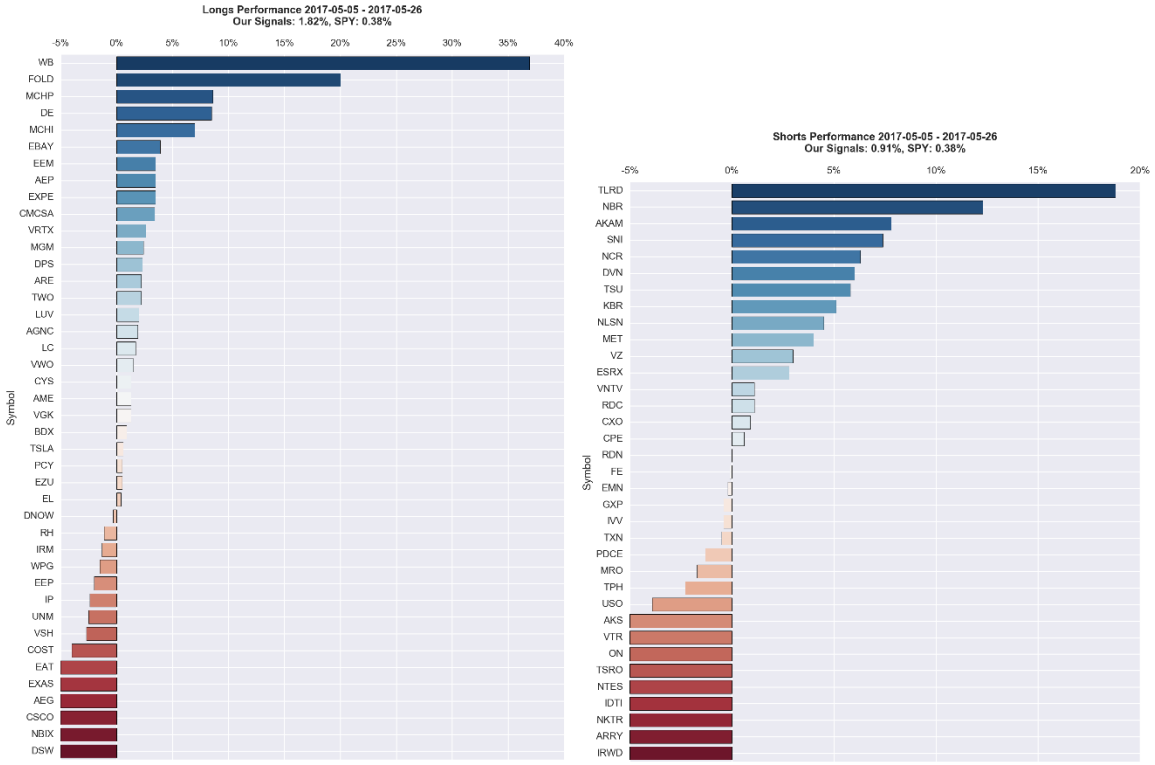
איתותי מכירה:

ENDP bearish signal (84%) for the next 25 days. Since 2016-12-05 we've made: 1.93 Open: 15.80 Close: 13.87 Change: 12.22%

UA bearish signal (89%) for the next 25 days. Since 2016-11-25 we've made: 3.87 Open: 30.71 Close: 26.84 Change: 12.60%

URBN bearish signal (88%) for the next 50 days. Since 2016-12-05 we've made: 4.65 Open: 31.67 Close: 27.02 Change: 14.68%

לאחר סינון המודלים נעבור לבחינת איכות המודלים על פי סטטיסטיקת הסימולציה של המסחר. כעת נבחן את איכות המודלים אפוסטריורית (כלומר אחרי שיצרנו את איתותי המכירה והקניה). איור 6.5 מציג השוואה של מספר איתותים היסטוריים אל מול המדד במשך אותה התקופה. ניתן לראות איתותי קניה (Long) ואיתותי מכירה (Short). השוואה מתבצעת מול מדד S&P500 [8]. כל העסקאות נחסמו בהפסד מקסימלי של 5%, כלומר ירידה של 5% נחשבת לאיתות שגוי. בדוגמה ניתן לראות כי מעל 60% מן האיתותים הם נכונים, והביצועים המשותפים של התיק הם מעל כפליים מן המדד.



איור 6.5 : ניתוח ביצועים היסטוריים והשוואה מול המדד

7 סיכום והצעות להמשך מחקר

במסגרת הפרויקט פותחה מערכת היברידית לאיתותי קניה ומכירה במכשירים פיננסיים שונים (מניות, תעודות סל וצמדי מט״ח). בפרויקט זה הצגנו שיטה כוללת לבניה של מודלים מנבאים ויצירה של איתותי מסחר. הפרויקט מהווה כלי תומך החלטה לסוחר הביתי. בחנו את ביצועי המערכת על מניות וצמדי מט״ח על ידי הרצה של סימולציית מסחר. התוצאות של הסימולציה היו חיוביות ומובהקות. זהו הבדל מהותי בין העבודות הקיימות לבין פרויקט זה. בנוסף לכך בחנו את המודלים לפי מטריקות של למידת מכונה.

המערכת מציגה ממשק נוח מהיר לאיתותי מסחר מבוססים על כריית מידע. זהו כלי עזר שיש בו צורך לסוחר הקמעונאי. המערכת מציגה מחזור חיים סטנדרטי בתחום כריית המידע והמסחר האלגוריתמי. המערכת מאפשרת הרחבה למסחר אלגוריתמי תוך חיבור מערכת האיתותים למנוע מסחר וביצוע העסקאות בזמן אמת.

חשיבות הפרויקט בהיבט של מדעי המחשב היא היכולת לספק ניתוח מתקדם של נתונים היסטוריים ואיתור תבניות בצורה אוטומטית. חשיבות הפרויקט בהיבט המסחר היא הקלה משמעותית בעבודתו של הסוחר אשר עמוס בנתונים וחוסר זמן. המערכת מאפשרת ניתוח רוחבי של מכשירים פיננסיים.

הפרויקט מהווה בסיס למחקרים במספר תחומים הן בהיבט כריית המידע והן בהיבט המסחרי. ניתן לאתר מודלים נוספים ליצירה של איתותי מסחר ושילוב מאפיינים נוספים כגון סחירות, נתוני זמן אמת, נתוני היצע וביקוש וכדומה. כיוון מחקר נוסף הוא שיפור המטה-מסווג על ידי הוספה של מסווגים חדשים או הסרה של מסווגים קיימים. בפן המסחר ניתן להרחיב את המערכת למכשירים פיננסיים חדשים כגון אופציות או מטבעות קריפטוגרפיים (כגון ביטקוין).

8 נספחים

8.1 פרסומים

- אלגוריתם חדשני לבניה של רשת בייסיאנית אופטימלית [1].
- סיווג אנסמבלי של תנודות בשוק המט״ח [2].

8.2 הוראות התקנה

1. התקנת java 8
2. התקנת aws cli
3. התקנת maven
4. העתקת הפרויקט לשרת
5. הרץ maven build install בתוך הפרויקט
6. התקנת Anaconda, Python 3.6

8.3 הוראות הרצה

1. יש להריץ את Scheduler.jar משורת הפקודה על ידי הרצת: java –Xmx1G –jar Scheduler.jar, בניית המודלים מדי יום אורכת כ-12 שעות.
2. בדפדפן יש לעבור לכתובת localhost/FXMiner

9 רשימת מקורות

[1] A. Kreimer, M. Herman, A Novel Structure Learning Algorithm for Optimal Bayesian Network: Best Parents. *Procedia Computer Science*. 96, pp. 43–52, 2016.

[2] A.A Baasher, M.W. Fakhr, Forex trend classification using machine learning techniques. *Recent Researches in Applied Informatics and Remote Sensing, ISBN: 978-1-61804-039*, *8*, pp. 41-47, 2012.

[3] C. Bielza, P. Larrañaga, Discrete Bayesian Network Classifiers: A Survey. *ACM Computing Surveys,47*(1), In press, 2014.

[4] V. Delage, C. Brandlhuber, K. Tuyls, G. Weiss, Multi-Agent based simulation of FOREX exchange market, *The Netherlands Teramark Technologies GmbH*, Maastricht University, Department of Knowledge Engineering, Munich, Germany, 2010.

[5] M. Hall, E. Frank, G. Holmes, B. Pfahringer, P. Reutemann, I. Witten, The WEKA data mining software: an update. *ACM SIGKDD explorations newsletter*, 11(1), 2009, pp. 10-18.

[6] T. Hastie, R. Tibshirani, J. Friedman, J. Franklin, The elements of statistical learning: data mining, inference and prediction. *The Mathematical Intelligencer*, *27*(2), p. 83-85, 2005

[7] F. Jin, N. Self, P. Saraf, P. Butler, W. Wang, N. Ramakrishnan, Forex-foreteller: Currency trend modeling using news articles. In *Proceedings of the 19th ACM SIGKDD international conference on Knowledge discovery and data mining*, ACM, pp. 1470-1473, August, 2013.

[8] B. Johnson, Algorithmic Trading & DMA: An introduction to direct access trading strategies. Vol. 200. 4, Myeloma Press, 2010.

[9] P. Larrañaga, H. Karshenas, C. Bielza, R. Santana, A review on evolutionary algorithms in Bayesian network learning and inference tasks. *Information Sciences,* 233, 2013, pp. 109-125.

[10] K. Lien, Day trading and swing trading the currency market: technical and fundamental strategies to profit from market moves. *Vol. 431. John Wiley & Sons,* 2008.

[11] N. Maknickienė, A. Maknickas, Application of neural network for forecasting of exchange rates and forex trading. In *The 7th international scientific conference Business and Management*, pp. 10-11, 2012.

[12] M. Mayo, Evolutionary data selection for enhancing models of intraday forex time series. *In Applications of Evolutionary Computation*, Springer Berlin Heidelberg, pp. 184-193, 2012.

[13] P.D. McNelis, Neural networks in finance: gaining predictive edge in the market, *Academic Press*, 2005.

[14] P.B. Myszkowski, A. Bicz, Evolutionary Algorithm in Forex trade strategy generation. In *Computer Science and Information Technology (IMCSIT), Proceedings of the 2010 International Multiconference*, *IEEE,* pp. 81-88, October, 2010.

[15] B. Schlossberg, Technical analysis of the currency market: classic techniques for profiting from market swings and trader sentiment. *Vol. 250. John Wiley & Sons, 2006.*

[16] L. Song, M. Kolar, E.P. Xing, Time-varying dynamic Bayesian networks. In *Advances in Neural Information Processing Systems*, 2009, pp. 1732-1740.

[17] H. Talebi, W. Hoang, M. L. Gavrilova, Multi-scale Foreign Exchange Rates Ensemble for Classification of Trends in Forex Market. *Procedia Computer Science*, *29*, 2065-2075, 2014.

[18] A. Tang, Dynamic Bayesian approach to forecasting. In *ICNC*, pp. 3933-3937, August, 2010.

[19] S. Villa, F. Stella, A continuous time Bayesian network classifier for intraday FX prediction. *Quantitative Finance*, *14*(12), 2079-2092, 2014.

[20] P. Wilmott, *Paul Wilmott on quantitative finance*. John Wiley & Sons, 2013.

[21] Finviz, Stock Screener, http://finviz.com/

[22] A. Kreimer, M. Herman, Ensemble Trend Classification in the Foreign Exchange Market Using Class Variable Fitting. HAIS 2017. *Lecture Notes in Computer Science*, vol 10334. Springer, Cham, 2017.

[23] Apache Spark, Lightning-Fast Cluster Computing, https://spark.apache.org/