Azrieli
College of Engineering
Jerusalem
עזריאלי
מכללה אקדמית להנדסה
ירושלים

המחלקה להנדסת תוכנה

פרויקט גמר - תשע"ח

פיתוח מגוון שיטות לזיהוי אנומליות וקבלת החלטות על פי הצבעת הרוב

Developing Variety of Methods for Identifying Anomalies and Making Decisions According to The Majority Vote

מאת

הדס בן מרדכי

קארין בנסון

מנחה אקדמי: דר' גיא לשם	**אישור:**	**תאריך:**
רכז הפרויקטים: מר אסף שפיינר	**אישור:**	**תאריך:**

Ramat Bet Hakerem, P.O.B 3566, Jerusalem 9103501 רמת בית הכרם, ת.ד. 3566, ירושלים מיקוד
Tel. +972-2-6588000 | Fax. +972-2-6588001 | Email: info@jce.ac.il | Web: www.jce.ac.il

מערכות ניהול הפרויקט:

#	מערכת	מיקום
1	מאגר קוד	https://github.com/karinbe/Developing-Variety-Of-Methods-For-Identifying-Anomalies-
2	יומן	https://calendar.google.com/calendar/embed?src=caki4u5vh65nckeb6gos39k8qc%40group.calendar.google.com&ctz=Asia%2FJerusalem
3	ניהול פרויקט (אם בשימוש)	
4	הפצה	
5	סרטון גרסת אלפא	https://drive.google.com/open?id=15Rjmhd7a3Tmx1-LQKO9XOrBfWIoPWhKw

תקציר

"כל המציל נפש אחת, כאילו הציל עולם ומלואו"

במסגרת סגירת התואר של החוג הנדסת תוכנה BS.c ב-"מכללת עזריאלי המכללה להנדסה ירושלים", פרויקט הגמר שלנו יעסוק בפיתוח שיטות לזיהוי אנומליות.

גילוי אנומליות בנתונים, על כל סוגיהם, נהפך להיות נושא מחקרי חשוב ופופולארי בעולם. זיהוי מצב חריג, בקרב קבוצת נתונים, מעיד כי משהו לא תקין התרחש ויש לדעת לזהות אותו ובמקרים מסוימים לעמוד על תיקונו.
כמו כן, הרשת הביאה נוחות לעולם בכך שהיא מאפשרת מעבר מהיר של נתונים אך בו במקביל היא חושפת אותנו ומאפשרת לנו להיות פגיעים.

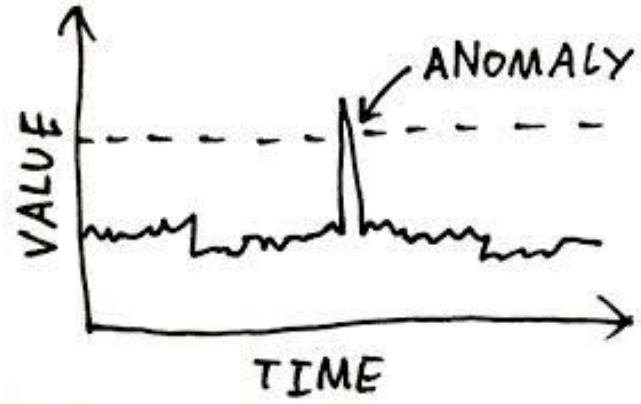
על אף ההתפתחויות הטכנולוגיות והכלים השונים להיענות הבעיה, עדין ישנו קושי לכסות בדיקות מקיפות שיאתרו את כל החריגות שישנן במלוא האחוזים, ואף בעת איתורן לדעת האם הם נכונים או לא. בעניין זה, במחקרנו אנחנו מציעות התבוננות עמוקה למודל של גילוי אנומליה, אשר מטרתו למקסם את הפתרון לבעיה תוף מינוף הסיכויים לאיתור מצב חריג, אנומליה, בזמן. לצורך המחשת הרעיון בחרנו לאמץ את הבדיקות שלנו על נתונים מעולם הרפואה, כשם יש חשיבות מאוד קריטית לדעת לעמוד מצבים של חריגות בקרב חולים ובריאים.

הצהרה

הפרויקט נעשה בהנחיית ד"ר גיא לשם במכללת עזריאלי המכללה להנדסה ירושלים, במחלקה להנדסת תוכנה. החיבור מציג את עבודתנו האישית ומהווה חלק בלתי נפרד מהדרישות לקבלת תואר ראשון בהנדסת תוכנה. העבודה מתבצעת בזוג נוכח היקף העבודה המחקרית הגובלת בנושא וכן מתקיימת חלוקת עבודה.

מבוא

אנומליה, או חריגה, פירושה דפוס התנהגות שאינו תואם לאירועים או דפוסים צפויים, כלומר תבנית החורגת מההתנהגות התקינה. קרי, על ידי פעולות ניטור המערכת, על נתונים, ניתן לתאר את המצב כ"נורמלי" ומנגד, בעת חריגה או פגם, כמצב "לא נורמלי" וזוהי למעשה האנומליה.



איתור אנומליה היא סוגיה חשובה בתחומים ומערכות רחבות כללים, ביניהם אבחון רפואי, זיוף זהות ביטוח, חדירה לרשת, פגמים בתכנות ועוד. לדוגמה, זיהוי דפוס תנועה חריג ברשת מחשב עשוי להיות סימן לכך שמחשב פרוץ שולח נתונים רגישים לגורם לא מורשה.
גילוי אנומליה מבוססת לרוב על שיטות של כריית נתונים - הפעלת אלגוריתם או תוכנת מחשב לצורך גילוי מידע הטמון בבסיסי נתונים קיימים, והסקת מסקנות מהצלבתו. בפועל, הדרך הפשוטה לאבחנת חריגות היא להגדיר תחילה התנהגות נורמלית ובעת התקלות בדפוסי נתונים לא צפויים הדבר יצביע על התנהגות בלתי תקינה -חריגה - וזוהי האנומליה.

הדפוסים הלא מתואמים האלה מכונים לעתים קרובות אנומליות, חריגות, חריגים, או הפתעות. טכניקות לזיהוי אנומליה פותחו במספר קהילות מחקר בעולם. חריגות החלו להיחקר כבר החל במאה ה-19 בקהילת הסטטיסטיקה. רבות מהטכניקות שבאו ליישם שיטות לזיהוי אנומליה הללו פותחו במיוחד עבור תחומים מסוימים של יישומים, בעוד שאחרות הן גנריות יותר.

בניגוד ל"מערכות מבוססות חתימה", אשר ביכולתן לזהות אנומליות שעבורן נוצרה בעבר חתימה, כאן מתאפשר זיהוי אנומליות חדשות. הסיווג מבוסס על כללים, ולא על דפוסים או על חתימות, ומנסה לאתר כל סוג של שימוש לא נכון הנופל מפעולת המערכת הרגילה.

כמו כן, איתור אנומליות ברשת באמצעות אופי התפלגות התעבורה נעשה יותר ויותר פופולרי על פני חיפוש חריגות בנפח התנועה בתעבורה.

היבטים שונים של בעיית זיהוי האנומליה

ניסוח ספציפי של הבעיה נקבע על ידי מספר גורמים שונים כגון אופי נתוני הקלט, הזמינות של תוויות והפלט המוחזר.

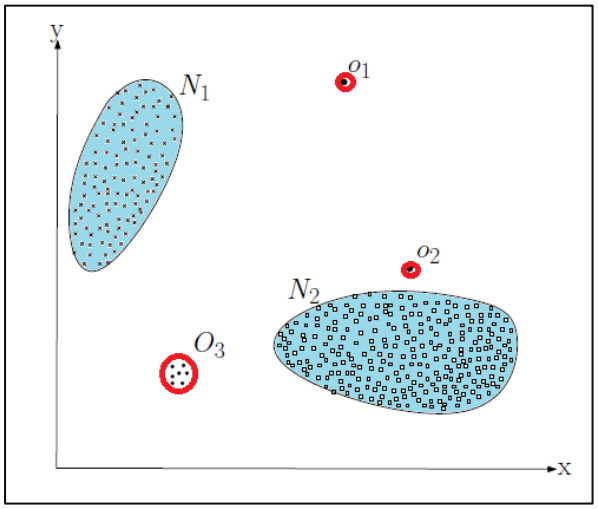
1. <u>נתוני קלט</u>

הקלט הוא בדרך כלל אוסף של מופעי נתונים (אירועים, רשומות, דפוסים , מדגמים, תצפיות, ישויות, אובייקטים). כל מופע יכול להיות מתואר באמצעות קבוצה של תכונות (תכונות, משתנים, מאפיינים). התכונות יכולות להיות מסוגים שונים כגון בינארי, קטגורי או רציף. כל מופע נתונים עשוי לכלול רק תכונה אחת (חד-פעמית) או תכונות מרובות (מרובות משתנים).

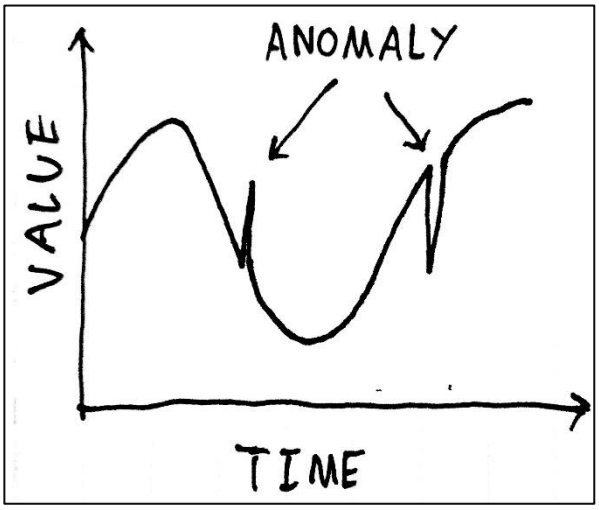
אופי התכונות קובע את הטכניקות לזיהוי האנומליה. לדוגמה, עבור טכניקות סטטיסטיות יש להשתמש במודלים סטטיסטיים לנתונים רציפים וקטגורים.

2. <u>סוג האנומליה</u>

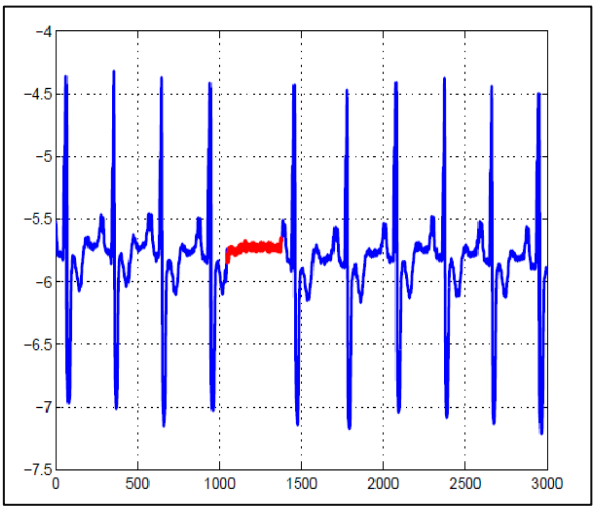
היבט חשוב של טכניקת זיהוי אנומליה היא טבעו של האנומליה הרצויה. אנומליות ניתן לסווג לשלוש קטגוריות הבאות:



2.2 "נקודות אנומליות" (Anomalies Point) - נתונים מבודדים אשר נמצאו כחריגים ביחס לשאר הנתונים, בסביבתם.



2.3 "אנומליית הקשר" (Contextual Anomalies / Conditional Anomaly) - מקרה בו מופע נתון נמצא חריג בהקשר ספציפי. כאן, כל תכונה למופע יכולה להיות התנהגותית או הקשרית.
לדוגמה, טמפרטורה של 35 מעלות בחודש אוגוסט הוא מצב שגרתי, אך אם אותה טמפרטורה תהייה גם בחודש חורפי, כגון פברואר, המצב יחשב כחריג.



2.4 "אנומליה קולקטיבית" (Collective Anomalies) - חריגה קולקטיבית הינו מצב בו אוסף נתונים מהווה מופע חריג, ביחס לסביבתו. חשוב לציין כי מופעים בודדים בקרב האנומליה הקולקטיבית לא נחשבים חריגים כשלעצמם, אך המופע שלהם יחד הוא חריג.

3. <u>תוויות נתונים</u>

בהתבסס על מידת הזמינות של נתונים, טכניקות זיהוי אנומליה יכולות לפעול בשלושה מצבים: פיקוח, חצי פיקוח וללא פיקוח.

3.1. פיקוח

טכניקות הפועלות במצב מפוקח מניחות את הזמינות של מערך נתוני האימון (training data) שבו המופעים מתויגים כנורמליים או אנומליים. גישה אופיינית במקרים כאלה היא לבנות מודל מנבא עבור מצב נורמלי ומצב של אנומליה, ולאחר מכן משווים כל מופע נתונים שרוצים לבדוק למודל כדי לבדוק האם הוא נורמלי או אנומלי.

ישנן שתי בעיות מרכזיות שעולות בזיהוי אנומליה בפיקוח:

א. בנתוני האימון יש הרבה פחות מופעים אנומליים על פני מופעים נורמליים.

ב. קבלת תווית מדויקת ומייצגת, במיוחד עבור המקרים האנומליים, היא מאתגרת בדרך כלל.

פרט לשני נושאים אלו, בעיית איתור אנומליה במצב מפוקח דומה לבניית מודלים מנבאים.

3.2. חצי פיקוח

טכניקות הפועלות במצב חצי מפוקח מניחות שבמערך נתוני האימון רק המופעים הנורמליים מתויגים כנורמליים ולא מניחות דבר על המופעים האנומליים. מכיוון שמצב זה אינו דורש תוויות עבור מופעים אנומליים, טכניקות כאלה ישימות יותר מאשר טכניקות בפיקוח. לדוגמה, באיתור בעיות בחלל, תאונה הוא תרחיש אנומלי שקשה לבנות לו מודל.

הגישה האופיינית במקרים כאלה היא לבנות מודל לקבוצת ההתנהגות הנורמלית, ולהשתמש במוזל זה כדי לזהות חריגות בנתוני הבדיקה (test data).

קיימת קבוצה מוגבלת של טכניקות זיהוי אנומליה המניחות שבנתוני האימון רק המופעים האנומליים מתויגים כאנומליים; אך טכניקות כאלה אינן נפוצות, בעיקר משום שקשה להשיג מערך נתוני אימון המכסה את כל ההתנהגות החריגה האפשרית בנתונים.

3.3. ללא פיקוח

טכניקות הפועלות במצב ללא פיקוח אינן דורשות נתוני אימון, ולכן הן ישימות ביותר. הטכניקות בקטגוריה זו מתבסוסות על ההנחה כי מקרים נורמליים שכיחים הרבה יותר מאשר אנומליות בנתוני הבדיקה. אם הנחה זו אינה נכונה, אז טכניקות כאלה סובלות שיעור אזעקות שווא גבוהות.

בענף הרפואה, לרב הנתונים הרפואיים מיוחסים למטופלים בריאים, ולכן רוב הטכניקות המאומצות כאן לצורך ניטור הנתונים יתבצעו בגישת החצי פיקוח.

4. <u>נתוני פלט</u>

היבט חשוב נוסף עבור כל טכניקת זיהוי אנומליה היא האופן שבו התוצאות מדווחות . ככלל, ניתן לסווג את התוצאה של זיהוי האנומליה בשני אופנים :

4.1. ציונים

הקצאת ניקוד בהתאם לרמה בו המופע נחשב אנומליה. לכן התפוקה של טכניקות כאלה היא רשימה מדורגת של אנומליות. אנליסט יכול לבחור לנתח מספר מצומצם של חריגות בולטות או להיעזר במדדי סף לבחינת האנומליה.

4.2. תוויות

הקצאת תווית – נורמלי או אנומליה - לכל מופע מבחן.

במחקרנו, נדון בכלים שניתן להשתמש בהם על-ידי מגיני רשת ומערכי נתונים, לצורך זיהוי אנומליה ברשת. בעניין זה, ניתן התייחסות עמוקה למודל משודרג של גילוי האנומליה עבור לטובת בריאות הציבור.

מבחינת נתונים רפואיים, זיהוי האנומליה נעשה עם רשומות חולה. נתונים יכולים להיות חריגים בשל מספר סיבות כגון מצב חולה חריג, שגיאות מכשור או שגיאות הקלדה. לכן זיהוי אנומליה היא בעיה קריטית מאוד בתחום זה ונדרשת רמה גבוהה של דיוק.

תיאור הבעיה

דרישות ואפיון הבעיה

חשיפת פגמים בנתונים רפואיים שלעצמם יכולה לאותת בפני מצב שיש להיערך לקראתו מבחינה רפואית ועל כן חשיבותה של זיהוי האנומליה בצורה מדויקת קריטית לחולה. לצערנו, טרם נרשמה הנוסחה לבעיה ולכן עודנה קיימת כמות גדולה של אנומליה בלתי מזוהה.

זיהוי מצב חריג בנתונים יכול להקדים תרופה למכה ברב המקרים או להתריע בפני תפנית חשודה שיש לבדוק לעומק.

כמו כן, ישנו גם עניין האבטחה על נתונים רפואיים - מתקיף עשוי לבצע נזק בכמה רמות - החל מחדירה לנתונים אישים, למשל של אישיות מפורסמת אשר מנסה להסתיר את מחלה (אפקט סטרייסנד) ועד לשינוי מינון של תרופה לילד חולה סרטן, או ניסיון פגיעה באדם הצורך תרופות רשומות.

הבעיה מבחינת הנדסת תוכנה

הבעיה שלעצמה קיימת וקשה להבטיח פתרון טכנולוגי שיכסה ויתריע מפני כל חריגה אפשרית. בפועל, כל ארגון או קבוצות ניסויי מאגד לעצמו מערכות לזיהוי חריגות המתאפיינות על ידי חוקים ושיטות, בכדי לזהות ולגלות תובנות עם מאפיינים חריגים.

מבחינת הנדסה, ייחודו של הפרויקט שלנו, הוא שהתבססנו על מגוון שיטות קיימות, ועליהם הלבשנו הליכי עבודה ותוספים חדשים.

מבחינת הנדסת תוכנה, ככל הנראה כעת, הקושי גובל בכתיבת הקוד אשר יבטא את שיטת הפתרון שמחקרנו מעוניין להציג. כמו כן, השאיפה היא להיעזר בתוכנות קיימות, כגון Machine Learning, MATLAB ועוד.

בנוסף לעובדה, שפתרוננו שואף לתת מענה מדויק על וידוא חריגה, עם כריית הנתונים, קושי נוסף עימו אנו עשויים להתמודד הוא מתן המסקנות בזמן מהיר יחסית.

תיאור הפתרון

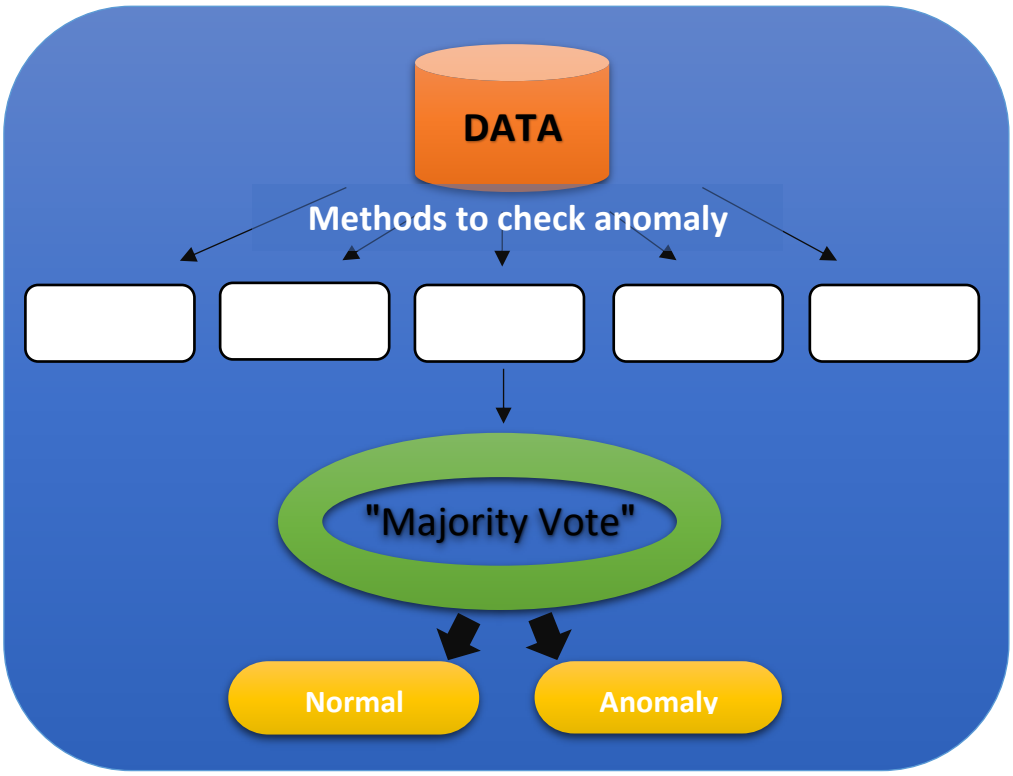
בעבר, אם נתבונן בהתנהלות ששררה בבתי החולים, כל המסמכים של החולה היו מתויקים תחת מספר דפים עם קליפס צמוד למיטת החולה. הרופא היה מוסיף הערות בכתב יד ובכך היה מסתכם "הדוח של החולה". כיום, בתי החולים עוברים למערכות מחשוביות, קרי, כל הרשומות הרפואיות של החולה נשמרות ומנוהלות בשרתי המחשוב.

מהי המערכת

המחקר שלנו מכוון לתת מענה לזיהוי אנומליה, תוך שיעור גבוה של הצלחה באיתור חריגות, בפיתוח של מכשירים רפואיים עתידים.

במסגרת חקר הספרות שעשינו, מצאנו שיטות שונות לזיהוי אנומליות. החל משיטות סטטיסטיות ועד לשיטות לוגריתמיות ושימוש ב - Machine Learning.

לפיכך החלטנו כי הפתרון שיוצג במחקרנו יהיה חקירת ואפיון נתונים עליהם נאמץ ונפתח שיטות שונות של זיהוי אנומליות, בתוכנת MATLAB, ננסה למקסם את שימושן על ידי חישובים, אשר טרם אומצו למערכות זיהוי אנומליות במערכות הרפואה. הדרך שלנו להגיע לסבירות גבוהה והכרעה היא על פי שיטת הצבעת הרוב.



תיאור פתרון מוצע

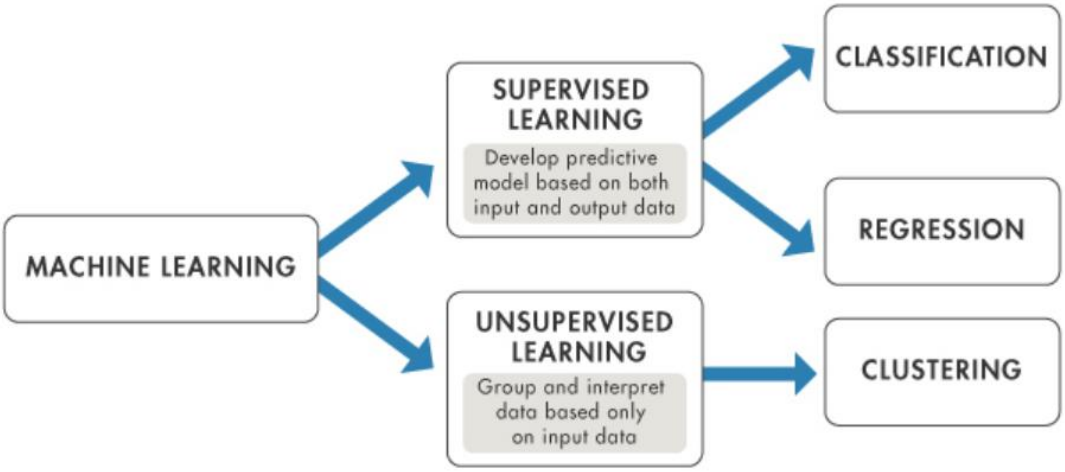
השיטות עמן נפעל על מנת לזהות אנומליה יהיו : למידת מכונה (ML), למפל זיו, אנטרופיה.

נציין כי בשלב זה של מחקרנו אנו בשלבי כתיבת קוד ב-MATLAB אשר יממש את הפרויקט ולכן אין עדיין תוצאות סופיות.

1. **למידת מכונה** (Machine Learning)

למידת מכונה, הוא תת תחום מחקרי שכיח למימוש בינה מלאכותית (AI - Artifical Intelligence). AI הוא למעשה תחום מחקר העוסק בדרכים אשר יאפשרו למחשב לבצע פעולות אשר כיום בני אדם מטיבים לבצע באופן שקול יותר. הגדרה שכיחה ומופשטת עבור למידת מכונה היא יכולת רכיב מכונה לשפר את הביצועים של עצמה וזאת באמצעות שימוש בתוכנה הכוללת בינה מלאכותית אשר מחקות את הדרך שבה בני אדם לומדים, כדוגמת ניסוי וטעייה.
כמו כן, למידת מכונה מתייחסת למחקר, תכנון ופיתוח אלגוריתמים המעניקים למחשב יכולת ללמוד וזאת כשהמחשב טרם תוכנת מראש. תחום זה מציג מספר מתודולוגיות המאפשרות למחשב לבצע משימות אינטליגנטיות בדומה לאדם, כגון חיזוי, זיהוי סיווג והכרה. נוסף על הכישרון שלה להתמקצע ולשפר ביצועים, מטרת למידת המכונה היא גם לזהות ולהתמודד עם פגיעות באבטחת מידע וכשלים פנים מערכתיים הנובעים מכשל אנושי או קוד, לכל הפחות בזמן הקרוב לזמן אמת.

יכולות AI מציעות פתרונות מעניינים לטובת זיהוי תרחישים לא רצויים. למעשה קיימים אלגוריתמי למידת מכונה אוטומטיים אשר ביכולתם לקבוע דפוסי "התנהגות נורמלית" על בסיס מקורות ידע ללא התערבות אנושית. חריגה מ"ההתנהגות הנורמלית" מאפשרת גילוי מוקדם של דפוסים אשר יכולים להעיד על ניסיון חריגה, חדירה, דליפה של מידע או כשל פנימי. כמו כן, עבודה נכונה של למידת מכונה עשויה לצמצם את מערך ה- False Positive Errors וגם את ה-False Negative Errors.



למעשה, שיטת העבודה של למידת מכונה, יכולה להתחלק לשתי שיטות:

- למידה בפיקוח" (Supervised Learning) - בשיטה זו המכונה למעשה נשענת על סדרה ידועה של נתונים (קלטים) שהושגו בצורות שונות ועל המכונה לבצע למידה שלהם. תוך קינפוג למחלקות והתנייה חלקית, המכונה מכשירה מודל ליצירת תחזיות סבירות כתגובה לנתונים חדשים. "למידה בפיקוח" משתמשת בטכניקות סיווג (Classification) ורגרסיה (Regression) כדי לפתח מודלים מנבאים.

Classification - סיווג נתונים לפי קטגוריות. למשל, אם דוא"ל הוא מקורי או דואר זבל, או אם הגידול הוא סרטני או שפיר.
Regression - טכניקות רגרסיה מנבאות תגובות מתמשכות, למשל, שינויים בטמפרטורה או תנודות בביקוש לחשמל.

באופן עקרוני, מאמצים אלגוריתם של למידה ומפעילים על מרבית הנתונים. לאחר מכן ממשיכים לבצע את הבדיקה על הנתונים הנותרים ובכך מוודאים האם הלמידה בוצעה כראוי.

- "למידה ללא השגחה" (Unsupervised Learning) - למעשה, בשיטה הראשונה יש סוג של ציפייה לגבי הנתונים והמשובים הרצויים בעוד שבשיטה השנייה המכונה מגדירה אותם תוך כדי ביצוע ניתוחים ובדיקות.

במקרה זה, הנתונים אינם מסודרים והמכונה מגלה דפוסים בנתונים ללא התניית סיווגים מוכנים מראש, אלא שהם מתרחשים תוך כדי הלמידה.

"קיבוץ אשכולות" (Clustering) היא טכניקת הלמידה הנפוצה ביותר בשיטת ללא השגחה. היא משמשת לניתוח נתונים, מציאת דפוסים מוסתרים או קיבוצים בנתונים. יישומים עבור אשכולות כוללים ניתוחים של רצף גנים, מחקר שוק ועוד.

במחקרנו החלטנו לאמץ את השיטה השנייה, Clustering, ובה בחרנו לעבוד עם אלגוריתם K-means, נקרא גם האלגוריתם של לויד. זוהי שיטה פופולרית עבור ניתוח אשכולות בכריית נתונים. מטרתה לחלק את התצפיות ל-k אשכולות לפי מרכזי כובד (k-means) כאשר כל תצפית משויכת לאחד מ"מרכזי הכובד". על ידי בחירה נכונה של מרכזי כובד ניתן לאתר את הקבוצות השונות.

בהינתן קבוצה של תצפיות (x_{1}, x_{2}, x_{3},…,x_{n}) כאשר כל תצפית היא וקטור ממשי היכול להיות בעל מספר ממדים, המודל שואף לחלק את n התצפיות ל-k אשכולות, על מנת למזער את סכום המרחקים בין התצפיות (הווקטורים) בתוך האשכול ובכך להתכנס למרכזי כובד מקומיים ואופטימליים. האלגוריתם מתחיל בשלב האתחול ונע בין שני שלבים לסירוגין:

שלב האתחול - חלוקה אקראית של אשכולות לכל תצפית ולאחר מכן מעבר לשלב העדכון.

שלב הקצאה - הקצאת ממוצע שכולל באשכול את סכום הריבועים, לכל תצפית (סכום ריבועים הוא מרחק אוקלידי בריבוע, זה באופן אינטואיטיבי מרכז הכובד "הקרוב ביותר").

שלב עדכון - חישוב מרכזי כובד חדשים כדי להיות במרכז הכובד הגאומטרי של התצפיות באשכולות החדשים.

האלגוריתם מתכנס כאשר לא ניתן לשנות עוד את הנתונים.

במקרה שלנו, k=2 מאחר ואנו מעוניינים להבחין בין שני מצבים, שתי אשכולות אוכלוסייה, בריאים ולא בריאים.

לצורך המחשה, לקחנו מדגם של נתונים אודות מספר מצומצם של אנשים - חולים ובריאים, אשר עשויים לחלות בסוכרת. החולים בסוכרת סובלים משני מצבי קצה הפוכים: מצב של עודף גלוקוז בדם, שנוצר לרוב כתוצאה מכמות לא מספקת של אינסולין בדם ומצב של מיעוט גלוקוז בדם, שנוצר לרוב כתוצאה מכמות גבוהה מדי של אינסולין בדם.

	A	B	C	D
1	ID	SugarBlood	Pulse	Age
2	1	50	80	20
3	2	80	70	50
4	3	90	88	41
5	4	130	90	40
6	5	120	83	50
7	6	70	82	21
8	7	79	81	22
9	8	118	85	55
10	9	150	89	57
11	10	78	69	23
12	11	158	71	24
13	12	88	73	25
14	13	89	77	19
15	14	68	78	18
16	15	72	87	51

בדוגמה שלנו ניקח נתונים רפואיים שונים כגון: רמת סוכר, גיל ודופק.

המטרה היא להריץ את המכונה הלומדת על נתונים אלה ולהסיק אילו מבין האישים המופיעים בטבלה חשודים להיות סובלי סכרת, כלומר חריגים. ברב המקרים, הקבוצה הקטנה היא הקבוצה החריגה (מאחר ורב האוכלוסייה מוגדרת כבריאה).
בעת ניתוח הנתונים, המכונה בוחנת כל שתי עמודות ומריצה עליהן את האלגוריתם ובכך למעשה גוזרת מסקנות לגבי הקבוצה החריגה בקרב הנתונים.

נדגים את הניתוחים שהתקבלו כתוצאה מהרצת האלגוריתם:

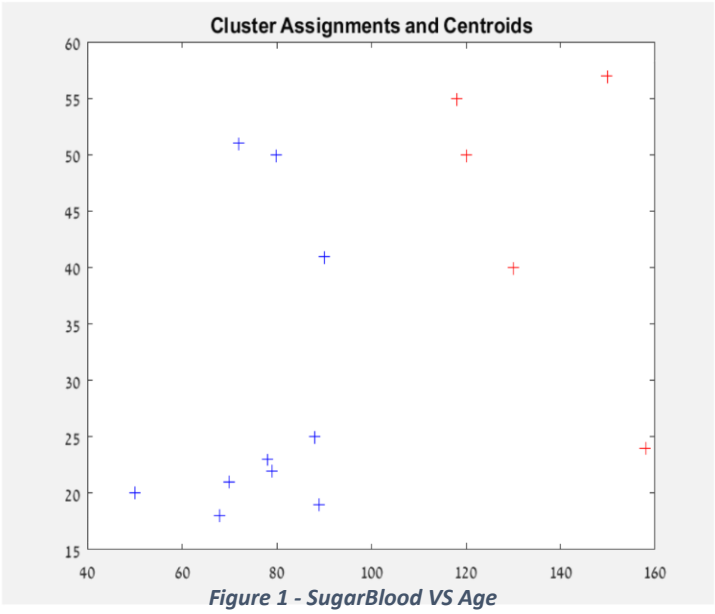


Figure 1 - SugarBlood VS Age

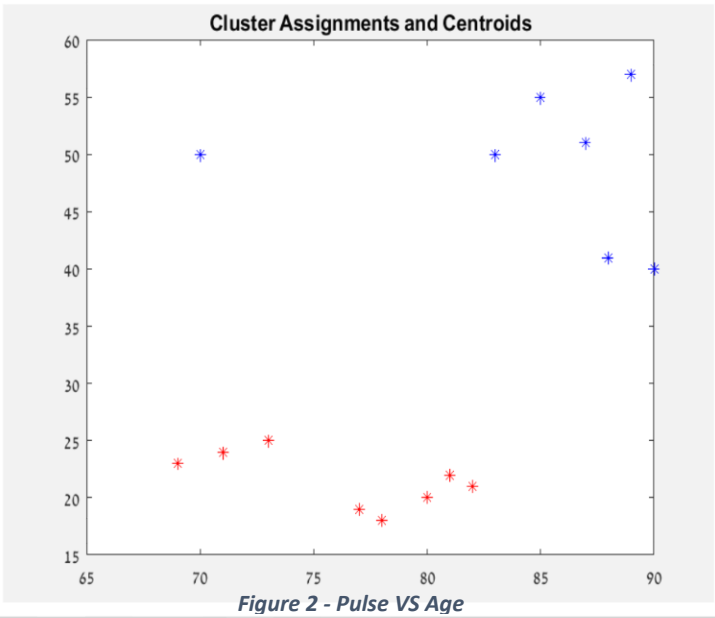


Figure 2 - Pulse VS Age

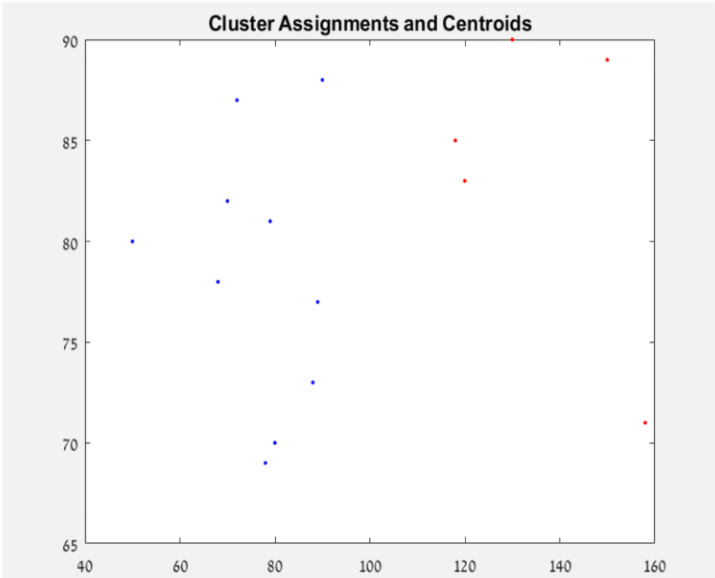


Figure 3 - Pulse VS SugarBlood

```
Command Window
>> ML
Replicate 1, 3 iterations, total sum of distances = 234.
Best total sum of distances = 234
Replicate 1, 2 iterations, total sum of distances = 296.
Best total sum of distances = 296
Replicate 1, 1 iterations, total sum of distances = 108.
Best total sum of distances = 108
4 is Anomaly.
5 is Anomaly.
8 is Anomaly.
9 is Anomaly.
11 is Anomaly.
fx >>
```

כתוצאה מבחינה זו, נמצאו כחשודים לחולים, חריגים, האישים עם ה - ID הבאים: 4,5,8,9 ו-11 נמצאו כחריגים.

	A	B	C	D
1	ID	SugarBlood	Pulse	Age
2	1	50	80	20
3	2	80	70	50
4	3	90	88	41
5	4	130	90	40
6	5	120	83	50
7	6	70	82	21
8	7	79	81	22
9	8	118	85	55
10	9	150	89	57
11	10	78	69	23
12	11	158	71	24
13	12	88	73	25
14	13	89	77	19
15	14	68	78	18
16	15	72	87	51

למעשה, מצאנו כי שיטה זו אכן הצליחה לאתר לנו את היחידים שהנחנו מראש שיהיו חשודים כחולים.

2. אלגוריתם למפל-זיו (Lempel-Ziv)

אלגוריתם למפל-זיו הינו אלגוריתם לדחיסת נתונים. הצורך לקודד מסר הוא מפני שאנחנו רוצים להתאים את המסר לצורה שניתן לטפל בה (מסר מעובד), לאחסן אותו ולהעביר אותו דרך ערוצי התקשורת.

במשך השנים התפתחו אלגוריתמים שונים על בסיס אלגוריתם למפל-זיו אשר שיפרו את הביצועים בצורה משמעותית והתגבשה משפחה של אלגוריתמים. הדחיסה היא מסוג דחיסה משמרת מידע, המאפשרת שיחזור המידע הדחוס במלואו (ללא עיוות). האלגוריתם מתבסס על חלוקת המחרוזת המקודדת לתתי-מחרוזות הנקראות פסקאות בתהליך המכונה פיסוק. כל פסקה מותאמת למחרוזת מעל א"ב סופי ונבנה מילון בתהליך דינמי. האלגוריתם הוא אוניברסלי, הדחיסה היא אסימפטוטית אופטימלית ולא נדרש ידע קודם של התוכן הנדחס.

למפל-זיו ידוע כאלגוריתם דחיסה אופטימלי, ולכן כדאי להשתמש בו. בהקשר של זיהוי אנומליה, ניתן להגדיר מודל סטטיסטי להתנהגות רגילה ולאחר מכן להציע מנגנון לבדיקת רצפים חדשים ולא ידועים תוך שימוש במודל זה.

אלגוריתם למפל- זיו ככלל ובגרסה LZ78 בפרט הוא שיטת דחיסה המבוססת על מילון: עבור רצף של סדרת סמלים נתונה, מילון של ביטויים מנותח מתוך רצף זה. הניתוח מצטבר כדלקמן - בהתחלה, המילון ריק. לאחר מכן, במהלך כל שלב של האלגוריתם, הקידומת הקטנה ביותר של סמלי נתונים עוקבים שטרם נראו - כלומר, שאינה קיימת במילון - מנותחת ומתווספת למילון; לפיכך, כל ביטוי הוא ביטוי ייחודי במילון, אשר עשוי להרחיב ביטוי שנצפה בעבר (ולכן קיים במילון) בסמל אחד. ייצוג נפוץ של המילון הוא עץ מושרש שבו כל ביטוי במילון מיוצג כנתיב מהשורש לצומת פנימי בעץ על פי קבוצת הסמלים מהן הביטוי מורכב.

הליך בניית העץ:

1 - מעבר על כל אות במחרוזת ובדיקה האם כבר קיימת בעץ,

1.2 במידה ולא - נוסיף אותה לעץ (למילון שלנו).

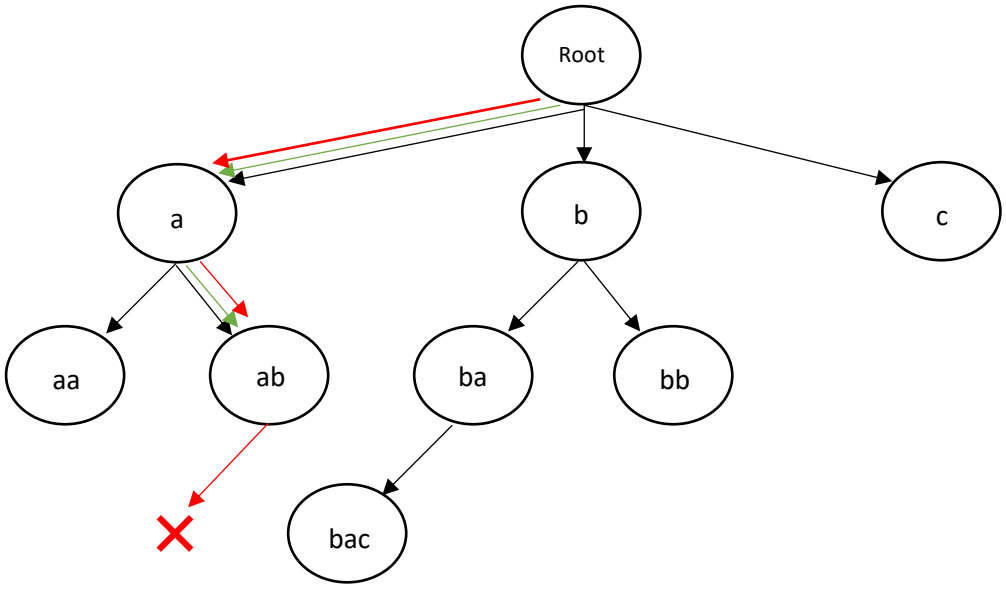
2.2 במידה וכן - נסתכל על תת המחרוזת הנוכחית עם האות הבאה ונבדוק האם השרשור מופיע בעץ.

2 - עם תת המחרוזת הקיימת כרגע נשוב לשלב 1 וחוזר חלילה, עד לסוף המחרוזת.

זמן בניית עץ למפל-זיו הינו לינארי.

הערה: נשים לב כי אלגוריתם הלמידה המבוסס על LZ דורש קלט אלפבית סופי - קיימים שיפורים והצעות כפתרון לבעיה זו, אך אנו לא נעסוק בזה כאן. לאחר בניית העץ, רצפים חדשים (חשודים) ייבדקו מהעץ בזמן לינארי באורך המחרוזת הנבדקת.

לדוגמה, עבור המחרוזת "abbacbaccbcabb", מתקבל העץ הבא:



בעוד מחרוזות כגון "ab" קיימות בעץ ומוגדרות כנורמליות, המחרוזת "aba" תאובחן כאנומליה.

במחקר שלנו, בשיטת גילוי האנומליה בעזרת עץ למפל זיו, החלטנו לעבוד בשני שלבים: בשלב הראשון, בניית עץ למפל זיו על פי רוב הטבלה והגדרת התנהגות נורמלית (נתוני אימון - training data), ובשלב השני חיפוש בעץ של יתר השורות וסווגן לנורמלי או אנומלי על סמך הגעתן לעלה בעץ או שאינן מופיעות בו, בהתאמה (נתוני בדיקה - test data).

גם כאן, כמו בשיטת למידת המכונה, לקחנו לצורך המחשה את אותו מדגם של נתונים, במטרה להריץ את השיטה הנוכחית על אותם נתונים בדיוק ולהסיק על פי שיטה זו אילו מבין האישים המופיעים בטבלה חשודים להיות חריגים; וכך בתום פיתוח שלב זה, נוכל להשוות בין השיטות השונות בתכליתן זו מזו ולהסיק מסקנות בהתאם.

חשוב להדגיש כי היות ובשלב זה של מחקרנו אנו בתהליכי מימוש הפרויקט, אנו עוד עומדים על פיתוחה של שיטה זו. לפיכך, נכון לעכשיו השלב השני טרם פותח לגמרי. עקב כך, ביצענו את

השלב הראשון כרגע על כלל הנתונים ולא רק על הנתונים המוגדרים כנתוני אימון, וזאת לצורך השוואה ובדיקה.

כאמור, הנתונים הרפואיים שלנו:

	A	B	C	D
1	ID	SugarBlood	Pulse	Age
2	1	50	80	20
3	2	80	70	50
4	3	90	88	41
5	4	130	90	40
6	5	120	83	50
7	6	70	82	21
8	7	79	81	22
9	8	118	85	55
10	9	150	89	57
11	10	78	69	23
12	11	158	71	24
13	12	88	73	25
14	13	89	77	19
15	14	68	78	18
16	15	72	87	51

שלב ראשון:

ראשית, המכונה מבצעת קוונטיזציה - כלומר, המרת הנתונים המספריים למחרוזת, כאשר כל תא בטבלה מומר לאות על פי טווחים. נדגים את תוצאת הקוונטיזציה על הנתונים לעיל:

```
FICIHFJIENJEMIFHICHICLIFPIFHGCPHCIHCIHBGHBHIF
```

לאחר היווצרות המחרוזת הארוכה, המכונה בונה את המילון, שמכיל למעשה את תתי המחרוזת שיהיו קדקודים בעץ; העץ ישורטט בעזרת יצירת מערך שמשייך כל בן לאבא שלו ולפיו נבנה העץ למפל-זיו, לפי הסדר הנדרש.

```
Dictionary:
  Columns 1 through 13

    'F'    'I'    'C'    'IH'    'FJ'    'IE'    'N'    'J'    'E'    'M'    'IF'    'H'    'IC'

  Columns 14 through 25

    'HI'    'CL'    'IFP'    'IFH'    'G'    'CP'    'HC'    'IHC'    'IHB'    'GH'    'B'    'HIF'
```

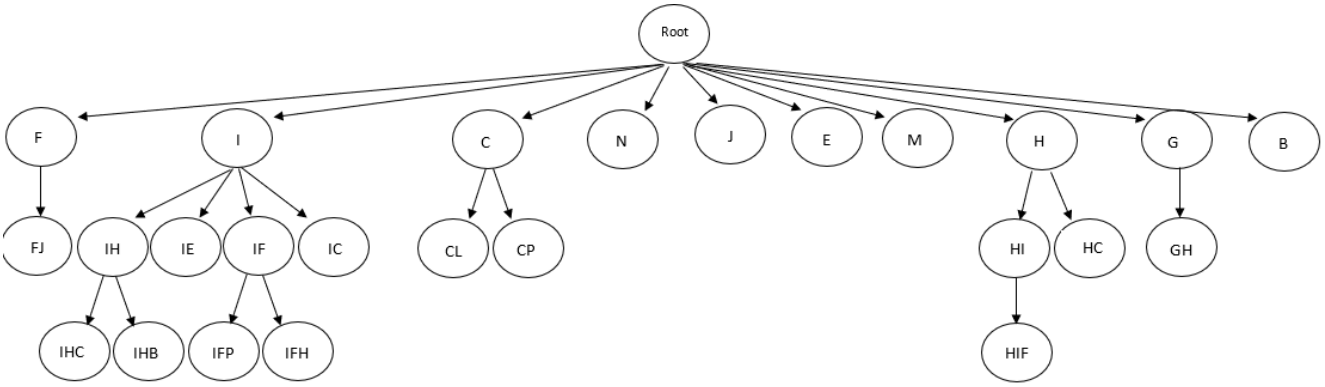
```
Lempel-Ziv Tree:
  Columns 1 through 13

    'F'    'I'    'C'    'N'    'J'    'E'    'M'    'H'    'G'    'B'    'FJ'    'IH'    'IE'

  Columns 14 through 25

    'IF'    'IC'    'CL'    'CP'    'HI'    'HC'    'GH'    'IHC'    'IHB'    'IFP'    'IFH'    'HIF'
```

שרטוט של העץ:



שלב שני:

השלב הבא בפיתוח השיטה - כאשר בניית העץ הושלמה, יש לבצע המרה של כל השורות שנותרו אותן נבחן למחרוזות - כל שורה תיוצג על ידי מחרוזת אחת. על כל מחרוזת נריץ חיפוש בעץ. שורה שהמחרוזת המייצגת אותה מגיעה לעלה בעץ תסווג כנורמלי, ושורה שהמחרוזת המייצגת אותה לא קיימת בעץ, תסווג כאנומלית.

3. אנטרופיה (Entropy)

המושג אנטרופיה הינו חלק בלתי נפרד מהחוק השני בתרמודינאמיקה (מונח בעולם הפיזיקה, נקבע כמושג על ידי רודולף קלאוזיוס) העוסק במעברי אנרגיה המניעים את היקום המוכר לנו. הסבר נוסף, המושג אנטרופיה הינו מדד כמותי המייצג את האנרגיה המשתחררת במערכת אשר מורכבת מחלקיקים רבים המייצרים אנרגיה.

החוק השני קובע שבאופן ספונטאני חום זורם מאזור בעל טמפרטורה גבוהה לאזור בעל טמפרטורה נמוכה. על מנת לגרום לחום לזרום בצורה הפוכה יש להשקיע אנרגיה חיצונית. **משמע, אנטרופיה של מערכת לעולם לא תקטן ללא התערבות חיצונית**. בנוסף, היות והחוק השני של התרמודינאמיקה קובע שהאנטרופיה אינה יכולה לקטון באופן ספונטאני, הרי שכל מערכת סגורה (חדר, מדינה, כדוה"א) הולכת ומתפזרת לאורך הזמן. פיזור זה ממשיך עד שהמערכת מגיעה למצב של שיווי משקל-מקסימום אנטרופיה.

דוגמא: החוק השני של התרמודינאמיקה קובע כי בהכנת תה חם, כמות חום מסוימת תעבור מהמים החמים אל שקיק התה ותחמם אותו. נניח שהכוס והשקיק מהווים מערכת סגורה, כלומר לא מושפעים מהעולם החיצוני. בהתאם להגדרת האנטרופיה, אנטרופיית המים תרד (היות וכמות החום המועבר הינה שלילית - חום יוצא מהמים), בעוד אנטרופיית השקיק תעלה מאותה סיבה. אולם, בגלל שטמפרטורת שקיק התה נמוכה מטמפרטורת המים, הרי שהאנטרופיה של השקיק תעלה יותר מאשר ירידת אנטרופיית המים. לכן, באופן כללי אנטרופיית המערכת (המים והשקיק) עלתה. מה שבעצם נוצר הוא שבתהליך ספונטאני (ללא התערבות חיצונית) האנטרופיה של המערכת גדלה.

הסבר נוסף למושג מגיע מתחום הסטטיסטיקה. נטען, כי האנטרופיה היא בעצם תופעה סטטיסטית המבטאת את חוסר הסדר של המערכת. עוד הוכח כי האנטרופיה גדלה ככל שלגוף מסוים יש יותר מצבים סטטיסטיים בו הוא יכול להימצא.

נניח שכוס התה שלנו נשפכת בשל תנועת יד לא זהירה, מולקולות המים שעד עתה היו מוגבלות בדפנות הכוס מגלות המגבלה הוסרה (באופן חלקי או מלא). היות והמולקולות נמצאות במצב של תנועה מתמדת, וכן הן נמצאות תחת פעילותו של כוח הכובד, סביר (סטטיסטית)

שהמולקולות שבתוך הכוס תדחפנה את חברותיהן הקרובות החוצה, אז התה יישפך מה שיגרום לכתם על השטיח ולגידול בחוסר הסדר בחדר.

מסקנה, ככל שמערכות מורכבות יותר כך גם האנטרופיה מורכבת וגבוהה יותר.

כעת נציג פירוש נוסף למושג מתוך עולם האינפורמציה.

האנטרופיה של שאנון

מוגדרת לפי הנוסחה הבאה :

$$H_s(X) = \sum_{i=1}^{n} p(x_i) \log_a \frac{1}{p(x_i)}$$

אנטרופיה כמדד לאי וודאות על קבוצת מצבים אפשריים X בהסתברויות p(X1),.....p(Xn)

קלוד שאנון (מתמטיקאי, מהנדס חשמל וקריפטוגרף אמריקאי, נחשב לאבי תורת האינפורמציה) אימץ את הרעיון של האנטרופיה לעולם האינפורמציה, לפיו האנטרופיה היא מדד לאי-הוודאות הקשורה למשתנה אקראי.

ככל שהמשתנה יותר רנדומלי כך האנטרופיה יותר גדולה ואותו דבר הפוך - ככל שיש לנו יותר מודעות באשר למשתנים כך האנטרופיה יותר קטנה.

המטרה שלנו בשימוש בשיטה זו ובפיתוחים שלה, היא לאמץ אותה ולהתאים אותה לעקרונות האנומליה. כאשר חוסר הוודאות יהיה קטן פרוש הדבר שאנטרופיה קטנה. אם נאמץ את הרעיון למערכת סגורה ומורכבת, שם יש הערכה לתקינות המשתנים, בעת זיהוי אנטרופיה גבוהה, נטען כי דפוס חריג זה מצביע על אנומליה.

תיאור הכלים המשמשים לפתרון

- הפרויקט יבוצע על המחשבים האישיים, כאשר יהיה שימוש בתוכנת MATLAB R2017B
- הקוד יימצא במאגר Github
- תיעוד ההתקדמות יתועד תחת היומן של GOOGLE - Google יומן

תכנית בדיקות

- ✓ בדיקות פונקציונליות:
  - האם המערכת יודעת לשאוב את נתוני קבצי xlsx למאגר נתונים של הכלי
  - האם הכלי יודע להתאים את קבצי xlsx הנבדקים אל מול השיטות השונות
  - האם המערכת מחזירה פלטים נדרשים בהתאם לשיטות השונות
  - האם התוכנית מצפה לקבל קלט
  - האם המערכת יודעת להתמודד עם נתונים לא צפויים

- ✓ בדיקות לא פונקציונליות:
  - מהירות (זמן):
    - זמן המתנה לטעינת טבלאות נתונים
    - זמן הרצת השיטות
    - זמן הריענון לעבר בדיקה מחודשת של טבלה נוספת
  - עומס:
    - מספר המשתמשים אשר יכולים לעבוד על הכלי הם כמספר המחשבים; אין בעיה בעבודה סימולטנית
  - זמינות:
    - הכלי יהיה זמין כל שעה שיש שימוש בכלי MATLAB (למעט מקרים שבהם תקרוס המערכת ללא התראה מוקדמת)
  - תחזוקה:
    - ניתן לשנות ולעדכן את השירותים למשתמש בזמן סביר, וללא צורך במשאבים רבים.
  - אילוצי פלטפורמת מימוש:
    - הכלי נכתב ונתמך על ידי מערכת MATLAB.

- ✓ בדיקות ממשק לקוח GUI:
  - האם קיים ממשק נגיש למשתמש
  - האם קיימת גישה להכניס נתונים לבדיקה

- ✓ בדיקות מערכת:
  - האם כל שיטה מבצעת את הנדרש ממנה לבצע כשורה
  - ווידוא אינטגרציה נכונה של השיטות יחדיו בכלי אחד
  - בדיקות מקיפות על הקובץ שהתקבל - האם מתקבלת הרצה תקינה של התוכנית ללא אזהרות \ תקלות
  - האם הריצה המחודשת של הקובץ אינה נופלת בשום מצב

- ✓ בדיקות תאימות:
  - האם המערכת כוללת את תוכנה ה- MATLAB עליה מריצים את הכלי

- ✓ בדיקות תחזוקה:
  - האם ניתן לעדכן או לתקן את הכלי במהלך חיי היישום
  - האם הקוד כתוב בצורה מודולרית
  - האם קיים תיעוד לקוד

סקירת עבודות דומות בספרות והשוואה

מסקירת ספרות שביצענו, נמצא כי סוגיית זיהוי האנומליה הוא תחום שנימצא בבחינה מתמדת בעשורים האחרונים. האנומליה הוא כלי חזק לאיתור חריגות ואכן יש מגוון שיטות ליישומו. ניתן לסווג את השיטות לפי קטגוריות שונות.

בפרט, מאחר ונתמקצע בענף הרפואה, ניתן להיווכח כי גם שם יש שיטות ייחודיות לאיתור חריגות על נתונים רפואיים.

הטבלה הבאה תמחיש את הקטגוריות לשיטות השונות לזיהוי אנומליות שחקרנו בענף הרפואה.

Categories / Technique Used	ML	Algorithm	Statistics
Bayesian Networks			✓
Neural Networks		✓	
Rule-based Systems:	✓		
Nearest Neighbor based Techniques		✓	
Parametric Statistical Modeling			✓

הייחוד במחקרנו הוא שאנו בוחרים לאמץ כלים ושיטות שונות לזיהוי אנומליות (מעולם הסטטיסטיקה, אלגוריתמיקה ולמידת מכונה) המוכרות בענף האנומליה אך שילובן יחד לצורך מענה לעולם הרפואה (עבודה על נתונים רפואיים) טרם נבחן.

כמו כן, מטרתנו למנף את רמת הדיוק של התוצאות לזיהוי חריגה לפי שיטת הצבעת הרוב, מה שטרם נמצא בשימוש. בנוסף, בשלב של כתיבת הקוד המטרה היא לתת הסקה מהירה של בחינת הנתונים.

נספחים

א. רשימת ספרות \ ביבליוגרפיה

- Anomaly Detection : A Survey
https://pdfs.semanticscholar.org/7b5a/c1fb5627addf92ad5804a6569a6cfa9385ac.pdf
- NETWORK ANOMALY DETECTION
https://pdfs.semanticscholar.org/964e/937e50bbcbd97b7d6c7205aa857919faa343.pdf
- Universal Anomaly Detection: Algorithms and Applications
https://arxiv.org/pdf/1508.03687.pdf
- Network Anomaly Detection: Methods, Systems and Tools
http://www.nr2.ufpr.br/~jefferson/pdf/Network_Anomaly_Detection-Methods,_Systems_and_Tools.pdf
http://www.mdpi.com/1099- - An Entropy-Based Network Anomaly Detection Method
4300/17/4/2367/htm
- Compression Algorithms: Huffman and Lempel-Ziv-Welch (LZW)
http://web.mit.edu/6.02/www/s2012/handouts/3.pdf
A Linear Programming Approach to Novelty Detection - http://papers.nips.cc/paper/1822-a-linear-programming-approach-to-novelty-detection.pdf
Cooperative Learning Virtual Reality-Based Visualization for Data Mining - https://www.researchgate.net/profile/Eric_Paquet2/publication/44052160_Cooperative_Learning_Virtual_Reality-Based_Visualization_for_Data_Mining/links/02e7e5322fea55e8c8000000/Cooperative-Learning-Virtual-Reality-Based-Visualization-for-Data-Mining.pdf
On Abnormality Detection in Spuriously Populated Data Streams - https://www.researchgate.net/profile/Charu_Aggarwal/publication/220907311_On_Abnormality_Detection_in_Spuriously_Populated_Data_Streams/links/0deec52415b18c0621000000.pdf
HOT SAX: Efficiently Finding the Most Unusual Time Series Subsequence - http://www.cse.cuhk.edu.hk/~adafu/Pub/icdm05time.pdf
Effect of Outliers and Nonhealthy Individuals on Reference Interval Estimation - http://clinchem.aaccjnls.org/content/47/12/2137.full
Detection of Outliers in Reference Distributions: Performance of Horn's Algorithm - http://clinchem.aaccjnls.org/content/51/12/2326.full
DAMAGE DETECTION IN MECHANICAL STRUCTURES USING EXTREME VALUE STATISTICS - http://permalink.lanl.gov/object/tr?what=info:lanl-repo/lareport/LA-UR-02-1891
Cybersecurity vulnerabilities in medical devices: a complex environment and multifaceted problem - https://www.ncbi.nlm.nih.gov/pmc/articles/PMC4516335/pdf/mder-8-305.pdf
STATE OF CYBERSECURITY & CYBER THREATS IN HEALTHCARE ORGANIZATIONS - http://blogs.harvard.edu/cybersecurity/files/2017/01/risks-and-threats-healthcare-strategic-report.pdf
Security and Privacy Issues in Wireless Sensor Networks for Healthcare Applications - https://link.springer.com/article/10.1007/s10916-010-9449-4
Contextual anomaly detection framework for big sensor data - https://journalofbigdata.springeropen.com/articles/10.1186/s40537-014-0011-y

- למידת מכונה - http://www.digitalwhisper.co.il/files/Zines/0x3B/DW59-1-ML-Security.pdf
- אנטרופיה - מתוך ספר של האוניברסיטה העברית "תורת האינפורמציה"
- אלגוריתם למפל זיו - https://he.wikipedia.org/wiki/%D7%90%D7%9C%D7%92%D7%95%D7%A8%D7%99%D7%AA_%D7%9D_%D7%9C%D7%9E%D7%A4%D7%9C-%D7%96%D7%99%D7%95

ב. תרשימים וטבלאות - הורד באישור המנחה

ג. תכנון הפרויקט

17.9.17	פגישת היכרות עם המנחה והגעת רעיונות למחקר
16.10.17	מציאת תחום מחקר – אנומליה
9.11.17	מציאת בעיה מחקרית במערכות רפואה
9.11.17	הגהת דרך לפתרון הבעיה המוצגת
19.11.17	חקירת השיטות המוצעות לזיהוי אנומליה
19.11.17	הגשת "שלב ההצעה"
20.11.17	מציאת מאגר נתונים בענף הרפואה – סכרת
17.12.17	MATLAB - מציאת סביבת עבודה והכרותה

אימוץ השיטות והתאמתן למאגר הנתונים שברשותנו - Machine Learning	9.1.18
אימוץ השיטות והתאמתן למאגר הנתונים שברשותנו - Lempel Ziv	15.2.18
אימוץ השיטות והתאמתן למאגר הנתונים שברשותנו - אנטרופיה	1.3.18
ממשק GUI למשתמש	1.4.18
תובנות ויעילות	1.5.18
שיפור ביצועים	1.6.18
הגשת מצע מחקרי	22.7.18

ד. טבלת סיכונים

#	הסיכון	חומרה	מענה אפשרי
1	קושי במציאת שיטות לזיהוי אנומליה	High	התייעצות עם גורם בקיא בנושא, שינוי שיטות מוכרות לצרכי המחקר
2	למידה של שימוש בכלי Matlab	High	למידה עצמית
3	קושי בהצלבת הנתונים אל מול השיטות המוצעות לזיהוי אנומליה	High	בחינת השיטות והתאמתן למאגר הנתונים שברשותנו, ייעוץ עם איש מקצוע
4	העדר ידע של סביבת ניהול מאגרי נתונים בענף הרפואה	Medium	ייעוץ עם המנחה ושימוש במנוע החיפוש של האינטרנט והספרות
5	כתיבת קוד בשפת תכנות לא מוכרת	Medium	למידה עצמית
6	קושי במציאת ידע אודות שיטות לזיהוי חריגות במערכות הרפואה	Medium	שינוי ההתמקדות של האנומליה במחקרנו
7	הערכה שגוייה של פריסת היקף המחקר	Low	בקרה מתמדת והגדרת דרישות ולוח זמנים ראלי
8	תחרות - מחקר דומה שנערך בו זמנית	Low	התייעצות עם המנחה וכן מעקב אחר ההתקדמות של המתחרה הפוטנציאלי וזירוז ניהול הפרויקט בבית

ה. רשימת \ טבלת דרישות

רשימת דרישות (User Requirement Document)

דרישות חומרה:

- המערכת תיתמך בסביבת העבודה MATLAB

דרישות תוכנה:

- המערכת תיכתב בשפת MATLAB
- פיתוח GUI מתאים לצורך תחילת הליך עבודה
- המערכת עבור המשתמש תהייה בשפה אנגלית
- המערכת תאפשר להכניס טבלאות נתונים כקלט
- המערכת תפעיל במקביל מספר שיטות ניתוח שונות
- המערכת תאפשר לקבל תמונות של ניתוחים לאחר הרצת שיטות בדיקה שונות
- המערכת תאפשר למשתמש לקבל תובנות לגבי הנתונים שהוזנו לשיטה
- המערכת תבצע ניתוח של מידע והצגת תוצאות בזמן קצר שלא יעלה על מספר שניות