פרויקט גמר – חוברת סופית

המחלקה להנדסת תעשייה וניהול

תשפ״ד

נושא הפרויקט:

Projects Effort Estimation

ארגון:

Amdocs

מנחה אקדמי:

ד״ר אלירן שרצר

מנחה תעשייתי:

אורלי קרמס

מגישים:

עמית פלאח - 318510070

לינוי מדלסי - 206073124

נועם סיטבון - 332391820

תוכן עניינים

1. תקציר מנהלים

1.1. רקע לפרויקט – תיאור המצב הקיים

חברת Amdocs הינה חברה בינלאומית המפתחת תוכנות ושירותים למערכות חיוב לקוחות (Billing) וניהול קשרי לקוחות (CRM). לחברה יש מרכזי פיתוח במגוון רחב של מדינות וכ20,000 עובדים (מתוכם כ5,000 בישראל). החברה מתעסקת בעיקר בלקוחות גדולים מעולם הטלקומוניקציה כדוגמת AT&T, Vodafone ועוד להם היא מספקת מנעד רחב מאוד של שירותים ומוצרים. מסיבה זו אמדוקס מתמודדת עם משימות שונות ומגוונות. לכל סוג משימה ישנה חטיבה ייעודית בתוך החברה שתפקידה להתמודד עם משימות ומוצרים מאותו הסוג. לכל מוצר מוגדר מתוך החטיבה מנהל מוצר, כחלק מתפקידו מנהל המוצר אחראי להגדיר מהו משך הפיתוח של הממשק החדש של המוצר (נמדד ב-Man Days), מה הדרישות לפיתוח ומי הם בעלי התפקיד שישתתפו בפיתוח בכל שלב. כל אחד מבעלי התפקיד (מהחטיבות השונות) מעביר הערכה לזמן הפיתוח בימי עבודה למנהל המוצר שאחראי לתת הערכה כוללת לזמן של כל תהליך הפיתוח. את ההערכה החזויה של משך הפיתוח ממלא מנהל המוצר במערכת הJira. מלבד ההערכה בימי פיתוח מנהל המוצר אחראי לתעד את אופי התהליך, שלבי הפיתוח, הכלים המשמשים לפיתוח ואנשי המקצוע בשפה חופשית במערכת הJira ולתייג את הפרויקט תחת נושא כללי (Digital, Data Science, Ecommerce ועוד..).

1.2. תיאור הבעיה

במצב הנתון, בגלל ריבוי המשימות ישנם המון פרויקטים שהערכת הזמן הראשונית שלהן אינה נכונה. השוני בין החלקים בתהליכי הפיתוח המגוונים גדול מדי מכדי שמנהל המוצר יצליח לאפיין בצורה נכונה את משך הפיתוח הדרוש לכל פרויקט. המון פעמים ההערכה הראשונית היא הערכה גסה שמוטה ע"י אופי המנהל ואופי העובדים ואינה מתייחסת לאופי הפרויקט או להיסטוריה של פרויקטים מאותו הסוג. בראיון עם מנהלי מוצר ב- Amdocs, הבהירו לנו המנהלים את הקושי בהערכת משך הפרויקטים בחברה גדולה ומגוונת כזו ואת ההפסדים שנגרמים בעקבות צווארי בקבוק בלתי צפויים, הן מבחינת הכנסה והן מבחינת ניהול תקציבים ומשאבים. המנהלים הדגישו כי זו לא רק בעיה שלהם, אלא גם בעיה עסקית רבת עניין המשפיעה על תחרותיות ויכולת גיוס לקוחות. בנוסף, הוסיפו המנהלים ואמרו כי עד כה לא הצליחו למצוא דרך להעריך בצורה מדויקת ויעילה את אורך הזמן של כל פרויקט. באמצעות מודל שמבוסס על NLP ו- Machine Learning נוכל לנתח את היסטוריית התיעודים במערכת הJira ולבנות באמצעותה מודל שייתן למנהלים כלים לחיזוי זמן ולתכנון יעיל יותר בצורה מושכלת.

1.3. הפתרון המוצע לבעיה

מטרת הפרויקט שלנו היא לאפשר למנהלי המוצר בחברה לחזות משך פיתוח של מוצרים ושל ממשקים הקשורים למוצרים בצורה יותר מדויקת. השיטה שפיתחנו במהלך הפרויקט מתבססת על שיטת NLP לניתוח תיאור הפרויקט ממערכת הJira ולאחר מכן ניתוח התוצאה וביצוע חיזוי למשך הפיתוח באמצעות מודל Machine Learning. שימוש בכלי חיזוי יאפשר למנהלי המוצר להחליט בצורה מושכלת על משך הפרויקט הצפוי מראש גם לפי התוצאה החזויה וגם לפי הידע המקדים שיש לכל מנהל.

2. תיאור הארגון

2.1. רקע

Amdocs היא חברה בינלאומית המתמחה בתוכנות ושירותים לספקי שירותי תקשורת, מדיה ובידור (BSB), נוסדה בישראל בשנת 1982 כחטיבה בשם "Aurec Information and Directory Systems". בשנת 1985, רכשה חברת SBC (Southwestern Bell Corporation) כ-50% מהבעלות על החברה ושינתה את שמה לAmdocs. לקוחות החברה מגיעים בעיקר מהסקטור הציבורי ומתחום הטלקומוניקציה (B2B) וכוללים חברות ענק כגון: AT&T, Vodafone, Sprint ו-T-Mobile. במשך הזמן הפכה החברה למפתחת מערכות ה-Billing הגדולה בעולם, כיום מוצריה משווקים בכ-60 מדינות והיא מספקת שירותים לכ-90% מחברות הטלפון בארצות הברית. על פי מידע עדכני מהחברה, Amdocs מעסיקה בישראל בערך כ-5000 עובדים, המהווים כ-25% ממצבת כוח האדם של החברה.

ב-Amdocs קיימות שלוש חטיבות עיקריות: חטיבת המוצר (Amdocs Technology), האחראית על פיתוח המוצר; חטיבת ה-(Customer Business Group) CBG, האחראית על המכירה ועל הניהול השוטף של צרכי הלקוח; וחטיבת ה-Services, המספקת שירותים שאינם מוצרים ללקוחות כגון נתונים ואפליקציות. כל חטיבה מתנהלת בפני עצמה עם עובדים שונים, ניהול נפרד ויעדים ייחודיים.

2.2. מטרות הארגון

מטרת העל של Amdocs היא לשמש כספק עיקרי לכל הצרכים והפיתוחים הטכנולוגיים של קהל לקוחותיה. החברה מתמקדת בפיתוח מוצרי תוכנה ושירותי תוכנה מתקדמים כגון Billing (חיוב) ו-CRM (ניהול קשרי לקוחות), המסייעים ללקוחותיה בתעשיות הטלקומוניקציה והמדיה לשפר את היעילות התפעולית, לשפר את חווית הלקוח ולהניע צמיחה בהכנסות באמצעות חדשנות מתמדת.

בכדי להגיע למטרה זו Amdocs שואפת לספק את כל הצרכים הטכנולוגיים הנוספים של לקוחותיה, מעבר לפיתוחים הקונבנציונליים, ולהציע פתרונות מתקדמים וחדשניים בכל התחומים שבהם עשוי להתעורר ביקוש טכנולוגי כמו פיתוח מוצרים ושירותים בתחום טכנולוגיות המידע (IT), חומרה ותחומים שונים נוספים.

בשנים האחרונות, מגמת ההתחדשות בפיתוחים של החברה סובבת סביב תחום ה- Generative AI (בינה מלאכותית יוצרת). בנוסף לפיתוחים החדשים החברה מנסה גם לשלב טכנולוגיה זו בכל המוצרים שכבר קיימים אצל קהל הלקוחות שלה. החברה מבססת את כל המוצרים על Generative AI מתוך שאיפה לספק פתרונות טכנולוגיים מתקדמים, חכמים ויעילים יותר ללקוחותיה.

3. הגדרת הבעיה

3.1. רקע על החלק בו הפרויקט ממוקד

ב- Amdocs קיימות 3 חטיבות ובתוכן מחלקות רבות, כשכל חטיבה עוסקת בתחום שונה וכל מחלקה בתוך כל חטיבה מתמקדת בהיבטים ייחודיים של התחום שלה. ניהול מוצרים הוא חלק מרכזי ומשמעותי בארגון. כל מנהל מוצר (Product Owner) אחראי על ניהול המוצרים בתחום עיסוק מסויים. במסגרת כל פרויקט פיתוח של מוצר / ממשק, מוקם צוות המורכב מבעלי תפקידים שונים, הכוללים מתכנתים מכל הסוגים הדרושים, מאפייני תהליכים, ארכיטקטים, אנשי Digital, אנשי מכירות, אלגוריתמיקאים, מתמטיקאים, מעצבים גרפיים, מהנדסים ועוד. בעלי התפקידים מגיעים מהחטיבות השונות בהתאם לתפקידם בפרויקט.

בעלי התפקידים השונים מחויבים לדווח בתחילת הפרויקט למנהל המוצר על משך הזמן הצפוי להם לפיתוח החלקים עליהם הם אחראים בתוך המוצר שנמצא במרכז הפרויקט. תפקידו של מנהל המוצר לבצע הערכה ראשונית של משך הזמן הכולל של הפרויקט, בהתבסס על הזמנים שנמסרו על ידי בעלי המקצוע המשתתפים בתהליך הפיתוח. בנוסף לכך, מנהלי המוצר אחראים לבצע תיעוד בשפה חופשית של תהליך הפרויקט בהתאמה למוצר, משמעות הפרויקט, אופי הפרויקט ואופי ההתאמה של צרכי הלקוח למוצר בתוך מערכת ה-Jira. תיעוד זה מסייע לשמור על שקיפות, לעקוב אחר התקדמות הפרויקט ולהבטיח שכל בעלי העניין מעודכנים בפרטי הפרויקט. ההערכה הראשונית היא קריטית עבור הארגון, שכן היא משפיעה ישירות על ניהול המשאבים, השקעת התקציבים והקצאת כוח האדם. בנוסף, ההערכה משפיעה על התקשורת עם הלקוחות בנוגע לציפיות ותוצאות.

3.2. בעיות במצב הקיים

אחד האתגרים המרכזיים בניהול פרויקטים בתור מנהלים הוא דיוק ההערכה הראשונית. כאשר ההערכה הראשונית של המנהל בפרויקט הפיתוח אינה מדויקת מספיק, עשויות להתרחש בעיות משמעותיות כמו אובדן לקוחות, הפסד הכנסות ובזבוז שעות עבודה. בעיות אלו נגרמות, בין היתר, מהסתמכות על הנחות: חוסר היכולת להעריך בדיוק את משך הפרויקט עשוי להוביל לכך שמנהלים יסתמכו על הנחות אופטימיות שונות לכל מנהל, שעשויות להתגשם בתוך מסגרת זמן ארוכה יותר או בעלות גבוהה יותר. לשיפור דיוק ההערכה הראשונית יש חשיבות רבה במניעת הפסדים אלו ובהבטחת הצלחת הפרויקטים. הפרויקט הנוכחי מתמקד בשיפור תהליך ההערכה הראשונית בניהול הפרויקטים, מתוך מטרה להגביר את הדיוק, לייעל את ניהול המשאבים ולשפר את שביעות הרצון של הלקוחות.

4. מטרות הפרויקט

4.1. מבנה הפתרון

הפתרון המוצע בפרויקט שלנו מבוסס על תחום למידת המכונה (Machine Learning) ומטרתו המרכזית היא שיפור הדיוק של ההערכה הראשונית של משך הפרויקטים. מבנה הפתרון כולל מספר שלבים מרכזיים:

1. איסוף נתונים: שימוש בתיעוד מקיף של הפרויקטים הקיימים במערכת ה-Jira של Amdocs. נתונים אלו כוללים תיאורים בשפה חופשית של הפרויקטים, משכי זמן ודיווחים אחרים.

2. עיבוד מקדים של הנתונים (Preprocessing): עיבוד הנתונים שנאספו כולל שימוש בעיבוד שפה טבעית (NLP) כדי לחלק את תיאורי הפרויקט למילות מפתח ולהשתמש בשיטות כמו TF-IDF לשקלול המילים החשובות ביותר מבחינת המוניטין והחשיבות שלהן לכל סוג של פרויקטים.

3. פיתוח מודלי חיזוי: בהתבסס על המשקלים שנקבעו בשלב הקודם, פיתחנו מודלי חיזוי שונים אשר מחשבים את משך הפרויקט לפי ההיסטוריה של פרויקטים מאותה קטגוריה שתיאורי הפרויקט שלהם היו דומים.

4. אופטימיזציה של המודלים: כל מודל חיזוי מייצר פלט חיזוי, ולאחר מכן מבצעים תהליך אופטימיזציה כדי לבחור את המודל הטוב ביותר מבין כל המודלים לצורך הפקת החיזוי הסופי.

5. בניית ממשק משתמש: המטרה הסופית היא לבנות ממשק שבו כל מנהל יוכל להכניס בשפה חופשית את תיאור הפרויקט כראות עיניו ולהוסיף את הקטגוריה של הפרויקט. הממשק יאפשר למנהל לקבל הערכה אופטימלית למשך הפרויקט על ידי שימוש במודלים שפיתחנו.

מבנה פתרון זה מאפשר לנו לנצל את הנתונים ההיסטוריים בצורה מיטבית ולשפר את הדיוק בהערכות הזמנים של הפרויקטים. באמצעות פתרון זה, אנו שואפים לייעל את ניהול המשאבים בארגון, לשפר את תכנון התקציבים והקצאת כוח האדם, ולהגביר את שביעות הרצון של הלקוחות.

4.2. מדדי הצלחה

מדדי ההצלחה הנדרשים על ידי חברת Amdocs מבוססים על מספר ימי העבודה לאדם הדרושים לפיתוח הפרויקט (Man Days). בגלל שהבסיס עליו יהיו בנויים מדדי ההצלחה הוא בסיס מספרי, המדדים הכוללים את Mean Squared Error (MSE), Mean Absolute Error (MAE), ו-Mean Absolute Percentage Error (MAPE) משמשים לבדיקת הדיוק של החיזויים בהתאם לנתונים הסטטיסטיים וההערכות שנקבעו בשלב המוקדם של הפרויקט. הבחירה גם ב-MSE תעזור להעריך את הטעות הריבועית הממוצעת ולשער את שונות התוצאות, בנוסף בחירה של מדדים עם ערכים יציבים כמו MAE ו-MAPE תעזור לנו להבין כמה השפעה יש לערכים חריגים (MSE רגיש לערכי קיצון) ויתרה מכך, נוכל להבין גם את הטעות וגם את אחוז הטעות המוחלטת.

בנוסף למדדים הכמותיים, מדד ההצלחה הוא גם שיעור האימוץ של ממשק הניבוי בקרב מנהלים ובעלי העניין באמדוקס. המדד מספק ניתוח של כמה נצליח להטמיע את הכלי בקרב המשתמשים ולהפוך אותו לפופולרי ושימושי

בארגון. שימוש במדד זה יאפשר לנו לעקוב אחרי ההשפעה של הפתרון כתלות בזמן ולקבוע את ההצלחה שלו בהטמעה רחבה יותר בעתיד.

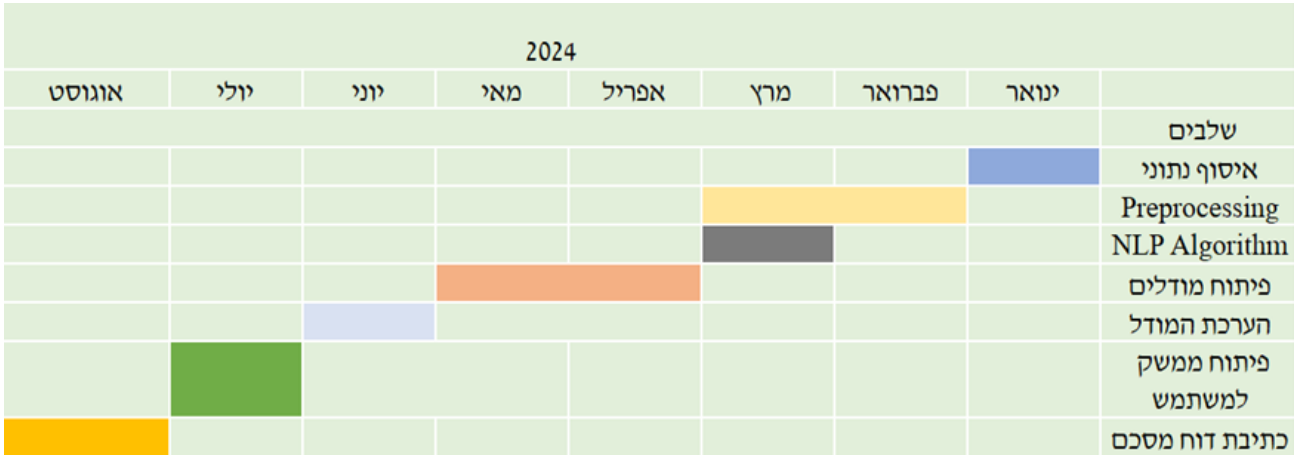
שיעור האימוץ הינו גם מדד שישמש בפרויקטים עתידיים על מנת לבחון את ההטמעה של ממשק הניבוי שפותח במסגרת פרויקט זה. על בסיס הנתונים שיצאו מהמדד, יהיה ניתן להעריך באופן מדויק את השפעתו של הכלי על תהליכי הפרויקט ועל חוויית המשתמשים בארגון.

5. תרשים גאנט

להלן תרשים גאנט המציג את השלבים של הפרויקט במהלך שנת 2024:

1. איסוף נתונים -התבצע בחודש ינואר.
2. שלב ה-Preprocessing התבצע בין פברואר למרץ.
3. שלב ה-NLP Algorithm התבצע בחודשים מרץ.
4. פיתוח מודלים -התרחש בחודשים אפריל ויוני.
5. הערכת המודל –המודלים הוערכו בחודש יוני.
6. פיתוח ממשק למשתמש -התרחש בחודש יולי.
7. כתיבת דוח מסכם התרחש בשבוע השני של חודש אוגוסט.

התרשים מראה את סדר השלבים ואת חלוקת הזמן עבור כל שלב בפרויקט.



איור 1- תרשים גאנט

6. סקר ספרות

הפרויקט נועד להציע פתרון עבור חברת Amdocs לחיזוי הערכת מאמץ בפיתוח פרויקטים באמצעות מודל המבוסס על למידת מכונה, בדגש על משך הזמן של הפרויקטים ובהתבסס על התיאורים שלהם בשפה טבעית (Natural Language Processing). חיזוי מדויק של משך הפרויקט הוא קריטי לניהול פרויקטים אפקטיבי והקצאת משאבים נכונה.

בסקירת ספרות זו, ניתן לראות על פי מאמר [1] ומאמר [5] שמודלים של רשתות עצביות מסוגלים ללמוד מנתונים היסטוריים של פרויקטי פיתוח תוכנה ולחלץ דפוסים קריטיים המאפשרים להם לחזות Effort Estimation בפרויקטים דומים.

עם זאת, בשל השונות בסוגי הפרויקטים, בטכניקות הניהול ובמבנים הארגוניים, יצירת מודל חיזוי מדויק לכל המקרים מהווה אתגר משמעותי. סקירת ספרות זו בוחנת את המחקרים והמתודולוגיות הבסיסיות שתמכו בפיתוח הפרויקט הזה.

ניתן לסווג את הסקירה הספרותית לתתי נושאים התומכים את שלבי הפרויקט לפי סדר העבודה:

1. מודלי למידת מכונה אל מול שיטות אחרות לחיזוי Effort Estimation.
2. טיפול בחריגים.
3. עיבוד שפה טבעית (NLP).
4. משקול מילים.
5. בחירת מודל אופטימלי ומדדי דיוק.

6.1 מאמר 1

"Using an Artificial Neural Network for Improving the Prediction of Project Duration"

Itai Lishner * and Shtub Avraham

מטרת המאמר

מאמר [1] עוסק בהתמודדות עם אתגר של חיזוי של משך הפרויקט בניהול הפרויקט לפי שיטת הנתיב הקריטי (Central Processing Unit - CPM). לאור השונות בסוגי פרויקטים, טכניקות ניהול ותרבויות ארגוניות, יצירת כלי חיזוי אוניברסלי לCPM הינה משימה קשה. המחברים מציעים כלי למידת מכונה (ML) דינמי המבוסס על רשת עצבית מלאכותית (ANN) המתאימה למערכי נתונים ושיטות חיזוי שונות. הכלי מותאם באמצעות אלגוריתם ANN לשיפור הדיוק. אימות עם נתונים ממשיים משני ארגונים מראה שיפור משמעותי בדיוק החיזוי למרות הבדלים בסוגי הפרויקטים ובמבני הנתונים.

שיטות מחקר

המתודולוגיה כללה שימוש בכלי דינמי (DPDP) לחיזוי משך פרויקט המתבסס על למידת מכונה מבוקרת ורשת עצבית מלאכותית (ANN). הכלי עיבד את הנתונים על ידי טעינה, ניקוי, קידוד נתונים לא מספריים ונרמול מאפיינים. מודל

ה-ANN עבר אופטימיזציה באמצעות מספר מחזורי אימון תוך שימוש בארכיטקטורות שונות ואלגוריתם בחר את התצורה הטובה ביותר. מערך הנתונים חולק לסט אימון, אימות ובדיקה (Train, Test, Validation) , והביצועים של המודל הוערכו באמצעות שגיאה ממוצעת מוחלטת (MAE), שורש ממוצע ריבועי השגיאה (RMSE) ושגיאת אחוז מוחלט ממוצע (MAPE). המודל הסתגל לנתונים הייחודיים של כל ארגון ושיפר את דיוק החיזוי.

תוצאות המחקר

התוצאות מראות שמודל למידת מכונה המבוסס על ANN משפר באופן משמעותי את דיוק חיזוי משך הפרויקט בהשוואה לשיטות מסורתיות כמו תרשימי גאנט ו-CPM.

הממצאים המרכזיים כוללים:

עבור ארגון 1: כלי ה- DPDP הפחית את השגיאה הממוצעת המוחלטת (MAE) מ-49.18 שבועות (בשיטה המסורתית) ל-24.25 שבועות, והשיג שיפור של 50.7%.

עבור ארגון 2: כלי ה- DPDP הפחית את שורש ממוצע ריבועי השגיאה (RMSE) מ-9.33 שבועות (בשיטה המסורתית) ל-4.64 שבועות, והשיג שיפור של 50.3%.

בשני המקרים, הדיוק המשופר של המודל מעולם למידת המכונה מדגיש את הפוטנציאל שלו לחיזוי אפקטיבי של משכי פרויקטים במערכי נתונים מגווני ביחס למודלים מסורתיים.

תרומה לפרויקט

מאמר [1] תרם באופן משמעותי לפרויקט על ידי מתן מסגרת מקיפה ליישום מודלים מעולם ה ML-בדגש על רשתות נוירונים מלאכותיות (ANNs) לחיזוי משכי פרויקטים. הדגש על התאמת כלי החיזוי למערכי נתונים ומבנים ארגוניים שונים התאים היטב למטרת הפרויקט לשיפור דיוק ההערכות. המתודולוגיה המפורטת, שכללה למידת מכונה מבוקרת, עיבוד נתונים ואופטימיזציה באמצעות אלגוריתם גנטי, הציעה גישה מובנית לבניית המודל.

בנוסף, המאמר הדגיש את החשיבות של אופטימיזציה ואימות צולב, שתרמו לנו לשיפור ביצועי המודל בפרויקט. למשל, הפחתת MAE ו RMSE-ביותר מ-50%, אישרו את היעילות של שיטות מבוססות ANN. הדבר העניק ביטחון ביישום טכניקות דומות בפרויקט, והבטיח הערכת מאמץ מדויקת ואמינה במערכי נתונים מגוונים. כמו כן, המאמר תרם להחלטה להשתמש במדדים כמו MSE ו MAPE-להערכת הדיוק.

6.2. מאמר 2

"Isolation-based Anomaly Detection"

Fei Tony Liu and Kai Ming Ting

מטרת המאמר

מאמר [2] שואף להציג ולאמת את שיטת iForest (Isolation Forest) לגילוי אנומליות וחריגים. שיטות גילוי אנומליות מסורתיות מסתמכות לרוב על התפלגות, מדדי מרחק או צפיפות, שיכולות להיות עתירות חישוב ופחות יעילות עבור נתונים רב-ממדיים. iForest מציגה גישה חדשנית על ידי בידוד נקודות נתונים באמצעות עצים

בינאריים שנבנו באקראי (iTrees). המטרה העיקרית היא להדגים כי iForest יכול לזהות אנומליות בצורה יעילה ומדויקת עם עלויות חישוב נמוכות יותר גם כשהנתונים לא מתפלגים לפי התפלגות מוגדרת ומובהקת, מה שהופך אותו למתאים למערכים גדולים של נתונים. המחברים גם שואפים להראות כי iForest עמידה בפני השפעות Masking ו-Swamping בעיות נפוצות בגילוי אנומליות.

שיטות מחקר

המתודולוגיה כוללת בניית מספר עצי בידוד (iTrees) באמצעות תתי-קבוצות אקראיות של הנתונים. כל עץ מחלק את הנתונים בצורה רקורסיבית על ידי בחירת תכונה וערך חלוקה באקראי עד שכל מופע מבודד. אנומליות, בהיותן מועטות ושונות, מבודדות קרוב יותר לשורש העץ, מה שמביא לאורך נתיב קצר יותר. אורך הנתיב הממוצע בכל העצים משמש לחישוב ציון אנומליה. השלבים המרכזיים כוללים:

- תת-דגימה: שימוש בתת-קבוצה קטנה ואקראית של הנתונים לבניית כל עץ.
- חלוקה רקורסיבית: חלוקה אקראית של הנתונים בכל צומת עד לבידוד.
- דירוג אנומליות: חישוב אורך הנתיב הממוצע לקביעת ציוני אנומליה.

תוצאות המחקר

התוצאות מראות כי iForest עולה באופן משמעותי על שיטות מסורתיות כמו ORCA, one-class SVM, LOF, ו - Random Forests מבחינת AUC (שטח תחת העקומה) וזמן עיבוד. ממצאים מרכזיים כוללים:

- יעילות iForest: משיג דיוק גילוי גבוה עם מספר קטן של עצים ותתי-קבוצות, מה שהופך אותו ליעיל מבחינת חישוב.
- השיטה יעילה בבעיות רב-ממדיות ועמידה בפני השפעות Masking ו-Swamping
- יכולת התרחבות iForest: מתאים עצמו היטב למערכי נתונים גדולים, תוך שמירה על מורכבות זמן ליניארית ודרישות זיכרון מינימליות.

תרומה לפרויקט

מאמר [2] היה משמעותי בהכוונה של גישת הפרויקט לטיפול בחריגים בנתוני משך הפרויקטים. הוא סיפק מסגרת חזקה לזיהוי אנומליות ללא הסתמכות על מדדי מרחק או צפיפות, ידיעה זאת קריטית בהתחשב בשונות הגבוהה ובמורכבות של נתוני הפרויקט (נתוני החיזוי לא הציגו התפלגות גאוסיאנית – מופיע). התובנות על תת-דגימה והיעילות של iForest במרחבים רב-ממדיים השפיעו על שלבי העיבוד המקדים של הפרויקט, והבטיחו שהמודל יוכל להתמודד עם מערכי נתונים גדולים ומורכבים בצורה יעילה. בנוסף, השיפורים המוכחים בדיוק הזיהוי ובזמן העיבוד אישרו את הבחירה ב iForest-כשיטה מתאימה לשיפור אמינות חיזוי משך הפרויקטים.

6.3. מאמר 3

"PROCESS ANALYSIS WITH AN AUTOMATIC MAPPING OF PERFORMANCE FACTORS USING NATURAL LANGUAGE PROCESSING"

Lauble, S., Zielke, P., Chen, H. & Haghsheno, S. (2023). Process analysis with an automatic mapping of performance factors using natural language processing (NLP). Proceedings of the 31st Annual Conference of the International Group for Lean Construction (IGLC31), 59–68. doi.org/10.24928/2023/0140

מטרת המאמר

מאמר [3] עוסק בפיתוח והערכה של שיטה אוטומטית למיפוי גורמי ביצועים היסטוריים למפרטי מכרזים של פרויקטים חדשים באמצעות עיבוד שפה טבעית (NLP). המטרה היא לשפר את ניתוח התהליכים והתכנון בפרויקטים על ידי שימוש יעיל בנתונים **לא** מובנים (טקסטואלים) שנאספו במהלך תהליכים אלה.

בפרוייקטים בתחום הבנייה, כמות גדולה של נתונים נאספת לטובת פירוט הדרישות של הפרוייקט ולכן יש צורך בשיטות יעילות על מנת להשתמש בנתונים שנאספו. המחקר משתמש במודל NLP כדי להשוות תיאורי תהליכים של מפרטי מכרזים לפרויקטים חדשים עם מסד נתונים ראשי. השוואה זו מזהה את הגורמים הביצועיים הנכונים ומחשבת את משך הזמן של כל תהליך, ובכך מסייעת בניתוח תהליכים וביצירת לוח זמנים מדוייק.

שיטות המחקר

במחקר משתמשים בשתי קבוצות נתונים :

- מאגר נתוני BKI הראשי, המכיל 3586 שלבי תהליך.
- נתוני מכרזים אמיתיים הכוללים 194 תיעודים.

מאגר הנתונים הראשי עבר ניקוי, שיפור והרחבה מ-1420 ל-14200 תיעודים על מנת לשפר את ביצועי מודלי למידת מכונה. הנתונים עובדו בקידוד תווית וחולקו לקבוצות Train ו-Test.

לצורך פיתוח המודל נעשה שימוש ב-GBERT[4] כלי בעל יכולות חזקות בעיבוד שפה טבעית (NLP) עבור טקסטים גרמניים. תיאורי תהליכים קצרים וארוכים שולבו יחד, וכמה פרמטרים הוגדרו מראש כגון Droppout Rate של 0.1, Epoch של 10, Batch (גודל אצווה) של 16 וקצב למידה של 2e-5.

ביצועי המודל הוערכו באמצעות דיוק (Precision), מקדם מתאם של מתיוס (MCC), שגיאה מוחלטת ממוצעת (MAE) ושגיאה באחוז ממוצעת מוחלטת (MAPE).

הדיוק מודד את היחס בין ערכים חיוביים אמיתיים לערכים חיוביים שחוזו, בעוד ש- MCC מעריך את הקורלציה בין ערכים אמיתיים לערכים שחוזו, MAE מחשב את השגיאה המוחלטת הממוצעת בין ערכים שחוזו לערכים בפועל, ו-MAPE מציין את דיוק החיזוי ביחס לערכים האמיתיים.

התוצאות מציגות דיוק ו -MCC גבוהים עבור נתוני הבדיקה, המעידים על ביצועים טובים בתנאים מבוקרים. עם זאת, נתונים מהעולם האמיתי מציגים דיוק ו-MCC נמוכים יותר, המדגישים את הצורך באימות ידנית להתמודדות עם ערכים חריגים ולשיפור אמינות החיזוי. גישה זו נועדה לצמצם תוספות מיותרות ובעיות, למנוע לחץ זמן ולשפר את התכנון בפרויקטי בנייה קלה על ידי שימוש בנתוני ביצועים מפורטים ומציאותיים.

תוצאות המחקר

המחקר בחן את ביצועי מודל ה-NLP למיפוי גורמי תפקוד בפרויקטים בעולם הבנייה הרזה, והשיג תוצאות מרשימות. בערכת הבדיקה מבסיס הנתונים הראשי, שכללה 2,806 ערכים, המודל השיג דיוק של 0.92 וערך MCC של 0.92, מה שמעיד על יכולת חזוי גבוהה ומדויקת. גם בערכת הנתונים מעולם המציאות, הדיוק של 0.65 וערך MCC של 0.64 מדגישים את הפוטנציאל של המודל, במיוחד בהתחשב במורכבות הנתונים. המדדים MAE ו-MAPE בערכת הנתונים מעולם המציאות היו 48.75 שעות ו-17.63% בהתאמה, עם הרוב הגדול של השגיאות פחות מ-8 שעות, מה שמצביע על איכות חיזוי כללית טובה. הממצאים מדגישים את יכולת המודל לייעל את תהליך התכנון, להפחית את זמן התכנון ולשפר את האיכות, ובכך להפוך את החיזויים לכלי עזר משמעותי בתכנון פרויקטים בנדל"ן.

תרומה לפרוייקט

מאמר [3] המדבר על עיבוד שפה טבעית (NLP) בתחום הבנייה סיפק תובנות קריטיות שהשפיעו ישירות על המודלים שלנו להערכת משך זמן פרויקט .המאמר הדגים כיצד NLP יכול למפות באופן אוטומטי גורמי ביצועי היסטוריים למפרטי פרויקט חדשים ,ובכך לשפר את הדיוק בחיזוי משך התהליכים. המתודולוגיה והמדדים שנידונו ,כגון דיוק, MAE ו- MAPE הכווינו את פיתוח המודל על ידי הדגשת החשיבות של איכות מאגר הנתונים **ואימון יעיל** של המודל .הבנה זו אפשרה לבצע Preprocessing על הנתונים ולייעל את האלגוריתם ובסופו של דבר לשפר את הדיוק והאמינות של הערכת המאמץ של פרויקטים באמצעות למידת מכונה.

6.4. מאמר 4

"A Comparative Study on TF-IDF feature Weighting Method and its Analysis using Unstructured Dataset"

Mamata Das, Selvakumar Kamalanathan and Pja Alphonse
NIT Trichy, 620015, Tamil Nadu, India

מטרת המאמר

מאמר [4] חוקר את האפקטיביות של שיטת שקלול תכונות TF-IDF ומשווה אותה לשיטת מיצוי תכונות N-Gram לסיווג טקסט. על ידי הערכת מודלים שונים של למידת מכונה על מערכי נתונים לא מובנים, המחקר שואף לקבוע איזו שיטה מספקת סיווג טקסט מדויק ואמין יותר. המחקר מדגיש את חשיבותן של טכניקות סיווג טקסט יעילות לעיבוד כמויות גדולות של נתונים לא מובנים, שהוא חיוני עבור יישומים שונים בעיבוד שפה טבעית.

שיטת המחקר

הניסוי כלל השוואה מפורטת בין שיטות חילוץ תכונות TF-IDF ו-N-Gram לסיווג טקסט. מערכי הנתונים שבהם השתמשו היו 50,000 ביקורות סרטים IMDB (מאוזנים בין סנטימנטים חיוביים ושליליים) ו-3,000 ביקורות של אמזון אלקסה. הנתונים עברו עיבוד מקדים, כולל שלבים כמו טוקניזציה, נורמליזציה, היווצרות, הלמטיזציה והסרה של מילות עצירה ורעש כדי להבטיח עקביות ודיוק.

לחילוץ תכונות (feature extraction) :

- N-Gram : ביגרמות (n=2) חולצו מהטקסט.

- TF-IDF : חושב כדי להעריך את חשיבות המילים במסמכים.

שיטות הסיווג שהופעלו כללו :

- Support Vector Machine (SVM)
- Logistic Regression
- Multinomial Naive Bayes (Multinomial NB)
- Random Forest
- Decision Tree
- k-nearest neighbours (KNN)

הביצועים של מסווגים אלה הוערכו באמצעות מדדים כמו דיוק, דיוק, זכירה וציון F1.

תוצאות המחקר

המחקר מצא כי שיטת חילוץ התכונות TF-IDF שיפרה באופן משמעותי את הביצועים של מודלים לסיווג טקסטים בהשוואה לשיטת N-Gram. תוצאות הסיווג הראו שהשימוש ב TF-IDF-הביא לדיוק גבוה יותר, עם תוצאות מצוינות עבור מודל ה Random Forest-שהשיג דיוק של 93.81%, דיוק חוזי של 94.20%, רגישות של 93.81% וציון F1 של 91.99%. בנוסף, המודלים Support Vector Machine (SVM) ו Random Forest-הציגו את הביצועים הכוללים הטובים ביותר מבחינת דיוק וציון F1. המחקר הדגיש גם את החשיבות של עיבוד מקדים נכון של הנתונים, אשר הוכיח כי הוא חיוני לשיפור הדיוק והאמינות של מודלים של למידת מכונה במשימות סיווג טקסט.

תרומה לפרויקט

המאמר סיפק תובנות חשובות לגבי שיטות שקלול תכונות לסיווג טקסט, והדגיש את היתרונות של שיטת TF-IDF על פני N-Grams. ממצאים אלו סייעו בפיתוח מודל הערכת משך הפרויקט על ידי הדגמת החשיבות של מיצוי תכונה יעילה בטיפול בנתונים לא מובנים. על ידי אימוץ שיטת TF-IDF, המודל שיפר את יכולתו לחזות במדויק את משכי הפרויקט בהתבסס על תיאורים טקסטואליים. המחקר הדגיש גם את הצורך בטכניקות עיבוד מקדים חזקות כדי להבטיח איכות נתונים, אשר שולבה בתהליך פיתוח המודל, ובסופו של דבר לשפר את המהימנות והדיוק של תחזיות משך הפרויקט.

6.5. מאמר 5

" Neural Network Prediction Model for Construction Project Duration "

Silvana Petruseva 1 , Vahida Zujo 2 , Valentina Zileska-Pancovska 3

1 Assis. Prof. PhD, University "St. Cyril and Methodius", Faculty of Civil Engineering, (Mathematics Department) Skopje, Macedonia, 2 . Prof. PhD, C.E., University "Džemal Bijedić" in Mostar, Faculty of Civil Engineering, Bosnia and Herzegovina, 3 Prof. PhD, C.E., University "St. Cyril and Methodius", Skopje, Faculty of Civil Engineering, Macedonia, (Department for Management of Construction)

מטרת המאמר

מאמר [5] מציג מודל של רשת נוירונים לחיזוי משך זמן של פרויקטי בנייה. במחקר נאספו נתונים על 75 מבנים שנבנו בפדרציה של בוסניה והרצגובינה באמצעות מחקרים בשטח. הנתונים כוללים מידע על זמן הבנייה המתוכנן והאמיתי, המחיר המתוכנן והאמיתי, והשימוש של המבנים השונים. מטרת המחקר הייתה לחזות את משך זמן הבנייה באמצעות מודלים שונים ולהשוות בין הדיוק שלהם. החוקרים השתמשו במודל רגרסיה לינארית ובמודל של רשת נוירונים מרובת שכבות (MLP-NN) כדי לבדוק איזה מודל מניב תוצאות מדויקות יותר.

שיטת המחקר

החוקרים השתמשו בשני מודלים עיקריים לחיזוי זמן הבנייה:

- מודל הרגרסיה הלינארית, המשתמש בקשר לינארי בין הזמן לבין מחיר הבנייה. לצורך בניית המודל בוצעו ראיונות וסקרים לאיסוף נתונים על המבנים כדי לאפיין משתנים מסבירים.
- מודל רשת הנוירונים - השתמשו ברשת נוירונים מרובת שכבות (MLP), רשת MLP היא סוג של רשת נוירונים המבצעת הזנה קדימה הממפה נתוני קלט לנתוני פלט באמצעות שכבות מרובות של נוירונים.

השיטות שנבדקו כללו השוואה בין הרגרסיה הלינארית, המותאמת למודלים לינאריים פשוטים, לבין רשתות נוירונים, המסוגלות להתמודד עם בעיות מורכבות ולא לינאריות.

תוצאות המחקר

התוצאות הראו כי מודל רשת הנוירונים מרובת השכבות (MLP-NN) סיפק דיוק גבוה יותר בחיזוי זמן הבנייה בהשוואה לרגרסיה הלינארית. בעוד שמודל הרגרסיה הלינארית הצליח להסביר כ-73% מהשונות בתוצאות, מודל רשת הנוירונים הצליח להגיע לרמת דיוק גבוהה יותר עם MAPE (Mean Absolute Percentage Error) של 10.35% עבור מודל הרגרסיה ו-9.21% עבור רשת הנוירונים. התוצאות מצביעות על יתרון ברור בשימוש ברשתות נוירונים לחיזוי משך זמן של פרויקטי בנייה, במיוחד כשמדובר בנתונים מורכבים ולא לינאריים.

תרומה לפרויקט

השימוש ברשתות נוירונים לחיזוי משך זמן של פרויקטי בנייה מציג יתרונות משמעותיים על פני שיטות למידת מכונה אחרות כמו רגרסיה לינארית בחיזוי משך זמן. רשתות נוירונים מסוגלות להתמודד עם בעיות מורכבות ולא לינאריות, ולספק תוצאות מדויקות יותר. בזכות מחקר הזה, החלטנו לשלב גם רשתות נוירונים בנוסף לשיטות למידת המכונה המסורתיות בפרויקטים נוספים. יתרון נוסף של רשתות נוירונים הוא יכולתן לזהות דפוסים מורכבים ולבצע אופטימיזציה של החיזויים, מה שהופך אותן לכלי יעיל יותר בתחומים שונים של הנדסה ואופטימיזציה לחזיוי זמן.

בפרויקט שלנו, כדי לספק מימד השוואה השתמשנו במודלים הכוללים LightGBM, Forest Regressor, SVM, Gradient Boosting Machine ובנוסף- Neural Networks. בין אלה, ביצענו שיטת אופטימיזציה לבחירת התוצאה הטובה ביותר. הרשת העצבית סיפקה את התוצאות הטובות ביותר, והדגימה את יכולתה המעולה ללכוד דינמיקה מורכבת בנתוני הפרויקט. כדי להעמיק את ההשוואה ולחזק את המסקנה שבחרנו את המודל הנכון בחרנו לנסות ליישם מודל רגרסיה פשוט בנפרד בדומה למאמר כדי לבצע השוואה בין כל התהליך הארוך שביצענו לבין פיתרון הרבה יותר קל שאולי יוכל להביא לתוצאות טובות יותר. גם בהשוואה זו רשת הנוירונים הציגה תוצאות מבטיחות בהרבה מהפיתרון הפשוט והנאיבי.

לסיכום, המחקר חיזק את הצורך במשאבי חישוב נאותים ליישום מודלים של למידת מכונה ביעילות. בדומה לממצאים במאמרים קודמים, המחקר שלנו הראה שכאשר מדובר בחיזוי משך פרויקט, מודלי רשתות נוירונים (NN) מצליחים להניב תוצאות טובות יותר לעומת מודלים פשוטים כמו רגרסיה ליניארית, למרות שבמקרים אחרים מודלים פשוטים יכולים להצליח יותר ממודלים מורכבים.

המחקר תמך בהתמקדות בהערכת זמן ועלויות כמדדים קריטיים להצלחת הפרויקט, והראה שמודלים חזויים וכלים סטטיסטיים יכולים לשפר משמעותית את תכנון הפרויקט והקצאת המשאבים. מסקנות אלה מדגישות את החשיבות של שימוש במודלים מתקדמים ומתודולוגיות חישוביות מתאימות לשיפור הדיוק והיעילות בתהליכי פיתוח תוכנה.

6.6. סיכום אינטגרטיבי

הפרויקט שלנו עוסק בחיזוי משך פרויקטים של פיתוח תוכנה בחברת אמדוקס, תוך שימוש במודלי למידת מכונה. מאמר [1] של Itai Lishner ו-Avraham Shtub סיפק את המסגרת ליישום מודלים מבוססי למידת מכונה, בדגש על רשתות נוירונים מלאכותיות (ANNs) לחיזוי משכי פרויקטים. מחקר זה הדגיש את היתרונות של ANN ואת היכולת לשפר את דיוק החיזוי בהשוואה לשיטות מסורתיות כמו תרשימי גאנט ו-CPM, דבר שהתאים למטרות הפרויקט שלנו. השלב הבא היה ניתוח תיאורים בשפה חופשית של פרויקטים ממערכת ה-Jira באמצעות עיבוד שפה טבעית (NLP). מאמר [3] תרם לפרויקט על ידי הצגת שיטה אוטומטית למיפוי גורמי ביצועים היסטוריים לתיאורי פרויקטים חדשים באמצעות NLP. טיפלנו בחריגים בנתונים באמצעות שיטת iForest (Isolation Forest) שהוצגה במאמר [2]. המאמר הציע גישה חדשנית לזיהוי אנומליות בנתונים רב-ממדיים באמצעות עצים בינאריים שנבנו באקראי, דבר שאפשר לנו להתמודד עם שונות גבוהה ומורכבות בנתונים בצורה יעילה ולהבטיח שהמודל שלנו יוכל להתמודד עם מערכי נתונים גדולים ומורכבים. בהמשך, בחרנו בשיטה TF-IDF לפי היתרונות של השיטה הניכרים במאמר [4], השתמשנו במשקול המילים שיצאו עם TF-IDF לצורך ניתוח המידע הטקסטואלי שהתקבל מהפרויקטים. לבסוף, ביצענו חיזוי באמצעות מספר מודלי למידת מכונה, כולל LightGBM, Forest Regressor, SVM, Gradient Boosting Machine ורשתות נוירונים. מאמר [5] הראה כי רשת נוירונים עדיפה למשימות חיזוי זמן בהשוואה למודלים לינאריים או מודלי למידת מכונה אחרים. בנוסף, הוא סיפק את ההשראה לשימוש במדדי הדיוק MSE, RMSE ו-MAPE ובמגוון המודלים שבחרנו כולל השוואה של התוצאה הסופית אל מול חיזוי במודל רגרסיה פשוט כדי לבחון את השימוש בכלי קל וזול יותר שעלול להוציא תוצאות טובות יותר. השימוש במודלים אלו, בשילוב עם אימות צולב כפי שהוצג במאמרים [1] ו-[5] ושיטת Grid Search לקביעת היפר-פרמטרים, אפשרו לנו לבחור את התוצאה האופטימלית ולשפר את דיוק החיזוי, תוך התבססות על תובנות ממאמרים שתמכו בכל שלב של הפרויקט.

<u>טבלה 1</u>- הטבלה מסכמת את התרומות הספציפיות של כל מאמר להיבטים שונים בפרויקט, החל מפיתוח מודלים וחילוץ תכונות ועד לטיפול בנתונים חריגים ושיפור דיוק החיזוי.

מאמר	תרומה לפרויקט
[1] "Using an Artificial Neural Network for Improving the Prediction of Project Duration "	סיפק תובנות לגבי עדיפות השימוש במודלי למידת מכונה לחיזוי משך פרויקט על פני מודלים מסורתיים ובשימוש באימות צולב כדי להוכיח את מובהקות התוצאה.
[2] "Isolation-based Anomaly Detection"	המאמר סיפק מודל המתאים לזיהוי אנומליות ללא הסתמכות על מדדי מרחק, צפיפות או התפלגות. דבר שעזר לנו להתמודד עם מורכבות התוצאות בתהליך החיזוי בהתחשב בשונות הגבוהה של נתוני הפרויקט.
[3] "Process Analysis with an Automatic Mapping of Performance Factors using Natural Language Processing"	המאמר מדגים כיצד עיבוד שפה טבעית יכול למפות באופן אוטומטי גורמי ביצועי היסטוריים למפרטי פרויקט חדשים, ובכך לשפר את הדיוק בחיזוי משך התהליכים.
[4] "A Comparative Study on TF-IDF Feature Weighting Method and its Analysis using Unstructured Dataset"	המאמר מראה את היתרונות במשימות חיזוי של TF-IDF
[5] " Neural Network Prediction Model for Construction Project Duration "	המאמר מדגיש את היתרונות של רשת נוירונים על פני מודלים לינאריים פשוטים בחיזוי זמן ולפיו בחרנו לבצע השוואה סופית עם מודל פשוט יותר. בנוסף המאמר תרם לבחירת מדדי דיוק לתהליך.

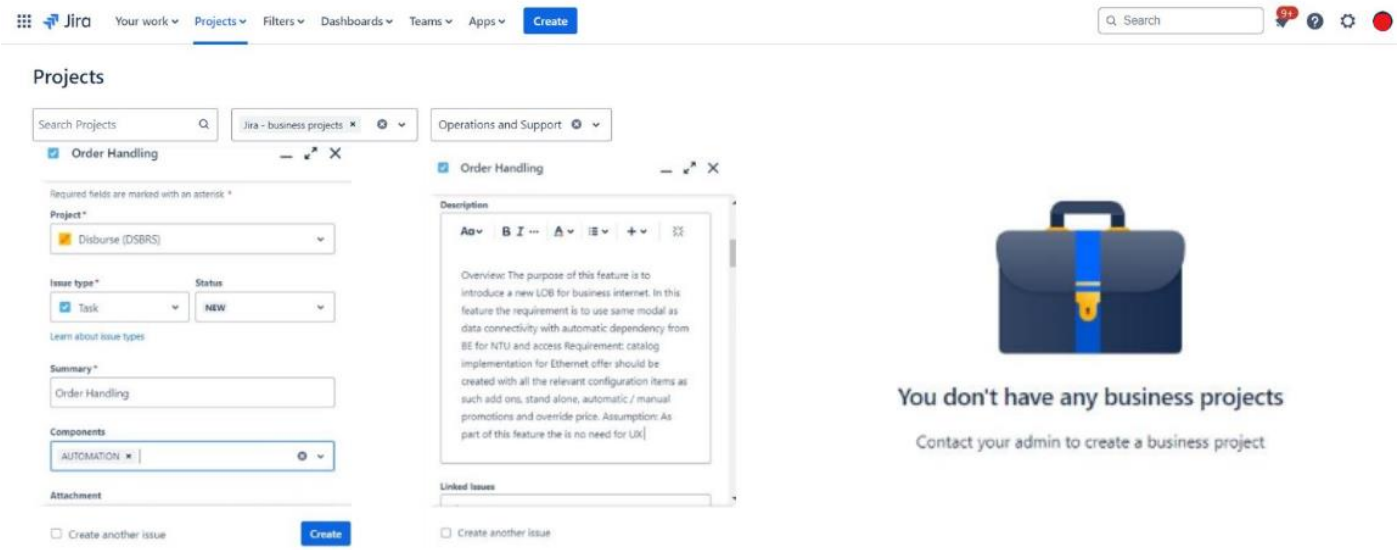
7. תיאור הנדסי של המצב הקיים

7.1. תיאור פורמלי של תהליכי העבודה בארגון

ב-Amdocs קיימות שלוש חטיבות עיקריות: חטיבת המוצר (Amdocs Technology) האחראית על פיתוח המוצר; חטיבת ה-(Customer Business Group) CBG האחראית על המכירה ועל הניהול השוטף של צרכי הלקוח; וחטיבת ה-Services המספקת שירותים שאינם מוצרים ללקוחות כגון נתונים ואפליקציות. כל חטיבה מתנהלת בפני עצמה עם עובדים שונים, ניהול נפרד ויעדים ייחודיים. במחקר שלנו עסקנו בפרויקטים המשלבים שיתוף פעולה בין חטיבת המוצר לחטיבת הCBG. בפרק זה נתאר תהליך הנדסי פורמלי של הגורמים והפעולות המעורבים בתהליך הפיתוח:

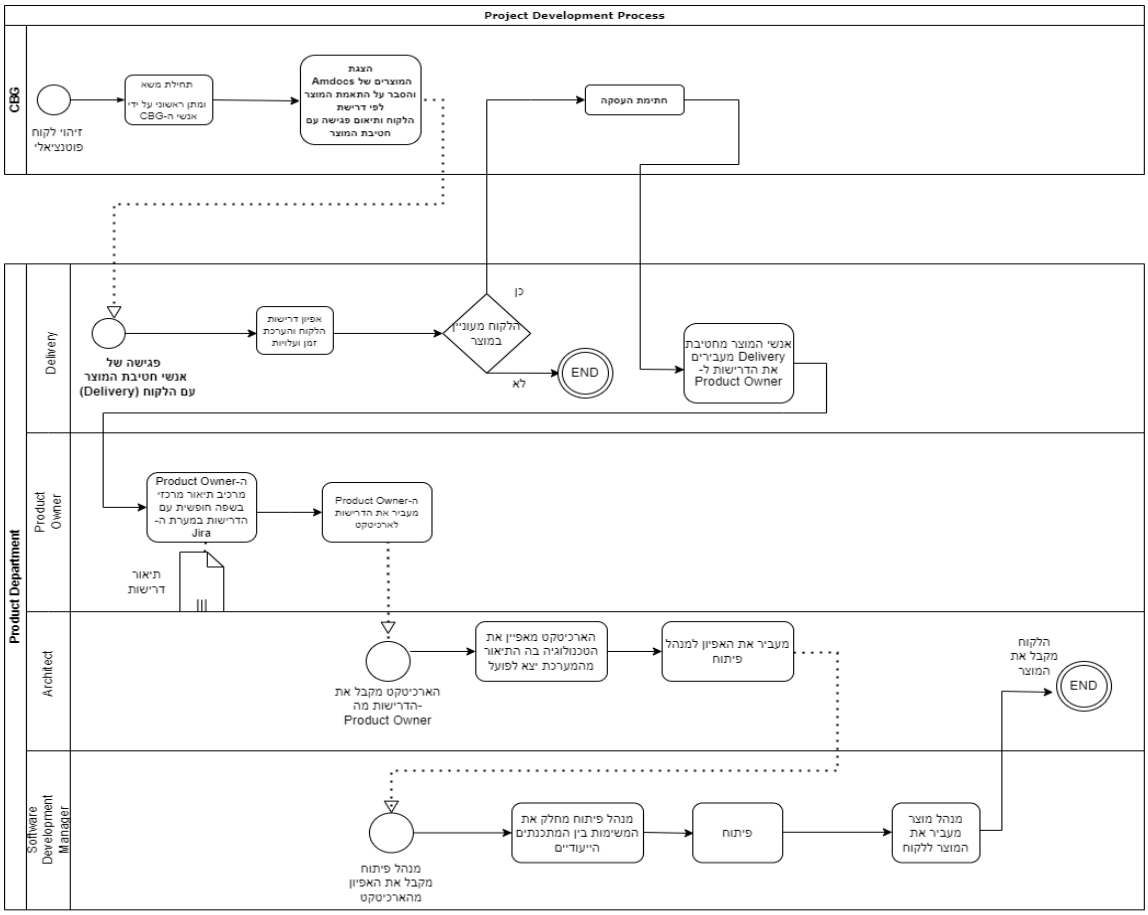
תהליך פרויקט לפיתוח מוצר

1. **זיהוי לקוח פוטנציאלי**
   - חטיבת ה-CBG מזהה לקוח פוטנציאלי ומתחילה במשא ומתן ראשוני.
   - אנשי ה- CBG מציגים את המוצרים של Amdocs ומסבירים על הפתרונות האפשריים להתאמת המוצר הייעודי לדרישות הלקוח.
2. **אפיון דרישות הלקוח**
   - כאשר יש כיוון לסגירת עסקה, אנשי חטיבת המוצר מקבוצת Delivery נפגשים עם הלקוח.
   - הם עובדים על אפיון דרישות הלקוח, השינויים שיש להכניס במוצר והערכת זמן ועלויות.
3. **חתימת עסקה**
   - מחלקת ה- CBG אחראית על תהליך חתימת העסקה.
   - במידה ויש צורך בשינויים נוספים, התהליך חוזר לאנשי המוצר לאפיון מחדש ותמחור נוסף.
   - לאחר סיום תהליך זה וחתימת העסקה, המוצר מותאם לדרישות הלקוח.
4. **העברת דרישות ל-Product Owner**
   - אנשי המוצר מחטיבת Delivery מסיימים להתאים את המוצר לדרישות הלקוח ומעבירים את הדרישות ל-Product Owner.
   - ה- Product Owner לוקח את האפיונים מחטיבת המוצר וקבוצת ה- Delivery ומרכיב במערכת ה-Jira תיאור מרכזי של הפרויקט בשפה חופשית עם כל הדרישות.



איור 2- תיאור של פרויקט במערכת ה- JIRA

5. **אפיון טכנולוגי על ידי הארכיטקט**
   - הארכיטקט מאפיין את הצורה הטכנולוגית שבה יצא לפועל הפרויקט שתואר במערכת ה-Jira.
   - הוא קובע באילו מערכות פיתוח ישתמשו, מה הכלים בהם ישתמשו ובנוסף את כיוון זרימת הנתונים בפיתוח.
6. **חלוקת משימות על ידי מנהל הפיתוח**
   - מנהל הפיתוח מקבל את האפיון ומחלק את המשימות בין המתכנתים הייעודיים.
   - כחלק מתפקידו הוא אחראי לדווח על צווארי בקבוק ולעדכן את ה- Product Owner במידת הצורך.
7. **עדכון אפיונים וחזרה על התהליך**
   - מבצעים חזרה על שלבים 4-7 במידה ויש צורך בעדכון הגדרות הפיתוח. לאחר אפיון מחדש, הנתונים והתיאורים עוברים שוב לארכיטקט לבדיקה ולאפיון טכנולוגי מחדש.
   - מנהל הפיתוח מקבל שוב את התיאור הטכנולוגי להמשך פיתוח.
   - במידה ופיתוח המוצר הסתיים הוא מועבר ללקוח והתהליך הנוכחי מסתיים.



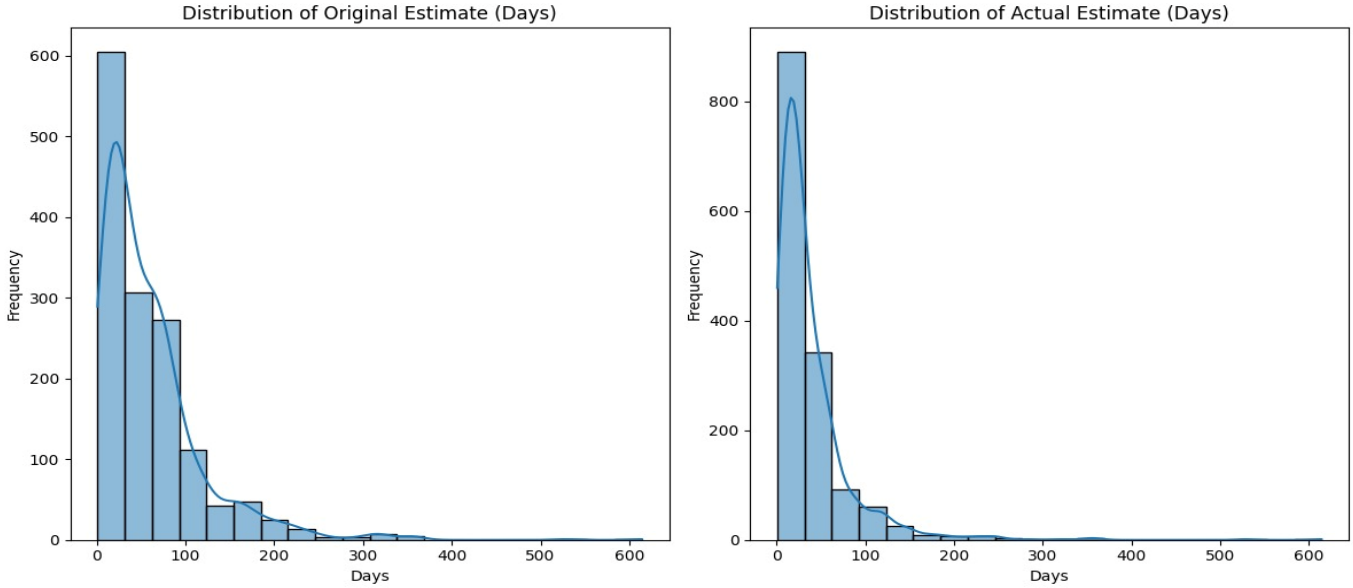
CBG-Customer Business Group- אחראים על המכירות של המוצרים ללקוחות

איור 3- תרשים תהליך BPMN- תרשים המתאר את התהליך שבו מתקבל פרויקט חדש בחברה

7.2. מדדים סטטיסטיים של המצב הקיים

7.2.1. ניתוח זמני פרויקט

תהליך הערכת המאמץ הנוכחי ב -Amdocs חשף מספר אי דיוקים בהערכות. על מנת להתמודד עם פערים אלו, ביצענו ניתוח מפורט של המצב הקיים. ניתוח זה כלל את ניתוח ההתפלגות של ההערכות בפועל וההערכות המקוריות לפי קטגוריות ה-"Release Train".

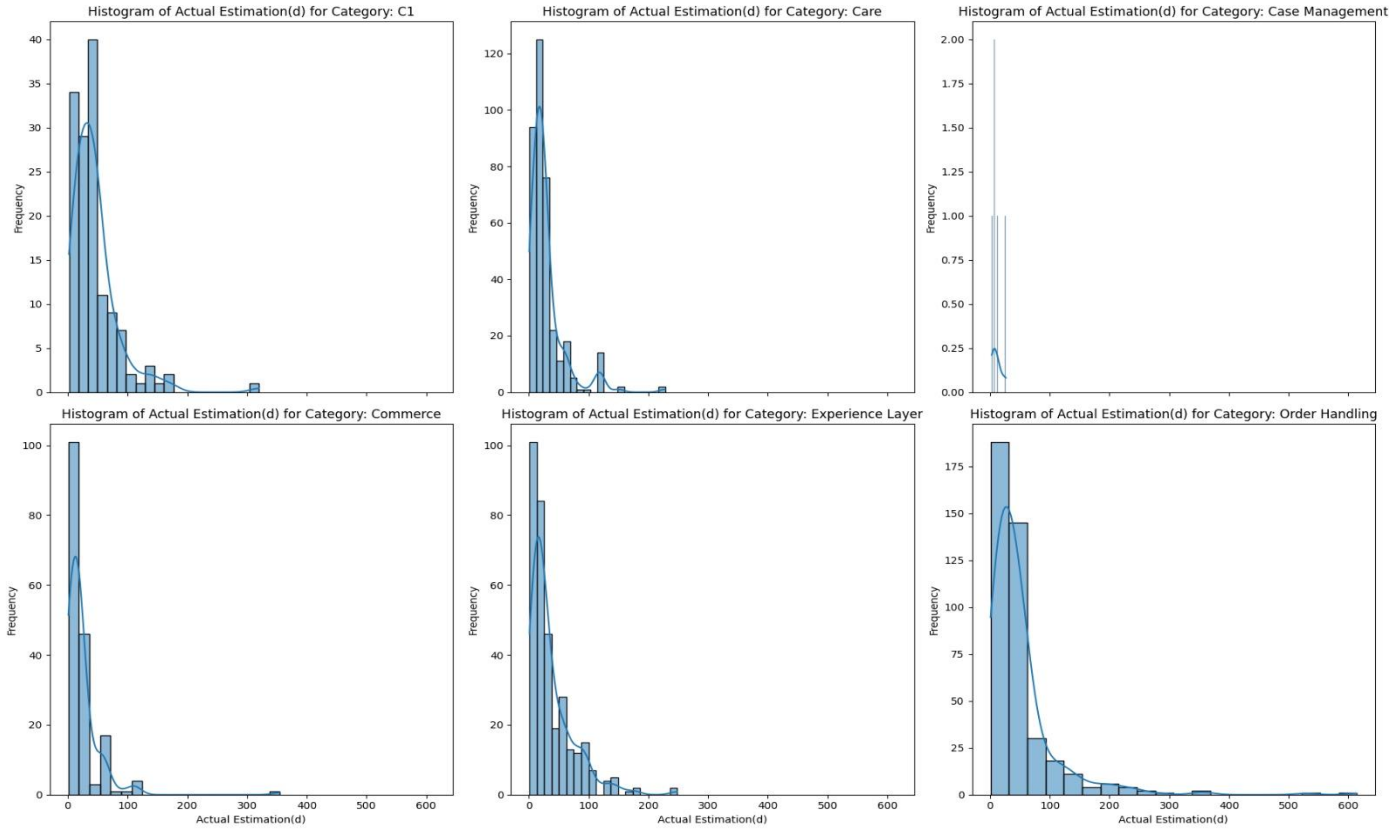


איור 4- גרף המתאר את התפלגות משך זמן חזוי (Original Estimate) ומשך זמן בפועל (Actual Estimate).

ניתן להבחין שהתפלגות משך זמן חזוי (Original Estimate) וגם התפלגות משך זמן בפועל (Actual Estimate) נראות כמו התפלגות מעריכית ואינן דומות להתפלגות גאוסיאנית מוגדרת. הדבר מצביע על כך שמשך פרויקט קצר הוא יותר נפוץ, בעוד שפרויקט ארוך יותר הוא פחות שכיח. הבנת ההתפלגות הפוטנציאלית הזו חשובה לצורך מידול וחיזוי מדויק של משכי פרויקטים, מכיוון שהיא מצביעה על כך שהנתונים עשויים להיות מוטים. זיהוי תבנית זו מסייע בבחירת טכניקות סטטיסטיות ולמידת מכונה מתאימות לשיפור דיוק ההערכה. בנוסף, הנחנו שנדרשים נתונים נוספים מכל קטגוריה על מנת לאשר את טיב ההתפלגות המדויקת. לאור ההתפלגות של הנתונים הקיימים לרשותנו בחרנו שיטות עבודה המתאימות לנתונים שאינם מוגדרים לפי התפלגות גאוסיאנית מובהקת (לדוגמא Isolation Forest).

7.2.2. ניתוח זמנים בפועל בכל קטגוריה

בעקבות המלצת המנחה האקדמי ומהנדסים נוספים מAmdocs, בחרנו לבחון חריגים לפי העמודה 'Release Train' המציגה את קטגוריות הפרויקטים. הנתונים מגלים שונות משמעותית בין קטגוריות הפרויקטים לפי העמודה 'Release Train', מה שמצביע ככל הנראה על הצורך בטיפול בחריגים לשיפור הדיוק והעקביות.



איור 5- התפלגות משך זמן <u>בפועל</u> (Actual Estimate) לפי קטגוריה מהעמודה 'Release Train'

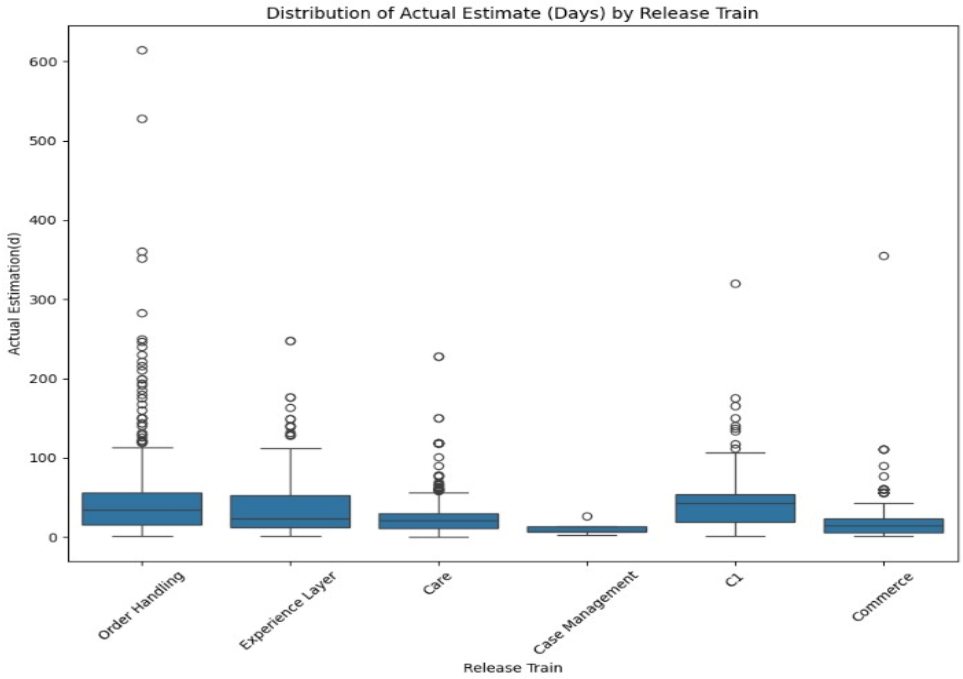
כדי להגיע למסקנות התחלתיות הצגנו את נתוני הסיכום הסטטיסטיים של כל קטגוריה:

	count	mean	std	min	25%	50%	75%	max
C1	140.0	45.197143	40.772444	2.0	18.500	42.6	54.00	320.0
Care	371.0	28.182210	30.131141	0.5	11.565	21.1	30.00	228.0
Case Management	5.0	11.300000	9.216832	3.0	7.000	7.0	13.00	26.5
Commerce	174.0	22.886667	33.763779	1.0	6.000	14.4	22.80	355.0
Experience Layer	339.0	37.042861	37.733446	1.0	12.000	23.4	52.85	248.0
Order Handling	413.0	50.673002	62.589440	1.0	16.000	34.0	55.80	614.0

איור 6- נתונים סטטיסטיים לכל קטגוריה מהעמודה 'Release Train'

מסקנות:

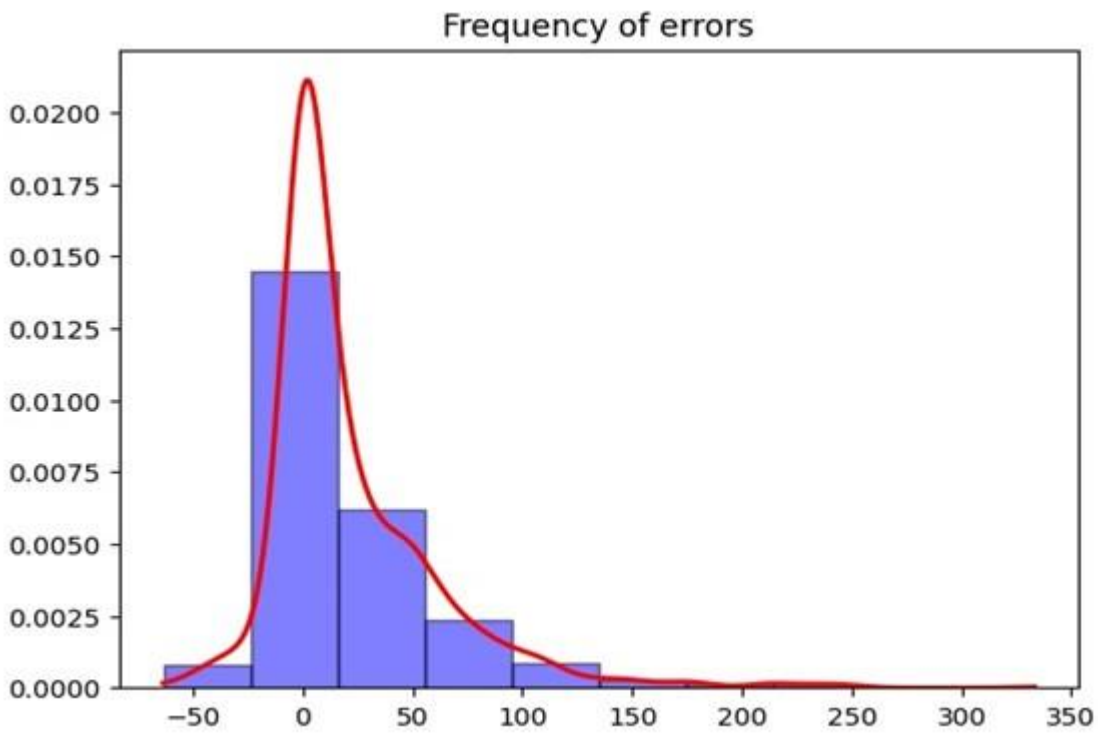
קטגוריית "Order Handling" מציגה את משך הזמן הממוצע הגבוה ביותר של 50.67 ימים ואת השונות הגדולה ביותר עם סטיית תקן של 62.59 ימים, אחוזון ה-75 של קטגוריה זו הינו 55.8 ימים, משמע 75% מהדגימות קטנות מ-55.8, ממצאים אלו מרמזים על נוכחות של דגימות חריגות באותה הקטגוריה.



איור 7- Boxplot המתאר את הזמן בפועל (Actual Estimate) לפי קטגוריה

קטגוריות כמו "Commerce", "Care", "C1" ו- "Experience Layer" מציגות שונות גבוהה במשכי הפרויקטים, עם סטיות תקן גבוהות ואחוזון 75 גבוה יחסית, מה שמצביע על חוסר עקביות במשכי הזמן. קטגוריית "Case Management" מראה שונות קטנה יותר, אך עשויה להיות מיוצגת באופן חסר. השונות הגבוהה נובעת מנוכחות חריגים ומדגישה את הצורך בטיפול בנקודות חריגות ובשיטות הערכה מדויקות ועקביות יותר.

7.2.3. ניתוח שגיאות



איור 8- התפלגות הטעויות

ההיסטוגרמה של הטעויות בין עמודות "Actual Estimations" לבין "Original Estimate" מראה התפלגות שנראית כמו התפלגות לוג נורמלית, מה שמעיד על כך שרוב השגיאות מרוכזות בצד השמאלי של ההתפלגות עם פיזור שאינו סימטרי. עם זאת, ישנו זנב ניכר בצד הימני של ההתפלגות. הזנב הימני המוטה מצביע על נוכחותם של מספר פרויקטים שבהם משך הזמן בפועל היה נמוך משמעותית ממשך הזמן המוערך. חריגים אלה עשויים להצביע על שגיאות בהערכה הראשונית כגון הערכת יתר של מורכבות המשימה או אי-התממשות של סיכונים משוערים. מצבים אלו עלולים להוביל להארכה של הזמנים המוקצבים למשימה לשווא. הבנת התפלגות זו היא קריטית לשיפור מודל ההערכה שלנו שכן היא מדגישה את הצורך לטפל גם בסטיות טיפוסיות וגם בשגיאות משמעותיות יותר. תובנה זו יכולה להנחות את פיתוחם של מודלי חיזוי עמידים יותר שיוכלו להתמודד טוב יותר עם אנומליות.

	count	mean	std	min	25%	50%	75%	max
Release Train								
C1	126.0	1.108730	3.811877	-0.50	0.000	0.0	0.000000	23.7
Care	334.0	12.832236	24.610628	-44.60	0.000	5.0	18.150000	150.0
Case Management	4.0	12.000000	8.746428	0.00	10.125	13.5	15.375000	21.0
Commerce	157.0	5.422420	13.725557	-17.00	0.000	1.2	4.700000	77.0
Experience Layer	305.0	35.189344	54.608981	-55.75	0.000	28.5	56.166667	333.3
Order Handling	372.0	35.127823	42.924289	-63.60	0.000	23.0	54.250000	210.0

איור 9- תיאור סטטיסטי של השגיאות עבור כל קטגוריה מהעמודה 'Release Train'

7.3. מדדי הערכה

התהליך הנוכחי ב-Amdocs תלוי במידה רבה בזמן המוערך שמספק מנהל המוצר אשר לרוב מבוסס על אינטואיציה וניסיון אישי ולרוב אינו מדויק. ההערכה הזו עוברת לאחר מכן השוואה לזמן האמיתי שנדרש לסיום הפרויקטים.

ביצענו השוואה בין כל המודלים לפי מדדי הערכה הבאים :

- **MSE (Mean Squared Error)** - מדד להערכת הדיוק, מחושב כהפרש המרובע הממוצע בין הערכים הנצפים לערכים החזויים.
- **RMSE (Root Mean Squared Error)** - השורש הריבועי שלMSE , מדד לפערים משמעותיים יותר בין התחזיות לערכים הנצפים.
- **MAE (Mean Absolute Error)** - ממוצע הערכים המוחלטים של הפרשי התחזיות מהערכים הנצפים.

	MSE	MAE	MAPE
0	2095.779772	26.268742	44.927717

איור 10- מדדי הערכה במצב הקיים

במצב הקיים, ביצועי החיזוי מציגים MSE של 2095.779, MAE של 26.268 ו-MAPE של 44.92%. ה-MSE מראה שיש הבדלים ריבועיים משמעותיים בין הערכים החזויים לערכים בפועל, מה שמדגיש את נוכחותן של שגיאות גדולות. ה-MAE, עם שגיאה ממוצעת של 26.268, משקף רמה מתונה של סטייה מהערכים בפועל. עם זאת, ה-MAPE הגבוה במיוחד של 44.92% מצביע על כך שהתחזיות של המודל נמצאות בממוצע בירידה משמעותית באחוזים, מה שמצביע על צורך בשכלול נוסף במודל או בעיבוד מוקדם של הנתונים כדי לשפר את הדיוק. מדדים אלו מצביעים על כך שהמודל עשוי לדרוש התאמות כדי להתמודד טוב יותר עם השונות במערך הנתונים ולשפר את דיוק החיזוי הכולל.

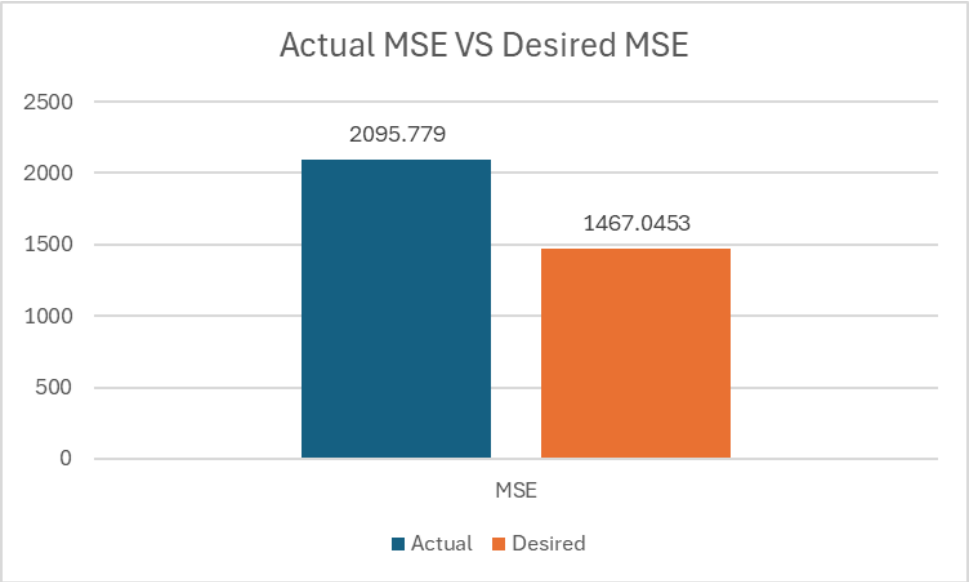
7.4. תחזיות במצב רצוי

כדי להבין את המצב הרצוי ערכנו ראיון [14.2] עם שלושה מנהלי פרויקטים ב- Amdocs שעתידים להשתמש במוצר כדי לאפיין את הגדרות המוצר כמה שיותר במדויק. שלב זה מתבסס על תשובותיהם לשאלות שנשאלו ועל הבקשות והדגשים שהם הציגו.

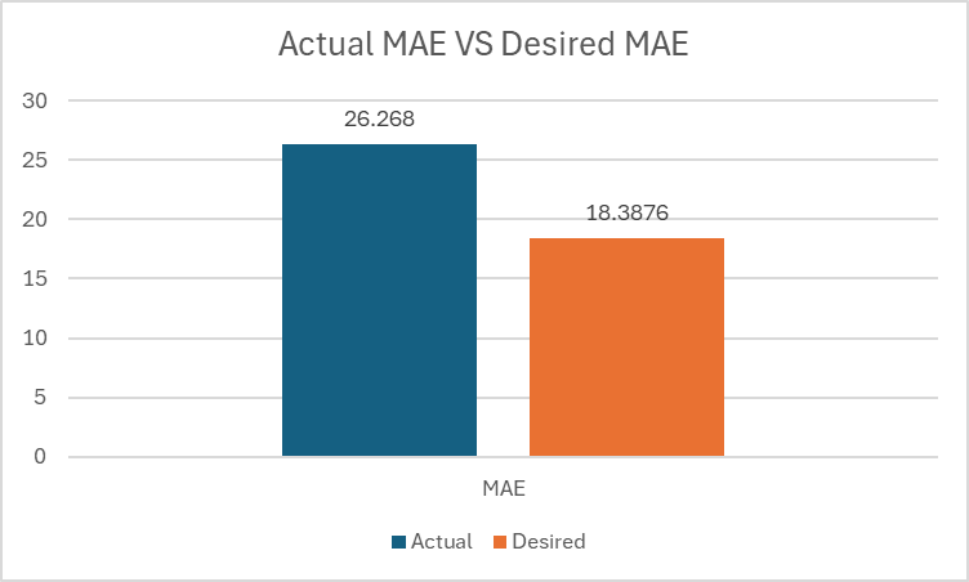
לאחר הטמעת המוצר, המצב הרצוי האידיאלי יכלול שיפורים משמעותיים בניהול פרויקטים והקצאת משאבים. הכלי יספק באופן עקבי תחזיות מדויקות יותר של משך הפרויקט ויפחית את השגיאה הממוצעת למינימום. הפחתת מדדי ההערכה כגון MSE, MAE ו-MAPE ב-30% לפחות תהווה שיפור משמעותי לפי הגורמים המקצועיים ב- Amdocs. רמת דיוק זו תאפשר למנהלי פרויקטים לתכנן בצורה יעילה יותר, להקצות משאבים ביעילות ולקבוע לוחות זמנים מציאותיים ובכך למזער את הסיכון לעיכובים בפרויקט וחריגות בתקציב.

במצב אופטימלי זה, הארגון יחווה תהליכי עבודה יעילים עם פחות מקרים של עיכובים בלתי צפויים או מחסור במשאבים. התחזיות האמינות שיספק המודל יהוו בסיס איתן לקבלת החלטות מושכלות ותאפשרנה התאמות יזומות בהתאם לשינויים בשוק . המודל יתאים את עצמו לסוגי פרויקטים שונים ולמורכבויות שונות תוך שמירה על הדיוק שלו על פני תרחישים שונים.

בנוסף, מחזור החיים הכולל של הפרויקט יהפוך לצפוי יותר, מה שיוביל לשיפור שביעות הרצון של הלקוחות עקב אספקה בזמן של הפרויקט. השיפור ביעילות יאפשר לארגון להרחיב את היקף הפעילות שלו ,לקחת על עצמו פרויקטים נוספים ולגדול בצורה יציבה, מבלי לפגוע באיכות או בלוחות הזמנים, מה שיוביל בסופו של דבר לניצול משאבים אופטימלי, רווחיות גבוהה יותר וליתרון תחרותי בשוק.



איור 11- מתאר את ה-MSE במצב הקיים ביחס למצב הרצוי



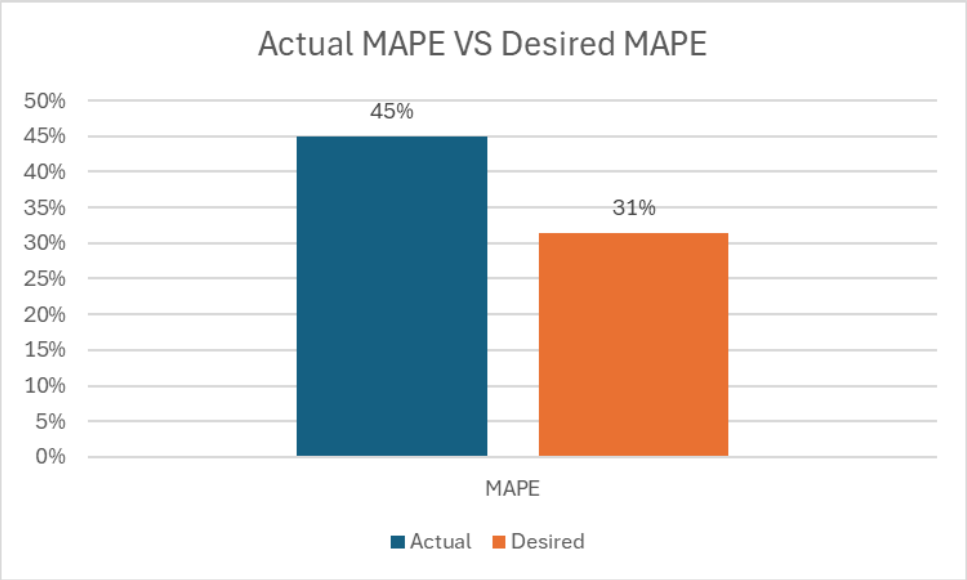
איור 12- מתאר את ה-MAE במצב הקיים ביחס למצב הרצוי

טכניקות Machine Learning צפויות להציג שיפור משמעותי בביצועים, עם ירידה ממוצעת של כ-30% לפחות בשגיאות התחזית כפי שצפוי לפי כל מדדי ההערכה.

MSE: צפויה ירידה מ-2095.779 ל-1467.04553, המצביעה על כך שהשגיאה הריבועית הכוללת בין התחזיות לערכים האמיתיים תצטמצם באופן משמעותי.

MAE: צפויה ירידה מ-26.268 ל-18.3876, מה שמעיד על כך שהטעות המוחלטת הממוצעת של המודל תתקרב יותר לאפס, כלומר התחזיות יהיו מדויקות יותר בממוצע.

MAPE: צפויה ירידה מ-45% ל-31%, המצביעה על שיפור משמעותי בדיוק היחסי של המודל.



איור 13- מדד MAPE במצב הקיים ביחס למצב הרצוי

לסיכום, השינויים הצפויים במדדי ההערכה מצביעים על כך שהמודל המעודכן יהיה מדויק ואמין יותר ויספק תחזיות קרובות יותר למציאות. שיפור זה יאפשר קבלת החלטות מושכלת יותר, תכנון יעיל יותר של פרויקטים והקצאת משאבים אופטימלית.

8. מתודולוגיה

בתהליך הפרויקט המורכב, נעשה שימוש במגוון רחב של כלים ואלגוריתמים על מנת לחזות את משך הפיתוח של פרויקטים בתעשיית התוכנה. כל שלב בתהליך כלל אפיון נתונים, עיבוד, מיצוי מידע, ושימוש במודלים מתקדמים לחיזוי. נציג את שלבי התהליך בצורה מפורטת, תוך הצגת הכלים והשיטות שנעשה בהם שימוש בכל שלב.

שלב 1: קבלת הנתונים ואפיון ראשוני של התהליך

בשלב הראשון, קיבלנו את מאגר הנתונים מאמדוקס, אשר כלל נתונים על פרויקטי פיתוח מוצרי תוכנה. הנתונים כללו מידע על נושא הפרויקט, סטטוס, תחום, ומשך זמן הפיתוח בפועל. לאחר קבלת הנתונים כדי להבין אותם לעומק ביצענו ראיונות עם עובדי החברה, תצפיות ומעקב אחרי כמה תהליכי פיתוח כדי לנסות לאפיין את צוואר הבקבוק בתהליך. לאחר בחינת הנתונים, נמצא שהעמודה "Description" מכילה מידע רב, אך ברוב המקרים בצורה בלתי מובנית. בשל כך, בדומה למאמר [3] הוחלט שבשביל לתרגם את נתוני הפרויקט למשימת חיזוי יש להתמקד בעיבוד הטקסט ותמצות מידע רלוונטי באמצעות כלים של עיבוד שפה טבעית (NLP).

שלב 2: עיבוד מקדים והכנת הנתונים

טיפול בחריגים :

בשלב הטיפול בחריגים, בחרנו להשתמש באלגוריתם Isolation Forest לאור ההתאמה שלו לנתונים שאינם מתאפיינים בהתפלגות גאוסיאנית, כמו הנתונים שלנו. האלגוריתם מצטיין גם בניהול נתונים שאינם מאוזנים ונתונים רב-ממדיים, המתאימים בדיוק לסוג הנתונים שבהם אנו עוסקים. בנוסף, במסגרת האפיון של הפרויקט, התבקשה התייחסות להטיות הנובעות מטעויות בהזנת נתונים למערכת. לכן, בהנחיית מנהלי הפרויקטים באמדוקס, נדרשנו להשתמש בשיטות להסרת חריגים בשלב זה כדי לשפר את דיוק המודלים ולמזער את השפעת הטעויות על התוצאות.

עיבוד שפה טבעית (NLP):

שימוש בכלי NLTK לטיפול בטקסטים שכלל הסרת מילות עצירה Lemmatization , ו- Stemming :

- **מילות עצירה** – נאספו על בסיס המלצות המחלקה בAmdocs ולפי מאגר קיים.
- **Lemmatization:** - הפחתת מילים לצורת הבסיס שלהן.
- **Stemming:** - הפחתת מילים לשורש המילה.

מיצוי תכונות ומשקול באמצעות TF-IDF:

לאחר עיבוד תיאור הפרויקט באמצעות NLP, השתמשנו בטכניקת IDF-TF (Term Frequency-Inverse Document Frequency) כדי למשקל את המילים בתיאורי הפרויקטים. טכניקה זו מסייעת בהבנת חשיבות של כל מילה במאגר ביחס לתדירות של אותה מילה באותו תיאור פרויקט ממנו היא הגיעה ובנוסף ביחס לייחודיות שלה ביחס למאגר תיאור הפרויקטים כולו. לאחר התייעצות עם המנחה התעשייתי בחרנו לייצא כ30 מילים מובילות לכל קטגוריה מהעמודה Release "Tarin" כדי לתת משקל שונה לכל תחום פיתוח שונה. לאחר משקול אסימוני המילים ובחירת משקולות מובילות סיימנו עם תהליך הכנת הנתונים לחיזוי.

$$w_{x,y} = tf_{x,y} \times \log\left(\frac{N}{df_x}\right)$$

איור 14- נוסחה ל- IDF-TF

שלב 3 : בחירת מודלים לחיזוי

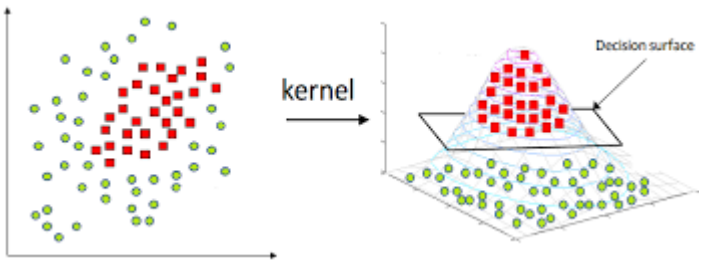
בשלב זה של הפרויקט בחלק הראשון בחרנו לממש כמה מודלים מתקדמים מעולם ה-ML. להלן פירוט על המודלים בהם השתמשנו :

- **Gradient Boosting Machine (GBM)** - משלב מספר רב של עצי החלטות חלשים ליצירת מודל חזק. האלגוריתם עובד באופן איטרטיבי, כאשר כל עץ חדש מתקן את השגיאות שנעשו על ידי העצים הקודמים עד הגעה לשגיאה מינימלית. מודל חזק זה יודע להתמודד עם א-ליניאריות בנתונים בדגש על נתונים רועשים בדיוק כמו הנתונים שלנו.

  הפרמטרים המרכזיים במודל זה כוללים את :

  1. **`n_estimators`** : קובע את מספר העצים הכוללים במודל. מספר גבוה עשוי לשפר את ביצועי המודל אך ידרוש יותר זמן אימון ויכול להוביל להתאמת יתר.
  2. **`learning_rate`** : קובע את גובה הצעד בכל הוספת עץ חדש. ערך נמוך דורש מספר גבוה יותר של עצים ויכול לשפר את איכות התחזיות. ערך גבוה מהיר יותר אך עשוי להקטין את הדיוק.
  3. **`max_depth`** : מגביל את עומק העצים כדי למנוע התאמת יתר על ידי מניעת יצירת עצים מורכבים מדי.
  4. **`min_samples_split`** : קובע את מספר הדגימות המינימלי הנדרש כדי לבצע פיצול בעץ. ערך גבוה יותר יכול למנוע התאמת יתר על ידי הגבלת פיצולים.
  5. **`subsample`** : קובע את החלק מהנתונים שישמשו לאימון כל עץ. ערך נמוך יותר יכול לשפר את הכללות של המודל אך עשוי להוביל לאימון פחות יציב.

- **Support Vector Machine (SVM)** – אלגוריתם המוצא את ההיפר-מישור שמפריד בצורה הטובה ביותר בין הקבוצות השונות בנתונים. כאשר הנתונים אינם ליניאריים SVM, משתמש בטרנספורמציות קרנל – במקרה שלנו, קרנל מסוג "RBF" (Radial Basis Function). קרנל RBF מתאים ללכידת קשרים שאינם ליניאריים, ומבצע מיפוי למרחב בעל ממדים גבוהים יותר שבו ניתן להפריד בין הקבוצות בצורה ליניארית. בקרנל מסוג זה פרמטר Gamma קובע את טווח ההשפעה של כל דוגמה. ערך גבוה של Gamma מצמצם את הטווח ויכול להביא להתאמת יתר, בעוד שערך נמוך מדי יכול להוביל להתאמת חוסר. פרמטר הקנס C קובע את העונש על שגיאות הסיווג, ערך גבוה של C מביא להתאמה מדויקת אך עם שוליים צרים, בעוד ערך נמוך של C יוביל לשוליים רחבים יותר עם יותר שגיאות.

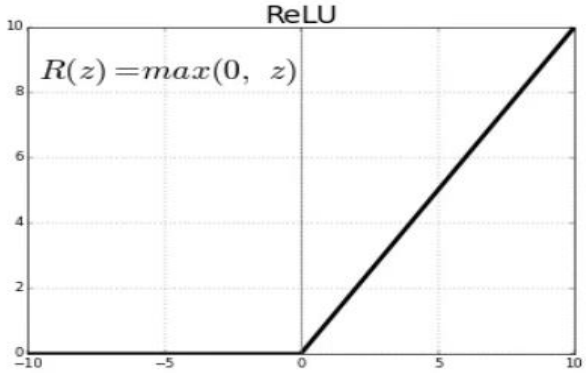


איור 15 – דוגמא לשימוש בקרנל במודל SVM

- **LightGBM -** מודל מבוסס על שיטת Boosting המיועדת לבניית מודלים חזקים על ידי שילוב של מספר עצי החלטה. המודל מתאפיין בדרך כלל במהירות גבוהה יותר מGBM ובגלל שידענו שבהמשך המוצר שנפתח יתמודד עם מערכי נתונים גדולים רצינו לתת אופציה למודל המתאים בדיוק למקרה כזה. עקרון העבודה של המודל הוא להוסיף עץ החלטה חדש בכל שלב של אימון כדי לתקן את השגיאות של העצים הקודמים.
  הפרמטרים המרכזיים במודל זה כוללים את:
  1. **`num_leaves`:** קובע את מספר העלים בעץ החלטה. ערך גבוה מספק למודל יכולת למידה מדויקת יותר אך עם סיכון להתאמת יתר. ערך נמוך עשוי לפשט את המודל אך לצמצם את הדיוק.
  2. **`learning_rate`:** קובע את גובה הצעד בכל הוספת עץ חדש. ערך נמוך דורש מספר גבוה יותר של עצים ויכול לשפר את איכות התחזיות. ערך גבוה מהיר יותר אך עלול להקטין את הדיוק.
  3. **`n_estimators`:** קובע את מספר העצים הכוללים במודל. מספר גבוה עשוי לשפר ביצועים אך ידרוש יותר זמן אימון ויכול להוביל להתאמת יתר.
  4. **`max_depth`:** מגביל את עומק העצים כדי למנוע התאמת יתר על ידי מניעת יצירת עצים מורכבים מדי.

- **Random Forest Regressor -** מודל מבוסס על עץ החלטה המשלב מספר עצי החלטה ליצירת מודל חיזוי חזק יותר. כל עץ בתוך היער לומד על תת-קבוצה שונה של הנתונים והמאפיינים, ותחזיותיו משולבות על ידי חישוב ממוצע של תחזיות כל העצים כדי להגיע לתוצאה הסופית. שיטה זו מספקת עמידות גבוהה נגד התאמת יתר ומביאה לשיפור בביצועים על פני עץ החלטה יחיד.
  הפרמטרים המרכזיים במודל זה כוללים את:
  1. **`n_estimators`:** קובע את מספר העצים ביער. מספר גבוה עשוי לשפר את ביצועי המודל ע"H יציבות בחיזוי אך ידרוש יותר זמן אימון ויכול להוביל להתאמת יתר.
  2. **'max_depth'**: מגביל את עומק העצים המרבי ביער. הגבלת העומק יכולה למנוע התאמת יתר על ידי מניעת יצירת עצים מורכבים מדי.
  3. **'min_samples_split'**: קובע את מספר הדגימות המינימלי הנדרש כדי לבצע פיצול בעץ. ערך גבוה יותר יכול למנוע התאמת יתר על ידי הגבלת פיצולים.
  4. **'min_samples_leaf':** קובע את מספר הדגימות המינימלי הנדרש עבור עלה בעץ. ערך גבוה יותר מונע יצירת עלים קטנים מדי שעשויים להיות מאוד מותאמים לנתוני האימון.
  5. **'max_features'**: קובע את המספר המירבי של מאפיינים לשקול כשבוחרים את הפיצול האופטימלי. ערך נמוך יגביר אקראיות, מה שעשוי לשפר את המודל בכללית ולמנוע התאמת יתר.

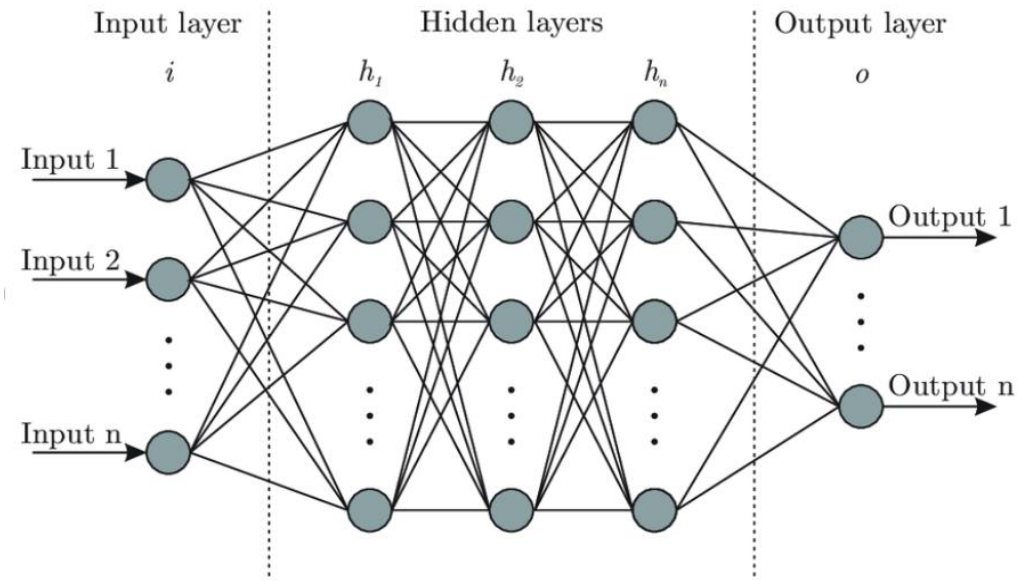
- (NN) Neural Networks - מודל מבוסס על מערכת של נוירונים מלאכותיים המחקים את הדרך שבה המוח האנושי פועל. הרשתות כוללות שכבות של נוירונים, כאשר כל נוירון מקבל משקלות מהשכבות הקודמות ומבצע חישובים באמצעות פונקציית אקטיבציה על מנת להעביר את המידע לשכבות הבאות. הפונקציה בה בחרנו היא פונקציה מסוג RelU שידועה ביעילותה החישובית ובכך שהיא מתאימה לנתונים חיוביים כמו חיזויי זמן. בחרנו להשתמש ברשת מסוג NN פשוטה בגלל שמשימת החיזוי היא משימה לחיזוי ערך יחיד ובתוכה פונקציית ההפסד לפי MSE בהתאמה לבעיות מספריות (בניגוד לבעיות סיווג). התהליך כלל מספר שכבות: שכבת קלט, מספר שכבות מוסתרות ושכבת פלט. המודל מתאים במיוחד לבעיות בהן יש נתונים עם קשרים מורכבים וא-ליניאריים, כמו חיזוי זמן פיתוח של פרויקטים.



איור 16- פונקציית אקטיבציה ReLU

הפרמטרים המרכזיים במודל זה כוללים את:

1. **'num_layers':** קובע את מספר השכבות הנסתרות ברשת. מספר גבוה יותר של שכבות מאפשר למודל ללמוד ייצוגים מורכבים יותר של הנתונים, אך יכול להוביל להתאמת יתר אם אינו מנוהל נכון.
2. **'units_per_layer'**: קובע את מספר הנוירונים בכל שכבה. ערך גבוה יותר בכל שכבה יכול לשפר את היכולת של המודל ללכוד דפוסים מורכבים, אך גם להגדיל את סיכון ההתאמת יתר ואת זמן האימון.
3. **'learning_rate'**: קובע את גובה הצעד בעדכון המשקלות של הנוירונים במהלך האימון. ערך נמוך דורש יותר צעדים לאימון אך יכול לשפר את איכות התחזיות, בעוד שערך גבוה מהיר יותר אך עלול להוביל לתוצאות פחות מדויקות.
4. **'batch_size'**: קובע את מספר הדגימות הנשלחות לרשת בבת אחת במהלך האימון. ערך גבוה יותר יכול להאיץ את התהליך אך עלול לדרוש יותר זיכרון, בעוד שערך נמוך עשוי לספק עדכונים מדויקים יותר למשקלות.
5. **'epochs'**: קובע את מספר הפעמים שהמודל עובר על כל מערך הנתונים במהלך האימון. מספר גבוה יותר של תקופות עשוי לשפר את ההתאמה של המודל, אך עשוי גם להוביל להתאמת יתר אם אינו מנוהל נכון.



איור 17- תיאור רשת נוירונים מלאכותית

שלב 4: אימות ובחירת המודל האופטימלי

לכל המודלים ביצענו אימות צולב לתשובות ובכל המודלים בחרנו היפר-פרמטרים באמצעות Grid Search. טכניקת האימות צולב ותהליך בחירת ההיפר-פרמטרים נועדו להבטיח אופטימליות לתוצאות המודלים בהתאמה לנתונים.

אימות צולב (Cross-validation) - שיטה זו מאפשרת להעריך את יציבות ביצועי המודל ע"י חלוקת הנתונים למספר חלקים (קבוצות), כאשר כל חלק בתורו משמש כקבוצת בדיקה והשאר כקבוצת אימון. התוצאה המתקבלת היא ממוצע הביצועים על כל הקבוצות, מה שמסייע למנוע התאמת יתר. האימות הצולב מבטיח שהמודל נבחן על נתונים שאינו נחשף להם במהלך האימון ובכך מונע הטיה ותוצאות שגויות.

Grid Search- שיטה זו משמשת לבחירת ההיפר-פרמטרים האופטימליים על ידי בדיקת כל האפשרויות האפשריות בתוך טווח מסוים. התוצאה הטובה ביותר נבחרת בהתאם למדד ההערכה הנבחר.

שלב 5: מדדי הערכה

על ידי ניתוח שלושה מדדים קריטיים אלו, ניתן להעריך את דיוק ומהימנות תהליך ההערכה הנוכחי:

1. **שגיאה ריבועית ממוצעת (MSE) :** מודד את ההבדלים הריבועיים הממוצעים בין הזמנים המוערכים לזמנים האמיתיים. ערך גבוה של MSE מצביע על פערים משמעותיים בין הזמנים המוערכים לזמנים האמיתיים של הפרויקטים, מה שמעיד על מקום לשיפור בתהליך ההערכה. יתרון נוסף הוא שהנגזרת של MSE רציפה ונוחה לחישוב, מה שהופך אותו למתאים לשימוש באלגוריתמי למידת מכונה המבוססים על גרדיאנט (לדוגמא רגרסיה ליניארית ורשתות נוירונים). יתרון משמעותי של MSE הוא בכך שהוא מעלה בריבוע את השגיאות, דבר המעניק משקל רב יותר לשגיאות גדולות. תכונה זו יכולה להיות מועילה במיוחד כאשר חשוב לזהות ולתקן שגיאות חריגות.

$$MSE = \frac{1}{n}\sum_{i=1}^{n}(y_i - \hat{y}_i)^2$$

איור 18- נוסחה לשגיאה ריבועית ממוצעת

2. **שגיאה מוחלטת ממוצעת (MAE):** מייצג את הטעות המוחלטת הממוצעת בין הזמנים המוערכים לזמנים האמיתיים. מדד זה מודד את ממוצע גודל ה"טעויות", כלומר ההבדל בין האומד לבין מה שנאמד, ה-MAE מדגיש את עקביות ודיוק התחזיות הראשוניות של מנהל המוצר. MAE הוא אינטואיטיבי וקל להבנה. בנוסף, בניגוד ל MSE- , MAEאינו מעלה בחזקה את השגיאות, ולכן הוא פחות רגיש לחריגות גדולות, מה שיכול להיות יתרון במקרים בהם יש הרבה חריגים.

$$MAE = \frac{1}{n}\sum_{i=1}^{n}|\hat{y}_i - y_i|$$

איור 19- נוסחה לשגיאה מוחלטת ממוצעת

3. **שגיאת אחוז מוחלט ממוצעת (MAPE) :** MAPE, או Mean Absolute Percentage Error, הוא מדד המשקף את ההבדל הממוצע באחוזים בין הערכים החזויים לבין הערכים בפועל. מדד זה מספק הבנה על כמה רחוקה התחזית מהערך הממשי באופן יחסי, ומציג זאת באחוזים. היתרון בשימוש ב-MAPE הוא היכולת להשוות בצורה נוחה בין

מערכי נתונים שונים או בין מודלים שונים, מכיוון שהוא מבטל את ההשפעה של יחידות המדידה ומספק פרספקטיבה אחידה על דיוק התחזיות.

$$MAPE = \frac{1}{n}\sum_{i=1}^{n}\left|\frac{\hat{y}_i - y_i}{y_i}\right| \times 100$$

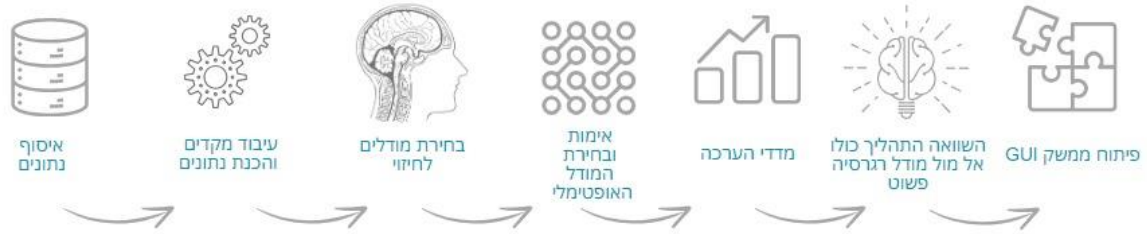
איור 20- נוסחה לשגיאת אחוז מוחלט ממוצעת

שלב 6: השוואת התהליך כולו אל מול מודל רגרסיה פשוט

לאחר שקיבלנו את התוצאות שהראו שרשת הנוירונים היא המודל האופטימלי לחיזוי, בדומה למאמר [5], ביצענו גם רגרסיה לינארית על מטריצת המשקלים הסופית שהתקבלה לאחר משקול המילים לפי קטגוריות. מטרת ביצוע מודל הרגרסיה הייתה לבדוק אם ניתן באמצעות מודל פשוט יותר להגיע לתוצאות טובות יותר לבעיה, כפי שלעיתים קורה בחיזוי בעיות מורכבות התלויות בזמן. הרעיון לבדיקה זו הגיע בהמלצת המנחה האקדמי ועל פי מאמר [5]. התוצאות הראו שרשת הנוירונים מדויקת יותר בחיזוי מאשר המודל הפשוט והנאיבי של רגרסיה לינארית, מה שחיזק את הבחירה ברשת הנוירונים כמודל המתאים ביותר לבעיה הנחקרת.

שלב 7: פיתוח ממשק GUI

עיצוב הממשק הגרפי (GUI) כלל שימוש ב-Tkinter ליצירת ממשק ידידותי למשתמש, הכולל תצוגת לוגו ורקע, אפשרות להזנת תיאור פרויקט ובחירה מקטגוריה מתוך תפריט נפתח, וכפתור חיזוי משך הפרויקט. לאחר לחיצה על כפתור החיזוי, המערכת מעבדת את הקלט דרך Pipeline ומציגה את תחזית משך הפרויקט על הממשק. המתודולוגיה משלבת את עיבוד הנתונים, חיזוי באמצעות מודל נוירונים מאומן מראש, וממשק גרפי אינטראקטיבי. תהליך זה מאפשר למשתמשים שאינם טכניים להשתמש במערכת בקלות.



איור 21-תרשים זרימה מתודולוגיית העבודה

9. הצגת חלופות

9.1. כללי

לטובת ניתוח החלופות בחנו שני פתרונות עיקריים: מודל רגרסיה ליניארית פשוט ומודלים מתקדמים של למידת מכונה. כל פתרון מציע יתרונות ואתגרים מובהקים, מה שהופך אותו חיוני להעריך אותם על פי קריטריונים מרכזיים כגון דיוק, מורכבות, יכולת הרחבה (Scalability) ועוד רבים אחרים.

9.2. תהליך הערכת החלופות

כל חלופה נבחנה על פי מספר קריטריונים, כאשר לכל קריטריון יינתן ציון בין 1 ל-5, כאשר 5 הוא הציון הגבוה ביותר. המשקל של כל קריטריון, המבטא את חשיבותו היחסית, יקבע מראש על ידי צוות המנהלים של Amdocs. הציון הסופי של כל חלופה יחושב על ידי חישוב ממוצע משוקלל של הציונים, כאשר כל ציון יוכפל במשקלו המתאים.

9.3. הצגת חלופות

9.3.1. המצב הקיים

הסתמכות על אינטואיציה ודעות סובייקטיביות כדי להעריך את הזמן הנדרש לפרויקטים. חלופה זו מבוססת על הניסיון והשיפוט של מנהלי פרויקטים, המבצעים תחזיות ללא תמיכה במודלים מונעי נתונים.

9.3.2. חלופה א

מודל הרגרסיה הליניארית מציע פשטות ויעילות חישובית, מה שהופך אותו לבחירה מצוינת עבור פרויקטים הדורשים פריסה מהירה ועלויות נמוכות. עם זאת, הדיוק המוגבל שלו עשוי להגביל את יכולתו לטפל במערכי נתונים מורכבים או הדורשים רמת דיוק גבוהה. לכן, מתאים יותר לתרחישים בהם הפשטות והמהירות עומדות בראש סדר העדיפויות.

9.3.3. חלופה ב

חלופה זו מסתמכת על מודלים מתקדמים מעולם ה-Machine Learning כגון SVM, LightGBM, Random Forest, GBM ורשתות נוירונים. מודלים אלו, הידועים ביכולתם להתמודד עם מערכי נתונים מורכבים וממדים גבוהים, מספקים דיוק תחזיתי משמעותית גבוה יותר. בעוד שהם דורשים משאבי חישוב ניכרים ומומחיות טכנית, הדיוק המתקבל הופך אותם לבחירה אידיאלית עבור סביבות בהן דיוק התחזיות קריטי.

9.4. השוואה בין חלופות

קביעת משקלים לפי חשיבות מול הארגון

משקלי הקריטריונים משקפים את חשיבותם היחסית להצלחת הפרויקט. הדיוק, המשוקלל ב-40%, הוא קריטי להערכת מדויקת של לוחות זמנים ולכן משקלו הוא הגבוה ביותר. עלות ומורכבות, המשוקללות ב-20% ו-15% בהתאמה, מאזנות בין הדיוק לבין אילוצי משאבים. יכולת הרחבה ותחזוקה, המשוקללות ב-10% כל אחת, חשובות לפיתוח ארוך טווח. התועלת, המשוקללות ב-5%, מדגישות את הצורך בשיפור מתמשך תוך שמירה על המטרות העיקריות.

דיוק: 5 מראה דיוק גבוה ו-1 דיוק נמוך

מורכבות: 5 מראה מורכבות נמוכה ו-1 מורכבות גבוהה

עלות: 5 מראה עלות נמוכה ו-1 עלות גבוהה

יכולת הרחבה: 5 מראה יכולת הרחבה גבוהה ו-1 נמוכה

תחזוקה: 5 מייצג תחזוקה נמוכה ו-1 תחזוקה גבוהה

<u>תועלת</u> : 5 מייצג תועלת גבוהה ו-1 תועלת נמוכה

Criteria	Weight (%)	Solution 1: Linear Regression	Solution 2: Advanced ML Models	Actual (Intuition)
Accuracy	40%	3	5	1
Complexity	15%	4	2	5
Cost	20%	4	2	5
Scalability	10%	3	5	1
Maintenance	10%	4	3	5
Benefits	5%	3	5	2
Total Weighted score	100%	3.45	3.75	2.85
Subjective Considerations	-	Quick to deploy, easy to understand	High potential for optimization and better long-term results	No change

טבלה 5- טבלת קריטריונים ממושקלים

9.5. קביעת החלופה האופטימלית

לאחר בחינה מעמיקה של שלוש האפשרויות שהוצגו, מתברר כי חלופה 2, המתבססת על מודלים מתקדמים מעולם ה- Machine Learning, עולה כבחירה המועדפת.

- **חלופה 2:** שימוש במודלים כגון SVM, LightGBM, Random Forest, GBM ורשתות נוירונים מבטיחים **דיוק גבוה ויכולות הרחבה משמעותיות** עם ציון משוקלל של 3.75 . יכולות אלה מאפשרות לטפל בנתוני פרויקטים מורכבים ולהשיג הערכות מאמץ מדויקות יותר. למרות שמורכבות הפתרון ועלויות היישום שלו גבוהות יותר, היתרונות העצומים שהוא מציע, כמו יכולת החיזוי המדויקת והאמינות הגבוהה, הופכים אותו לבחירה האידיאלית עבור פרויקטים הדורשים רמת דיוק גבוהה.

- **חלופה 1:** מודל הרגרסיה הליניארית מקבל ציון משוקלל של 3.45 . חלופה זו מציעה פשטות ויעילות, אך סובלת מ**דיוק מוגבל**. היא מתאימה יותר לפרויקטים קטנים ופשוטים, או למצבים בהם המשאבים מוגבלים. בנוסף, בהמשך כשהמודל יאלץ להתמודד עם מערכי נתונים גדולים יותר הרגרסיה צפויה להוציא תוצאות פחות מדויקות מכיוון שהיא אינה מותאמת למערכי נתונים גדולים כמו שצפוי.

- **המצב הקיים:** הסתמכות על אינטואיציה ושיפוט סובייקטיבי היא הגישה הפחות אמינה. היא חסרת **דיוק, מדרגיות ואמינות**, מה שמוביל להערכות מאמץ לא מדויקות ופוגע ביעילות ניהול הפרויקט.

איור 22- תוצאות השוואות החלופות

לסיכום ,בעוד שחלופה 2 דורשת השקעה גדולה יותר ,היא מציעה את התועלת הגבוהה ביותר מבחינת דיוק ,יכולות הרחבה ואמינות. בחירה בפתרון זה תאפשר לנו לשפר משמעותית את יכולתנו לחזות מאמצים בפרויקטים ,לקבל החלטות מושכלות יותר ולנהל את הפרויקטים שלנו בצורה יעילה יותר.

10. מימוש הפתרון

לאחר שבחרנו בחלופה של מודלים מתקדמים מעולם ה - Machine Learning התחלנו במימוש הפתרון בהתאם למתודולוגיה שהוגדרה בפרויקט. מטרת העל הייתה ליצור מודל חיזוי אופטימלי ונוח לשימוש למטרות החיזוי שהוגדרו.

10.1 איסוף נתונים

במהלך שלב איסוף הנתונים, ביצענו ראיונות עם מנהלי מוצר ממחלקת ה-Product ב-Amdocs. במסגרת הראיונות זוהו נקודות מפתח קריטיות וצווארי בקבוק, ובהתאם לכך הוחלט למקד את הפיתוח בנתוני פרויקטים מעולם ה-Digital של לקוחות Amdocs. לשם כך, נאספו נתונים ממערכת ה-JIRA של 1,505 פרויקטים הקשורים לאותו תחום.

	Business Priority	Key	Summary	Status	Reporter	Assignee	Created	Updated	Due Date	Issue Comments	Last Comment	Description	Original Estimate (Hours)	Original Estimate (Days)	Effective Estimation(d)	Actual Estimation(d)	Release Train	Product Line
0	NaN	NaN	Features for extract (GenAI pilot)	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN
1	8750.0	DE-275485	[ATT NFT Issue-Scheduler] Blue\Green for Sched...	Ready for NFT	Manish Sehgal	NaN	12/1/2023	3/26/2024	NaN	NaN	NaN	There's a critical gap w.r.t Blue\Green for Sc...	80:00:00	10:00	10.0	10.0	Order Handling	NaN
2	NaN	DE-274762	[Generic] Customer Engagement Framework (solut...	Cancelled	Tamar Sadeh	NaN	11/29/2023	11/30/2023	NaN	NaN	NaN	Purpose of this feature is to provide solution...	NaN	0:00	0.0	0.0	Experience Layer	NaN
3	7700.0	DE-273471	[B2B Sales] [HVO] Amend/ Amend Cancel of singl...	Feature Review Done	Amos Krayef	NaN	11/20/2023	1/23/2024	NaN	NaN	NaN	Overview: The purpose of this feature is to su...	NaN	0:00	0.0	0.0	Order Handling	NaN
4	7960.0	DE-273047	[B2B Sales] [HVO] Promotion Entitlement for SM...	Feature Review Done	Amos Krayef	NaN	11/15/2023	1/23/2024	NaN	NaN	NaN	For SMB handle not to qualify benefit from Pi ...	NaN	0:00	0.0	0.0	Order Handling	NaN
5	54.0	DE-272756	[New Billing Care] Business Components Alignme...	Cancelled	Milena Guryan	NaN	11/9/2023	11/15/2023	NaN	NaN	NaN	Purpose The purpose of this feature is to alig...	NaN	0:00	0.0	0.0	Experience Layer	NaN

איור 23- מערך הנתונים המקורי

10.2. עיבוד מקדים והכנת הנתונים

10.2.1.בחירת משתנים

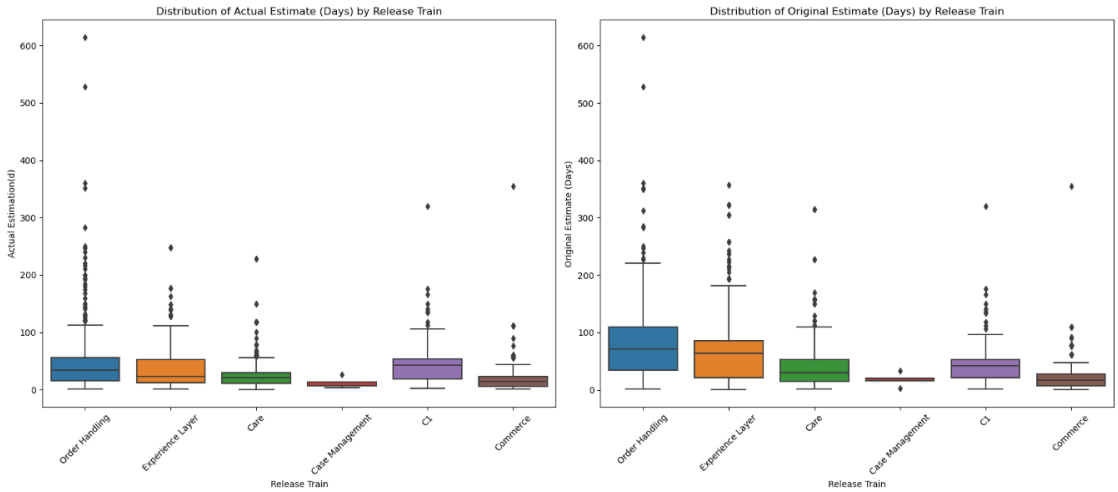
בשלב העיבוד המקדים והכנת הנתונים, התמקדנו בעמודות שהוגדרו כרלוונטיות על ידי החברה מתוך רשימת המאפיינים הקיימים בנתונים. שאר העמודות הכילו ערכים חד-חד ערכיים וחסרי משמעות לחיזוי, ולכן, בהתאם לבקשת החברה, הוחלט שלא לעבד עמודות אלו. העמודות הרלוונטיות שנבחרו כוללות: 'Key' (מזהה פרויקט), 'Description' (תיאור הפרויקט בשפה חופשית), 'Original Estimate (Days)' (הערכת משך הפרויקט), 'Actual Estimation(d)' (משך הפרויקט בפועל), ו-'Release Train' (קטגוריית הפרויקט).

	Description
0	B2B Sales System Decoupling from OM Domain: This feature aims to decouple the B2B sales process from the OM domain to reduce BOM costs and support accounts using external fulfillment systems. The solution involves separating the Proposal & Agreement processes from OM, enabling direct execution of actions such as capturing templates, creating draft orders, and setting agreement statuses without OM. It also supports configurations for handling agreements with or without OM, ensuring that only necessary components are processed. Successful decoupling of B2B sales from OM will be validated through system configurations, process execution without OM, and comprehensive testing of agreement processes.
1	Integrate customer feedback into the development process with a new collection module and dashboard for real-time analysis and action. [JIRA-445566]
2	Enhance user activity mapping by automating the process to map reasons to SRM. This includes handling scenarios for amendments and cancellations, and selecting the primary reason in cases with multiple reasons. Ensure the solution integrates smoothly with current systems and does not disrupt existing processes.
3	Multi-Locale Support for Billing Dashboard: Enhance the Billing Dashboard to support multiple locales, including English and Spanish. This feature will enable the dashboard to dynamically display labels and financial information based on the selected locale, ensuring accurate representation of billing data across different regions. Successful implementation will include verification of label translations, locale-specific date formats, and accurate display of billing information in both English and Spanish. [JIRA-334455]

איור 24- דוגמא לתיאורי פרויקט מהעמודה "Description"

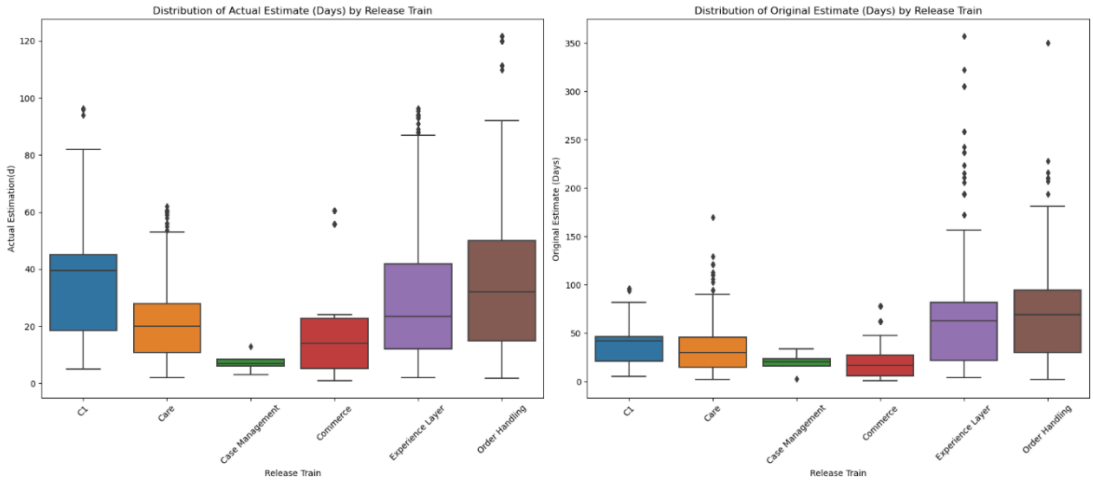
10.2.2.זיהוי חריגים

לאחר הגדרת המשתנים הרלוונטיים, אמדוקס ביקשה שנבצע זיהוי חריגים בשלב זה כדי למנוע הטייה בנתונים שעלולה להשפיע על דיוק החיזוי [14.2]. בהתאם לבקשתם, ביצענו מחקר EDA לזיהוי חריגים לפי קטגוריות מתוך העמודה 'Release Train', עבור העמודות 'Actual Estimation(d)' ו-'Original Estimate (Days)'.



איור 25 – גרף Boxplot לזיהוי חריגים לפי 'Release Train' לפני שימוש ב-Isolation Forest

כמו שצוין בפרק המתודולוגיה ובסקר הספרות, בחרנו ביחד עם המנחה התעשייתי להשתמש באלגוריתם Isolation Forest לסינון החריגים בשלב הנוכחי. לאחר סינון החריגים, נותרו 1,441 שורות נתונים סך הכל.



איור 26 – גרף Boxplot לזיהוי חריגים לפי 'Release Train' אחרי השימוש ב-Isolation Forest

10.2.3.עיבוד שפה טבעית (Natural Language Processing)

בשלב זה, עיבדנו את הטקסט מתוך העמודה 'Description' באמצעות ספריית NLTK. חילקנו אותו למילים, וביצענו Lemmatization ו-Stemming להכנת הנתונים לתהליך המשקול. לאחר ייצוא המילים המובילות בעזרת NLP, שקללנו את האסימונים באמצעות TF-IDF בהתאם לקטגוריות 'Release Train'. בהנחיית המנחה האקדמי והתעשייתי, בחרנו לייצא משקלים של 30 המילים המובילות מכל קטגוריה, כך שבמטריצת המשקלים הסופית לחיזוי ישנן 160 עמודות. כל עמודה במטריצה מייצגת את שם הקטגוריה והמילה המובילה לאחר עיבוד, והערכים משקפים את המשקל שניתן למילה במשימת החיזוי.

באיור 7 ניתן לראות את מטריצת המשקלים שקיבלנו בתור פלט לאחר שימוש בשיטת NLP ושקלול של המילים לפי קטגוריות מהעמודה "Release Tarin" .

[40]: combined_data

[40]:

	C1_continu	C1_persona	C1_duplic	C1_sync	C1_widget	C1_contract	C1_code	C1_filter	C1_msb	C1_site	...	Order Handling_sitelist	Order Handling_site	Order Handling_continu	Order Handling_export
0	0.814174	0.781076	0.755281	0.708667	0.70424	0.699955	0.692241	0.669296	0.667682	0.665947	...	0.000000	0.000000	0.000000	0.000000
1	0.814174	0.781076	0.755281	0.708667	0.70424	0.699955	0.692241	0.669296	0.667682	0.665947	...	0.000000	0.000000	0.000000	0.000000
2	0.814174	0.781076	0.755281	0.708667	0.70424	0.699955	0.692241	0.669296	0.667682	0.665947	...	0.000000	0.000000	0.000000	0.000000
3	0.814174	0.781076	0.755281	0.708667	0.70424	0.699955	0.692241	0.669296	0.667682	0.665947	...	0.000000	0.000000	0.000000	0.000000
4	0.814174	0.781076	0.755281	0.708667	0.70424	0.699955	0.692241	0.669296	0.667682	0.665947	...	0.000000	0.000000	0.000000	0.000000
...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...
1291	0.000000	0.000000	0.000000	0.000000	0.00000	0.000000	0.000000	0.000000	0.000000	0.000000	...	0.694519	0.684076	0.681638	0.680429
1292	0.000000	0.000000	0.000000	0.000000	0.00000	0.000000	0.000000	0.000000	0.000000	0.000000	...	0.694519	0.684076	0.681638	0.680429
1293	0.000000	0.000000	0.000000	0.000000	0.00000	0.000000	0.000000	0.000000	0.000000	0.000000	...	0.694519	0.684076	0.681638	0.680429
1294	0.000000	0.000000	0.000000	0.000000	0.00000	0.000000	0.000000	0.000000	0.000000	0.000000	...	0.694519	0.684076	0.681638	0.680429
1295	0.000000	0.000000	0.000000	0.000000	0.00000	0.000000	0.000000	0.000000	0.000000	0.000000	...	0.694519	0.684076	0.681638	0.680429

1296 rows × 181 columns

איור 27- מטריצת המילים המובילות מכל קטגוריה לאחר משקול עם TF-IDF

10.3. מודלי חיזוי

בשלב החיזוי, פיתחנו פונקציה המיישמת תחזיות באמצעות מספר מודלים קלאסיים של למידת מכונה מספריית Scikit-learn, כולל SVM, יער רנדומלי, GBM, ו-LGBM. בתוך הפונקציה. בסופו של תהליך, הפונקציה מוציאה פלט בצורת טבלה המראה את תוצאות החיזוי של כל מודל, המוערכות על פי מדדי דיוק שונים, ומספקת השוואה בין ביצועי המודלים השונים.

בנוסף, בנינו פונקציה לחיזוי באמצעות רשת נוירונים פשוטה (ANN), שנבחרה בהתייעצות עם המנחה האקדמי שמתמחה בתחום. רשת הנוירונים כוללת שכבה חבויה אחת, והיא מיושמת באמצעות PyTorch. תהליך בניית הרשת כולל הכנת הנתונים, חלוקתם לסטי אימון ובדיקה, והגדרת הארכיטקטורה של הרשת. במהלך האימון, הקוד משתמש בפונקציות מטרה כמו MSELoss ו-L1Loss כדי לעקוב אחר שגיאות החיזוי ולבצע התאמה דינמית של שיעור הלמידה לאורך הepoch. בסיום תהליך האימון, הרשת מוערכת על סט הבדיקה כדי לבדוק את דיוקה, ומספקת תובנות לגבי יעילותה בחיזוי ערכים רציפים.

10.4. אימות ובחירת מודל אופטימלי

בשלב האימות ובחירת המודל האופטימלי, ביצענו מספר בדיקות כדי להבטיח שהמודלים שנבחרו יהיו מדויקים ורלוונטיים לתהליך החיזוי. השתמשנו באופטימיזציה של היפר-פרמטרים באמצעות Grid Search, שהייתה הכרחית למציאת השילוב האופטימלי של הפרמטרים עבור כל מודל. כדי לוודא שהמודלים מתפקדים היטב ולא סובלים מהטיה או בעיות אחרות, ביצענו אימות צולב (cross-validation) במהלך תהליך האימון. גישה זו אפשרה לנו לבדוק את היכולת של המודלים הכלליים לפעול גם על נתונים חדשים ולא רק על סט הנתונים שאיתו אומנו.

כחלק מהניסיון לשפר את יעילות התהליך, בחנו גם את האפשרות לבצע הורדת מימדים באמצעות Kernel PCA. המטרה הייתה לבדוק אם ניתן להפחית את מורכבות הנתונים ולהשיג חיזוי מדויק יותר מבחינה חישובית. עם זאת, תוצאות ה-PCA לא הראו שיפור משמעותי בדיוק המודל (איור 33), ולכן לאחר התייעצות עם נציגי אמדוקס והמנחה האקדמי, החלטנו לשמור על המימדים המקוריים. הסיבה לכך היא שמדובר בתהליך ניסיוני, וייתכן שקשרים בין מימדים מסוימים יתגלו רק בהמשך. בנוסף, גם החברה וגם המנחה האקדמי הסכימו שדיוק המודל הוא הקריטריון החשוב ביותר בשלב זה, והפחתת מורכבות חישובית אינה מהווה גורם מכריע בהחלטה. בהתאם לכך, בחרנו לא להשתמש בהורדת המימדים באמצעות ה-PCA ולא לוותר על מידע שעשוי להיות רלוונטי להמשך הפיתוח והדיוק של המודלים.

10.5. מדדי הערכה

המודלים נמדדו באמצעות מדדי הדיוק MSE, MAE ו-MAPE, והשווינו את התוצאות כדי לבחור את המודל האופטימלי. עבור המודלים SVM, Random Forest, GBM ו-LGBM, התוצאות היו מגוונות: ה-MSE המיטבי נמדד במודל GBM עם ערך של 373.07, ה-MAE המיטבי נמדד במודל SVM עם ערך של 14.41, וה-MAPE המיטבי נמדד גם במודל SVM עם ערך של 115.99.

	MSE	MAE	MAPE
Support Vector Machine	392.279992	14.393710	116.401526
Random Forest	373.308442	14.778081	141.197094
Gradient Boosting Machine	373.073678	14.751310	140.298542
LightGBM	374.139495	14.822696	142.040736

איור 28- מדדי דיוק למודלים הנ"ל

בהשוואה, רשת הנוירונים סיפקה את התוצאות הבאות: MSE של 365.05 ו-MAE של 14.62.

```
Epoch [94/100], Learning Rate: 0.000005, Test MSE: 365.0489, Test MAE: 14.6152
Epoch [95/100], Learning Rate: 0.000005, Test MSE: 365.0536, Test MAE: 14.6157
Epoch [96/100], Learning Rate: 0.000005, Test MSE: 365.0482, Test MAE: 14.6155
Epoch [97/100], Learning Rate: 0.000005, Test MSE: 365.0511, Test MAE: 14.6162
Epoch [98/100], Learning Rate: 0.000005, Test MSE: 365.0455, Test MAE: 14.6154
Epoch [99/100], Learning Rate: 0.000005, Test MSE: 365.0526, Test MAE: 14.6158
Epoch [100/100], Learning Rate: 0.000005, Test MSE: 365.0506, Test MAE: 14.6158
(365.0506354437934, 14.615813785129124)
```

איור 29 – תוצאות מדדי הדיוק של 5 ה-epoch האחרונים במודל ANN

מכיוון שה-MSE המיטבי נמדד ברשת הנוירונים, החלטנו לבחור במודל זה בתור המודל האופטימלי, היות והוא הניב את תוצאות הדיוק הטובות ביותר לפי מדד ה-MSE .

10.6. השוואת התוצאות אל מול מודל חיזוי נאיבי

כדי להצדיק את השימוש במודלים מתקדמים של Machine Learning לחיזוי משך פרויקט, ביצענו השוואה בין המודל המיטבי הנבחר, רשת הנוירונים (ANN), לבין מודל פשוט יותר, רגרסיה לינארית רבת משתנים. ההשוואה התבצעה בהתאם לשיטה המוצגת במאמר [5], אשר משווה בין מודל נאיבי למודל מתקדם. תוצאות החיזוי באמצעות הרגרסיה הליניארית הפשוטה הראו MSE של 431.26, MAE של 15.49, ו-MAPE של 144.57. מדדי הדיוק שנמדדו לא השתוו לאלו שהושגו באמצעות רשת הנוירונים, מה שמחזיר את המסקנה שהרשת הנוירונים מספקת תוצאות מדויקות יותר. בהתאם למאמר [5], **רשת הנוירונים נשארה המודל האופטימלי לחיזוי המשך הפרויקט.**

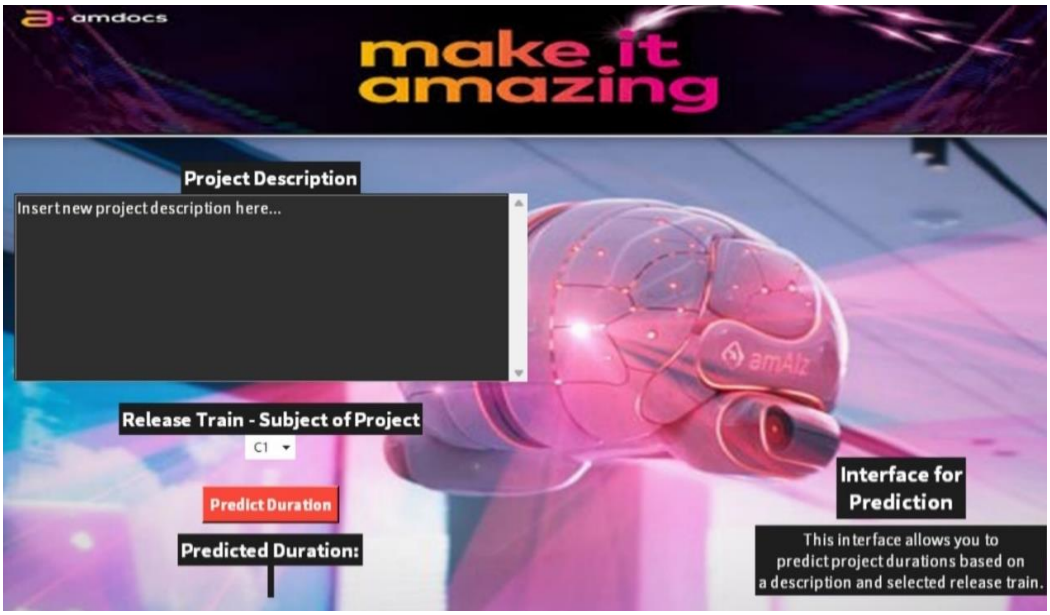
	MSE	MAE	MAPE
0	431.268304	15.495144	144.576691

איור 30 – תוצאות מדדי הדיוק של המודל הנאיבי

10.7. פיתוח ממשק GUI

כדי לעודד שימוש של המוצר בקרב מנהלי המוצר בAmdocs ובעקבות נכונות להמשך פיתוח המוצר לאחר סיום הפרויקט, פיתחנו ממשק GUI פשוט המאפשר חיזוי פרויקטים בצורה אינטואיטיבית ויעילה. הממשק עובד בכמה שלבים פשוטים:

1. הזנת תיאור הפרויקט: המשתמש מכניס לתיבה המיועדת לכך תיאור חופשי של הפרויקט שברצונו לבצע עבורו חיזוי (איור 31).



איור 31-המסך הראשי של הממשק

2. בחירת נושא הפרויקט: המשתמש בוחר אחת מתוך שש האפשרויות הקיימות עבור עמודת Release Train, המייצגת את נושא הפרויקט (איור 32).

איור 32-בחירה Subject והכנסה תיאור לממשק

3. ביצוע החיזוי: המשתמש לוחץ על כפתור "Predict Duration", שמבצע את פעולת החיזוי על פי הנתונים שהוזנו.
4. קבלת התוצאה: הממשק מספק את משך הפרויקט הצפוי ביחידות של ימים כפלט סופי (איור 33).

איור 33-תוצאה משך זמן פרויקט

11. הערכת הפיתרון

בשלב הערכת הפתרון, ביצענו הערכה של תוצאות מדדי הדיוק של כל המודלים והשווינו את התוצאות על מנת לבחור את המודל האופטימלי.

11.1. הערכה על פי מדדים כמותיים

עבור המודלים SVM, Random Forest, GBM ו-LGBM, מדדי הדיוק הצביעו על תוצאות שונות: ה-MSE המיטבי נמדד במודל GBM עם ערך של 373.07, ה-MAE המיטבי נמדד במודל SVM עם ערך של 14.41, וה-MAPE המיטבי נמדד גם במודל SVM עם ערך של 115.99. בהשוואה, רשת הנוירונים סיפקה MSE של 365.05 ו-MAE של 14.62.

לאור התוצאה הטובה ביותר שהתקבלה במדד ה- MSE ובהתאמה לתוצאות מאמר [5] ומאמר [1] **בחרנו במודל רשת הנוירונים בתור המודל האופטימלי.** מדד ה-MSE (Mean Squared Error) נחשב למתאים לסוג הנתונים שלנו לאחר סינון הערכים החריגים מאוד. היתרון המרכזי של ה-MSE הוא בכך שהוא מעניק משקל גבוה יותר לטעויות גדולות, מה שמסייע בהערכת דיוק החיזוי כאשר נוכחים ערכים חריגים שנועדו להיות מסוננים. בהתחשב בכך שהנתונים שלנו עברו סינון מוקדם לערכים חריגים בהתאם לדרישות חברת Amdocs, מדד ה-MSE אינו מוטה על ידי דגימות גדולות מדי ואינו נותן להן משקל קיצוני. בנוסף, ניתן להגיד שבראייה קדימה כשמערך הנתונים יגדל מודל מסוג NN יתמודד בצורה מיטבית עם מערכי נתונים גדולים בעלי מימדים רבים.

11.2. ניסיון להורדת מימדים

במהלך הניסיונות לשפר את יעילות התהליך, בחנו את השימוש ב-Kernel PCA להורדת מימדים כדי להפחית את מורכבות הנתונים ולהשיג חיזוי מדויק יותר. עם זאת, התוצאות לא הצביעו על שיפור משמעותי בדיוק המודל. לאחר התייעצות, הוחלט לשמור על המימדים המקוריים, מכיוון שדיוק המודל נחשב לקריטריון החשוב ביותר בשלב זה, והפחתת מורכבות חישובית אינה מהווה גורם מכריע.

	MSE	MAE	MAPE
Support Vector Machine	394.044457	14.406592	115.990641
Random Forest	374.010550	14.834917	142.423319
Gradient Boosting Machine	373.073678	14.751310	140.298542
LightGBM	386.596388	15.284840	149.032582

איור 34 – מדדי דיוק לאחר הורדת מימדים באמצעות Kernel PCA

11.3. השוואת המצב אליו הוביל השימוש בממשק למצב הקיים

בהשוואה למצב הקיים, שבו ה-MSE המקורי נמדד (איור 25), השימוש ברשת הנוירונים הפחית את ה-MSE בכ-83%. תוצאה זו מצביעה על שיפור משמעותי בחיזוי, בדומה לתוצאות שנמדדו במאמרים [5] ו-[1]. לכן, נוכל לקבוע שהפתרון שלנו שיפר בצורה משמעותית את החיזוי והשיג את מטרת הפרויקט.

```
[45]: data_new['mse_Amdocs'] = (data_new['Original Estimate (Days)']-data_new['Actual Estimation(d)'])*
      data_new['mse_Amdocs'].mean()

[45]: 2182.0416541838135
```

איור 35 – MSE לפני החלת הפתרון.

11.4. הטמעת הפתרון ומשוב מהארגון

בשל גודלה של Amdocs, הטמעה והערכת ההשפעה של ממשק חדש עשויה להיות מאתגרת. לכן, הצגנו את הממשק בפגישה עם קבוצות מהנדסים ומנהלי מוצר מהחברה. התגובות מהגורמים המקצועיים היו חיוביות מאוד. מעבר לדיוק של הממשק בחיזוי משך פיתוח פרויקט, המהנדסים הביעו התלהבות רבה מהעיצוב שלו, והקבוצה הסכימה פה אחד שהממשק פשוט לשימוש וקל להבנה גם עבור משתמשים שאינם טכנולוגיים. כתוצאה מכך, הוסכם כי יהיה קל להטמיע את הממשק כמוצר ניסיוני ולהמשיך לפתח אותו בהתאם לצרכים העתידיים. החברה אף הביעה נכונות להמשיך ולחקור את השימוש במוצר כדי לפתח בעתיד ממשק ייעודי שיוכל לשמש את כל עובדי החברה. מאחר שהחיזוי נאמד ב"Man Days", ניתן להשתמש בממשק לחיזוי פרויקטים מכל סוג (תגובה רשמית ומלאה של אמדוקס מצורפת בנספח 4).

11.5. הצעות לשיפור

בפגישה עם המהנדסים עלו מספר הצעות לשיפור הממשק כדי לשפר את תהליך החיזוי ולהפוך את השימוש בממשק לאינטואיטיבי ונגיש יותר. ההצעות נועדו לייעל את פעולת החיזוי ולסייע בהטמעת הממשק ככלי מרכזי באמדוקס.

1. **בניית מאגר מילים מובילות:**

הוצע לבנות מאגר מילים מובילות עבור כל סוג פרויקט העמודה 'Release Train' , אשר יוכל לייעל את תהליך החיזוי ולהפחית את עומס החישוב. באמצעות מאגר זה, ניתן יהיה לדייק את משקול המילים ולהתאים את החיזוי לסוג הפרויקט הרלוונטי.

2. **עיבוי מערך הנתונים לאימון המודל:**

כדי לאפיין מאגר מילים אפקטיבי, נדרש מערך נתוני אימון רחב יותר. מערך נתונים גדול יותר יעזור לדייק את החיזוי ולהגדיל את מובהקות התוצאה. בנוסף, יהיה ניתן להעמיק את הקשרים בין המאפיינים המשפיעים על משך פרויקטים.

12. דיון ומסקנות

הפרויקט התמקד בחיזוי משך פיתוח פרויקטים בחברת Amdocs, במטרה לשפר את דיוק החיזוי בלפחות 30%, לפי דרישת החברה, כאשר האמידה בוצעה לפי Man Days. לצורך כך, נעשה שימוש ב-NLP לתרגום תיאורי פרויקטים ממערכת Jira ולחיזוי משך הפרויקט על בסיס מילים מובילות בכל קטגוריית פרויקטים. במהלך הפרויקט, יישמנו שיטות Machine Learning ובחרנו ברשת נוירונים (ANN) כשיטה האופטימלית לחיזוי. את מודל החיזוי שילבנו בתוך ממשק ייעודי לProduct Managers בחברה, תוך דגש על נגישות וקלות הפעלה כדי להקל על תהליך ההטמעה. בפרק זה, נציג את המסקנות מהפרויקט ואת ההמלצות להמשך עבור חברת Amdocs בהתאמה לממצאים שהתקבלו.

12.1. ניתוח ממצאי הפרויקט והשלכות

לאחר השלמת הפרויקט והערכת הפתרון, ניתוח הממצאים מראה כי השימוש במודל חיזוי מבוסס רשתות נוירונים (ANN) תרם לשיפור של כ-80% בדיוק בהערכת משך הפרויקט. השימוש בשיטות Machine Learning לשיפור הדיוק התברר כיעיל ביותר, עם פוטנציאל לחסוך לחברה משאבים רבים, לצמצם צוואר בקבוק בתהליכי פיתוח ולהגביר את ערכה בעיני הלקוחות, בכך שהיא תיתפס כחברה איכותית ומהימנה יותר. הממשק שבנינו שומר על אופן שימוש דומה לזה של ה-Product Managers במערכת, מה שמבטיח קלות ונוחות בהטמעתו. בנוסף, מימשנו את הבקשות והדגשים שקיבלנו במהלך הראיונות עם המנהלים והמהנדסים בחברה, ולכן, לאור התגובות החיוביות, ניתן להניח שישנו פוטנציאל להמשך פיתוח והטמעה של הממשק בעתיד.

12.2. המלצות לארגון

על בסיס התוצאות ניתן לשקול מספר המלצות שהארגון יכול לשקול בהמשך עם הממשק שפיתחנו :

1. לעבות את מערך נתוני האימון:

עיבוי מערך נתוני האימון יכול לשפר את הדיוק של המודל על ידי הגברת המגוון והייצוגיות של הנתונים. בעזרת נתונים מגוונים יותר, המודל יוכל ללמוד דפוסים נוספים ולהתמודד טוב יותר עם מצבי קיצון.

2. לאפיין מאגר מילים מוכרות וחשובות לכל סוג פרויקט:

בניית מאגר מילים מובילות עבור כל סוג פרויקט תייעל את תהליך החיזוי, תקטין את עומס החישוב, ותשפר את דיוק המשקלים שניתנים למילים. מהלך זה יאפשר התאמה מדויקת יותר של החיזוי לאופי הפרויקט הרלוונטי.

3. הטמעת הממשק במערכות קיימות לקבוצה ניסיונית:

מומלץ לבחור קבוצת משתמשים מצומצמת שתבחן את השימוש בממשק החדש, כדי להעריך את היתרונות והאתגרים בשימוש בו. תוצאות הניסוי יכולות לסייע בקבלת החלטה לגבי המשך פיתוח והטמעה רחבה יותר בארגון.

12.3. תובנות ולקחים

- **עבודת צוות ושיתוף פעולה:**

הפרויקט הצריך עבודה משותפת של שלושה חברי קבוצה, וכל אחד הביא עימו סט כישורים וניסיון ייחודיים. למדנו כמה חשוב שיתוף פעולה אפקטיבי, תקשורת פתוחה וחלוקת תפקידים ברורה להצלחת הפרויקט. חילקנו את המשימות לפי תחומי החוזקה של כל אחד מאיתנו, מה שאפשר לנו לעבוד ביעילות ולהשלים את הפרויקט בצורה מיטבית.

- **תכנון וזמן ניהול:**

אחד האתגרים הגדולים בפרויקט היה ניהול הזמן והתכנון לטווח הארוך. למדנו עד כמה חשוב להקצות מספיק זמן לשלבים המוקדמים של איסוף דרישות, תכנון ולימוד הנושא, כדי להימנע מבעיות בלוחות הזמנים בהמשך. תכנון נכון בתחילת הפרויקט עזר לנו להישאר ממוקדים ולעמוד בזמנים שנקבעו.

- **עבודה עם גורמים חיצוניים:**

במהלך הפרויקט עבדנו בשיתוף פעולה עם צוותי מקצועיים באמדוקס. למדנו כיצד לשלב בין הדרישות והציפיות של הלקוח לבין המטרות האקדמיות שלנו. יצירת קשרים מקצועיים ושמירה על דיאלוג פתוח ומתמשך עם הגורמים החיצוניים הייתה חשובה להצלחת הפרויקט.

- **למידה מתמשכת וגמישות:**

הפרויקט דרש מאיתנו ללמוד נושאים חדשים ולעדכן את הידע שלנו לאורך הדרך. התמודדנו עם אתגרים שלא צפינו, מה שדרש מאיתנו להיות גמישים ולהתאים את הגישה שלנו תוך כדי תנועה. למדנו שהיכולת ללמוד במהירות ולהסתגל לשינויים היא חיונית להצלחת פרויקט מורכב.

- **השפעתו של פרויקט על הארגון:**

העבודה על הפרויקט באמדוקס גרמה לנו להבין את ההשפעה הפוטנציאלית של המודל שפיתחנו על הארגון. הבנו עד כמה כלי ניתוח מבוססי נתונים יכולים לתרום לשיפור תהליכים ולייעול העבודה בחברות גדולות, ומהי החשיבות של עבודה על פרויקטים שיכולים להשפיע באופן משמעותי על תפקוד הארגון.

12.4. תרומת חברי הפרויקט

במהלך הפרויקט עבדנו בשיתוף פעולה כך שפיזור המשימות התחלק בצורה יחסית אחידה מה שאפשר לנו לחלק בנינו משימות אך גם לעבוד יחדיו בכל שלבי הפרויקט. כל חברי הקבוצה נגעו בכל שלבי הפרויקט החל מהאפיון והתכנון ועד לבניית המודל כולל גם כתיבת הדו״ח הסופי כאשר לכל אחד היה תחום אחריות משלו בחלק מהדברים.

לאורך כל הפרויקט נעזרנו במנחה האקדמי ד״ר אלירן שרצר, בנוסף הגענו מספר פעמים במהלך הסמסטר למפגשים עם ד״ר אילן לאופר כדי לדייק את הכתיבה האקדמית ואת הפרויקט.

לינוי מדלסי:

במהלך הפרויקט לינוי לקחה חלק באיסוף הדרישות, אפיון המערכת, מחקר נרחב על חברת אמדוקס ועל הצורך שלה בבניית מודל לחיזוי משך הפרויקט ובהבנת האילוצים שמשליכים על הפרויקט, כל אלו בוצעו ע״י מפגשים משותפים של חברי הקבוצה עם הצוות המקצועי מטעם אמדוקס וריאיון שלהם, שיחות עם מספר עובדי חברה רלוונטיים ומחקר נרחב באינטרנט.

בכתיבת הדו״ח הסופי הייתה אחראית על הפרקים של תקציר מנהלים, תיאור הארגון והגדרת הבעיה, וגם על תמלול הריאיון שעשינו עם הצוות המקצועי של אמדוקס. היא השתתפה בכתיבת סקר הספרות ומצאה חלק מהמאמרים הרלוונטיים. כמו כן, לקחה גם חלק באסתטיקה של הדו״ח הסופי ובחלקים נוספים בדוח.

בחלק של בניית המודל והעבודה על הנתונים לינוי לקחה חלק בכתיבת הקוד – בהכנת שלבי עיבוד הדאטה המקדים, שיפור הדיוק והערכת המודלים לחיזוי.

לינוי הייתה אחראית על בניית ממשק הGUI שהוא בעצם השלב הסופי שחתם את פיתוח המודל שלנו ובכך הביאה לידי ביטוי את העבודה הקשה שלנו לאורך הפרויקט באמצעות בניית ממשק גרפי למשתמש שבו מתבצע החיזוי.

נועם סיטבון:

נועם מילא תפקיד מרכזי הן בפיתוח הטכנולוגי והן בתיעוד. מבחינה טכנית ,נועם פיתח אלגוריתם עיבוד שפה טבעית (NLP), הטמיע את שיטת TF-IDF כדי להקצות משקולות למילים, תוך המרת נתוני טקסט גולמיים לתכונות מספריות. בנוסף ,נועם פיתח מודל רגרסיה לינארית נאיבי כמודל בסיסי לחיזוי , ששימש כנקודת ייחוס להשוואה עם מודלים מורכבים יותר.

בפן התיעודי ,נועם הגדיר את מטרת הפרויקט ,תוך הדגשת הצורך בהערכת מאמץ מדויקת יותר .הוא תיאר את המצב הקיים ואת האתגרים הקשורים בשיטות הקיימות .כמו כן ,נועם השווה בין גישות שונות ,לרבות מודלים פשוטים ומורכבים של למידת מכונה . העבודה המקיפה של נועם סיפקה בסיס מוצק לפרויקט ,הכולל יעדים ברורים ותיעוד מפורט .

בנוסף, נועם היה אחראי בדוח על הפרקים מטרות הפרויקט, תיאור הנדסי של המצב הקיים, הצגת חלופות ותרשים הגאנט.

עמית פלאח:

תפקידו של עמית בפרויקט היה ריכוז העבודה מבחינת חלוקת משימות בפיתוח המודל והממשק ובנוסף להיות הגורם המקשר בין חברי הפרויקט לבין Data Scientists שליוו את התהליך.

בפרק העיבוד המקדים תפקידו של עמית לחקור את התפלגות המשתנה המוסבר ולנסות מגוון בדיקות סטטיסטיות כדי לאפיין קשרים מובהקים בין משתנים מסבירים. בנוסף, מעבר לתפקיד שהוגדר עמית ניסה עוד מגוון שיטות לזיהוי אשכולות שלא צלחו בגלל דלילות הנתונים.

לאחר מכן, בשלב ייצוא הנתונים, עמית היה נציג הקבוצה בדיונים מול החברה ומול המנחה האקדמי בקשר לקביעת מספר המילים המובילות שישמשו לחיזוי מכל קטגוריית פרויקט ובהמשך לבחירת שיטת משקול אופטימלית.

בשלב החיזוי עמית היה אחראי לכתיבת הפונקציות של המודלים (כולל רשתות הנוירונים), כתיבת הפונקציות כללה מחקר מקדים על המודלים, על שיטות האימות ובחירת ההיפר-פרמטרים. בהמשך עמית גם היה אחראי על האופטימיזציה של המודלים ועל הניסיונות לשיפור התוצאות (לדוג' Karnel PCA).

בשלב כתיבת הדוח עמית היה אחראי על חלוקת המשימות. הפרקים אותם עמית כתב בפועל הם – מתודולוגיה, סקר ספרות, מימוש הפתרון, הערכת הפתרון ודיון ומסקנות.

13. ביבליוגרפיה

[1] Lishner, Itai, and Avraham Shtub. "Using an artificial neural network for improving the prediction of project duration." *Mathematics* 10.22 (2022): 4189.

[2] Liu, Fei Tony, Kai Ming Ting, and Zhi-Hua Zhou. "Isolation-based anomaly detection." *ACM Transactions on Knowledge Discovery from Data (TKDD)* 6.1 (2012): 1-39.

[3] Lauble, S., Zielke, P., Chen, H. & Haghsheno, S. (2023). "Process analysis with an automatic mapping of performance factors using natural language processing (NLP)." *Proceedings of the 31st Annual Conference of the International Group for Lean Construction* (IGLC31), 59–68.

[4] Das, Mamata, and P. J. A. Alphonse. "A comparative study on tf-idf feature weighting method and its analysis using unstructured dataset." *arXiv preprint arXiv: 2308.04037* (2023).

[5] Petruseva, Silvana, Vahida Zujo, and Valentina Zileska-Pancovska. "Neural network prediction model for construction project duration." *International Journal of Engineering Research & Technology (IJERT)* 2.11 (2013): 1646-1654.

14. נספחים

14.1. נספח 1 – קטעי קוד וצילומים מהמחברת

14.1.1.התפלגות זמנים

קוד המתאר את התפלגות משך זמן חזוי (Original Estimate) ומשך זמן בפועל (Actual Estimate).

```
plt.figure(figsize=(12, 6))


plt.subplot(1, 2, 1)
sns.histplot(data_new['Original Estimate (Days)'], bins=20, kde=True)
plt.title('Distribution of Original Estimate (Days)')
plt.xlabel('Days')
plt.ylabel('Frequency')

plt.subplot(1, 2, 2)
sns.histplot(data_new['Actual Estimation(d)'], bins=20, kde=True)
plt.title('Distribution of Actual Estimate (Days)')
plt.xlabel('Days')
plt.ylabel('Frequency')

plt.tight_layout()
plt.show()
✓ 1.0s
```

14.1.2.טבלה סטטיסטית של הזמנים לפי קטגוריה

קוד התפלגות משך זמן <u>בפועל</u> (Actual Estimate) לפי קטגוריה וטבלה של תיאור סטטיסטי של הנתונים עבור כל קטגוריה.

```
def analyze_column_by_category(data, column_name, category_column='Release Train'):
    grouped = data.groupby(category_column)
    descriptions = {}
    num_categories = len(grouped)
    num_cols = 3
    num_rows = math.ceil(num_categories / num_cols)
    fig, axes = plt.subplots(num_rows, num_cols, figsize=(num_cols * 6, num_rows * 6), sharex=True)
    axes = axes.flatten()

    for ax, (category, group_data) in zip(axes, grouped):
        category_data = group_data[column_name]
        descriptions[category] = category_data.describe()
        sns.histplot(category_data, kde=True, bins=20, ax=ax)
        ax.set_title(f'Histogram of {column_name} for Category: {category}')
        ax.set_xlabel(column_name)
        ax.set_ylabel('Frequency')

    for i in range(num_categories, len(axes)):
        fig.delaxes(axes[i])
    plt.tight_layout()
    plt.show()
    description_df = pd.DataFrame(descriptions).T
    print(description_df)
✓ 0.0s
```

14.1.3.ספריות

ספריות בהן השתמשנו במהלך הפרויקט.

```
import pandas as pd
import math
import regex as re
import matplotlib.pyplot as plt
import seaborn as sns
import numpy as np
import nltk
from nltk.tokenize import word_tokenize
from nltk.corpus import stopwords
from nltk.probability import FreqDist
from nltk.tokenize import RegexpTokenizer
from nltk.stem import WordNetLemmatizer
from sklearn.feature_extraction import text
from sklearn.feature_extraction.text import TfidfVectorizer
import re
from nltk.stem import PorterStemmer
from sklearn.cluster import KMeans
from scipy import stats
from scipy.stats import norm, lognorm, expon,poisson
from scipy.stats import kstest
from sklearn.ensemble import IsolationForest
from pyod.models.cblof import CBLOF
from pyod.models.iforest import IForest
from pyod.models.ocsvm import OCSVM
from pyod.models.lof import LOF
from sklearn.model_selection import train_test_split,GridSearchCV
from sklearn.metrics import classification_report, confusion_matrix, accuracy_score,mean_absolute_error, mean_absolute_percentage_error,mean_squared_error, r2_score
from sklearn.model_selection import train_test_split
from sklearn.ensemble import RandomForestRegressor
from sklearn.ensemble import GradientBoostingRegressor
from sklearn.preprocessing import StandardScaler
from sklearn.svm import SVR
from xgboost import XGBRegressor
from lightgbm import LGBMRegressor
from sklearn.preprocessing import StandardScaler
from sklearn.decomposition import KernelPCA
import torch
import torch.nn as nn
import torch.optim as optim
from torch.utils.data import DataLoader, TensorDataset
```

14.1.4. Boxplot של הזמנים

קוד Boxplot המתאר את משך זמן בפועל לפי קטגוריה

```
pfig, axes = plt.subplots(1, 2, figsize=(18, 8))

sns.boxplot(ax=axes[0], x='Release Train', y='Actual Estimation(d)', data=data_new)
axes[0].set_title('Distribution of Actual Estimate (Days) by Release Train')
axes[0].tick_params(axis='x', rotation=45)


sns.boxplot(ax=axes[1], x='Release Train', y='Original Estimate (Days)', data=data_new)
axes[1].set_title('Distribution of Original Estimate (Days) by Release Train')
axes[1].tick_params(axis='x', rotation=45)

plt.tight_layout()
plt.show()
✓ 0.7s
```

14.1.5. טעויות במצב הקיים

קוד התפלגות הטעויות במצב הקיים

```
from scipy.stats import gaussian_kde
plt.hist(y_test-y_actual,alpha=0.5, color='blue', edgecolor='black', density=True)
plt.title("Frequency of errors")
kde = gaussian_kde(y_test-y_actual)
x_values = np.linspace(min(y_test-y_actual), max(y_test-y_actual), 100)
plt.plot(x_values, kde(x_values), color='red', linewidth=2)
✓ 0.1s
```

14.1.6. טבלה סטטיסטית של הטעויות במצב הקיים

קוד טבלה תיאור סטטיסטי של הטעויות המקוריות

```
data_3=data_new
data_3["error"]=data_new["Original Estimate (Days)"]-data_new["Actual Estimation(d)"]
data_3.groupby("Release Train")["error"].describe()
✓ 0.0s
```

NLP Algorithm & TF-IDF .14.1.7

חלוקת תיאור הפרויקט מהעמודה "Description" לפי שיטת NLP תוך שימוש במגוון מתודות. משקול התוצאות לפי IDF-TF של 30 המילים המובילות מכל קטגוריה מהעמודה "Release train".

```
ignore_words = set(['https', 'dd', 'OU', 'IoT', 'OM','http'])

stemmer = PorterStemmer()
lemmatizer = WordNetLemmatizer()

def preprocess_text(text):
    # Convert text to lowercase
    processed_text = text.lower()
    processed_text = re.sub(r'\b\d+\b|\b\w*\d\w*\b', '', processed_text)
    tokens = nltk.word_tokenize(processed_text)
    filtered_tokens = [token for token in tokens if token.lower() not in stopwords.words('english') and token.lower() not in ignore_words]
    stemmed_tokens = [stemmer.stem(lemmatizer.lemmatize(token)) for token in filtered_tokens]
    return ' '.join(stemmed_tokens)
```

```
tfidf_results = {}

for category in data_new['Release Train'].unique():
    category_df = data_new[data_new['Release Train'] == category]
    tfidf_vectorizer = TfidfVectorizer(preprocessor=preprocess_text)
    tfidf_matrix = tfidf_vectorizer.fit_transform(category_df['Description'])
    feature_names = tfidf_vectorizer.get_feature_names_out()
    tfidf_scores = tfidf_matrix.max(axis=0).toarray()[0]
    tfidf_results[category] = pd.DataFrame({'Word': feature_names, 'TF-IDF Score': tfidf_scores})

for category, result_df in tfidf_results.items():
    print(f"Status: {category}")
    print(result_df.sort_values(by='TF-IDF Score', ascending=False).head(30))
    print()

joblib.dump(tfidf_vectorizer, 'trained_vectorizers.pkl')
```

```
num_projects = len(data_new)
num_features = 30

feature_matrix = pd.DataFrame(index=range(num_projects))
for category, result_df in tfidf_results.items():
    top_words_df = result_df.sort_values(by='TF-IDF Score', ascending=False).head(num_features)
    top_words = top_words_df['Word'].tolist()
    for word in top_words:
        feature_matrix[f'{category}_{word}'] = np.nan

for idx, project_row in data_new.iterrows():
    category = project_row['Release Train']
    if category in tfidf_results:
        top_words_df = tfidf_results[category].sort_values(by='TF-IDF Score', ascending=False).head(num_features)
        top_words = top_words_df['Word'].tolist()
        top_scores = top_words_df['TF-IDF Score'].tolist()
        for word, score in zip(top_words, top_scores):
            feature_matrix.loc[idx, f'{category}_{word}'] = score

feature_matrix.fillna(0, inplace=True)
combined_data = pd.concat([feature_matrix, data_new['Actual Estimation(d)']], axis=1)
```

ML Model .14.1.8

פונקציית חיזוי המכילה את המודלים למידת מספריית Scikit-learn, כולל SVM, יער רנדומלי, GBM, ו-LGBM כולל בחירת היפר-פרמטרים ואימות צולב (לפני הורדת ממדים).

```
def evaluate_models(data, target_column):
    if not isinstance(data, pd.DataFrame):
        raise ValueError("Input data must be a pandas DataFrame.")
    if target_column not in data.columns:
        raise ValueError(f"Target column '{target_column}' not found in the DataFrame.")
    X = data.drop(columns=[target_column])
    y = data[target_column]
    X_train, X_test, y_train, y_test = train_test_split(X, y, test_size=0.2, random_state=42)
    models = {
        'Support Vector Machine': (SVR(kernel='rbf'), {
            'C': [0.1, 1, 10, 100],
            'gamma': [0.001, 0.01, 0.1, 1]
        }),
        'Random Forest': (RandomForestRegressor(), {
            'n_estimators': [100, 200, 500],
            'max_depth': [None, 10, 20, 30]
        }),
        'Gradient Boosting Machine': (GradientBoostingRegressor(), {
            'n_estimators': [100, 200, 500],
            'learning_rate': [0.01, 0.1, 0.2],
            'max_depth': [3, 5, 7]
        }),
        'LightGBM': (LGBMRegressor(verbose=-1), {
            'n_estimators': [100, 200, 500],
            'learning_rate': [0.01, 0.1, 0.2],
            'num_leaves': [31, 50, 100]
        })
    }
    results = {}
    for model_name, (model, param_grid) in models.items():
        grid_search = GridSearchCV(estimator=model, param_grid=param_grid, cv=5, scoring='neg_mean_squared_error')
        grid_search.fit(X_train, y_train)
        best_model = grid_search.best_estimator_
        y_pred = best_model.predict(X_test)
        mse = mean_squared_error(y_test, y_pred)
        mae = mean_absolute_error(y_test, y_pred)
        mape = np.mean(np.abs((y_test - y_pred) / y_test)) * 100
        results[model_name] = {'MSE': mse, 'MAE': mae, 'MAPE': mape}
    results_df = pd.DataFrame(results).T

    return results_df
```

```
results = evaluate_models(combined_data, 'Actual Estimation(d)')
results
```

PCA Model.14.1.9

פונקציית חיזוי המכילה את המודלים למידת מספריית Scikit-learn, כולל SVM, יער רנדומלי, GBM, ו-LGBM כולל בחירת היפר-פרמטרים ואימות צולב אחרי הורדת ממדים באמצעות Karnel PCA.

PCA Karnel

```
def evaluate_models(data, target_column):
    if not isinstance(data, pd.DataFrame):
        raise ValueError("Input data must be a pandas DataFrame")
    X = data.drop(columns=[target_column])
    y = data[target_column]
    scaler = StandardScaler()
    X_scaled = scaler.fit_transform(X)
    kpca = KernelPCA(kernel='rbf', n_components=20)
    X_kpca = kpca.fit_transform(X_scaled)
    X_train, X_test, y_train, y_test = train_test_split(X_kpca, y, test_size=0.2, random_state=42)
    models = {
        'Support Vector Machine': (SVR(kernel='rbf'), {
            'C': [0.1, 1, 10, 100],
            'gamma': [0.001, 0.01, 0.1, 1]
        }),
        'Random Forest': (RandomForestRegressor(), {
            'n_estimators': [100, 200, 500],
            'max_depth': [None, 10, 20, 30]
        }),
        'Gradient Boosting Machine': (GradientBoostingRegressor(), {
            'n_estimators': [100, 200, 500],
            'learning_rate': [0.01, 0.1, 0.2],
            'max_depth': [3, 5, 7]
        }),
        'LightGBM': (LGBMRegressor(verbose=-1), {
            'n_estimators': [100, 200, 500],
            'learning_rate': [0.01, 0.1, 0.2],
            'num_leaves': [31, 50, 100]
        })
    }

    results = {}
    for model_name, (model, param_grid) in models.items():
        grid_search = GridSearchCV(estimator=model, param_grid=param_grid, cv=5, scoring='neg_mean_squared_error')
        grid_search.fit(X_train, y_train)
        best_model = grid_search.best_estimator_
        y_pred = best_model.predict(X_test)
        mse = mean_squared_error(y_test, y_pred)
        mae = mean_absolute_error(y_test, y_pred)
        mape = np.mean(np.abs((y_test - y_pred) / y_test)) * 100
        results[model_name] = {'MSE': mse, 'MAE': mae, 'MAPE': mape}
    results_df = pd.DataFrame(results).T

    return results_df
```

```
results = evaluate_models(combined_data, 'Actual Estimation(d)')
results
```

14.1.10. ANN

תהליך מימוש מודל רשתות נוירונים (ANN) מספריית PyTorch.

חלק א -

```
# Prepare the data
X = combined_data.drop(columns=['Actual Estimation(d)'])
y = combined_data['Actual Estimation(d)']
X = np.array(X)
y = np.array(y)
X_train, X_test, Y_train, Y_test = train_test_split(X, y, test_size=0.2, random_state=42)

class SimpleNN(nn.Module):
    def __init__(self, input_size, hidden_size, output_size):
        super(SimpleNN, self).__init__()
        self.fc1 = nn.Linear(input_size, hidden_size)
        self.relu = nn.ReLU()
        self.fc2 = nn.Linear(hidden_size, output_size)

    def forward(self, x):
        out = self.fc1(x)
        out = self.relu(out)
        out = self.fc2(out)
        return out

def train_model(model, criterion_mse, criterion_mae, optimizer, scheduler, train_loader, num_epochs, test_loader):
    for epoch in range(num_epochs):
        model.train()
        for inputs, targets in train_loader:
            optimizer.zero_grad()
            outputs = model(inputs)
            loss_mse = criterion_mse(outputs, targets)
            loss_mse.backward()
            optimizer.step()
        scheduler.step()

        avg_test_loss_mse, avg_test_loss_mae = evaluate_model(model, test_loader, criterion_mse, criterion_mae)

        print(f'Epoch [{epoch+1}/{num_epochs}], Learning Rate: {scheduler.get_last_lr()[0]:.6f}, '
              f'Test MSE: {avg_test_loss_mse:.4f}, Test MAE: {avg_test_loss_mae:.4f}')

def evaluate_model(model, test_loader, criterion_mse, criterion_mae):
    model.eval()
    total_loss_mse = 0
    total_loss_mae = 0
    with torch.no_grad():
        for inputs, targets in test_loader:
            outputs = model(inputs)
            mse_loss = criterion_mse(outputs, targets)
            mae_loss = criterion_mae(outputs, targets)
            total_loss_mse += mse_loss.item()
            total_loss_mae += mae_loss.item()
    avg_loss_mse = total_loss_mse / len(test_loader)
    avg_loss_mae = total_loss_mae / len(test_loader)
    return avg_loss_mse, avg_loss_mae

# Convert data to torch tensors
X_train = torch.tensor(X_train, dtype=torch.float32)
Y_train = torch.tensor(Y_train, dtype=torch.float32).view(-1, 1)
X_test = torch.tensor(X_test, dtype=torch.float32)
Y_test = torch.tensor(Y_test, dtype=torch.float32).view(-1, 1)
```

חלק ב –

```
# Create datasets and loaders
train_dataset = TensorDataset(X_train, Y_train)
test_dataset = TensorDataset(X_test, Y_test)
train_loader = DataLoader(train_dataset, batch_size=32, shuffle=True)
test_loader = DataLoader(test_dataset, batch_size=32, shuffle=False)

# Define model, loss functions, and optimizer
input_size = X_train.shape[1]
hidden_size = 64
output_size = 1
num_epochs = 100
learning_rate = 0.005

model = SimpleNN(input_size, hidden_size, output_size)
criterion_mse = nn.MSELoss()
criterion_mae = nn.L1Loss()
optimizer = optim.Adam(model.parameters(), lr=learning_rate)

# Define learning rate scheduler
def lr_lambda(epoch):
    if epoch >= 75:
        return 0.1 ** 3
    elif epoch >= 50:
        return 0.1 ** 2
    elif epoch >= 25:
        return 0.1
    else:
        return 1

scheduler = optim.lr_scheduler.LambdaLR(optimizer, lr_lambda)

# Train the model
train_model(model, criterion_mse, criterion_mae, optimizer, scheduler, train_loader, num_epochs, test_loader)

# Final evaluation of the model
evaluate_model(model, test_loader, criterion_mse, criterion_mae)
torch.save(model.state_dict(), 'simple_nn_model.pth')
```

14.1.11. Linear Regression

פונקציה לחיזוי באמצעות מודל רגרסיה נאיבי להשוואה מול מודלים מעולם ה-Machine learning.

```
def simple_model(data,target_column):
    X = data.drop(columns=[target_column])
    y = data[target_column]
    X_train, X_test, y_train, y_test = train_test_split(X, y, test_size=0.2, random_state=42)
    model = LinearRegression()
    model.fit(X_train, y_train)
    y_pred = model.predict(X_test)
    mse = mean_squared_error(y_test, y_pred)
    mae = mean_absolute_error(y_test, y_pred)
    mape = np.mean(np.abs((y_test - y_pred) / y_test)) * 100
    errors=pd.DataFrame({"MSE":[mse],"MAE":[mae],"MAPE":[mape]})
    return errors
```

14.2. נספח 2 – ראיון עם מנהלי פרויקטים ובעלי תפקידים בAmdocs

נספח: ריאיון עם מנהלי פרויקטים בAmdocs - שלב אפיון המערכת

שאלה 1: מהן הבעיות המרכזיות שאתם נתקלים בהן כיום בניהול פרויקטים והקצאת משאבים?

תשובה: הבעיה המרכזית שאנחנו נתקלים בה היא חוסר דיוק בתחזיות משך הפרויקטים. ישנה סטייה משמעותית בין ההערכות הראשוניות לבין הביצוע בפועל, מה שמוביל לעיכובים, חריגות תקציביות, ותכנון לא מיטבי של משאבים. כמו כן, קיימים קשיים בהקצאת משאבים יעילה לפרויקטים מורכבים, מה שגורם לבזבוז משאבים ולעיתים גם לאי עמידה בלוחות הזמנים.

שאלה 2: איך לדעתכם ניתן לשפר את תהליך ניהול הפרויקטים בעזרת הכלי החדש?

תשובה: הכלי החדש, שמבוסס על טכניקות Machine Learning, יכול לספק תחזיות מדויקות יותר של משכי הפרויקטים, מה שיאפשר לנו לתכנן טוב יותר ולמזער את הסיכון לעיכובים. אם נצליח להפחית את השגיאה הממוצעת ב-MSE, MAE ו-MAPE בכ-30% לפחות, כפי שאנחנו מצפים, זה יוביל לשיפור משמעותי בתכנון ובהקצאת המשאבים.

שאלה 3: מהם האתגרים המרכזיים עם הנתונים שבהם אתם משתמשים היום?

תשובה: אחד האתגרים המשמעותיים שאנחנו מתמודדים איתם הוא הימצאותם של נתונים חריגים במערכות שלנו. לעיתים אנחנו לא מזינים את משך הפרויקט שנראה לנו מדויק או ישנם מקרים שבהם אנו נותנים הערכות חוזרות ומוגזמות או לא מספקות. לדוגמה, במקרים מסוימים אנחנו מעריכים את משך הפרויקט כהערכה יתרה, כדי להבטיח שלוחות הזמנים יהיו בטוחים, ובמקרים אחרים אנחנו מעריכים הערכת חוסר בגלל אילוצי זמן או לחץ לסיים את ההערכה מהר. נתונים אלו יכולים לגרום להטיה משמעותית בתוצאות החיזוי.

שאלה 4: איך אתם מציעים להתמודד עם הבעיה של נתונים חריגים בתהליך החיזוי?

תשובה: אנחנו מאמינים שהכלי החדש צריך להיות מסוגל לזהות ולהסיר את הנתונים החריגים כבר בשלב המוקדם של התהליך. חשוב לנו שתהיה בדיקה ראשונית של הנתונים כדי לאתר הערכות שאינן מדויקות או חריגות, ולהסיר אותן לפני תחילת תהליך החיזוי. רק כך נוכל להבטיח שהמודל עובד על נתונים אמינים, מה שיוביל לתחזיות מדויקות יותר. :

שאלה 5: מהם הפיצ'רים המרכזיים שעליהם תרצו להתמקד בתהליך החיזוי?

תשובה: עבורנו, הפיצ'רים המרכזיים שצריכים להיות בחיזוי הם ה-Original Estimate, שזה הערך המקורי שאנחנו מעריכים בתחילת הפרויקט, ה-Actual Estimate, שזה מה שהיה בפועל בסיום הפרויקט, ה-Release Train שמייצג את מחזור השחרור שבו הפרויקט מתבצע, ו-Description, שמתאר את הפרויקט עצמו. אנחנו מאמינים שפיצ'רים אלו הם החשובים ביותר להשגת תחזיות מדויקות.

שאלה 6: מה יהיו ההשפעות הישירות של המודל על עבודתכם היום-יומית?

תשובה: השפעה ישירה תהיה שיפור משמעותי ברמת הדיוק של התחזיות, מה שיאפשר לנו לתכנן פרויקטים באופן ריאלי ומבוסס יותר. זה יאפשר הקצאה נכונה יותר של משאבים והתאמה טובה יותר של לוחות זמנים למציאות, מה שיקטין את הסיכויים לעיכובים וחריגות. בנוסף, תחזיות אמינות יאפשרו לנו להתאים את עצמנו לשינויים בשוק בצורה מהירה ויעילה יותר.

שאלה 7: איך אתם מעריכים שהכלי ישפיע על הלקוחות שלכם?

תשובה: בעזרת המודל החדש, לקוחותינו ייהנו ממחזור חיים צפוי יותר של פרויקטים, עם פחות עיכובים בלתי צפויים, מה שיוביל לשיפור שביעות רצון הלקוחות. הארגון יוכל לספק פרויקטים בזמן ובאיכות גבוהה יותר, מה שיבנה אמון מול הלקוחות ויתרום לשימורם לאורך זמן.

שאלה 8: כיצד אתם רואים את השפעת הכלי על צמיחת הארגון בטווח הארוך?

תשובה: הכלי יאפשר לארגון להרחיב את היקף הפעילות שלו בצורה יציבה, תוך שמירה על איכות ולוחות זמנים. שיפור היעילות והדיוק בתחזיות יאפשר לנו לקחת על עצמנו יותר פרויקטים מורכבים ולגדול בצורה מבוקרת, מבלי לפגוע באיכות או לגרום לעיכובים.

שאלה 9: מהי החשיבות של הסרת נתונים חריגים בעיניכם לפני תחילת תהליך החיזוי?

תשובה: הסרת נתונים חריגים היא קריטית מבחינתנו. כפי שציינו, לעיתים אנחנו מזינים נתונים שאינם מדויקים, בין אם בגלל הערכות יתר או חוסר. נתונים אלו עלולים לעוות את התחזיות ולגרום לסטיות משמעותיות. לכן, אנחנו מבקשים שהכלי יבחן ויבטל נתונים חריגים כבר בשלב הראשון של התהליך, כדי להבטיח שהמודל יעבוד עם נתונים מדויקים ואמינים בלבד.

שאלה 10: האם אתם רואים צורך בהתאמות נוספות או דרישות מיוחדות מהכלי?

תשובה: כפי שציינו, חשוב לנו שהכלי יסיר מראש נתונים חריגים ויבדוק אותם לפני תחילת התהליך, כדי להבטיח תחזיות מדויקות יותר. בנוסף, הכלי צריך להיות גמיש מספיק כדי להתאים לסוגי פרויקטים שונים ולמורכבויות שונות. חשוב לנו שהוא יוכל לשמור על הדיוק שלו על פני תרחישים שונים ויוכל להשתפר ולהתעדכן בהתאם לשינויים שוטפים בשוק ובדרישות הלקוח.

הנספח מתאר את השיחה שנערכה עם מנהלי הפרויקטים באמדוקס, אשר חיזקה את ההבנה לגבי החשיבות הרבה של הדיוק בתחזיות, ניהול נתונים חריגים, ובחירת הפיצ'רים המרכזיים לתהליך החיזוי.

14.3. נספח 3 – קטעי קוד חשובים מממשק הGUI

14.3.1. ספריות ממשק

ספריות שהשתמשנו במימוש ממשק הGUI

```
import tkinter as tk
from tkinter import messagebox
from tkinter import ttk
from PIL import Image, ImageTk
import pandas as pd
import pywinstyles
```

14.3.2.פונקציית בנאי של ממשק

פונקציית בנאי של ממשק הGUI.

```
#Build the GUI
class ProjectDurationPredictorApp:
    def __init__(self, root):
        self.root = root
        self.root.title("Project Duration Predictor")
        self.root.geometry("1000x600")

        # Apply the font to all widgets
        default_font = ("Source Sans Pro", 12)
        self.root.option_add("*Font", default_font)

        # Load and display logo
        self.logo_frame = tk.Frame(self.root, bg="#1e1e1e", height=150)
        self.logo_frame.pack(fill="x")
        self.load_logo()

        # Create canvas for the rest of the UI
        self.canvas = tk.Canvas(self.root, width=1000, height=450)
        self.canvas.pack(fill="both", expand=True)
        self.load_background_image()

        # Create frames on top of the canvas
        self.right_frame = tk.Frame(self.canvas, bg="#f3f6f4")
        self.canvas.create_window(850, 300, window=self.right_frame, anchor="n", width=300, height=250)
        pywinstyles.set_opacity(self.right_frame, value=1, color="#f3f6f4")

        self.left_frame = tk.Frame(self.canvas, bg="#f3f6f4")
        self.canvas.create_window(260, 20, window=self.left_frame, anchor="n", width=480, height=420)
        pywinstyles.set_opacity(self.left_frame, value=1, color="#f3f6f4")

        # Add elements to the left frame
        self.add_left_frame_elements()

        # Add elements to the right frame
        self.add_right_frame_elements()
```

14.3.3.פונקציות ויזואליזציה 1

פונקציות להגדרת גודל תמונת הרקע והלוגו וטעינתן לממשק.

```
def load_logo(self):
    try:
        logo_image = Image.open("title.jpg")
        logo_image = logo_image.resize((1000, 130), Image.LANCZOS)
        logo_photo = ImageTk.PhotoImage(logo_image)
        logo_label = tk.Label(self.logo_frame, image=logo_photo, bg="#1e1e1e")
        logo_label.image = logo_photo
        logo_label.pack(fill="both", expand=True)
    except Exception as e:
        print(f"Failed to load logo image: {e}")

def load_background_image(self):
    try:
        bg_image = Image.open("reka brain.jpg")
        bg_image = bg_image.resize((1000, 480), Image.LANCZOS)
        bg_photo = ImageTk.PhotoImage(bg_image)
        self.canvas.create_image(0, 0, anchor="nw", image=bg_photo)
        self.canvas.lower("all")
        self.canvas.image = bg_photo
    except Exception as e:
        print(f"Failed to load background image: {e}")
```

14.3.4.פונקציה ויזואליזציה 2

פונקציה להגדרת אגף שמאל של הממשק.

```
def add_left_frame_elements(self):
    label_font = ("Source Sans Pro", 14, "bold")

    # Project Description
    tk.Label(self.left_frame, text="Project Description", bg="#1e1e1e", fg="#ffffff", font=label_font).pack(pady=(5, 0))

    # Adding Scrollbar to Text widget
    text_frame = tk.Frame(self.left_frame, bg="#2e2e2e")
    text_frame.pack(pady=(0, 15), fill="both", expand=True)
    self.description_text = tk.Text(text_frame, height=7, width=55, bd=2, relief="groove", bg="#2e2e2e", fg="#ffffff")
    self.description_text.pack(side="left", fill="both", expand=True)
    scrollbar = tk.Scrollbar(text_frame, command=self.description_text.yview, bg="#2e2e2e")
    scrollbar.pack(side="right", fill="y")
    self.description_text.config(yscrollcommand=scrollbar.set)

    # Placeholder text functionality
    self.description_text.insert("1.0", "Insert new project description here...")
    self.description_text.bind("<FocusIn>", self.clear_placeholder)
    self.description_text.bind("<FocusOut>", self.add_placeholder)

    # Release Train - Subject of Project
    tk.Label(self.left_frame, text="Release Train - Subject of Project", bg="#1e1e1e", fg="#ffffff", font=label_font).pack(pady=(5, 0))
    self.release_train_options = ["C1", "Care", "Case Management", "Commerce", "Experience Layer", "Order Handling"]
    self.release_train_var = tk.StringVar(self.left_frame)
    self.release_train_var.set(self.release_train_options[0])

    style = ttk.Style()
    style.configure('TMenubutton', background='#ffffff', foreground='#1e1e1e')
    self.release_train_menu = ttk.OptionMenu(self.left_frame, self.release_train_var, *self.release_train_options, style='TMenubutton')
    self.release_train_menu.pack(pady=(0, 20))

    # Predict Duration button
    tk.Button(self.left_frame, text="Predict Duration", command=self.predict_duration, bg="#fb473a", fg="white", font=("Source Sans Pro", 12, "bold"),
              bd=2, relief="raised").pack(pady=(5, 5))

    # Predicted Duration
    tk.Label(self.left_frame, text="Predicted Duration:", bg="#1e1e1e", fg="#ffffff", font=label_font).pack(pady=(5, 0))
    self.result_text = tk.Label(self.left_frame, text="", bg="#1e1e1e", fg="#ffffff", font=("Source Sans Pro", 14, "bold"))
    self.result_text.pack(pady=(0, 5))
```

14.3.5.פונקציות ויזואליזציה 3

פונקציה להגדרת אגף ימין של הממשק.

```
def add_right_frame_elements(self):
    tk.Label(self.right_frame, text="Interface for\nPrediction", bg="#1e1e1e", fg="#ffffff", font=("Source Sans Pro", 16, "bold")).pack(pady=3)
    tk.Label(self.right_frame, text="This interface allows you to\npredict project durations based on\na description and selected release train.",
             bg="#1e1e1e", fg="#ffffff").pack(pady=3)
```

14.3.6.פונקציות תיבת התיאור

פונקציות להגדרת ההוראות מעל תיבת התיאור.

```
def clear_placeholder(self, event):
    if self.description_text.get("1.0", "end-1c") == "Insert new feature description here...":
        self.description_text.delete("1.0", tk.END)
        self.description_text.config(fg="#ffffff")

def add_placeholder(self, event):
    if not self.description_text.get("1.0", "end-1c"):
        self.description_text.insert("1.0", "Insert new featur description here...")
        self.description_text.config(fg="#777777")
```

14.3.7.פונקציית חיזוי

פונקציה שלוקחת את תיאור הפרויקט שהוכנס לתיבה הייעודית ואת בחירת נושא הפרויקט ומבצעת חיזוי.

```
def predict_duration(self):
    description = self.description_text.get("1.0", tk.END).strip()
    release_train = self.release_train_var.get().strip()

    if not description or not release_train or description == "Insert new project description here...":
        messagebox.showwarning("Input Error", "Please provide both project description and release train.")
        return

    new_project_data = pd.DataFrame({
        'Description': [description],
        'Release Train': [release_train]
    })

    predicted_duration = process_and_predict(new_project_data)
    self.result_text.config(text=f"{predicted_duration[0][0]:.2f} days")
```

14.3.8.פונקציית אקטיבציה

קוד לאקטיבציית הממשק.

```
if __name__ == "__main__":
    root = tk.Tk()
    app = ProjectDurationPredictorApp(root)
    root.mainloop()
```

14.4. נספח 4 – משוב רשמי מחברת Amdocs

<u>משוב על עבודת הגמר של הסטודנטים</u>

ברצוננו להודות לכם, שלושת הסטודנטים: עמית לינוי ונועם, על העבודה המעמיקה והמקצועית שביצעתם במסגרת הפרויקט שלכם. פיתחתם כלי חדשני המתבסס על ניתוח שפה טבעית (NLP) ומודלים של חיזוי מבוססי Machine Learning, שמאפשר למנהלי מוצר באמדוקס לחזות באופן מדויק יותר את משך הפיתוח של מוצרים וממשקים הקשורים להם.

הכלי שפיתחתם מציב סטנדרט חדש עבור ניהול הפרויקטים בחברה, בכך שהוא מעניק למנהלי המוצר יכולת לקבוע את משך הפרויקט הצפוי בצורה מושכלת ומבוססת נתונים, תוך שילוב של התוצאה החזויה יחד עם הידע המקדים של המנהלים, ומשחרר אותם מהתלות במנהלי פיתוח לצורך הערכת משך הפרויקט.

אנו מתכננים להשיק את השימוש בכלי כפיילוט בכמה קבוצות על מנת לבחון את הצלחתו ואחרי כן להטמיע אותו ככלי עבודה לכלל הקבוצות.

השפעת הכלי צפויה במגוון היבטים:

שיפור בתכנון הפרויקטים: הכלי יאפשר למנהלי המוצר לבצע תכנון מדויק יותר של משך הפרויקטים, דבר שיוביל לתיאום ציפיות טוב יותר עם לקוחות ובעלי עניין.

ייעול תהליך קבלת ההחלטות: יכולת חיזוי מבוססת נתונים מסייעת למנהלים לקבל החלטות מושכלות ולשפר את ניהול המשאבים במהלך הפרויקט.

שיפור בתהליכי העבודה: הכלי מספק תובנות חדשות שמשפיעות על תהליכי העבודה, מה שמוביל לשיפור ביצועים לאורך זמן.

הצעות לשיפור:

שיפור אלגוריתמי החיזוי: מומלץ לשקול שימוש במודלים נוספים או שילוב טכניקות נוספות להעמקת הדיוק של החיזוי.

הרחבת המודל: כדאי לשקול הרחבה של הכלי כך שיוכל לנתח סוגי פרויקטים שונים ומגוונים מעבר לפרויקטים הקיימים במערכת Jira.

אינטגרציה עם מערכות נוספות: ניתן לחשוב על אינטגרציה של הכלי עם Jira כך שחווית המשתמש תהיה הוליסטית, מנהל המצור יכניס את התאור בJira ויקבל באותו ממשק את החיזוי למשך הפרויקט.

אנו צופים כי הפרויקט ישפר בצורה ניכרת את היכולת של הארגון לנהל פרויקטים בצורה מדויקת ויעילה יותר. היכולת לחזות משך פרויקטים מאפשרת להימנע מעיכובים בלתי צפויים ולייעל את תהליך קבלת ההחלטות, מה שמוביל לשיפור מתמשך בביצועים.

שוב, תודה על עבודתכם המסורה ועל התרומה שלכם להצלחת הארגון.

אורלי קרמס, מנהלת מוצר, אמדוקס.