דוח מסכם – פרויקט גמר קורס Data Science:

תקציר

הפרויקט שבחרנו מתעסק בתאונות דרכים, נושא רגיש וכואב לכלל האוכלוסייה. בימים אלו אנו עדים לבעיות תחבורה רבות הן במדינת ישראל והן בעולם כולו. במהלך ההבנה העסקית, התעניינו בשתי שאלות מחקר שונות. הראשונה, בעיית רגרסיה, בה רצינו לחזות את משך זמן התאונה. השנייה הינה בעיית סיווג, בה רצינו לסווג את חומרת התאונה. בחרנו תחילה בבעיית הרגרסיה, משום שהיא עניינה אותנו יותר. שאלה זו הצליחה לאתגר אותנו מאוד, עודדה אותנו לחקור לעומק את מודלי הרגרסיה השונים ולפתח יצירתיות. לאכזבתנו, כאשר הגענו לשלב ההערכה, קיבלנו תוצאות לא טובות שנבעו בין היתר מרעש בנתונים. לאחר ניסיונות ממושכים לשיפור והתייעצות עם מרצה הקורס, בחרנו לשנות את שאלת המחקר לסיווג חומרה. שינוי שאלת המחקר, גררה שינויים שונים בכל אחד משלבי ה-CrispDM. חלוקת העבודה בינינו הייתה שיוונית וממוקדת מטרה. אחד התמקד בחיפוש אחר מאמרים ומחקרי עבר, השני ספג מידע ורעיונות מאנליזות ועבודות שנעשו בנושא זה בKaggle. לאורך הפרויקט עצמו, כל אחד מחברי הצוות לקח חלק בכל אחד משלבי ה-CrispDM הן בשלב הביצועי והן בשלב הסקת המסקנות. בשלב ה-Modeling ו-Data Understanding, בצענו חלוקת עבודה בינינו למשימות של מימוש מודלים וביצוע אנליזות שונות. כתיבת הדוח הייתה משותפת ונעשתה לאורך כל העבודה על הפרויקט.

סקירה ספרותית

על פי מאמרם של Yuexu Zhao, Wei Dang (2021)-
Prediction in Traffic Accident Duration Based on Heterogeneous Ensemble Learning
שפורסם בכתב העת Applied Artificial Science עוסק במטלת רגרסיה כדי לחשב משך זמן תאונת דרכים בהתבסס על זמן, מיקום גאוגרפי, מזג אוויר ותנאי דרך. במאמר השתמשו בסט נתונים זהה לזה שאנו משתמשים, סט זה מכיל מידע על תאונות דרכים בארה"ב בין השנים 2016-2020.
במאמר הציעו אנסמבל של מודלי Boosting שונים. לטענת כותבי המאמר, האנסמבל שהרכיבו מניב ביצועים טובים יותר לעומת מודלים שבהם השתמשו בעבר (למשל Logistic Regression). בנוסף, נעשתה עבודה מקיפה על בנייה והוספה של פיצ'רים משמעותיים חדשים והשלמת ערכים חסרים של פיצ'רים קיימים באמצעות KNN ו Random Forest.
כותבי המאמר פירטו שלושה מדדים שונים בהם השתמשו על מנת להשוות את כל אחד מהמודלים לאנסמבל: MSE,MAE,MAPE והשתמשו בעץ החלטה סטנדרטי כ- Baseline. עץ ההחלטה הגיע לביצועים הנמוכים ביותר בכל המדדים שצוינו. מודל LightGBM הגיע לתוצאות הטובות ביותר מבין המודלים שמהם הורכב האנסמבל עם MAPE של 35.45%, MAE של 31.14 דקות ו MSE של 4,314 דקות. האנסמבל הגיע לתוצאות האופטימליות מבחינת כל המדדים: MAPE -35.6%, MAE

- 30.7 דקות וMSE של 4,252 דקות.

לסיכום, לאנסמבל שהרכיבו כותבי המאמר היו את הביצועים הטובים ביותר, אם כי רק במעט מעל LightGBM. יחד עם זאת, יש לציין כי ביצועי כלל מודלי הBoosting דומים, אך טובים בהרבה משל עץ ההחלטה הבסיסי. בנוסף, במאמר השתמשו בערכי SHAP על מנת להבחין בגודל ההשפעה של המאפיינים השונים על משימת הרגרסיה ונמצא כי למאפיינים של תאריך ומיקום משמעות רבה.

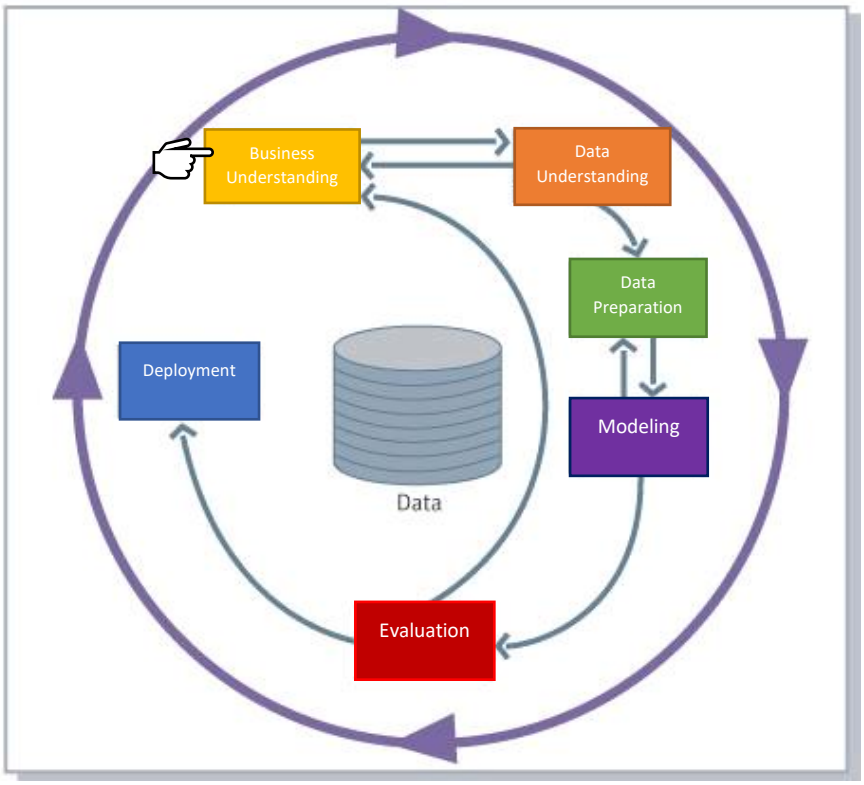
על פי מאמרם של Labib et. al (2019), Road Accident Analysis and Prediction of Accident Severity by Using Machine Learning in Bangladesh שפורסם בכתב העת International Conference on Smart Computing & Communications עוסק בניתוח תאונות דרכים בבנגלדש וסיווג של חומרת התאונה לארבע קטגוריות חומרה שונות, שזוהי גם שאלת המחקר שלנו. מאגר נתונים המכיל כ-43 אלף תאונות דרכים בין השנים 2001-2015 באמצעות שימוש בלמידת מכונה. השיטה בה פעלו החוקרים היא במימוש ארבעה מודלים בסיסים (KNN, Naïve Bayes, Decision Tree, AdaBoost) בשני ניסויים שונים. בניסוי הראשון, החוקרים ניסו לחזות את סיווג חומרת התאונה לפי ארבעת קטגוריות החומרה:
(Fatal / Grievous /Simple Injury/ Motor Collision) , הם גילו שהמודלים שהכי הצליחו לדייק בניסוי זה הוא AdaBoost עם כ-80% דיוק במדד הAccuracy ובין 70-75% דיוק במדדי Precision ו-F1 וכן מודל Naïve Bayes עם 80% ו60-65% בהתאמה. המודלים שפחות צלחו בניסוי הראשון היו KNN ו Decision Tree. הניסוי השני כלל 2 קטגוריות בלבד, Fatal וכל שאר דרגות החומרה וזאת משום שרוב המוחלט של התאונות בעלות חומרת Fatal לכן בוצע מיזוג.
תוצאות ניסוי 2 העידו כי מדדי המודלים שהצליחו בניסוי הראשון נשארו ללא שינוי, ואילו נצפתה עלייה במדדי הדיוק עבור שני המודלים שהצליחו פחות בניסוי הראשון. במחקרם, מימשו שלושה אלגוריתמים שונים (Univariate Feature Selection, Recursive Feature Elimination, and Feature Importance) לביצוע Feature selecting כאשר המאפיינים שנבחרו אלו המאפיינים החופפים ב- TOP 15 של כל אחד מהאלגוריתמים.
בסיכום הכללי, המודל הטוב ביותר לפי מחקר זה הוא הAdaBoost. מחקר זה תורם משמעותית להבנת שניתן לסייע בהתמודדות עם תאונות דרכים בבנגלדש על ידי כלים בלמידת מכונה.

על פי מאמרם של חאלד חאמד ושות' (2020), Prediction Incident Duration Using Random Forest שפורסם בכתב עת Transportmetria A: Transport Science עוסק במטלת רגרסיה לחישוב משך זמן תאונת דרכים. השיטה בה פעלו החוקרים הוא במימוש Random Forest תוך אופטימיזציה של הפרמטרים והשוואת תוצאותיו אל מודל נוסף שמימשו (ANN) ואל תוצאות העבר. סט הנתונים בו השתמשו הכיל כ-140 אלף רשומות ו- 52 פיצ'רים שונים אשר כל רשומה היא אירוע

של תאונות דרכים. תרומתו העיקרית של המחקר הוא הוכחת האפקטיביות בשימוש בRF לטובת סוג משימות אלו. תוצאות המחקר הראו כי עבור כלל הנתונים שעמדו לרשותם, מדד הMAE בRF הצביע על משך זמן תאונה של 36.9 דקות. כאשר הקטינו את מאגר הנתונים בכ-15% (ע"י הורדת קצוות) לטובת השוואה מול מחקרי עבר, גילו נתון מחזק אף יותר של מדד MAE עם 14.979 דקות למשך תאונה. החוקרים השוו מול המודל השני שמימשו, הביצועים של ANN היו טובים יותר בכ-0.24%, אך RF הצליח לזהות את הפרמטרים המשפיעים ביותר על משך זמן התאונה וכן תוצאותיו היו יציבים יותר. לסיכום, החוקרים השיגו את מטרתם בהוכחת אפקטיביות השימוש בRF.

הקדמה

- **שאלת המחקר:** סיווג חומרת התאונה על פי פרמטרים כמו מזג אוויר, תנאי דרך ומשך התאונה.
  - **חומרת תאונה** מדורגת מ-1 (קלה ביותר) ל-4 (קשה ביותר).
- מאגר הנתונים בו השתמשנו נלקח מאתר **Kaggle** הוא טבלאי ומכיל כ-1.5 מיליון רשומות על תאונות דרכים שהתרחשו בארה"ב בין השנים 2016-2020. המידע נאסף בזמן אמת ממגוון גופים שונים כמו משרדי התחבורה האמריקאים, רשויות אכיפת החוק, מצלמות וחיישני תנועה. קיים גיוון בסוגי השדות אשר מכילים מידע קטגוריאלי, נומרי, חותמות זמן וערכים בוליאניים.
- **שיטת המחקר** שלנו מבוססת על מודל **CrispDM**, כעת נפרט את הדרך שעשינו בכל אחד מהשלבים במודל:



Business Understanding : בשלב זה התמקדנו בהבנת מטרות, דרישות ויעדי הפרויקט מנקודת מבט עסקית. לאחר מכן, תרגמנו את כל הידע העסקי שצברנו לידי הגדרת שאלת מחקר אותה ניתן לפתור באמצעות כלים שרכשנו בקורס והגדרנו תכנון ראשוני כדי להשיג את מטרות אלו.

הגדרת שלוש נקודות עיקריות:

- **Business objective** – אנו עוסקים בתאונות דרכים, נושא קשה וכואב שגודע באופן יום יומי חיי אדם רבים סביב העולם. בעשור האחרון, יש עלייה משמעותית באחוז תאונות הדרכים בארה"ב. לפי נתונים שפרסם מינהל הבטיחות בדרכים האמריקאי (NHTSA). מגפת הקורונה הפחיתה את מספר המכוניות בכביש מה שגרם לעלייה בנהיגה במהירות גבוהה ועוד התנהגויות מסוכנות שנהפכו לנפוצות יותר.
- **Data Mining Objective** – המיקוד העיקרי שלנו בפרויקט אינו בצמצום מספר תאונות הדרכים אלא בלמידה כיצד ניתן לייעל את הטיפול והתמודדות בזמן אמת על פי דיווח של תאונות דרכים.
  - מטרת העל – היערכות טובה יותר לתאונות דרכים כתלות בחומרת התאונה/משך התאונה.

בפרויקט זה, תחילה בחרנו להתמקד בשאלת חיזוי של משך זמן תאונה. לאחר שלב הערכה, הבנו שעלינו להחליף את שאלת המחקר ולכן בחרנו בסיווג חומרת התאונה שיאפשר לנהל בצורה חכמה יותר את כמות צוותי החירום הנדרש, לייעל את התנועה בכבישים וכן להצביע על אזורים בעייתיים שחומרת התאונה באזורים אלו גבוהות יותר בממוצע.

ידיעת חומרת התאונה הצפוי לתאונה יכול לסייע לאפליקציות ניווט לשנות את מסלול המומלץ לנהגים שמסלולם המתוכנן עובר דרך התאונה, סביר שככל שהתאונה חמורה יותר, היא תשפיע יותר על עומס התנועה. דבר אשר ימנע עיכוב לנהגים שמתוכננים לעבור במסלול התאונה ובכך לשחרר את העומס בנתיב המוביל לתאונה, מה שיאפשר לצוותי החירום להגיע במהירות רבה יותר.

- **Success Criteria** – מדד ההצלחה היא קבלת תוצאות טובות יותר ממדד דיוק הנקבע על פי "חוק הרוב" בשימוש בדאטה המעובד. התוצאות יקבעו על פי מדד $Accuracy$.

Data Understanding: בשלב זה הבנו מה הדאטה שיש לרשותנו בעזרת אנליזה וחקירה. ניסינו להבין מה מתוך המידע הקיים יכול לסייע לנו לענות על שאלת המחקר. על מנת להבין זאת בצורה הטובה ביותר, ניסחנו שאלות מנחות:

- התשובות מופיעות במחברת תחת הקישור המתאים בתוכן העניינים.

1) מהן עשרת המדינות המסוכנות ביותר בארה"ב?
2) מהן עשרת הערים המסוכנות ביותר בארה"ב?
3) מהם הכבישים המסוכנים ביותר בארה"ב?
4) מתוך המדינות המסוכנות ביותר, מהן חמשת הערים המסוכנות ביותר עבור כל מדינה וכמה מהן נכללות בקטגוריית עשרת הערים המסוכנות בארה"ב?
5) כמה תאונות מכל רמת חומרה היו בכל עיר מסוכנת במדינה?
6) כמה תאונות היו בכל כביש/עיר/מדינה לפי שעה ביום ומהן השעות המסוכנת ביותר?
7) מה היה מזג האוויר בשעות המסוכנות ביותר?
8) מה היו תנאי הדרך(מהמורות/צומת/כיכר וכו') בשעות המסוכנות ביותר?
9) מהי התפלגות כמות התאונות לאורך השנים?

מסקנות מתהליך הבנת הנתונים:

- כלל המידע מתחלק ל:
  - מידע גאוגרפי - מיקום התאונה ברזולוציות שונות.
  - מידע על מזג אוויר – מהירות הרוח, טמפרטורה וכו'.
  - תנאי דרך בקרבת התאונה – מהמורות, פסי האטה, כיכר, תמרור עצור וכו'.
- כ-95% מהתאונות מתרכזות ל-20 מהכבישים, כאשר רובם חוצי מדינות.
- קליפורניה היא המדינה בעלת מספר התאונות הגבוה ביותר.
- שלוש מתוך עשר הערים המסוכנות בארה"ב נמצאות בקליפורניה.
- העיר המסוכנת ביותר בקליפורניה היא לוס אנג'לס.
- כ-90% מהתאונות בערים המסוכנות ביותר בקליפורניה הן בדרגת חומרה 2, כאשר העיר **סן חוזה** בעלת הפיזור המגוון ביותר עם 89%.
- משך זמן התאונה נמצא ביחס הפוך לכמות התאונות (כלומר, כמות התאונות שנמשכו כ-24 שעות קטנה בהרבה מכמות התאונות שנמשכה פחות משעה).
  - כמות התאונות שנמשכו **בדיוק** 6 שעות היא גבוהה באופן חריג ולא מתאימה למגמה שתיארנו לכן הוסרו מהרשימה.
- ב-80% מהתאונות בקליפורניה היה מזג אוויר נאה (Fair).
- 74% מהתאונות מתרכזות בשנתיים האחרונות (2019-2020).

- מצאנו כי עבור כל אחת מקטגוריות משכי זמן התאונה, קיימת התפלגות דומה של תנאי מזג אוויר.
- כ-18% מהתאונות בקליפורניה מתרחשות בצמוד לצמתים.
- מצאנו כי עבור כל אחת מקטגוריות משכי זמן התאונה, קיימת התפלגות דומה של הימצאות רמזור, מעבר חצייה או צומת.

Data Preparation: בשלב זה כחלק מהכנת הנתונים, עשינו מספר פעולות כמו בחירה, ניקיון, בנייה, עיצוב ואינטגרציה לנתונים על מנת להכין אותם לקראת שלב ה**Modeling**.
במהלך פעולות אלו, נתקלנו במספר תהיות:

1) **רזולוציה של מיקום התאונה** – בדאטה קיימות מספר שכבות מידע אשר מתארות את מיקום התאונה ברזולוציות שונות: מדינה > עיר > כביש > נקודה גאוגרפית. על מנת להחליט באיזה רזולוציה נרצה להתמקד, היינו צריכים לחזור לשלב business understanding, כדי להבין מה ייתן את הערך העסקי הגבוה ביותר. מבחינה עסקית, החלטנו שנכון יותר לפקס את הפרויקט ברזולוציה של עיר וזאת משום שהנגשת תוצאות המחקר לראשי הערים תוכל לסייע לקבל החלטות מתאימות כמו הגברה מקומית של צוותי החירום.
2) **מיזוג רמות חומרה –** בדומה למאמר שקראנו (Labib et al, 2019) בו אחד מהניסויים שבצעו היה למזג רמות חומרה בעלות תדירות נמוכה, כאשר תוצאות הניסוי העידו על שיפור באחוזי הדיוק של חלק מהמודלים. גם במאגר הנתונים שלנו יש רמת חומרה מובילה (2) וכן רמות חומרה (1,3,4) בעלות תדירות נמוכה. לכן עלתה בנו תהייה, האם נכון לבחון זאת גם אצלנו כאשר יש אפשרות לפצל לתאונות בעלות חומרה 2 וכל השאר. החלטנו לא לבצע זאת, כיוון שחלוקת החומרה לארבע רמות מגדירה את הקצאת המשאבים הנדרשים לטיפול בתאונה ולכן איחוד בין רמות חומרה עלול לפגוע בדיוק הקצאת משאבים מתאימה.
3) **סיווג משך זמן התאונה –** לצורך ניתוח הנתונים ועל מנת להבין טוב יותר מאפיינים שונים של תאונות לפי משך זמנם, במקום להסתכל ברמת השעה, בחרנו לצמצם את הטווח ולבצע דיסקרטיזציה לשלושה טווחים שונים: "מהיר", "בינוני", "איטי". כדי לקבוע מה יהיה טווח השעות עבור כל אחת מהקטגוריות, החלטנו לדון בשאלה באמצעות מומחי תוכן ומקורות מידע שונים. המסקנה שעלתה היא שאין תשובה חד משמעית וניתן להגדיר חלוקת טווחים בהתאם לצרכים שונים ולחומרת התאונה בפרט. לאור הנאמר לעיל, התמקדנו בחלוקה לטווחים על פי כמות התאונות, כך שבכל קטגוריה תהיה כמות תאונות זהה.

שלב הכנת הנתונים כולל בתוכו שישה שלבים אותם בצענו:

Data selection – בחירת הנתונים:

- ראשית בחרנו להתמקד במדינה המסוכנת ביותר בארה"ב והיא קליפורניה.
- העיר שבחרנו היא **סן חוזה** שבקליפורניה וזאת משום שגילינו כי היא חלק מחמשת הערים המסוכנות ביותר בקליפורניה וגם יש בה את פיזור התפלגות חומרת התאונה המגוון ביותר. החלטה זו גררה הסרה של רשומות שאינן שייכות לסן חוזה והסרת עמודות בעלות ערך זהה. כמו למשל: State, City, Time zone.

Data Cleaning – ניקוי הנתונים:

- **<u>טיפול בערכים חסרים:</u>**
  - **ערכים דיפולטיביים –** במהלך הבנת הנתונים, נכחנו לגלות כי יש מעל ל-20% רשומות אשר משך זמן התאונה הוא בדיוק 6 שעות. הדבר העלה בנו את החשד והחשש שנתון זה יכול להשפיע על התוצאות, לאחר בחינה עמוקה וצפייה בהערות בפורום המתאים ב-Kaggle, ניתן להניח כי מדובר בערכים דיפולטיביים עבור תאונות שמשך זמנם לא דווח, על כן בחרנו להסיר רשומות אלו על מנת שלא יטעו את אמינות התוצאות.
  - **הסרה אנכית**: ניתחנו את אחוז הערכים החסרים עבור כל פיצ'ר. פיצ'רים בעלי אחוז ערכים חסרים גדול מ-8%, הוסרו מהמאגר.
  - **הסרה מאוזנת**: הסרנו את כל הרשומות אשר הכילו ערך חסר באחת או יותר מהפיצ'רים.
- **<u>טיפול בערכים חריגים:</u>**
  - **ערכי קיצון (Outliers)** - השתמשנו בשיטות שונות לטיפול בערכי קיצון במשתנים קטגוריאליים ובמשתנים נומריים.
    - עבור משתנים **נומריים**, אימנו מודל מסוג IsolationTree שתפקידו לאבחן נקודות קיצון במרחב. את המודל אימנו על קבוצות מאפיינים שונות. המודל לא הטיב עם התוצאות ולכן בחרנו לא להשתמש בו.
    - עבור משתנים **קטגוריאליים** בעלי מספר רב של קטגוריות עם מספר מופעים נמוך, בדקנו את התפלגות כמות התאונות שבה מופיעה כל קטגוריה, כאשר איחדנו אוסף קטגוריות נדירות תחת קורת גג אחת (other). פעולה זו אפשרה לערכים אלו לקבל משקל גדול יותר ולהשפיע לטובה על התוצאה.

Data Construction– בניית נתונים: בשלב זה גזרנו שדות חדשים מתוך שדות קיימים.

- **Feature Construction:**

<u>שדות מקוריים</u>:

- *Weather_Measurement_Proximity_Hours* – כמה זמן עבר מהדגימה האחרונה של מזג אוויר (*Weather_Timestamp*) לבין מועד תחילת התאונה (*Start_Time*).
- *Has_Road_Conditions_Info* – ערך בוליאני, מקבל *true* כאשר באחד מהשדות שירדו בשלב ה*data selection* כחלק מסעיף של הורדת עמודות בעלות רוב של מעל ל-95% ערך זהה, היה את ערך המיעוט (קטן מ-5%).
- *Geohash_Start* – נבנה משדות של קווי אורך ורוחב. הבנו כי שימוש בשדות אלו באופן גולמי יכול להטעות את מודל רגרסיה משום שקו אורך 34 לא בעל משמעות רבה יותר מקו אורך 33 ולכן השתמשנו ב*Geohash*. *Geohash* ממפה לקטגוריות אזורים שנתחמים על פי קווי אורך ורוחב (ערכי *hash*), את ערכים אלו נמיר לערכים נומריים.
  - עשינו ניסיון להפיק ערכי *GeoHash* שונים מקווי האורך והרוחב של תחילת וסיום התאונה שנמצאו בנתונים. חששנו שווקטורי ה-*GeoHash* יהיו זהים ולכן לא יוסיפו מידע. ביצענו מבחן סטטיסטי (קולמוגרוב-סמירנוב) על מנת לאשש השערה זו ומצאנו כי היא נכונה
(*P-value < 0.05,alpha=0.05*), לכן נותרנו רק עם וקטור *GeoHash* אחד.

<u>שדות שהשתמשו במאמר</u> <u>(2021, Yuexu Zhao, Wei Dang):</u>

- *Accident_Duration_Hours* – שדה זה הוא משך זמן התאונה שמחושב על ידי $End\_Time - Start\_Time$.
- *Year / Month / Hour* – שדות שנגזרו מחותמת הזמן של *Start_Time*.
- *Season* – עונה בשנה חושב על פי *Day_of_Year* שגם הוא שדה חדש שיצרנו בהתבסס *Start_Time* אך החלטנו בהמשך להסירו כי תרומתו אינה משמעותית.
- *Is_Holiday* – האם תאריך התאונה (*Start_Time*) הוא ביום חג (לוח שנה אמריקאי).
- *Traffic_Peak_Status* – שדה בוליאני, האם התאונה התרחשה בזמן שעות העומס בתנועה בלוס אנג'לס על פי גוגל (07:00-09:00, 17:00-19:00).
- *Are_Dangerous_Hours* – שדה בוליאני, האם התאונה התרחשה בזמן שעות המסוכנות ביותר על פי הנתונים שלרשותנו.

- **Feature Aggregation:**

שדות שהשתמשו במאמר (2021, Yuexu Zhao, Wei Dang):

- *Street_Monthly_ Sum / Min / Max / Median / Mean / CumCount*
- *Zipcode_Monthly_ Sum / Min / Max / Median / Mean / CumCount*
  - במאמרים שקראנו ייחסו חשיבות רבה למיקום גאוגרפי. עם זאת, עם בחירתנו להתמקד בעיר סן חוזה, ויתרנו על מידע גאוגרפי רב הטמון במדינה, מחוז ואיזור זמן. על מנת לחלץ משמעות נוספת ממשתנים ברזולוציה גאוגרפית במימד נמוך יותר, כמו כביש או מיקוד, בחרנו לבצע מספר פעולות אגרגטיביות על משתנים אלו:
    - יש לציין שמשתנים אלו היו קטגוריאליים ולכן נדרשנו לבצע קידוד לפני פעולות האגריגציה. בחרנו ב-*Target Encoding* מכיוון שהיו לכל אחד מהמשתנים קטגוריות רבות ו-*One Hot Encoding* פחות התאים למשימה זו.
    - בנוסף, כפי שראינו בעבודה 5 וכפי שצוין במאמר, ביצענו את האגריגציות בחלונות זמן של חודש אחורה.

Data Transformation– טרנספורמציית נתונים:

- **Zip code –** בשלב הבנת הנתונים, גילינו כי מאגר הנתונים מכיל שני סוגים של מיקוד – בסיסי בעל 5 ספרות ומורחב בעל 9 ספרות. בשלב הניקיון, צמצמנו את כולם לפורמט אחיד של 5 ספרות.
- **נרמול הנתונים** – בצענו נרמול על מנת לשנות את ערכי העמודות הנומריות לטווח משותף, תוך שמירה על הפרופורציות ללא עיוות ההבדלים בין הטווחים המקוריים. בנוסף, בחרנו בנרמול לפי ערכי מינ' ומקס' מכיוון שנרמול לזה מתאים למשתנים שונים בעלי מספר התפלגויות שונות אשר אליהם נחשפנו במהלך ניתוח הנתונים.

Data Integration – אינטגרציה לנתונים:

הנתונים בהם השתמשנו נלקחו משלושה מקורות שונים ועברו אינטגרציה לפני שהועלו ל*Kaggle*.

חלק מההשלכות לכך התבטאו, בין היתר, בערכי ברירת מחדל לשעת התחלה וסיום של התאונה, שהובילו לכך שמספר רב של תאונות נמשך בדיוק 6 שעות, כפי שפירטנו בשלב ה-*Cleaning*.

Data Reduction – צמצום הנתונים:

- **פיצ'רים של תנאי דרך** הן מסוג בוליאני, עמודות שהיו עם מעל 95% ערך זהה כמו False, בחרנו להסיר אותן וזאת משום שכמות גדולה מדי של פיצ'רים יעלו את זמני הריצה ועלולות לגרום ל-Overfitting.
- **פיצ'ר Description** בעלת טקסט חופשי ומכילה בעיקר מידע גאוגרפי על מיקום התאונה. על מנת להשתמש בנתון זה נדרש שלב של Parsing ובחרנו שלא לבצע זאת, משום שגילינו שרוב המידע אותו מכילה מונגש דרך עמודות שונות.
- **פיצ'ר Accident Duration Hour –** הורדנו שדה זה, מכוון שאימון מודל תוך שימוש בפיצ'ר שידיעתו אינה בזמן אמת של התאונה סותרת את הרעיון המחקר שלנו.
- **פיצ'רים של מיקום:**

  City, Timezone, County, State, Country, Start_Lng/Lat, End_Lng/Lat

  מכיוון שבחרנו בעיר בודדת, שדות המתארים מדינה, מחוז וכו' יהיו בעלי אותו הערך. אם כך, הם חסרי משמעות ובחרנו להוריד אותם. הפקנו משמעויות מקווי אורך ורוחב באמצעות GeoHash ולכן בחרנו להוריד אותם.

Modeling:

לאור מחקרים רבים בנושא סיווג חומרת תאונות, מעבר למודלים המקובלים המטיבים בסיווג חומרה שאותם מימשנו. בחרנו לנסות לממש ארכיטקטורה חדשה שתפצל את הדאטה ותחלק את התאונות לאשכולות באמצעות מודל K-Means. חשוב לציין, על מנת לשמר את הסדר הכרונולוגי, מיינו את האשכולות לפי תאריך. המטרה בשימוש בארכיטקטורה זו היא למצוא את המודל שיטיב עם כל אחד מהאשכולות השונים ובכך לשפר את דיוק הסיווג.

שלב הModeling התחלק לארבעה תתי שלבים:

1. **בחירת מודלים** בהתבסס על מאמרים שקראנו וידע שרכשנו בקורס:
   a. Random Forest Classifier – שימוש ביער אקראי מתאים כאשר מערך הנתונים גדול וראינו הצלחה שלו במאמר של חאלד חאמד ושות' (2020).
   b. Logistic Regression –מודל זה מתאים כאשר ה-Target הוא קטגוריאלי.
   c. AdaBoost – מודל זה עובד טוב במשימות סיווג וכן נחל הצלחה רבה במאמר של Labib et. al (2019)
   d. SVM – מודל ורסטילי שעובד טוב על מאגרי נתונים שונים.
2. **חלוקת הדאטה - Train, Test, Validation sets:**
   a. **ציר זמן** – בכל אחת מהחלוקות שנתאר מטה, התחשבנו בציר הזמן. קרי, סט המבחן יורכב מתצפיות שהתרחשו בסדר כרונולוגי מאוחר יותר מסט האימון.

b. **חלוקה לפי שנים** – בגלל שמרבית התאונות התרחשו בשנת 2020, אז סט המבחן יהיה מאמצע שנת 2020.

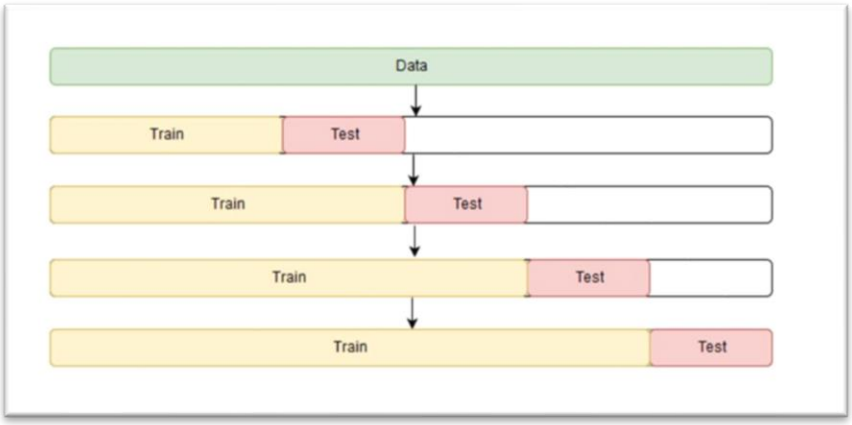
i. סט אימון: 1/2016-6/2020

ii. סט מבחן: 7/2020-12/2020

c. **Cross validation** – כפי שראינו בעבר במטלה 5, לא השתמשנו ב K-Fold Cross Validation מכיוון שעלינו לשמר את הסדר הכרונולוגי בין התאונות.

לכן, בחרנו להשתמש בCross Validation באופן הבא: בכל איטרציה, בחרנו אוסף מצומצם של נתוני עבר ובאמצעותם חזינו נתוני עתיד. באיטרציה הבאה, נתוני העתיד יתווספו לנתוני העבר וישמשו לחיזוי של נתונים נוספים בעתיד.

להלן דיאגרמה, המסבירה זאת:



3. **בחירת היפר פרמטרים:**

a. לכל אחד מהמודלים, הרצנו חיפוש ממצה על אוסף מצומצם של היפר-פרמטרים וערכים מתאימים שנבחרו בקפידה באמצעות מקורות מידע שונים באמצעות GridSearchCV. בעזרת הפלט שקיבלנו, בחרנו את הפרמטרים הטובים ביותר לכל אחד מהמודלים.

4. **הרצה והערכה ראשונית:**

הערכה הראשונית תהיה על מול "חוק הרוב" אותו הגדרנו כ-Success Criteria בשלב ההבנה העסקית.

a. **Baseline –** אימנו את כל אחד מהמודלים (הלא מאומנים) על סט האימון ובצענו סיווג על סט המבחן. את ה תוצאות שקיבלנו ממדדו באמצעות מדד Accuracy.

b. **K-Means** - מכיוון שעבדנו עם כמות רשומות מוגבלת, ביצענו חלוקה של הדאטה כאשר K=2,3. על כל אחד מהדאטה סטים שנוצרו אימנו את כל המודלים ובחרנו את המודל בעל הביצועים הטובים ביותר. בדקנו האם נבחרו מודלים שונים וגם האם אחד המודלים שנבחרו היו בעלי הביצועים הטובים ביותר גם בסיווג ה- Baseline.

Evaluation:

	Measure		Majority Rule	Random Forest Classifier	Logistic Regression	AdaBoost
Baseline	**Accuracy**		87.51%	87.51%	87.51%	**87.97%**
K-Means	**Accuracy**	**Cluster 1**	89.75%	89.75%	89.84%	**89.93%**
		Cluster 2	**82.89%**	**82.89%**	82.55%	76.74%

מסקנות:

- בBaseline, מודל AdaBoost עבר במעט את חוק הרוב על פי מדד ה- Accuracy, שאר המודלים השוו תוצאות לחוק הרוב. אנו מעריכים שהסיבה לכך היא שמכיוון שרוב התאונות הן ברמת חומרה 2, המודלים התקשו לזהות את התאונות בעלות חומרות 1, 3 ו4 ולכן הגיעו לתוצאות דומות לשל חוק הרוב.
- ניתן לראות שאחוזי הדיוק של חוק הרוב משתנים בהתאם לקלאסטרים שנוצרו. עם זאת, באופן דומה ל-baseline, פרט ל-AdaBoost שהתעלה על חוק הרוב בקלאסטר 1, שאר המודלים מתקשים לעלות על ביצועי חוק הרוב ומקבלים תוצאות זהות או קטנות מחוק הרוב.
- יש לשים לב שהמודל בעל הביצועים הטובים ביותר על הBaseline היה AdaBoost. לעומת זאת, תחת הK-Means קיבלנו מודלים אופטימליים שונים עבור קלאסטרים שונים, אם כי רק הביצועים של המודלים המנצחים דומים.
- לסיכום, במצב הנוכחי, לא היינו בוחרים להשתמש בארכיטקטורת החלוקה לאשכולות מכיוון שהיא מורכבת יותר ולא מספקת יתרון משמעותי מבחינת ביצועים.

Future work:

1) כפי שציינו לעיל, ניתן להניח כי אי היכולת להגיע לביצועים טובים יותר מחוק הרוב נובעת מכך שמרבית התאונות (להכניס מספר סטטיסטי באחוזים) הן תחת אותה חומרה ולכן המודלים מתקשים לסווג את התאונות האחרות.

   ניתן לנסות להתמודד עם אתגר זה באמצעות סיווג בשני שלבים:

   - **בשלב הראשון**, סיווג של תאונות ברמת חומרה 2 או לא ברמת חומרה זו. אנו צופים כי משימה זו תנחל הצלחה לא פחות מאשר המשימה הנוכחית.
   - **בשלב השני,** סיווג נוסף לתת הקבוצה "לא ברמת חומרה 2". להשערתנו, קבוצה זו תהיה מאוזנת יותר ועשויה להניב ביצועים טובים בחיזוי.

2) במשימת הרגרסיה בה ניסינו לחזות את משך זמן התאונה - הפעלנו מספר מודלים שתוצאותיהם היו קרובות ל-0 על פי מדד R2 ומשמעותית פחות טובות מחוק הרוב. לכן בחרנו להחליף את שאלת המחקר. בעתיד, נרצה לחזור למשימה זו, על מנת להתגבר על מכשול שעמד בפנינו, ולהצליח להגיע לתוצאות טובות יותר.