US ACCIDENTS

המטרה

חיזוי השפעת התאונה על התנועה (חזקה/חלשה),
ע״י מציאת משתני מפתח ומודל

השימוש

פיתוח אסטרטגיות ניהול תנועה, בקרה, בטיחות,
ניהול סיכונים וביטוח, תכנון ערים

DATASET

500k sampled

נתונים גאוגרפיים

קווי אורך ורוחב, מדינה, עיר

נתונים סביבתיים

חנות נוחות, תחנה, צומת, כיכר, עיקול

נתוני זמן

שעת התחלה וסיום, יום לילה

נתונים דרך

רמזורים, תמרורים, פס האטה, מעבר חצייה

מדדים אקלימיים

טמפרטורה, רוח, לחות, ראות, משקעים, תיאור

נתונים נוספים

מזהה תאונה, חומרה

US ACCIDENTS

AMIR NAVON

BIU DS17

שינוי טיפוסי משתנים

Bool to Int, Object to String

חילוץ נתונים ויצירת חדשים

חודש, יום, שעה, יצירת משך Duration

השמטה של נתונים

מיקוד, שעת ערביים, זמן התחלה וסיום, שדה תעופה מדווח ועוד
נותרו 29 מתוך 46 שבמקור

מיקוד בשנה מסוימת

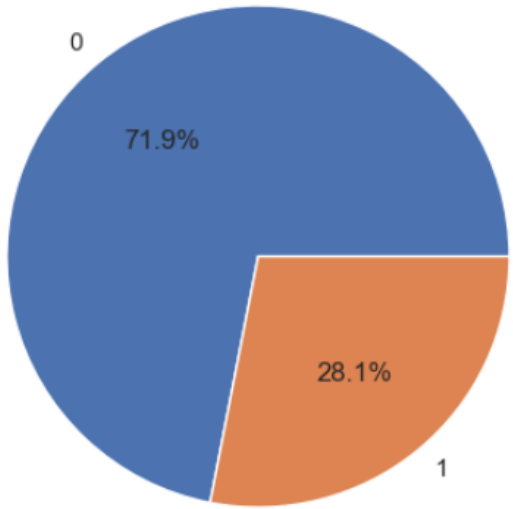
מבין 2016-2023 נבחרה שנת 2019, 61,852 תצפיות

TARGET VALUE

משתנה מטרה - Severity

חומרה מתאר עד כמה התאונה השפיע על התנועה
הנתונים התקבלו לפי 4 דרגות

נמוכה - 16 תצפיות
בינונית - 44480 תצפיות
חזקה - 15518 תצפיות
חזקה מאוד - 1838 תצפיות



שינוי משתנה המטרה לבינארי

0 - השפעה נמוכה 1 - השפעה חזקה
Not Imbalanced

US ACCIDENTS
AMIR NAVON
BIU DS17

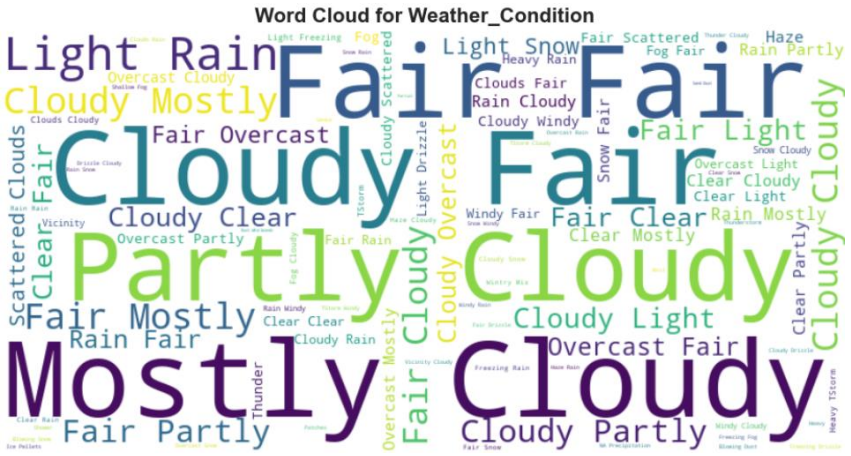
EDA

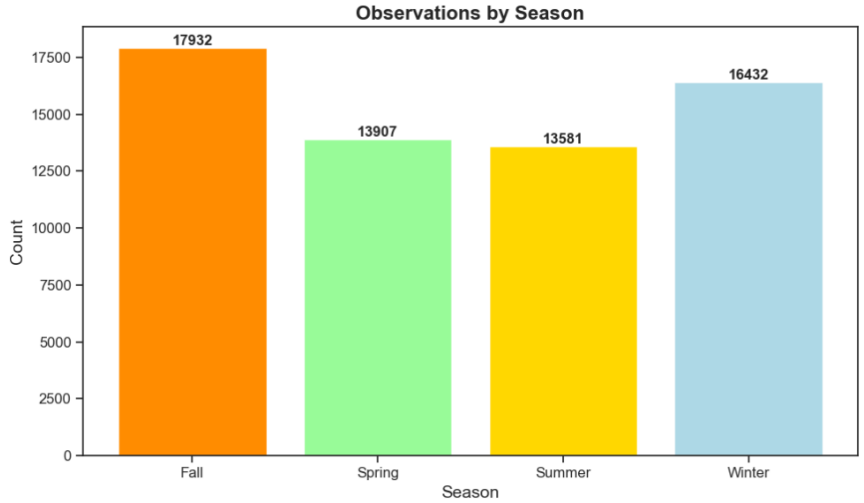
FEATURE-

ENGINEERING

הצגה ויזואלית EDA

גרפים ידניים ואוטומטיים. יצירת עונות (אגרגצית חודשים)
ריבוי תאונות לקראת סוף שנה (סתיו: ספט'-אוק'-נוב'),
בתחילת החודש ובתחילת היום ובעיקר במזג אוויר מעונן





יצירת משתנה Region

בעקבות הגילוי לגבי ההשפעה של האקלים, מצאתי לנכון ליצור משתנה "אזור" מתוך נתונים חיצוניים: צפון, מרכז, דרום (מדינות לפי מיקום קווי אורך רוחב). התמונה הנ"ל נוצרה ב-AI המחשת אקלים ממוצע בסתיו בארה"ב

US ACCIDENTS

AMIR NAVON

BIU DS17

OUTLIERS

MISSING

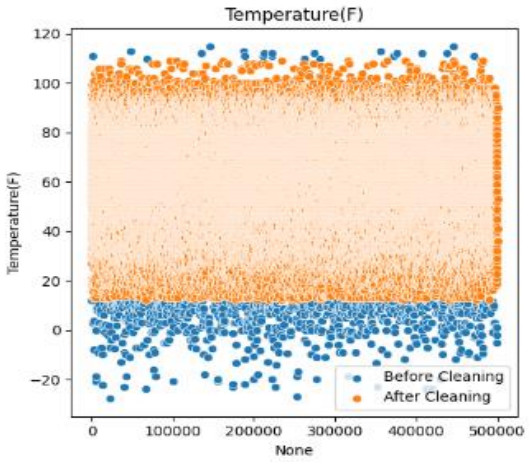
IMPUTATION

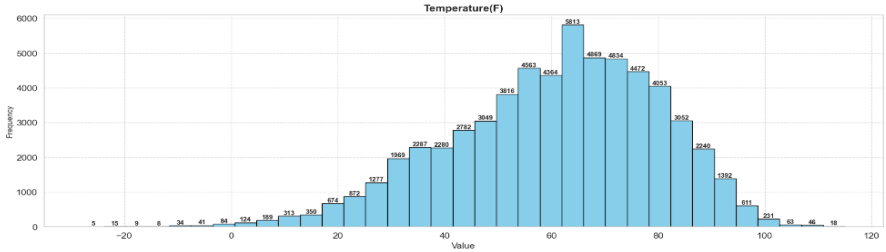
ENCODING

חריגים

טיפול בנומריים ע"י IQR ובנורמליים ע"י Z-score

"משך" לא הוסר (משפיע על קורלציה וגם התפלגות)



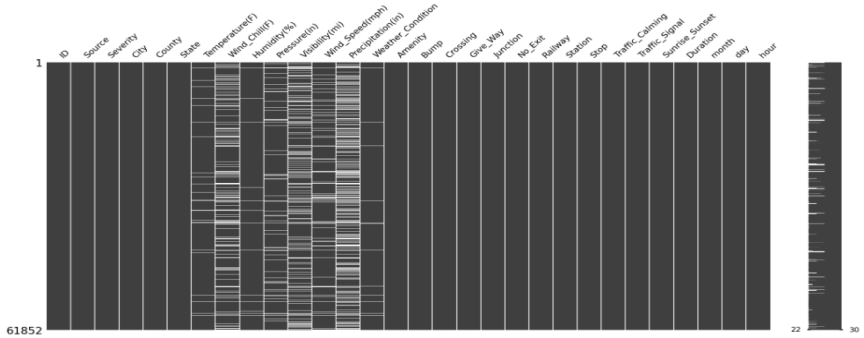


השלמת נתונים חסרים

עיר ומזג אוויר = "אחר"

יום/לילה = לפי השעה

השלמה MICE



Label Encoding

7 משתנים מתוך 29

US ACCIDENTS

AMIR NAVON

BIU DS17

Feature Selection

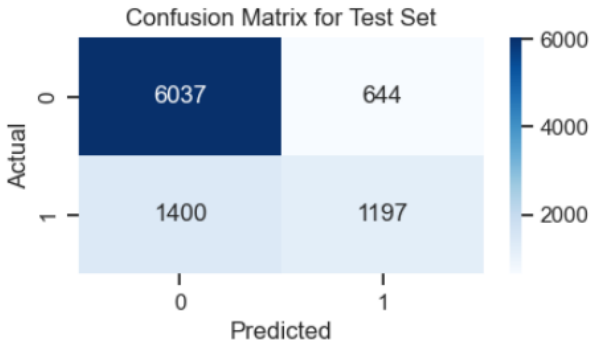
באמצעות טבלת Feature Importance (ציון מעל 3)
Lasso, LinearSVC (SVM), GradientBoosting, RandomForestClassifier
מתוך 29 נותרו 17 משתנים

Train Dev Test

70%, 15%, 15%

Model Selection

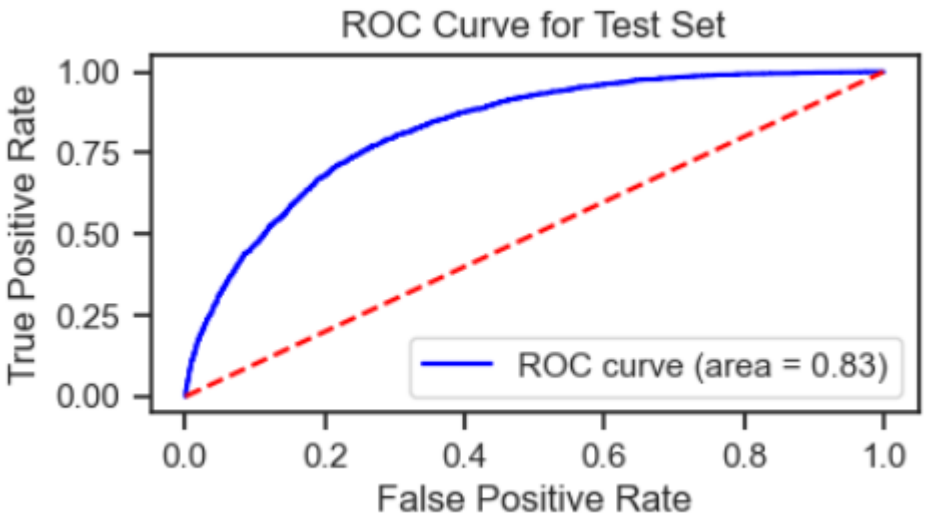
בחינת מטריקות Accuracy, Precision, Recall, F1-Score
בחינת Confusion Matrix
הרצת המודלים: LogisticRegression, DecisionTreeClassifier,
RandomForestClassifier, Support Vector Machine (SVC), GaussianNB,
KNeighborsClassifier (KNN), XGBClassifier



XGBoost

המודל XGBoost הצליח לנבא טוב יותר מאחרים
קרוב אליו היה RandomForrest

Accuracy	Precision	Recall	F1-Score	True Positives	True Negatives	False Positives	False Negatives
0.779694	0.766529	0.779694	0.766828	1197	6037	644	1400



END

תודה רבה!

US ACCIDENTS

AMIR NAVON

BIU DS17