**פרויקט כריית מידע**

**2015- Flights delay and cancellation**

**מרצה: ד"ר גונן זינגר**

**רקע**

ה-Data Set שעליו בחרנו לעבוד הינו: "2015 Flights delays and Cancellations" שפורסם באתר Kaggle.  
המידע פורסם על ידי מחלקת התעבורה של ארה"ב ( (DOT-Department Of Transportationוסופק על ידי מחלקת נתונים סטטיסטיים שעוקבת בזמן אמת על טיסות פנימיות בארה"ב שמופעלות על ידי מספר רב של חברות תעופה.  
קישור למאגר:

<https://www.kaggle.com/usdot/flight-delays>

**תיאור הנתונים המקוריים**

מאגר הנתונים המקורי מורכב משלושה קבצי CSV כאשר שניים מהם הם קבצים מסבירים.

המאגר העיקרי בו השתמשנו (flights) מורכב מ-31 עמודות ומכיל למעלה ממיליון רשומות:

|  |  |  |  |
| --- | --- | --- | --- |
| שם המשתנה | תיאור | סוג המשתנה | הערות |
| YEAR | שנה | Integer |  |
| MONTH | חודש | Integer | 1-3, ינואר-מרץ |
| DAY | תאריך יומי | Integer | 1-31 |
| DAY\_OF\_WEEK | יום בשבוע | Integer | 1-7, ראשון-שבת |
| AIRLINE | חברת תעופה | String |  |
| FLIGHT\_NUMBER | מזהה טיסה | Integer |  |
| TAIL\_NUMBER | מזהה מטוס | String |  |
| ORIGIN\_AIRPORT | השדה ממנו הטיסה המריאה | String |  |
| DESTINATION\_AIRPORT | השדה בו הטיסה נחתה | String |  |
| SCHEDULED\_DEPARTURE | שעת המראה מתוכננת | Integer |  |
| DEPARTURE\_TIME | שעת המראה בפועל | Integer | WHEELS\_OFF – TAXI\_OUT |
| DEPARTURE\_DELAY | איחור כולל בהמראה | Integer |  |
| TAXI\_OUT | הזמן בין מעבר הנוסעים ב-gate להמראה | Integer |  |
| WHEELS\_OFF | הרגע בו המטוס עוזב את הקרקע | Integer |  |
| SCHEDULED\_TIME | הזמן המתוכנן למשך הטיסה | Integer |  |
| ELAPSED\_TIME | הזמן מ-gate בשדה בו ממריאים ועד ה-gate בשדה בו נוחתים | Integer | AIR\_TIME+TAXI\_IN+TAXI\_OUT |
| AIR\_TIME | הזמן שבין הרגע בו המטוס ממריא לרגע בו הוא נוחת | Integer |  |
| DISTANCE | מרחק בין השדה ממנו ממריאים לשדה בו נוחתים | Integer |  |
| WHEELS\_ON | הרגע בו המטוס נוגע בקרקע | Integer |  |
| TAXI\_IN | הזמן בין הנחיתה להגעת הנוסעים ל-gate | Integer |  |
| SCHEDULED\_ARRIVAL | זמן הגעה מתוכנן | Integer |  |
| ARRIVAL\_TIME | זמן הגעה ל-gate בשדה בו נוחתים | Integer | WHEELS\_ON+TAXI\_IN |
| ARRIVAL\_DELAY | האיחור/ההקדמה בנחיתה | Integer | ARRIVAL\_TIME-SCHEDULED\_ARRIVAL |
| DIVERTED | המטוס נחת בשדה אחר מהמתוכנן | Integer |  |
| CANCELLED | האם הטיסה בוטלה | binary | 1. אין ביטול 2. טיסה בוטלה |
| CANCELLATION\_REASON | סיבת הביטול | String | 1. חברת התעופה 2. מזג אויר 3. מערכות אויר 4. ביטחון |
| AIR\_SYSTEM\_DELAY | עיכוב בגלל מערכות אויר | binary |  |
| SECURITY\_DELAY | איחור עקב בעיות ביטחון | binary |  |
| AIRLINE\_DELAY | עיכוב בגלל חברת התעופה | binary |  |
| LATE\_AIRCRAFT\_DELAY | עיכוב בגלל איחור אחר | binary |  |
| WEATHER\_DELAY | עיכוב בגלל מזג האויר | binary |  |

טבלת חברות תעופה (Airlines):

|  |  |  |  |
| --- | --- | --- | --- |
| שם המשתנה | תיאור | סוג המשתנה | הערות |
| IATA\_CODE | מזהה חברת תעופה | String |  |
| AIRLINE | השם המלא של שדה התעופה | String |  |

טבלת שדות תעופה (Airports):

|  |  |  |  |
| --- | --- | --- | --- |
| שם המשתנה | תיאור | סוג המשתנה | הערות |
| IATA\_CODE | מזהה חברת תעופה | String |  |
| AIRPORT | שם שדה התעופה | String |  |
| CITY | עיר | String |  |
| STATE | מחוז | String |  |
| COUNTRY | מדינה | String |  |
| LATITUDE | קואורדינטת קו רוחב | Double |  |
| LONGITUDE | קואורדינטת קו אורך | Double |  |

**תהליך pre process**

**Data Transformation**

בחרנו להתמקד בטיסות שיצאו ממדינת קולורדו. לאחר סינון הרשומות נותרו כ-43,000.

ערך המטרה (Delay Type) חושב ע"י העמודה DEPARTURE\_DELAY.

בחרנו לבצע חלוקה לשלושה סוגי איחורים על-פי משך האיחור בהתבסס על המדיניות המונהגת בחברות התעופה:

1. N- אין איחור.
2. A- איחור קטן- עד שעה.
3. B- איחור בינוני- בין שעה ל-3 שעות.
4. C- איחור גדול- מעל 3 שעות.

בחרנו להתמקד באיחורים מסוג B ו-C מכיוון שהמשמעות הכלכלית שלהם (הפיצויים שחברת התעופה תאלץ לשלם ללקוחות) הם הגבוהים ביותר.   
איחור מסוג C גורר בממוצע תשלום פיצויים של כ-450$ לנוסע. דמי הפיצויים מורכבים מעלויות לינה, הסעות מהשדה וחזרה אליו ואש"ל (הנתונים נלקחו מאתרי חברות התעופה).

**Data Cleaning**

מחקנו רשומות בהן היו חסרים נתונים לגבי האיחורים.  
העמודות אותן בחרנו למחוק הן:

|  |  |
| --- | --- |
| שם העמודה | סיבת הסרה |
| YEAR | כל הנתונים נאספו בשנת 2015 |
| TAIL\_NUMBER | נתון זה גורם ל-Over fitting (מכיוון שאותו מטוס מבצע מספר טיסות) |
| DEPARTURE\_TIME | נתון זה גורם ל-Over fitting |
| DEPARTURE\_DELAY | השתמשנו בעמודה זו ליצירת ערך המטרה |
| TAXI\_OUT | לא רלוונטי |
| WHEELS\_OFF | לא רלוונטי |
| SCHEDULED\_TIME | לא רלוונטי |
| ELAPSED\_TIME | לא רלוונטי |
| AIR\_TIME | לא רלוונטי |
| DISTANCE | לא רלוונטי |
| WHEELS\_ON | לא רלוונטי |
| TAXI\_IN | לא רלוונטי |
| SCHEDULED\_ARRIVAL | לא רלוונטי |
| ARRIVAL\_TIME | לא רלוונטי |
| ARRIVAL\_DELAY | לא רלוונטי |
| DIVERTED | מספר קטן מאוד של רשומות מלאות |
| CANCELLED | מספר קטן מאוד של רשומות מלאות |
| CANCELLATION\_REASON | מספר קטן מאוד של רשומות מלאות |
| AIR\_SYSTEM\_DELAY | מספר קטן מאוד של רשומות מלאות |
| SECURITY\_DELAY | מספר קטן מאוד של רשומות מלאות |
| AIRLINE\_DELAY | מספר קטן מאוד של רשומות מלאות |
| LATE\_AIRCRAFT\_DELAY | מספר קטן מאוד של רשומות מלאות |
| WEATHER\_DELAY | מספר קטן מאוד של רשומות מלאות |

**עמודות לאחר הpre process**

|  |  |  |  |
| --- | --- | --- | --- |
| שם המשתנה | תיאור | סוג המשתנה | הערות |
| MONTH | חודש | Integer | 1-3, ינואר-מרץ |
| DAY | תאריך יומי | Integer | 1-31 |
| DAY\_OF\_WEEK | יום בשבוע | Integer | 1-7, ראשון-שבת |
| AIRLINE | חברת תעופה | String |  |
| FLIGHT\_NUMBER | מזהה טיסה | Integer |  |
| ORIGIN\_AIRPORT | השדה ממנו הטיסה המריאה | String |  |
| DESTINATION\_AIRPORT | השדה בו הטיסה נחתה | String |  |
| SCHEDULED\_DEPARTURE | שעת המראה מתוכננת | Integer |  |
| Delay\_Type | סוג האיחור | String | N- אין איחור  A- איחור קטן  B- איחור בינוני  C- איחור גדול |

# תוצאות האלגוריתמים

לשם טיוב המודל הגדרנו כי לאורך הרצת כל האלגוריתמים Cross validation עם Folds=10 . כלומר השיטה מחלקת את המידע ל-10 חלקים שווים ומבצעת 10 איטרציות כאשר בכל איטרציה יש 9 חלקים המשמשים כ Training data ו Fold אשר מתחלף מאיטרציה לאיטרציה ומשמש כTesting data.

## **עץ החלטה**

עצי החלטה מחלקים את הרשומות ע"י מיונם לעומק העץ מהשורש כל צומת בעץ מציין חוק של תכונה מסוימת של הרשומה, וכל ענף שיוצא מהצומת מתאים לאחד הערכים האפשריים לתכונה זו. התהליך מתחיל בשורש העץ שהוא תכונת האב, בוחנים את החוק הנמצא בצומת זה ומחליטים לפי הענפים לאן לנווט את התכונה למטה בעץ. תהליך זה חוזר על עצמו שוב ושוב עד שמגיעים לצומת עלה התחתונה ביותר האפשרית שבה מקבלים את התשובה לשאלה הנשאלת.

השתמשנו באלגוריתם J48 בתכונת ה-weka. אלגוריתם זה משמש ליצירת עץ החלטה והינו הרחבה של אלגוריתם ID3 .

עץ החלטה יכול לשמש כמודל חיזוי, הממפה תצפיות על פריטים ויוצר מסקנות על ערך היעד של הפריט החדש היכול לעזור בקבלת החלטות שונות.

על מנת לקבל את עץ ההחלטות היה עלינו לשנות את הערכים של confidenceFactor בקפיצות של 0.05 ושל minNumObj בקפיצות של 5 וכל פעם לבדוק את הערכים המתקבלים של ROC Area הקשורים לערכים של B (איחור בינוני) ו-C (איחור גדול).

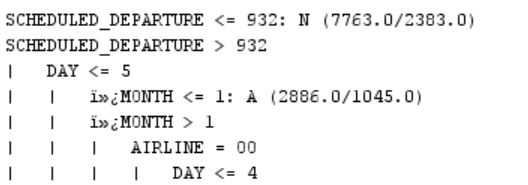
הרצנו את confidenceFactor מ 0.1 עד 0.7 ו- minNumObj מ25 עד 80,

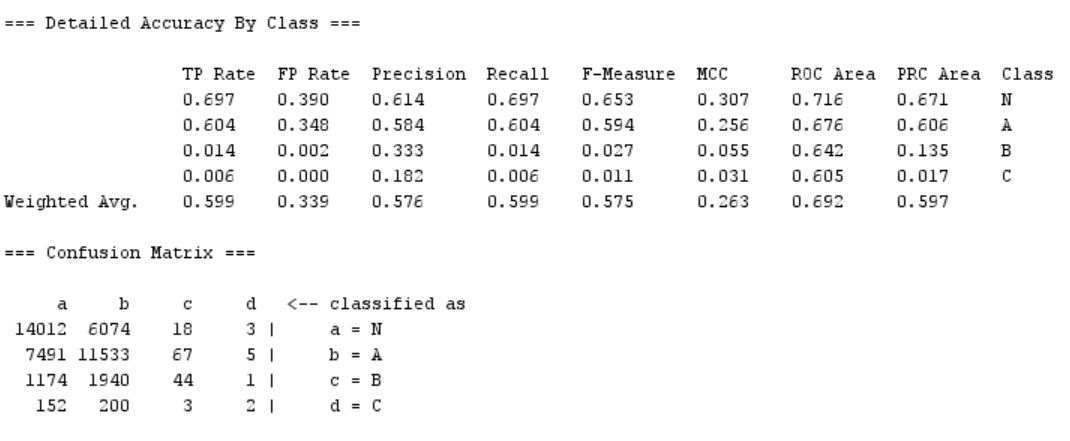
התחלנו 0.1= confidenceFactor ואז ביצענו שינוי רק בפרמטר minNumObj מ25 עד 80. כאשר תוצאת Roc Area בעמודת הסיכום ירדה או לא השתנתה מהתוצאה הקודמת הפסקנו ושיננו את 0.15= confidenceFactor והתחלנו לשנות את- minNumObj וכך הלאה.

כאשר הערכים של ה-Roc עלו אבל נשארו בטווח בין 0.6-0.8 החשבנו כתוצאה טובה. ברגע שהערכים הפסיקו לעלות עצרנו את הבדיקות.

מצאנו את השינוי שהציג את התוצאה הטובה ביותר כאשר:

* confidenceFactor = 0.7
* minNumObj = 30
* Roc Area B=0.642
* Roc Area C=0.605



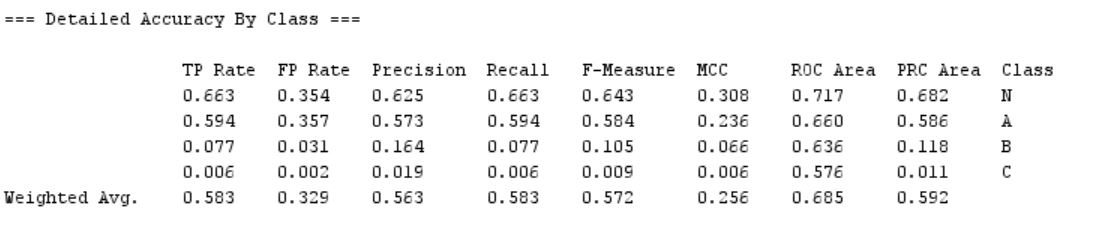
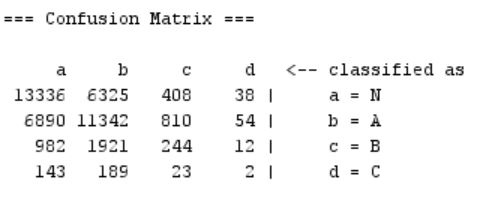


**Random Forest**

אלגוריתם שהינו supervised learning, יוצר יער באופן אקראי, כלומר עצי החלטה מרובים וממזג אותם יחד, לשם בעיות סיווג רגרסיה.

הרצנו את המודל עם הערכים הפרמטרים הדיפולטיבים מלבד מספר ה-Folds שהגדרנו ל-6 מאחר וקיבלנו שגיאה עבור חלוקה גדולה יותר.

תוצאות: B ROC = 0.636

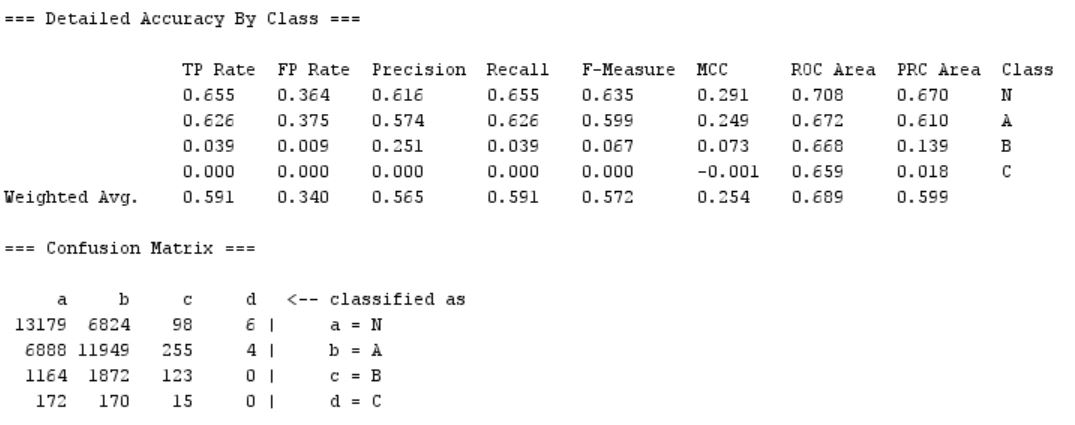
C ROC = 0.576  


**Bayes Net**

אלגוריתם זה מבוסס על ההנחות של האלגוריתם הבייסאיני הנאיבי. אך השוני בו הוא שאלגוריתם זה מתמודד עם תלות בין משתנים במידה וקיימת כזו.

תוצאות: B ROC = 0.668

C ROC = 0.659



**דיון ומסקנות:**

במחקר זה בחרנו את שלושת האלגוריתמים המתאימים ביותר לפתרון הבעיה הקיימת:  
1. J48

2. Bayesnet

3. RandomForest

נציג את השילוב של שלושת המודלים שבחרנו, בערכים דיסקרטים נבחרים על גבי גרף אחד, לצורך השוואה למציאת אלגוריתם הטוב ביותר לנקודת העבודה הנבחרת.

השוואה בין תוצאות האלגוריתמים השונים נעשתה לפי מדד ROC.  
לשם השוואה בחרנו את הערך TPR=  
בשני הגרפים יש להתמקד בהערך בציר Y (מסמל את ה-Sensitivity) . המטרה שלנו היא לתפוס כמה שיותר איחורים וזה מה שה-Sesitivity מסמל. נבחר נק' שהערך שלה בציר ה-Y יחסית גבוהה אך עם זאת לשים לב שהערך בציר ה-X לא יהיה גבוה מידי כי זה מסמל את הניחושים הלא נכונים.

איחורים מסוג B: (אלונה)

בגרף B המסמל את האיחור מסוג B נבחר את הנקודות Y=0.7 x=0.45 כלומר נרצה ש70% מהטיסות שאמרנו שיאחרו באמת איחרו ומכיוון שאיחור מסוג B הוא האיחור עם הכמות הגדולה משני האיחורים שבחרנו להתמקד נרצה לתפוס כמה שיותר איחורים מסוג זה.

הפיצויים לאיחורים מסוג זה עומדים באזור 300 דולר ונרצה להקטין את כמות הפיצויים שהחברה תשלם ואת הנזק שאיחור של טיסה יכול לגרום לנוסעים כמו לדוגמא לפספס טיסת המשך

.

**איחורים מסוג C: (איתי)**

כאמור, איחורים מסוג C הינם האיחורים הגורמים לנזק הכלכלי הגדול ביותר עבור חברות התעופה וגם עבור הלקוחות. מבדיקה שעשינו בשלב ה-Preprocess לחלוקת האיחורים לסוגים גילינו כי הפיצוי הכספי הממוצע בעבור איחור מסוג C הינו כ-400$, לעיתים נלווה לפיצוי זה לינה במלון בצירו, הסעות ואש"ל.   
ברצוננו לבחור ערך באלגוריתם Bayes Net בו ה-Sensitivity גבוה יחסית ל-TPR, זאת משום שאנחנו רוצים "לתפוס" כמה שיותר איחורים ואילו.  
על כן, נבחר FPR=0.35 שעבורו הרגישות הינה 0.6.

**הוספתי גם את החלק הראשון של העץ באלגוריתם J48 (גונן ביקש)**

**תוצאות**

**בחרנו 2 ערכים בהם ה-ROC היה הגבוה ביותרB,C.**

**בשני הגרפים המוצגים, האלגוריתם הטוב ביותר הינו Bayes Net. ההשוואה בין האלגוריתמים נעשתה באמצעות מדד הAUC- שמציין את השטח מתחת לגרף הROC. באלגוריתם Bayes Net שטח זה היה הגדול ביותר.** טבלה מסכמת את הAUC:

|  |  |  |  |
| --- | --- | --- | --- |
| **Bayes Net** | **Random Forest** | **J48** |  |
| **0.6683** | **0.6362** | **0.6422** | **B** |
| **0.659** | **0.5762** | **0.6053** | **C** |