מבוא ללימוד מכונה

364-1-1811

פרויקט חלק א' - קבוצה 15

WeatherAustralia Dataset



**צעדים עיקריים שנדב ביצע:**

1. כתיבת תוכן עניינים
2. הוספת פרק הגדרת הבעיה
3. הוספת סעיף מקור הנתונים
4. עריכה קלה של טבלה 1.2
5. עריכה קלה של הפרק על נתונים אפריורים
   1. הכנסת הנתונים לטבלה (לדעתי כדאי אבל אפשר לוותר אם לא אוהבות)
   2. וותור פירוט על העמודה - CloudsinJakarta, זו עיר באינדונזיה שהמתרגל הכניס בתור נתון סרק (לא מופיע בדאטה סט המקורי באינטרנט. לדעתי אין קשר בין עננות באינדוזניה לגשם באוסטרליה) – אבצע מבחן סטטיסטי כדי לוודא זאת.
6. הוספת ה-2 סנט שלי בפרק 2.2 – סט הנתונים מייצג את המציאות
7. פרק משתנים מסבירים – LOCATION. רוצה להתייעץ איתכן
8. עבודה נוכחית: כעת אני עובד כותב על פרק 2.5. מתכוון לבדוק סטטיסטית אם יש נתונים שלא משפיעים על משתנה המטרה. בדיקה דומה לבדיקה שהמתרגל הציע במייל אתמול.
9. לגבי חלוקת המשתנים לקטגוריות – אני נוטה לחשוב שכן צריך לבצע. במהלך היום אחדד את הבדיקה בנושא זה ואם צריך אכתוב קוד שעושה את זה

תוכן עניינים

[WeatherAustralia Dataset 1](#_Toc168498954)

[הגדרת הבעיה 3](#_Toc168498955)

[הבנת הנתונים: 4](#_Toc168498956)

[1. תיעוד מקורות הנתונים ומשמעותם 4](#_Toc168498957)

[1.1 מקור הנתונים במאגר: 4](#_Toc168498958)

[הנתונים במאגר הם למעשה מדידות יומיות שנמדדו ב-7 תחנות מטאורולוגיות שונות באוסטרליה לאורך תקופה ארוכה (10 שנים לפי התיאור המופיע בסט הנתונים המקורי באתר Kaggle). הנתונים נמדדו באמצעות חיישנים מסוגים שונים ומפורסמים לשימוש הציבור הרחב החל משנת 2010 באישור תאגיד המטאורולוגיה של אוסטרליה. 4](#_Toc168498959)

[1.2 משמעות המשתנים וסוגם: 4](#_Toc168498960)

[2. הסתברויות אפריוריות וקשרים בין מאפיינים: 5](#_Toc168498961)

[2.1 הסתברויות אפריוריות: 5](#_Toc168498962)

[2.2 סט הנתונים מאוזן ומייצג מציאות: 8](#_Toc168498963)

[2.3 קשרים בין מאפיינים: 9](#_Toc168498964)

[3. איכות הנתונים: 11](#_Toc168498965)

[3.1 נתונים חסרים: 11](#_Toc168498966)

[3.2 נתונים לא הגיוניים: 12](#_Toc168498967)

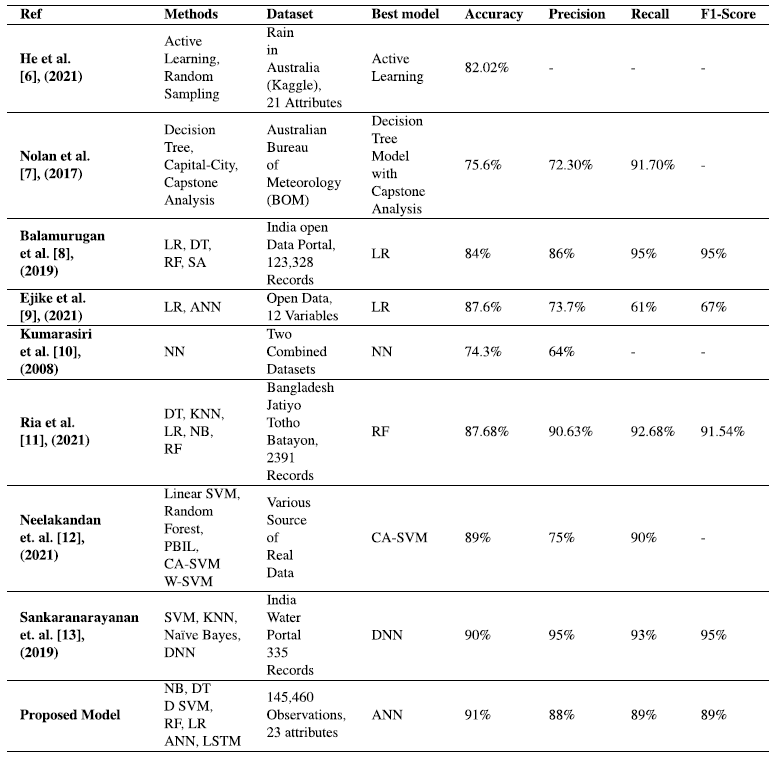
[הכנת הנתונים: 13](#_Toc168498968)

# הגדרת הבעיה[[1]](#footnote-1)

סט הנתונים שברשותנו מתאר מאפיינים שונים אודות מזג האוויר באוסטרליה שנאספו במספר תחנות מטאורולוגיות שונות לאורך עשור. הסט מכיל תצפיות יומיות של מספר רב של מאפיינים כגון: כמות גשם שירדה, מהירות הרוח שנמדדה, אחוזי לחות שנמדדו, טמפרטורה שנמדדה בזמנים שונים ביום בנוסף למשתנה מטרה בוליאני נוסף שמתאר האם ירד 1 מ"מ גשם או יותר ביום המדידה העוקב.

על ידי אימון מודלים של סיווג על נתונים אלו, ניתן לפתח מערכת שתחזה את הסבירות של ירידת גשם ביום הבא. המטרה היא לפתח מערכת אוטומטית שתוכל לעזור לאנשים להיערך למזג אוויר עתידי. למשל, לדעת אם כדאי לקחת מטריה או ללבוש מעיל גשם כשיוצאים מהבית.

מחקרים קודמים: מחקר בנושא פיתוח שיטות חיזוי מזג אוויר באמצעות כלים סטטיסטיים החל עוד בסוף המאה הקודמת כאשר שימוש בכלים מעולמות למידת המכונה החל להיחקר באופן נרחב מאמצע העשור הקודם. מאז פורסמו בנושא עשרות רבות של מאמרים. במסגרת כך, יושמו מגוון רחב של שיטות שהשתכללו לאורך השנים. למשל במאמר Machine Learning-Based Rainfall Prediction: Unveiling Insights and Forecasting for Improved Preparedness משנת 2023 הושוו מספר שיטות שיושמו על סט הנתונים – RainAustralia, סט דומה לסט אותו קיבלנו במסגרת הפרויקט.



ממעבר ראשוני על המאמר, ניתן לראות כי השיטות שהניבו אחוזי הצלחה גבוהים במיוחד כללו תהליך מקדים לניקוי המידע מרעשים וחריגות וכן, שילוב של שיטות שונות של למידה קלאסית (למשל שיטת בייס שעליה למדנו בהרצאה) ולמידה עמוקה (למשל שיטת LSTM).

# הבנת הנתונים

## תיעוד מקורות הנתונים ומשמעותם

### מקור הנתונים במאגר:

### הנתונים במאגר הם למעשה מדידות יומיות שנמדדו ב-7 תחנות מטאורולוגיות שונות באוסטרליה לאורך תקופה ארוכה (10 שנים לפי התיאור המופיע בסט הנתונים המקורי באתר Kaggle). הנתונים נמדדו באמצעות חיישנים מסוגים שונים ומפורסמים לשימוש הציבור הרחב החל משנת 2010 באישור תאגיד המטאורולוגיה של אוסטרליה.

### משמעות המשתנים וסוגם:

|  |  |  |  |
| --- | --- | --- | --- |
| שם המשתנה | משמעות המשתנה | סוג משתנה | טווח ערכים  בסט הנתונים |
| MinTemp | טמפרטורה מינימלית ביום (מעלות צלזיוס) | רציף | 5.8-30.2- |
| MaxTemp | טמפרטורה מקסימלית ביום (מעלות צלזיוס) | 4.1-46.4 |
| Rainfall | כמות הגשם שירד ביום (מ"מ) | 3-119.4- |
| Evaporation | אידוי מים ב-24 שעות עד 9 בבוקר (מ"מ)  (Class A pan evaporation) | 0-20.2 |
| Sunshine | מספר שעות השמש הבהירות ביום | 0-14.2 |
| WindSpeed9am | מהירות הרוח ממוצעת ב- 10 הדקות לפני 9 בבוקר (קמ"ש) | 2-69 |
| WindSpeed3pm | מהירות הרוח ממוצעת ב- 10 הדקות לפני 15 בצהריים (קמ"ש) | 2-65 |
| Humidity9am | לחות ב-9 בבוקר (אחוזים) | 7-105 |
| Humidity3pm | לחות ב-15:00 בצהריים (אחוזים) | 3-100 |
| Pressure9am | הלחץ האטמוספרי בגובה פני הים הממוצעים ב- 9 בבוקר (הקטו-פסקל) | 982.2-1040.4 |
| Pressure3pm | הלחץ האטמוספרי בגובה פני הים הממוצעים ב- 15 בצהריים (הקטו-פסקל) | 977.1-1038.9 |
| Temp9am | טמפרטורה ב-9 בבוקר (מעלות צלזיוס) | 0.7-36.5- |
| Temp3pm | טמפרטורה ב-15 בצהריים (מעלות צלזיוס) | 3.7-45.4 |
| WindGustSpeed | מהירות משב הרוח החזק ביותר ב-24 השעות עד חצות (קמ"ש) | 11-124 |
| Location | שם מיקום התחנה המטאורולוגית בה בוצעו המדידות | קטגוריאלי שלא ניתן לסידור | sydneyAirport Hobart  Melbourne  Sydney  Canberra  Bendigo  Uluru |
| WindGustDir | כיוון משב הרוח החזק ביותר ב-24 השעות עד חצות | בדיד | N,W,S,E |
| WindDir9am | כיוון משב הרוח ב-9 בבוקר | N,W,S,E |
| WindDir3pm | כיוון משב הרוח ב-15 בצהריים | N,W,S,E |
| Cloud9am | שבריר שמים מוסתר על ידי ענן בשעה 9 בבוקר (שמיניות) | קטגוריאלי שניתן לסידור | 0-8 |
| Cloud3pm | שבריר שמים מוסתר על ידי ענן בשעה 15 בצהריים (שמיניות) | 0-8 |
| CloudsinJakarta | סולם 1-7 של עננות בעיר ג'קרטה (אינדונזיה) ביום | 1-7 |
| RainToday | משתנה בוליאני המציין אם כמות המשקעים ב-24 השעות עד 9 בבוקר גדולה מ-1 מ"מ או לא | 0,1 |
| RainTomorrow | משתנה המטרה: משתנה בוליאני המציין אם כמות המשקעים ב-24 השעות העוקבות עד 9 בבוקר תהיה גדולה מ-1 מ"מ או לא | קטגוריאלי שניתן לסידור | 0,1 |

## 

## הסתברויות אפריוריות וקשרים בין מאפיינים:

### 2.1 הסתברויות אפריוריות:

|  |  |  |
| --- | --- | --- |
|  | | **משתנה המטרה**  **RainTomorrow**  האם יירד מחר גשם. משתנה זה הוא בינארי אשר מקבל את הערך 1 במידה וירד מחר יותר ממ"מ אחד של גשם ו-0 אחרת. המשתנה המוסבר, נרצה לראות את כמות הצפיות כתלות במשתנים השונים. ניתן לראות בתרשים העמודות 6727 תצפיות של ימים לא גשומים. תצפיות אלו מהוות 75% מכלל התצפיות. |
|  | תמונה שמכילה תרשים, טקסט, עלילה  התיאור נוצר באופן אוטומטי | **MinTemp**  טמפרטורה מינימלית שנמדדה ביום המדידה במעלות צלזיוס. משתנה רציף שערכו בסט הנתונים נע בין -5.8 ל-30.2. ערכו הממוצע הוא 11.7, סטיית התקן היא 5.6 והחציון הינו 11.5. הנתונים מתפלגים באופן סימטרי והממוצע קרוב לסטיית תקן -התפלגות נורמלית. |
|  |  | **Evaporation**  אידוי במ"מ ב-24 שעות האחרונות עד 9 בבוקר. משתנה רציף שערכו נע בין 0 ל- 20.2. ערכו הממוצע הוא 4.6, סטיית התקן היא 2.9 והחציון הינו 4. צורת ההיסטוגרמה מראה התפלגות נורמלית א-סימטרית, כאשר חלה ירידה חדה בתדירות אחרי המקסימום והזנב הימני מתמשך. |
|  |  | **Sunshine**  מספר שעות השמש הבהירות ביום. משתנה רציף שערכו בסט הנתונים נע בין 0 ל-14.2. ערכו הממוצע הוא 6.9, סטיית התקן היא 3.7 והחציון הינו 7.7. ההיסטוגרמה מתאימה להתפלגות נורמלית א-סימטרית עם שני נתונים מקסימליים ב-0 וב-10 שעות שמש. |
| תמונה שמכילה טקסט, צילום מסך, תרשים, קו  התיאור נוצר באופן אוטומטי |  | **WindSpeed15am**  מהירות הרוח ממוצעת ב- 10 הדקות לפני 15 בצהריים (קמ"ש). משתנה רציף שערכו נע בין 2 ל-65. ערכו הממוצע הוא 21.1, סטיית התקן היא 9.1 והחציון הינו 20. ההתפלגות מראה כי המהירות הנפוצה ביותר היא 20 קמ"ש. ההתפלגות נראית נורמלית מוטה מעט ימינה מה שמראה שמהירויות גבוהות נפוצות פחות בשעה 15:00. |
| תמונה שמכילה צילום מסך, טקסט, מלבן, קו  התיאור נוצר באופן אוטומטי |  | **Humidity15am**  לחות (אחוזים) בשעה 15:00. משתנה רציף שערכו נע בין 3 ל-100. ערכו הממוצע הוא 52.06, סטיית התקן היא 17.9 והחציון הינו 52. ההתפלגות נראית סימטרית ונורמלית והממוצע קרוב לסטיית תקן עם הטייה קלה לימין. |
| תמונה שמכילה טקסט, צילום מסך, מלבן, תרשים  התיאור נוצר באופן אוטומטי |  | **Pressure15am**  הלחץ האטמוספרי בגובה פני הים הממוצעים בשעה 15 בצהריים (הקטו-פסקל). משתנה רציף שערכו נע בין 977.1 ל-1038.9. ערכו הממוצע הוא 1015.2, סטיית התקן היא 7.88 והחציון הינו 1051.6. ההתפלגות נורמלית והממוצע קרוב לסטיית התקן עם זנב קל שמאלה. |
|  |  | **Temp15am**  טמפרטורה שנמדדה בשעה 15 בצהריים (מעלות צלזיוס). משתנה רציף שערכו נע בין 3.7 ל-45.4. ערכו הממוצע הוא 19.7, סטיית התקן היא 5.9 והחציון הינו 19.3. ניתן לראות כי ההתפלגות נורמלית בעלת זנב ימני קל- טמפרטורות של מעל 40 מעלות לא נפוצות. |
| תמונה שמכילה מלבן, צילום מסך, קו, טקסט  התיאור נוצר באופן אוטומטי |  | **WindGustSpeed**  המהירות של משב הרוח החזק ביותר ב-24 השעות עד חצות (קמ"ש). משתנה רציף שערכו נע בין 11 ל-124. ערכו הממוצע הוא 44.3, סטיית התקן היא 15 והחציון הינו 43. התפלגות נראית נורמלית מוטה ימינה. |
|  | | **Location**  שם מיקום תחנת מזג האוויר בה בוצעו המדידות. משתנה קטגוריאלי שלא ניתן לסידור. הקטגוריות: sydneyAirport, Hobart , Melbourne, Sydney, Canberra, Bendigo, Ulur. ניתן לראות שמספר הדגימות הגדול ביותר בוצע בשדה התעופה בסידני עם 2202 תצפיות ומספר המדידות הנמוך ביותר בוצע בעיקר אולורו עם 178 תצפיות. |
|  | | **WindDir3pm**  כיוון משב הרוח בשעה 15. משתנה קטגוריאלי שלא ניתן לסידור (אין סדר טבעי בין הקטגוריות). הקטגוריות:N ,W,S,E. ניתן לראות כי בצהרי היום הרוח נושבת יותר לכיווני צפון ודרום. |
|  | | **Cloud3pm**  שבריר שמים מוסתר על ידי ענן בשעה 13 בצהריים (שמיניות). משתנה קטגוריאלי שניתן לסידור. קטגוריות - 0-8. רוב התצפיות מעידות על עננות יחסית גבוהה (7) בשעה 15 או עננות נמוכה מאוד (1) |
|  | | **RainToday**  משתנה בינארי המתאר אם כמות המשקעים ב-24 השעות ביום המדידה עד 9 בבוקר עולה על 1 מ"מ. אם כן אז ערכו 1, אחרת 0. ניתן לראות כי כמות התצפיות של ימים לא גשומים מהווה 74% כמו במשתנה המוסבר מה שיכול להעיד על קשר חזק בניהם. |

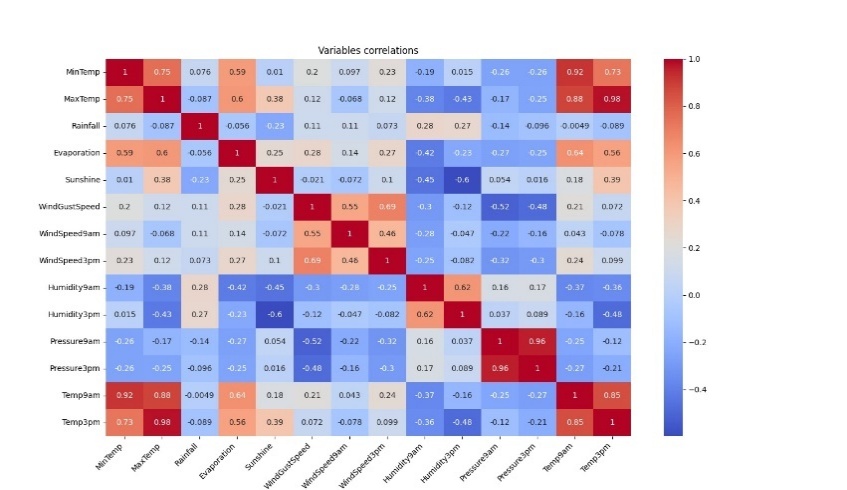
### 2.2 סט הנתונים מאוזן ומייצג מציאות:

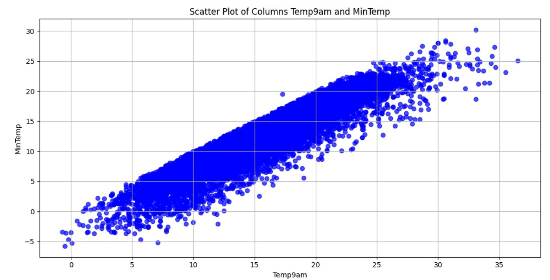
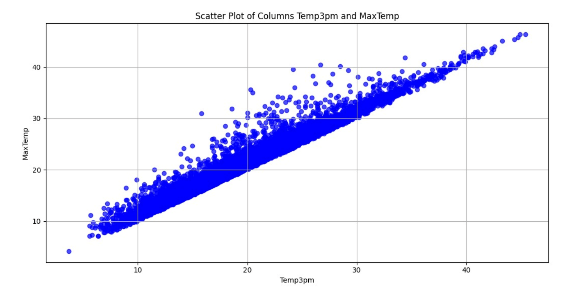
1. **התפלגות הנתונים הרציפים:** ניתן לראות כי הנתונים הרציפים מתפלגים באופן כללי בצורה נורמלית. לכן ניתן לומר באופן כללי כי הנתונים הללו מייצגים את המציאות. חלק מההתפלגויות מכילות זנבות ותצפיות חריגות, יש לבחון בשלב הבא של הפרויקט האם ניפוי תצפיות אלו מסט האימון משפר את תוצאות הלימוד.

2. **כמות תצפיות בקטגוריות:** כמות התצפיות במשתנה Location לא מאוזנת בין תחנות המדידה השונות. כמו כן, נראה שהתחנות לא מכסות את כלל באיזורים באוסטרליה היות ומדובר ביבשת גדולה יחסית. כמו כן, כמות התצפיות בין איזורים גשומים יותר ואיזורים גשומים פחות לא מאוזנת.

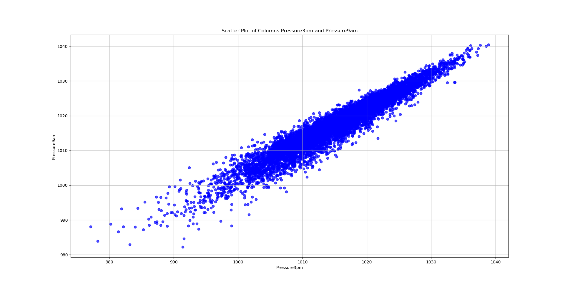
3. **נתונים מוטים והעדר נתונים**: יש נתונים לא הגיוניים וחסרים במשתנים שונים בסט הנתונים. על מנת שהנתונים יהיו מאוזנים יותר ויסבירו את המציאות בצורה טובה יותר יש צורך לטפל בחריגים אלו בשלב הכנת הנתונים.

### 2.3 קשרים בין מאפיינים:

ניתן לראות במפת החום הבאה את הקשרים בין המשתנים הרציפים.

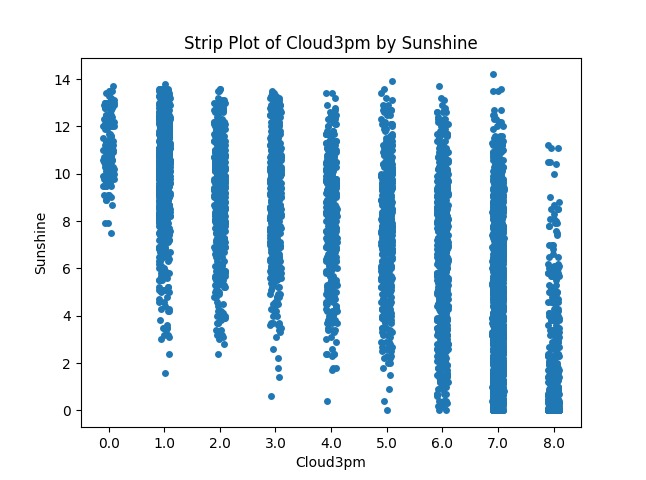
**הקשר בין MinTemp לTemp9am ובין MaxTemp לTemp3pm:** ניתן לראות בגרפי הפיזור כי יש קשר לינארי עולה בין שני המשתנים. בשעות הבוקר הטמפרטורות יחסית נמוכות, לכן ניתן להסיק כי קיים קשר בין הטמפרטורה המינימלית ביום לבין הטמפרטורה בשעות הבוקר. הטמפרטורות בשעות הצהריים גבוהות יחסית לשאר שעות היום לכן ניתן לראות קשר בין הטמפרטורה בשעה בצהריים 15:00 לבין הטמפרטורה המקסימלית ביום. בשני הגרפים ניתן לראות כי ככל שהטמפרטורה באותו חלק ביום עולה גם הטמפרטורה המינימלית/מקסימלית עולה.

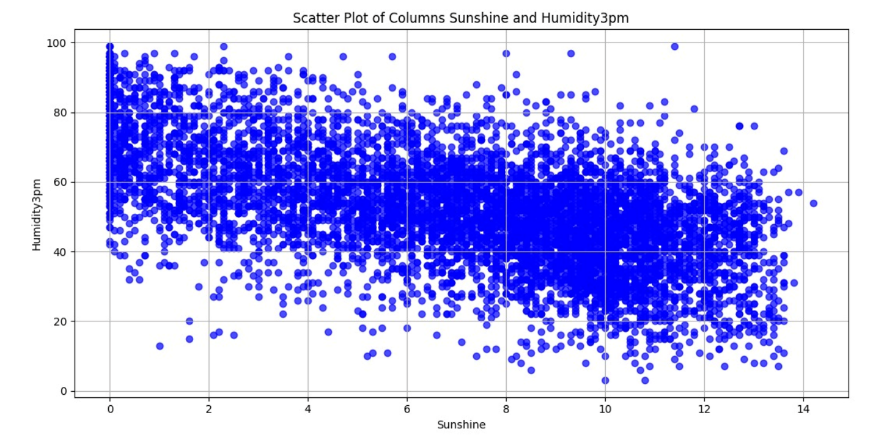
**הקשר בין Pressure3pm ו Pressure9am :** ניתן לראות כי קיימת קורלציה חזקה בין הלחץ האטמוספרי הנמדד ב-9 בבוקר והלחץ האטמוספרי הנמדד ב-15 בצהריים.

****

תמונה שמכילה טקסט, תרשים, עלילה, קו

התיאור נוצר באופן אוטומטי**הקשר בין WindGustSpeed לWindSpeed3pm:** המהירות (קמ"ש) של משב הרוח החזק ביותר ב-24 השעות ולבין המהירות בשעה 15:00 נמצאות בקשר לינארי עולה. ככה שמהירות רוח המקסימלית במהלך היום עולה גם המהירות ב15:00 עולה.

**הקשר בין Cloud3pm לSunshine:** בגרף הפיזור ניתן לראות את הקשר בין המשתנה הקטגוריאלי של שבריר שמים מוסתר על ידי ענן (שמיניות) בשעה 9 בבוקר לבין מספר שעות השמש הבהירות ביום. מהגרף ניתן לראות כי ככל שהשמיים מעוננים יותר כך שעות השמש הבהירות יורדות.

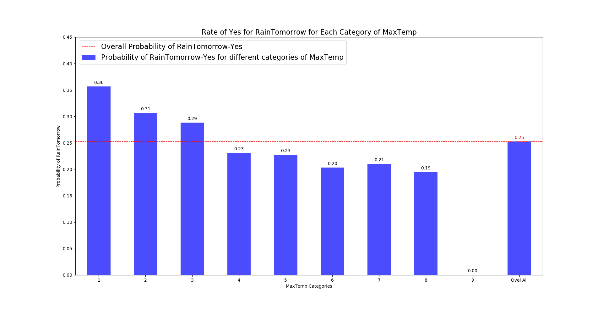
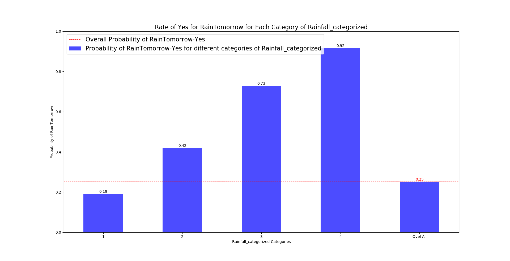
****

**הקשר בין Sunshine לHumidity3pm:**

בגרף הפיזור ניתן לראות קשר בין שני המשתנים הרציפים של כמות שעות השמש הבהירות ביום לבין הלחות בשעה 15:00. מהגרף ניתן לראות כי ככל שכמות שעות השמש הבהירות גבוה יותר כך הלחות באותו יום הייתה נמוכה יותר-כלומר ככל שיש שמש חזקה לאורך יותר זמן ביום מזג האוויר יבש יותר.

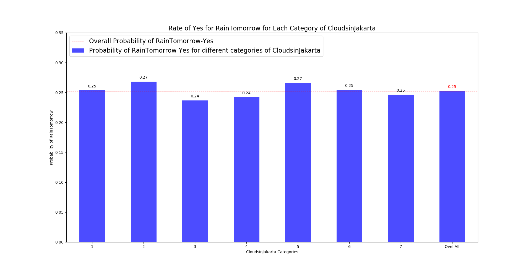
2.4 המאפיינים שניתן לחשוד בהם כבעלי השפעה על משתנה המטרה:

עבור מספר מאפיינים לדוגמא, נראה את גרף ההסתברות המותנית של הסיכוי לירידת גשם ביום העוקב כתלות בערך המשתנה. על מנת להמחיש את התוצאה בצורה טובה חילקנו את ערכי המשתנים ל-10 קטגוריות מהערך הנמוך ביותר לערך הגבוה ביותר.

* **MinTemp וMaxTemp-** טמפרטורות מינמליות ומקסימליות יכולות להשפיע על דפוסי מזג האוויר וירידת גשם. בגרף הבא חילקנו את הטמפרטורה המקסימלית שנמדדה ל-8 קטגוריות. ניתן לראות כי ההסברות ליום גשום ביום העוקב עולה ככל שהטמפרטורה המקסימלית ביום המדידה הייתה קטנה יותר.
* **Rainfallו-RainToday-** במערכות מזג אוויר ישנה המשכיות עונתית - במידה ויש תחזית לגשם ביום מסוים עשוי להיות שגם למחרת נראה גשם (שעשוי להמשך גם למספר ימים נוספים). בגרף הבא חילקנו את את כמות המשקעים שירדה ביום המדידה ל-4 קטגוריות. ניתן לראות באופן כללי שהסיכוי לרידת גשם ביום העוקב עולה ככל שכמות הגשם שירדה ביום הנוכחי גדולה יותר.
* **Sunshine-** שעות השמש הבהירות יכולות להשפיע על מזג האוויר בקשר הפוך- ככל שמספר שעות השמש הבהירות נמוך יותר כך קיימת יותר עננות שתגרום לירידת גשם. כלומר, סביר להניח שמשתנה זה הינו בקורלציה הפוכה עם ירידת גשם; פחות שעות שמש לרוב מעידות על עננות וסיכוי לגשם.
* **WindGustSpeed, WindSpeed3pmוWindSpeed9am-** ככל שמהירות הרוח עולה היא נושאת לחות אל אזורים יבשים ויכולה לגרום לגשמים.
* **Humidity9am וHumidity3pm-** כאשר יש לחות גבוהה באוויר מתקיימים תהליכי אידוי והתפתחות עננים ובטמפרטורות נמוכות יש סיכוי לגשם
* **Pressure3pm וPressure9am-** לחץ נמוך מאפשר לאוויר לעלות למעלה ולהתעבות לאדי מים. התעבות מתמשכת יכולה להוביל לגשם.
* **Temp9am וTemp3pm-** כאשר הטמפרטורה נמוכה בדר"כ הסיכוי לגשם עולה.
* **Location-** למשתנה מיקום יש השפעה על ירידת גשם- יש אזורים גשומים יותר ופחות. באוסטרליה יש מגוון רחב של אקלימים, חלקים מסוימים גשומים פחות (קאנברה, בנדיגו ואולורו) ואחרים גשומים יותר (סידני, הוברט ומלבורן).
* **Cloud3pm וCloud9am-** עננות גבוהה יכולה להעיד על סיכוי גבוה לגשם.
* **Evaporation-** התאדות לבדה אינה מספיקה כדי לגרום לגשם. נדרשים גם תנאים של טמפרטורה נמוכה שיגרמו לעיבוי ולגשם.

2.5 יש שני סוגי משתנים שהיינו רוצים להסיר:

1. משתנים שאינם משפיעים על משתנה המטרה: משתנים אלו עלולים להוסיף רעש למודל הלימוד וכתוצאה מכך להאריך את זמן הלימוד ולהפחית את דיוק ניבוי משתנה המטרה. אנו חושדים שמידת העננות בעיר ג'קורה לא משפיעה על משתנה המטרה בפרויקט. על מנת להוכיח נקודה זו נצייר את גרף ההסתברות המותנית של הסיכוי לירידת גשם ביום העוקב בהינתן מידת העננות בג'קורה - כאשר בעמודה הימנית ביותר צירפנו את הסיכוי הממוצע לירידת גשם בכל התצפיות יחד.



מהגרף ניתן לראות כי הסיכוי לירידת גשם ביום העוקב קבוע בקירוב כתלות במידת העננות בעיר ג'קורה. לפיכך, ניתן להסיק כי מידת העננות בעיר ג'קורה לא משפיעה על הסיכוי לירידת גשם.

1. משתנים עם קורלציה חזקה: זוגות משתנים עם קורלציה חזקה בניהם (חיובית או שלילית) מספקים את אותו מידע למודל ולכן ניתן להסיר אחד מהם מהליך הלימוד בלי לפגוע בביצועים. הסרת מידע כפול מסט הלימוד מפשטת את המודל ומקלה על פרשנות תוצאות החיזוי. במסגרת כך נשקול להוריד משתנה אחד מכל זוג:
   * MinTemp},{Temp9am
   * {MaxTemp, לTemp3pm}
   * { Pressure9am,Pressure3pm}

קיימים זוגות משתנים נוספים בעלי קורלציה פחות מובהקת. למשל: מידת העננות בשעה 15 בצהריים וכמות שעות השמש הבהירות (פרטנו על כל הזוגות בפרק הקודם). בשלב הבא של הפרויקט נרצה לבחון כיצד הסרה של אחד מבני הזוג משפיעה על סיכויי החיזוי של משתנה המטרה. במקרה זה אנו סבורים כי בגלל שהקורליציה לא מובהקת אז שני המשתנים יתרמו לסיכויי הניבוי.

## איכות הנתונים:

### 3.1 נתונים חסרים:

במסגרת הניתוח המקדים של הנתונים נמצאו הרשומות הריקות הבאות סה"כ מתוך 9000 ימי מדידה עליהן סט הנתונים מבוסס:

|  |  |  |  |
| --- | --- | --- | --- |
| שם המשתנה | כמות תצפיות חסרות | שם המשתנה | כמות תצפיות חסרות |
| MaxTemp | 1 | Pressure9AM | 1 |
| Rainfall | 1 | Pressure3PM | 1 |
| Evaporation | 1239 | Cloud9AM | 1 |
| Sunshine | 1850 | Cloud3PM | 1 |
| WindGustSpeed | 1 | Temp9AM | 1 |
| WindSpeed9AM | 1 | Temp3PM | 1 |
| WindSpeed3AM | 1 | Pressure9AM | 1 |
| Humidity9AM | 1 | Pressure3PM | 1 |
| Humidity3PM | 1 |  |  |

את התצפיות החסרות מילאנו בהסתברות המותנת מומצעת של המשתנה כתלות במשתנה המטרה.

### 3.2 נתונים לא הגיוניים:

**Rainfall:** כמות הגשם שירד ביום במ"מ חייבת להיות כמות חיובית או אפסית. ניתן לראות כי בסט הנתונים יש תצפית בה נמדד ערך שלילי של כמות הגשם שלא הגיונית -3. בדקנו את הערך של RainToday באותה תצפית וראינו כי הערך של המשתנה הבינארי היה 0, כלומר ירדו פחות מ-1 מ"מ משקעים ולכן בחרנו לשים את הערך 0 במשתנה זה בתצפית זו.

כמו כן, נבדקה התאמת הנתונים בין RainfallלRainToday-. ציפינו לראות כי כאשר כמות הגשם גדולה מ-1 מ"מ אז הערך הבינארי המתאים ב- RainToday הוא. הבדיקה שערכו העלתה כי הנתונים מתאימים והגיוניים.

**תמונה שמכילה טקסט, צילום מסך, תרשים, מלבן

התיאור נוצר באופן אוטומטיתמונה שמכילה טקסט, צילום מסך, תרשים, מלבן

התיאור נוצר באופן אוטומטיCloud9AM:** משתנה קטגוריאלי שערכו בין 0-8. ראינו תצפית חריגה שבה הערך הוא 999. בדקנו את התצפית וראינו כי המדידה של התצפית התרחשה ב-Hobart. כמו כן, באותה תצפית נראה שלא ירד גשם. בשל הקשר בין המשתנים Cloud9AM ו - Cloud3PM בדקנו מה הערך הנפוץ ביותר בתצפיות הדומות לתצפית בעלת הערך החסר. בתצפיות מ-hobart ללא גשם ניתן לראות שהערכים של Cloud3PM בין 5-7, ושמעל ל-2/3 תצפיות מתאימות לכך שגם Cloud9AM בין 5-7. רוב הערכים בתצפיות הנמדדות הן 7 ולכן בחרנו לשים בערך Cloud9AM את הערך 7.

**WindDir3pm**: נצפה ערך חריג של zzzzzzzzzz…... מבדיקתנו התצפית מ-Canberra ביום ללא גשם. כיוון הרוח ב WindDir9am הוא N ומבדיקתנו נראה כי יש התאמה כלשהי בין WindDir9am ל .WindDir3pm כאשר WindDir9am הוא לכיוון צפון 56% מהערכים של WindDir3pm הם גם לכיוון צפון לכן בחרנו בערך N.

# הכנת הנתונים:

* 1. השמטת מאפיינים "רועשים" או חסרי חשיבות (נדב)
  2. . השמטת מאפיינים בעלי איכות נמוכה מדי )שגיאות, ערכים חסרים וכו'(. (נדב)
  3. השמטת תצפיות בעלות חוסר רב (עדן)
  4. השלמת מושכלת של ערכים חסרים במידה ואפשרי. (עדן)

2 על פי הצורך, תנו טיפול פרטני במאפיינים (נדב)

• דיסקרטיזציה של משתנים רציפים, למשל בצורה שמייצגת את ההתפלגות או

בצורה שמייצגת דרישות של עולם התוכן.

o דוגמה: משתנה רציף "גיל" ניתן להפוך למשתנה בדיד ע"י הגדרת

סיפים כגון ,6 ,18 ,21 65 וכו'.

• גזירת מאפיינים חדשים )פונקציות של משתנים קיימים(.

1. חלק מההסבר מבוסס על הסבר מאתר Kaggle - עבור DataSet : Rain Australia. הנחנו באופן סביר שמדובר באותו DataSet. [↑](#footnote-ref-1)