חיזוי מזג אוויר ליום הבא בעיר מדריד ע"פ נתונים היסטריים

משתתפים

ברק זן, 305634487, barakzan@campus.technion.ac.il

אופיר שומרון, 201450574, ofirshomron@campus.technion.ac.il

מנחה: ליאון ורדי.

מרצה:

מספר קורס: 236502.

מבוא

בעיית חיזוי מזג האוויר ידועה בתור בעיה לא פשוטה. על מנת לחזות את מזג האוויר חזאים ואנשי מטאורולוגיה משתמשים במודלים מטאורולוגים מסובכים ביותר המתבססים על מידע רב כמו: שקעים ברומטריים, כיווני רוחות, זרמים בים ומקורות מידע מורכבים נוספים. בנוסף לכך, עומדים לרשותם שנים של ניסיון וידע שנצבר בתחום.

אנו רוצים לבדוק, האם בעזרת בינה מלאכותית ושימוש במידע זמין ופשוט כמו מדידות יומיות בתחנה מטאורולוגית – ניתן לחזות את הטמפרטורה הממוצעת ביום הבא בצורה מהירה ומדוייקת יחסית.

לצורך נקודת ייחוס, בנינו מודל חיזוי שחוזה את הטמפרטורה של היום הקודם. במהלך הפרויקט ננסה לבנות חזאי העולה לפחות על מודל זה.

תוצאות החיזוי של חזאי זה על דוגמאות ממאגרי המידע של הערים מדריד ואוסטין, בהם השתמשנו:

מדריד –

- חיזוי מדוייק – 24.05%.
- חיזוי עם סטייה מקסימלית של ± 1 מעלות – 58.07%.
- חיזוי עם סטייה מקסימלית של ± 2 מעלות – 79.15%.
- חיזוי עם סטייה מקסימלית של ± 3 מעלות – 91.38%.

אוסטין –

- חיזוי מדוייק – 21.7%.
- חיזוי עם סטייה מקסימלית של ± 1 מעלות – 52.5%.
- חיזוי עם סטייה מקסימלית של ± 2 מעלות – 68.74%.
- חיזוי עם סטייה מקסימלית של ± 3 מעלות – 78.38%.

תיאור מפורט של הפתרון המוצע לבעיה

הפתרון חולק למספר שלבים עיקריים:

איסוף המידע:

- השגת מדדים מטאורולוגיים נבחרים מה-20 שנה האחרונות בתחנת מדידה מסויימת בעיר מדריד (ספרד).
- השגת מדדים מטאורולוגיים נבחרים מה-4 שנים האחרונות בתחנת מדידה מסויימת בעיר אוסטין (טקסס) על מנת לבצע ניסויים בהמשך.

מבנה המידע עליו אנו מתבססים:

עבור כל אחת מהערים, השתמשנו בטבלה עם המאפיינים הבאים עבור כל יום – טמפרטורה מקסימלית, טמפרטורה ממוצעת, טמפרטורה מינימלית, נקודת הטל, נקודת הטל ממוצעת, נקודת הטל מינימלית, לחות מקסימלית, לחות ממוצעת, לחות מינימלית, לחץ מקסימלי בגובה הים, לחץ ממוצע בגובה הים, לחץ מינימלי בגובה הים, ראות מקסימלית, ראות ממוצעת, ראות מינימלית, מהירות רוח מקסימלית, מהירות רוח ממוצעת, משקעים אטמוספריים, כיסוי עננים וכיוון הרוח.

הכנת המידע לעיבוד ראשוני:

- הסרת מדדים שהמידע בהם חסר או לא עקבי מספיק.
- במדדים בהם המידע היה חסר באופן חלקי מאוד – השלמנו את המידע לפי שתי שיטות – העתקת המידע מהיום הקודם, ולפי אלגוריתם closest fit .

אגריגציה ועיבוד ראשוני של המידע:

יצירת מאפיינים נוספים המבוססים על המידע הקיים. לכל מאפיין במערכת יצרנו את המאפיינים הבאים:

- שורש ממוצע הריבועים של המאפיין בX ימים האחרונים.
  נוסחא : $RMS = \sqrt{\frac{\sum_{i=1}^{x} f_i^2}{x}}$
- ריבוע ממוצע השורשים של המאפיין בX ימים האחרונים.
  נוסחא : $SMR = \left(\frac{\sum_{i=1}^{x} \sqrt{f_i}}{x}\right)$
- ממוצע המאפיין בX ימים האחרונים.
  נוסחא : $MEAN = \frac{\sum_{i=1}^{x} f_i}{x}$

כאשר X מיצג את מספר הימים אחורה שהמאפיין החדש נבנה על פיהם (בדקנו ל1 עד 14 ימים אחורה).

חלוקת המידע:

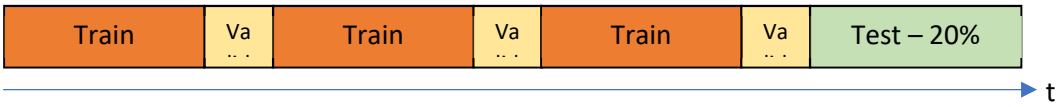
כיוון שאנו בונים מערכת המבצעת חיזוי של אירוע עתידי על סמך העבר – חשוב לשים דגש על חלוקה נכונה של הדוגמאות וסדר כרונולוגי נכון בין קבוצות המבחן, ההערכה והאימון. לכן, כאשר חילקנו את הדוגמאות – סט האימון היה מורכב מ60% הקדום ביותר (כרונולוגית) של המדידות, סט ההערכה היה ה20% שבאו אחריהם, וקבוצת המבחן הייתה מורכבת מ20% המדידות המאוחרות ביותר.

אימון מודל החיזוי:

המודלים בהם השתמשנו לבניית החזאים הם עץ החלטה, יער החלטה רנדומלי, KNN, רגרסיה לינארית. עבור כל חזאי ביצענו כיוונון פרמטרים כללי הרלוונטי לכל החזאים, הכולל בחירת מספר הימים אחורנית בו החזאי משתמש, אופן השלמת המידע, בחירת אגריגציה.

בנוסף לכל חזאי התבצע כיוונון פרמטרים בהתאם למודל. למשל, למודל KNN כוונן מספר השכנים, לעץ החלטה וליער כוונן עומק העץ, מספר העלים, בסיס סטטיסטי, וכו.

בתחילה, אומנו החזאים בשיטה הסטנדרטית, אימון על סט האימון ובדיקה על סט הולידציה. לאחר קבלת תוצאות, כדי למנוע במקרים מסויימים התאמת יתר של החזאי רצינו להתשמש ב cross validation. אך, מכיוון שאנו בונים חזאי אשר יש בו חשיבות לסדר הכרונולוגי של המדידות, לא ניתן להשתמש באלגוריתם הסטנדרטי. במקום זאת, חילקנו את הסטים אימון+ולידציה ל-K חלקים שכל אחד מהם מכיל חלק אימון ולאחריו מיד חלק ולידציה, למשל עבור K=3 נקבל את השרטוט הבא.



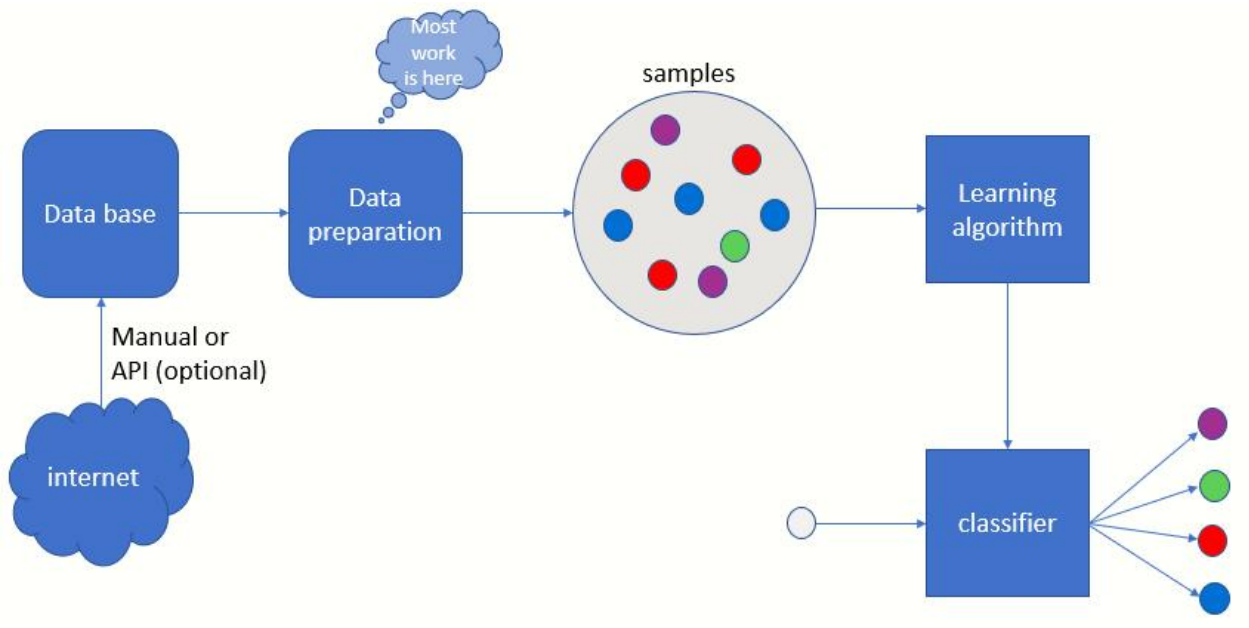
הערכת החזאים:

הערכת טיב החזאים נעשתה לפי חיזוי הטמפרטורה הממוצעת ביום הבא על פי ארבע רמות דיוק

- חיזוי מדוייק.
- חיזוי עם סטייה מקסימלית של ± 1 מעלות.
- חיזוי עם סטייה מקסימלית של ± 2 מעלות.
- חיזוי עם סטייה מקסימלית של ± 3 מעלות.

בשלב כוונון הפרמטרים הערכת החזאים התבצעה על פי התוצאות שלהם על סט הולידציה, כאשר ב cross validation משתמשים בממוצע התוצאות. ושלב ההערכה הסופי הערכת החזאים התבצעה על פי התוצאות שלהם על סט המבחן, בו לא נתקלו החזאים מעולם.

תיאור המערכת ששימשה למימוש הפתרון



באופן מעשי, המערכת התחלקה לשלושה חלקים עיקריים:

<u>הכנת המידע:</u>

זה החלק העיקרי והמשמעותי ביותר בפרוייקט, בחלק זה אנחנו לוקחים את המידע הגולמי שמצאנו באינטרנט, מורידים תכונות לא רלוונטיות על פי ידע אישי או ידע של מומחים, משלימים מידע חסר באחת משתי השיטות שבחנו, מבצעים אגריגציה ליצירת מאפיינים חדשים. שלב זה התבצע על המידע שאספנו על העיר מדריד, שבה אנו רוצים לחזות את המזג אויר. ובנוסף, על המידע של העיר אוסטין, בו נשתמש לניסויים עתידיים.

באופן מעשי, כדי לעשות את כל התהליך אוטומטי, בנינו מודל המקבל את הפרמטרים שיטת השלמת מידע, מאפיינים חדשים אותם יש ליצור ומספר הימים אחורנית עליו יתבסס החיזוי. לאחר מכן בזמן האימון, ניתן להכניס למודל את הפרמטרים כדי לבחור את סט המידע המתאים.

<u>אימון מודל החיזוי:</u>

בשלב זה, אימנו את כל החזאים שאותם רצינו לבדוק וביצענו את כיוונון הפרמטרים. כלומר, אימנו כל חזאי מספר פעמים על כל סט פרמטרים (הספציפי לחזאי) , כל סט דגימות ובשתי שיטות האימון.

<u>הערכת החזאים:</u>

הערכת טיב החזאים נעשתה לפי חיזוי הטמפרטורה הממוצעת ביום הבא על פי ארבע רמות דיוק

- חיזוי מדוייק
- חיזוי עם סטייה מקסימלית של ± 1 מעלות.
- חיזוי עם סטייה מקסימלית של ± 2 מעלות.
- חיזוי עם סטייה מקסימלית של ± 3 מעלות.

בשלב כוונון הפרמטרים הערכת החזאים התבצעה על פי התוצאות שלהם על סט הולידציה, כאשר ב cross validation משתמשים בממוצע התוצאות. ושלב ההערכה הסופי הערכת החזאים התבצעה על פי התוצאות שלהם על סט המבחן, בו לא נתקלו החזאים מעולם.

מתודולוגיה ניסויית

הדרך הכללית שבה ערכנו את הניסויים היא:

לאחר שלב הכנת המידע, אימנו מספר חזאים שונים, לפי המידע אותו רצינו להעריך ולהשוות. המסווגים נבדלים ביניהם במידע עליו הם אומנו (עיר\ערים, מספר ימים, שיטת השלמת מידע), סוג המסווג (עץ, יער, KNN, רגרסיה לינארית).

לאחר אימון המסווגים, בדקנו את תוצאות החיזוי של כל המסווגים על הפרמטרים שאותם רצינו לחזות. הפרמטרים לחיזוי הם הטמפרטורה הממוצעת ביום למחרת בסטיה של [0-3] מעלות.

לאחר חישוב תוצאות החיזוי, השווינו בין התוצאות על מנת לקבל את המידע אותו רצינו לגלות בכל ניסוי. בנוסף בדקנו האם הניסוי עמד בציפיות שלנו.

במהלך הניסויים, עלו שאלות ותהיות איך ניתן להגיע לתוצאות טובות יותר. כך הגענו לניסויים חדשים בהם אנו בודקים שיטות אחרות של עריכת המידע וכוונון פרמטרים.

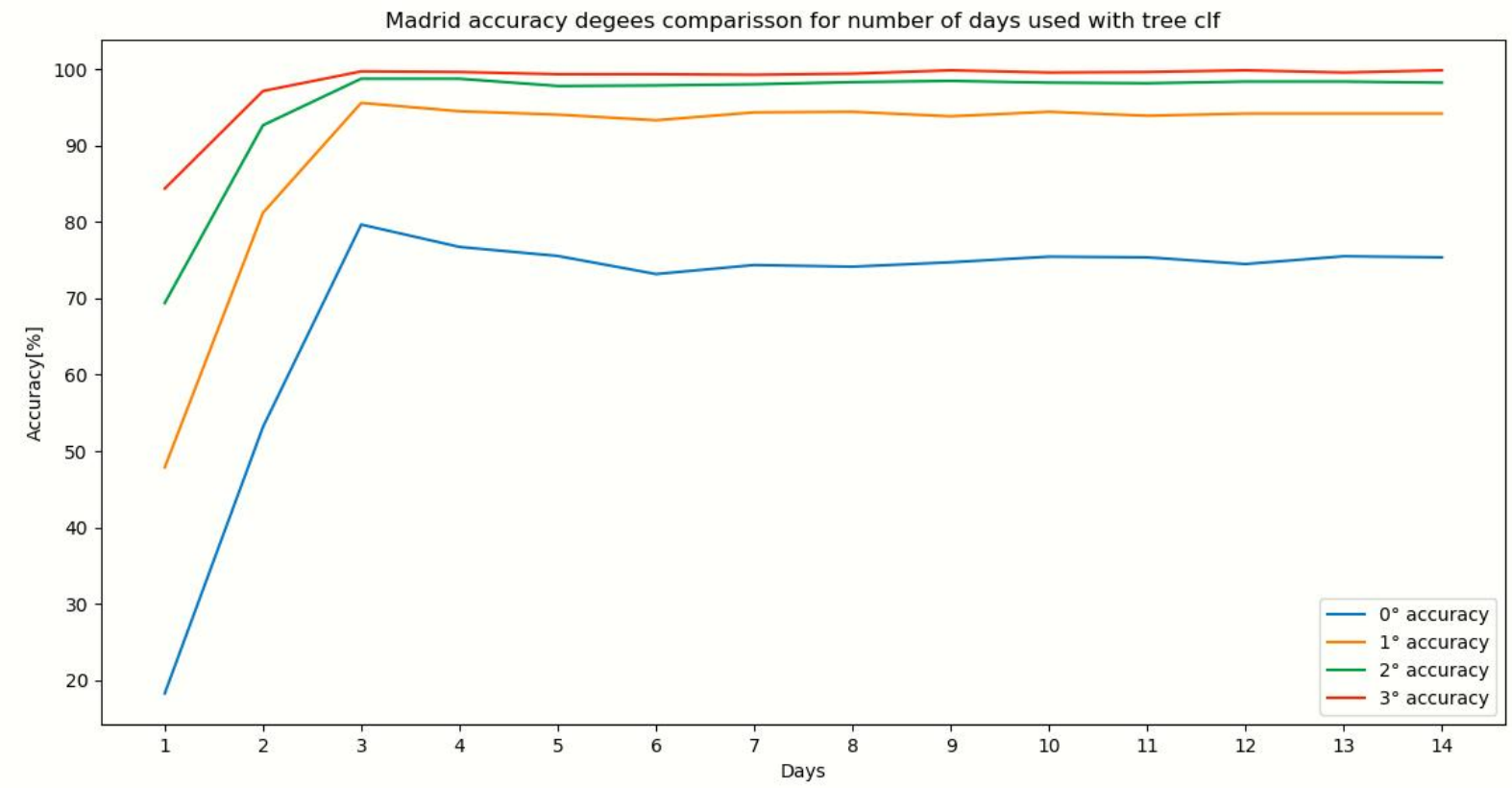
תיאור הניסויים תוצאות ומסקנות

ניסוי ראשון: כמות המידע שהמסווג צריך על מנת לתת חיזוי איכותי

תיאור הניסוי: מטרת ניסוי זה היא בדיקת השפעת מספר הימים שעליהם מתבססים בהרכבת המידע על איכות החיזוי.

בניסוי זה השוונו בין מספר חזאים הנבדלים ביניהם במספר הימים אחורנית שעליהם התאמנו.

בגרף מוצגים ארבעה קווים, אחד לכל רמת דיוק של המסווג – כתלות במספר הימים אחורנית.



מסקנות: בכל רמות הדיוק, ניתן לראות בבירור שהתוצאות הטובות ביותר מתקבלות כאשר מסתכלים שלושה ימים אחורה.

כאשר הסתכלות רק יום או יומיים אחורה אינה מביאה מידע מספק ומביאה תוצאות גרועות משמעותית.

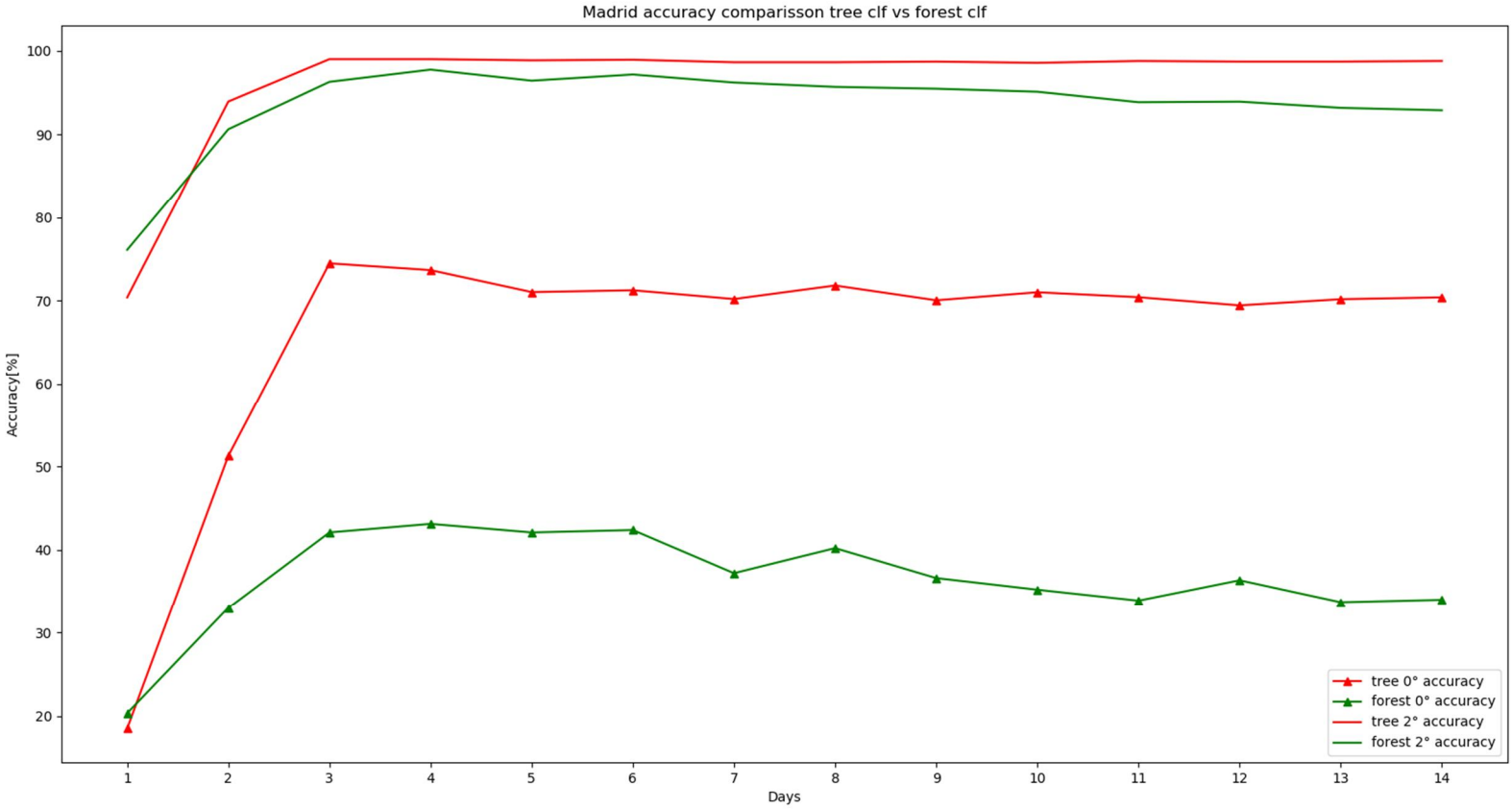
ואילו הסתכלות של יותר מ3 ימים אחורנית אינה מועילה ואף מוסיפה רעש שלעיתים פוגע קלות בביצועים. להערכתנו, מצב זה מתרחש עקב תוספת תכונות שאינן נבדלות יותר מדי אחת מהשניה.

<u>ניסוי שני:</u> השוואת מסווגים – עץ החלטה מול יער החלטה רנדומלי

<u>תיאור הניסוי:</u> ידוע כי עץ החלטה נוטה לעיתים להגיע למצב של התאמת יתר, לכן בניסוי זה רצינו לבדוק האם שימוש ביער החלטה רנדומלי יתן תוצאות עדיפות, מכיוון שמסווג זה מתבסס על ועדת עצי החלטה ופחות רגיש לבעיית התאמת יתר.

לצורך ניסוי זה, אימנו חזאים מסוג עץ החלטה, ויער רנדומלי המתבססים על מספר ימים אחורנית שונים.

בגרף מוצג ביצועי העץ לעומת היער, עבור דיוק של 0 מעלות, ועבור דיוק של 2 מעלות.



<u>מסקנות:</u> באופן כללי, עצי ההחלטה נתנו תוצאות מדויקות בהרבה. במקרה החריג של הסתכלות רק על יום אחד אחורה, היה ליער יתרון קטן מאוד כיוון שבמקרה זה לא קיים מידע רב להתבסס עליו.

לעומת זאת, כאשר יש מידע רב ומגוון עצי ההחלטה יודעים להפריד בין המידע החשוב לפחות חשוב, לעומת יער החלטה רנדומלי שביצועיו נפגמים מריבוי תכונות לא חשובות.

להערכתנו, זה מתקבל עקב ריבוי התכונות, חלקן דומות או לא חשובות, אשר ביער יכולות להיות בעלות השפעה חשובה בועדת העצים. ועץ ההחלטה הבודד מצליח להתעלם מהם מכיוון שבכל שלב בוחר את התכונה החשובה ביותר.

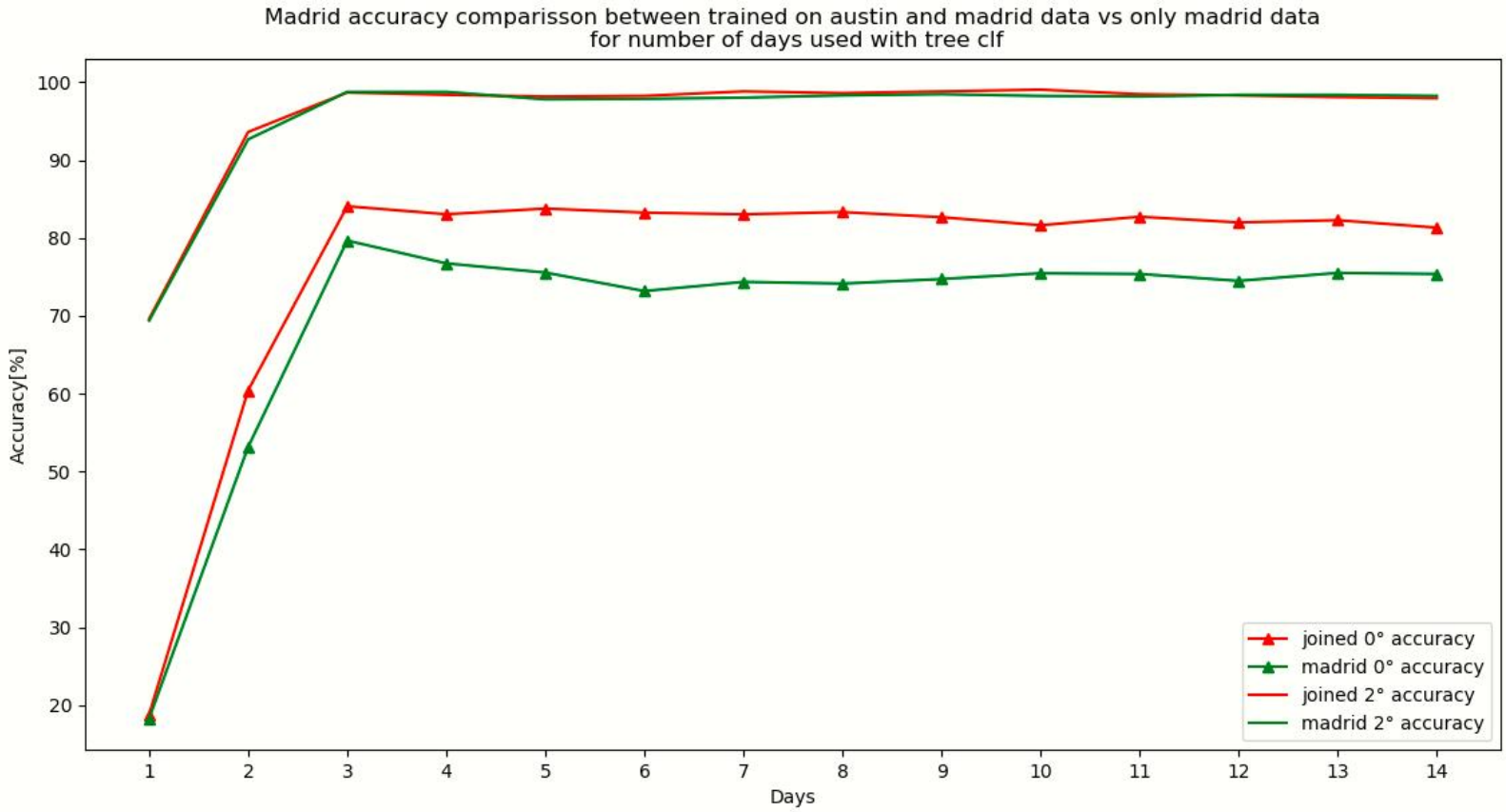
ניסוי 3: התמודדות המסווג עם רעש

תיאור הניסוי: בניסוי זה אנו מעוניינים לבדוק את השפעת למידה מסט דוגמאות מורחב, הכולל עיר נוספת על ביצועי המסווג.

עבור ניסוי זה אימנו מסווגי עץ החלטה על דוגמאות של שתי הערים, מדריד ואוסטין. שוב עבור מספר ימים שונה אחרונית ובדקנו את יכולות החיזוי על העיר מדריד.

מסווג מסוג עץ החלטה

בגרף מוצגים ביצועי מסווגי עץ החלטה משני סוגים: כאלה שאומנו על סט דוגמאות של מדריד בלבד וכאלה שאומנו על סט דוגמאות מעורב של מדריד ואוסטין ביחד. מוצגים הביצועים עבור דיוק של 0 מעלות, ועבור דיוק של 2 מעלות.



מסקנות: באופן כללי, ניתן לראות שאימון על מידע משתי הערים מביא לתוצאות עדיפות עבור כל מספר ימים.

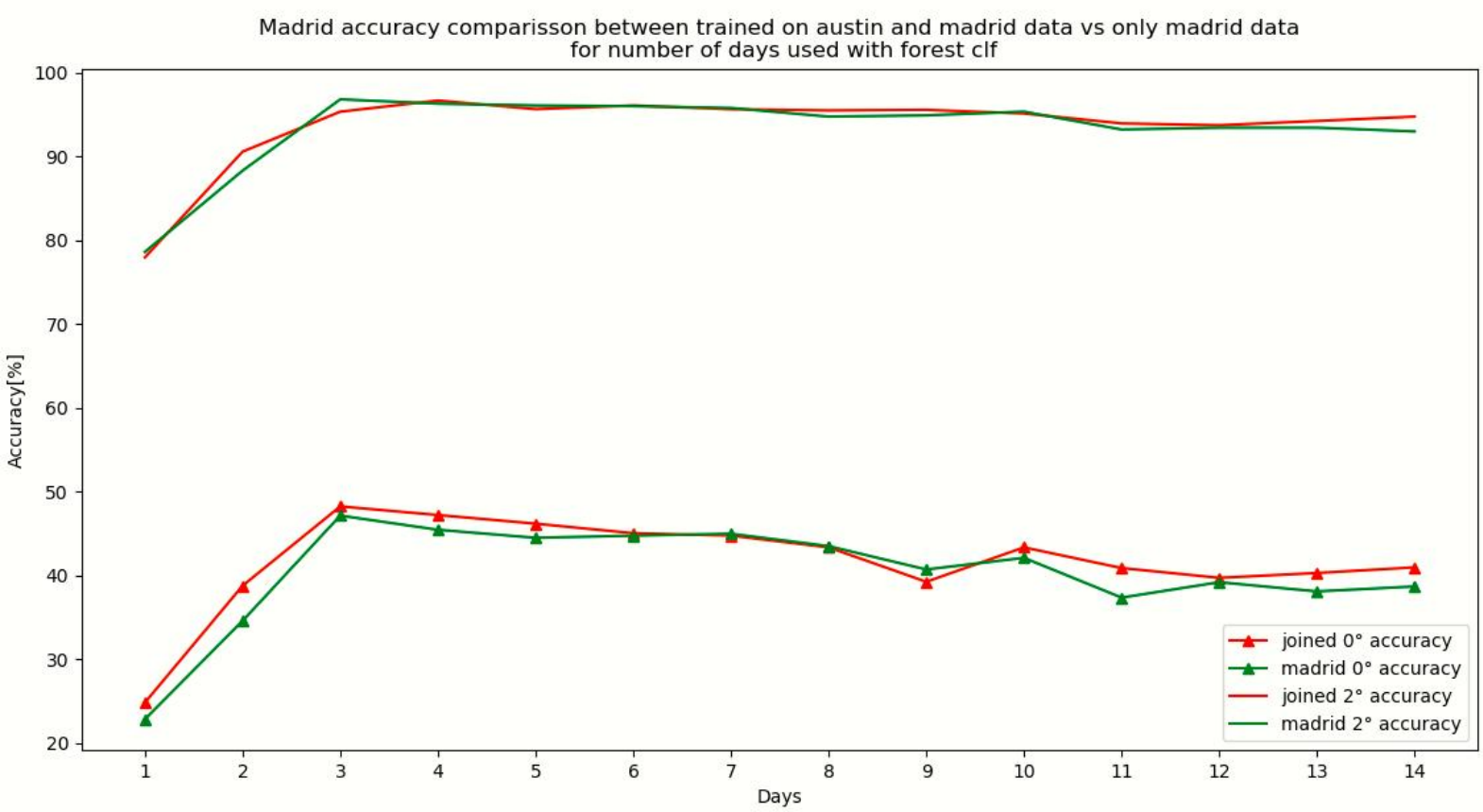
אנו ציפינו שאימון על מידע של שתי הערים יביא לירידה בתוצאות החיזוי עקב הכנסת רעש לדוגמאות האימון. באופן מפתיע, ביצועי עצי ההחלטה שלמדו על שתי הערים השתפרו באופן משמעותי (~10%) במתן תוצאות חיזוי מדויקות (סטיה של 0 מעלות). ההערכה שלנו היא כי השיפור בתוצאות נובע מכך שהמערכת בשלב האימון נחשפה לדוגמאות מגוונות יותר (התווספו הדוגמאות מאוסטין) ולכן הפכה ליותר גמישה ומיודעת.

תיאור הניסוי: בניסוי זה אנו מעוניינים לבדוק את השפעת למידה מסט דוגמאות מורחב, הכולל עיר נוספת על ביצועי המסווג.

עבור ניסוי זה אימנו מסווגי יער רנדומלי על דוגמאות של שתי הערים, מדריד ואוסטין. שוב עבור מספר ימים שונה אחרונית ובדקנו את יכולות החיזוי על העיר מדריד.

מסווג מסוג יער החלטה רנדומלי

בגרף מוצגים ביצועי מסווגי יער רנדומלי משני סוגים: כאלה שאומנו על סט דוגמאות של מדריד בלבד וכאלה שאומנו על סט דוגמאות מעורב של מדריד ואוסטין ביחד. מוצגים הביצועים עבור דיוק של 0 מעלות, ועבור דיוק של 2 מעלות.



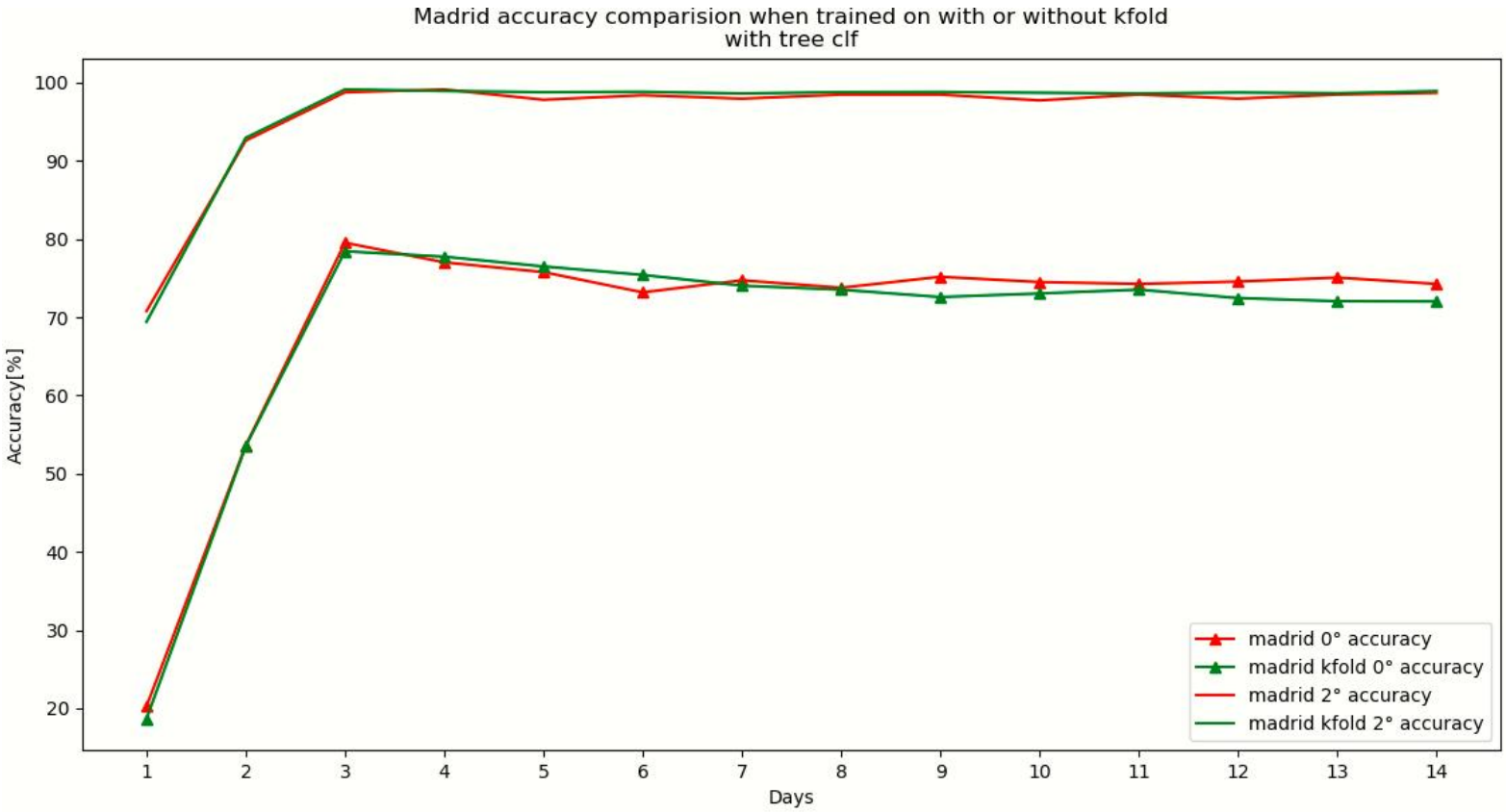
מסקנות: במקרה זה, בשונה משימוש בעץ החלטה רגיל, לא ראינו שיפור כמעט בכלל, אך עדיין אין הרעה בביצועים למרות הכנסה של דוגמאות מעיר בעלת אופי מזג אוויר שונה מאוד (מצורפת היסטגרמה בסוף).

להערכתינו זה נובע מכך שמראש יער החלטה רנדומלי פחות סובל מהתאמת יתר, ולכן הכנסה של דוגמאות למידה פחות רלוונטיות – לא שיפרו את ביצועיו.

<u>ניסוי 4:</u> אימון כרונולוגי לעומת שימוש בK FOLD

<u>תיאור הניסוי:</u> בניסוי זה אנו מעוניינים לבדוק את ההשפעה של שיטת האימון הכרונולוגית לעומת k-fold

עבור ניסוי זה אימנו שני מסווגים מסוג עץ החלטה על דוגמאות של העיר מדריד, שוב עבור מספר ימים שונה אחרונית, רק שהפעם ההבדל בין המסווגים היה שאחד אומן רק על הדוגמאות הכי ישנות כרונולוגית, ואילו המסווג השני אומר בשיטת Kfold במקטעים שתוארה במבוא. בדקנו את יכולות החיזוי על העיר מדריד.



<u>מסקנות:</u> ניתן לראות שהתוצאות בשתי השיטות זהות, כלומר לשיטת הKFOLD, שאמורה להתמודד היטב עם מצבים של high bias, high variance ולתת תוצאות מהימנות יותר של יכולות החזאי אין עדיפות על פני השיטה הרגילה.

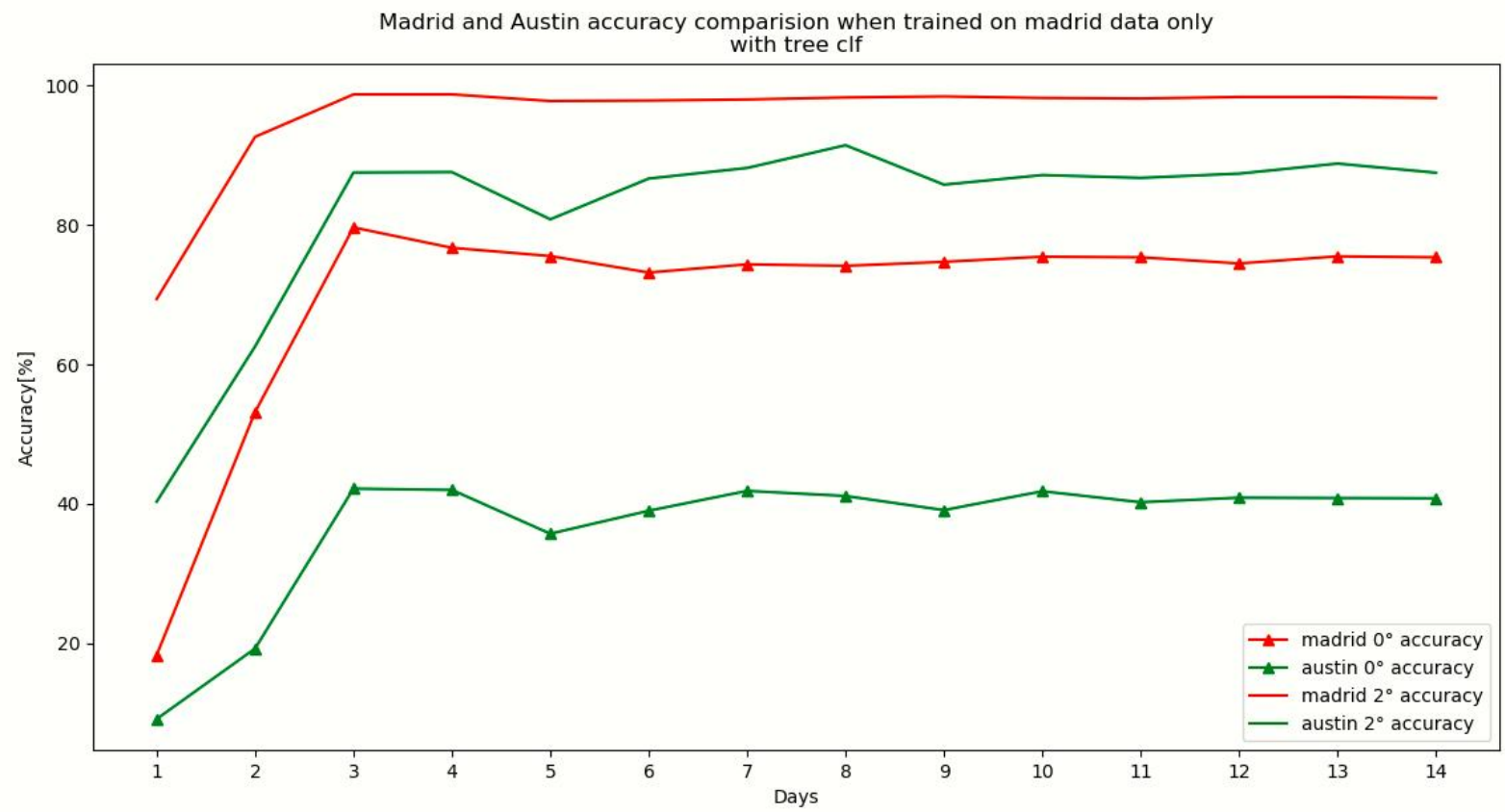
ניסוי 5: יכולת הכללת המסווג

תיאור הניסוי: בניסוי זה אנו מעוניינים לבדוק האם מסווג שאומן על מדידות בעיר אחת מסוגל לחזות באופן איכותי מדידות של עיר שונה.

לצורך ניסוי זה אימנו חזאים על מידע של העיר מדריד ובדקנו את תוצאות החיזוי על מידע של שתי הערים, מדריד ואוסטין.

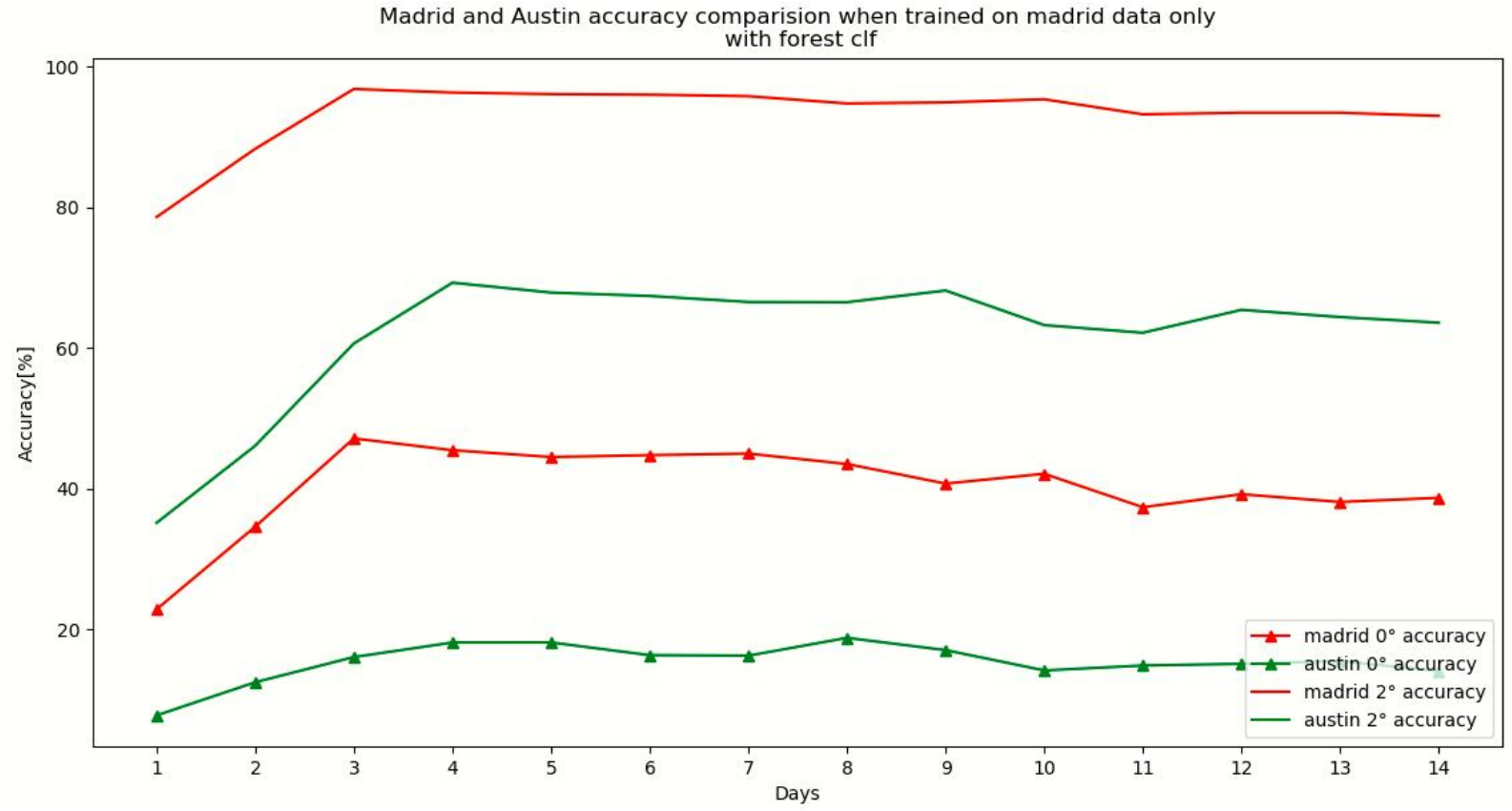
מסווג מסוג עץ החלטה

בגרף מוצגים ביצועי מסווגי עץ החלטה משני סוגים שאומנו על דוגמאות מהעיר מדריד בלבד. ההבדל הוא שבחצי מהמסווגים סט המבחן הוא דוגמאות מאוסטין ובחצי השני הדוגמאות הם ממדריד. מוצגים הביצועים עבור דיוק של 0 מעלות, ועבור דיוק של 2 מעלות.



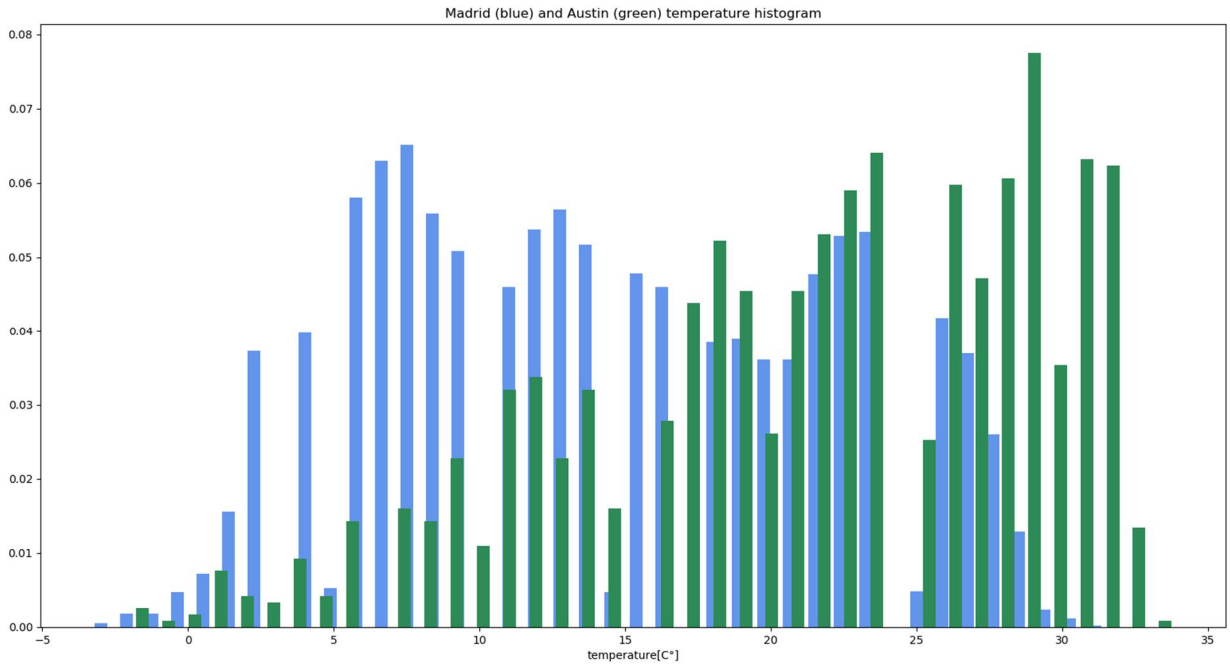
מסווג מסוג יער החלטה רנדומלי

בגרף מוצגים ביצועי מסווגי יער החלטה רנדומלי משני סוגים שאומנו על דוגמאות מהעיר מדריד בלבד. ההבדל הוא שבחצי מהמסווגים סט המבחן הוא דוגמאות מאוסטין ובחצי השני הדוגמאות הם ממדריד. מוצגים הביצועים עבור דיוק של 0 מעלות, ועבור דיוק של 2 מעלות.



מסקנות: כצפוי, בכל המקרים תוצאות החיזוי עבור העיר מדריד היה יותר טוב. אך, ניתן לראות כי גם תוצאות החיזוי של העיר אוסטין סבירות ביותר, אף על פי שהמסווגים לא אומנו על דוגמאות מעיר זאת כלל. כלומר, ניתן להעריך כי יש יכולת לאמן מסווג כמודגם בעבודה על עיר אחת בעולם ולהכליל את המסווג לעיר אחרת.

ניתן להסביר את ההבדלים תוצאות על ידי הסתכלות על היסטוגרמת הטמפרטורות של שתי הערים. ניתן לראות שיש הבדל משמעותי של הממוצע בין שתי הערים, כלומר נצפה שסט הדוגמאות של שתי הערים יהיו שונים מאוד. כלומר, שוני זה מסביר את הקושי של מודל חיזוי שאומן על עיר אחת לחזות היטב טמפרטורות של העיר השניה.



ניסוי נוסף

במהלך ההתקדמות בפרויקט, עלה הרעיון של שימוש בחזאי רגרסיה לינארית. חזאי זה ידוע בהיותו בעל תוצאות יחסית טובות למרות המודל הפשוט שלו. החסרון העיקרי של החזאי, שלגביו חששנו, הוא הקושי שלו להתמודד עם מאפיינים רבים ודומים, המצב הקיים במידע שלנו. אכן, לאחר כמה ניסויים עם החזאי גילינו שכאשר מתשמשים במידע המופק ממספר רב של ימים אחורנית תהליך הלמידה נכשל והחזאי אינו מצליח להתכנס.

סיכום

לאחר שיקלול כל הניסויים, החלטנו שהחזאי הטוב ביותר יהיה מסוג עץ החלטה עם מידע מהשלושה ימים האחרונים בלבד, המאומן בשיטת 10FOLD עם עומק עץ מקסימלי 13.

אלה התוצאות שחזאי זה נותן על סט המבחן (20% הדוגמאות האחרונות בסדר כרונולוגי שלא נגענו בהן עד עכשיו):

madrid 10 fold precision:

zero degree accuracy = 83.88

one degrees accuracy = 96.82

two degrees accuracy = 99.45

three degrees accuracy = 99.91

נזכיר כי רצינו להשוות את החזאי שלנו לחזאי הפשוט שחוזה לפי הטמפרטורה ביום הקודם, בעל התוצאות האלו:

yesterday precision austin:

zero degree accuracy = 21.7

one degrees accuracy = 52.5

two degrees accuracy = 68.74

three degrees accuracy = 78.38

ניתן לראות בבירור מהתוצאות שגם ללא שום ידע מקדים במטארולוגיה, הצלחנו, בעזרת שימוש בבינה מלאכותית, ליצור חזאי היודע להעריך את הטמפרטורה הממוצעת ביום שלמחרת בדיוק גבוה מאוד.

הצעות להמשך המחקר:

כפי שראינו, שימוש במידע של עיר נוספת, הביא לשיפור בביצועים. על כן, יכול להיות מעניין להשתמש במידע ממספר רב של ערים – האם יביא לשיפור או לדעיכה בדיוק.

כיוון נוסף לשיפור הדיוק יכול להיות שימוש במידע מערים שכונת תוך התחשבות בעוצמת וכיוון הרוח.