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

משתתפים

ברק זן, 305634487, [barakzan@campus.technion.ac.il](mailto:barakzan@campus.technion.ac.il)

אופיר פיצי שומרון, 201450574, [ofirshomron@campus.technion.ac.il](mailto:ofirshomron@campus.technion.ac.il)

מנחה:

מרצה:

מספר קורס:

**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.

במהלך הפרוייקט נבדוק קשרים בין ערים שונות, להוסיף

**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 ניסיים בהמשך.

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 – לפי מה שהיה יום לפני, ולפי אלגוריתם closest fit

אגריגציה ועיבוד ראשוני של המידע:

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

* ריבוע מגמת שינוי המאפיין בX ימים האחרונים.
* שורש מגמת שינוי המאפיין בX ימים האחרונים.
* ממוצע המאפיין בX ימים האחרונים.

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

חלוקת המידע:

כיוון שאנו בונים מערכת המבצעד חיזוי של אירוע עתידי על סמך העבר – חשוב לשים דגש על חלוקה נכונה של הדוגמאות וסדר כרונולוגי נכון בין קבוצות המבחן, ההערכה והאימון. לכן, כשאשר 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.

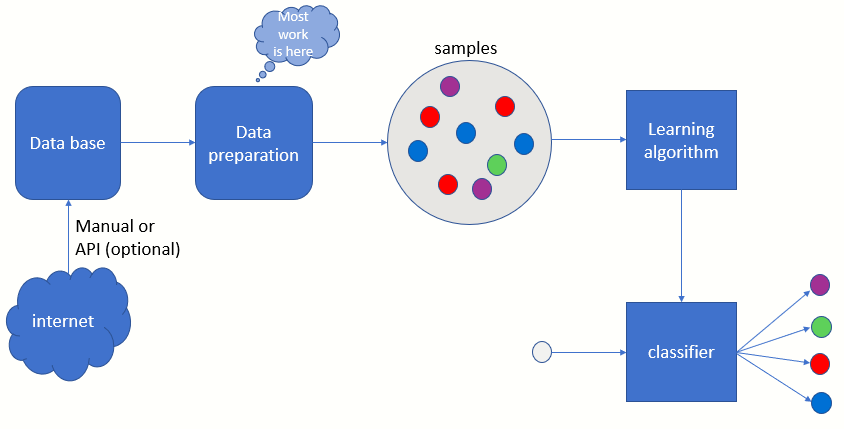
אימון המסווג:

המסווגים בהם השתמשנו הם עץ החלטה ויער החלטה רנדומלי. עבור כל סוג של מסווג – יצרנו מספר רב של מסווגים שונים הנבדלים בניהם במספר הימים אחורה עליו הם לומדים עבור כל מאפיין ובדרך בא המידע החסר הושלם (הסתכלות על יום אחורה או שימוש ב closest fit).

הערכת המסווגים:

שימוש במסווגים השונים על מנת לסווג את קבוצת המבחן והשוואה טיב הביצועים.

**תיאור המערכת ששימשה למימוש הפתרון**



באופן מעשי, המערכת התחלקה לשלושה חלקים עיקריים:

הכנת המידע:

זה החלק העיקרי והמשמעותי ביותר בפרוייקט, בחלק זה אנחנו לוקחים את המידע הגולמי שמצאנו באינטרנט, מורידים תכונות לא רלוונטיות, משלימים מידע חסר, מבצעים אגריגציות ויוצרים מאפיינים חדשים. את כל זה ביצענו 28 פעמים, ויצרנו 28 מאגרי מידע חדשים הנבדלים בניהם בצורת השלמת המידע ומספר הימים אחורנית מהם יצרנו את התכונות החדשות.

יש לציין שעשינו זאת גם עבור העיר אוסטין ע"מ לבצע ניסיים מעניינים בהמשך.

אימון המסווגים:

בשלב זה של הפרוייקט, לאחר מספר בדיקות וניסויים, ראינו שהתוצאות היו דומות מאוד בין השלמת מידע על פי יום קודם או בשיטת closest fit היו מאוד דומות עם עדיפות קלה ליום קודם ולכן בחרנו להמשיך משלב זה רק עם שיטת השלמה על פי יום קודם.

בשלב זה היו לנו 14 מאגרי מידע שונים, כשאשר התכונות בכל אחד מהם נוצרו על ידי התסתכלות על מספר ימים שונה אחורה.

לכל מאגר מידע יצרנו שני מסווגים: עץ החלטה ויער החלטה רנדומלי ולכן יש לנו 28 מסווגים שונים לבצע עליהם בדיקות וניסויים על מנת לנסות לקבל מסקנות מעניינות.

הערכת המסווגים וביצוע ניסויים:

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

* חיזוי מדוייק
* חיזוי עם סטייה מקסימלית של מעלות.
* חיזוי עם סטייה מקסימלית של מעלות.
* חיזוי עם סטייה מקסימלית של מעלות.

**מתודולוגיה ניסויית**

הדרך הכללית שבה ערכנו את הניסויים היא:

לאחר שלב הכנת המידע, אימנו מספר מסווגים שונים, לפי המידע אותו רצינו להעריך ולהשוות. המסווגים נבדלים ביניהם במידע עליו הם אומנו (עיר\ערים, מספר ימים), סוג המסווג (עץ, יער).

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.

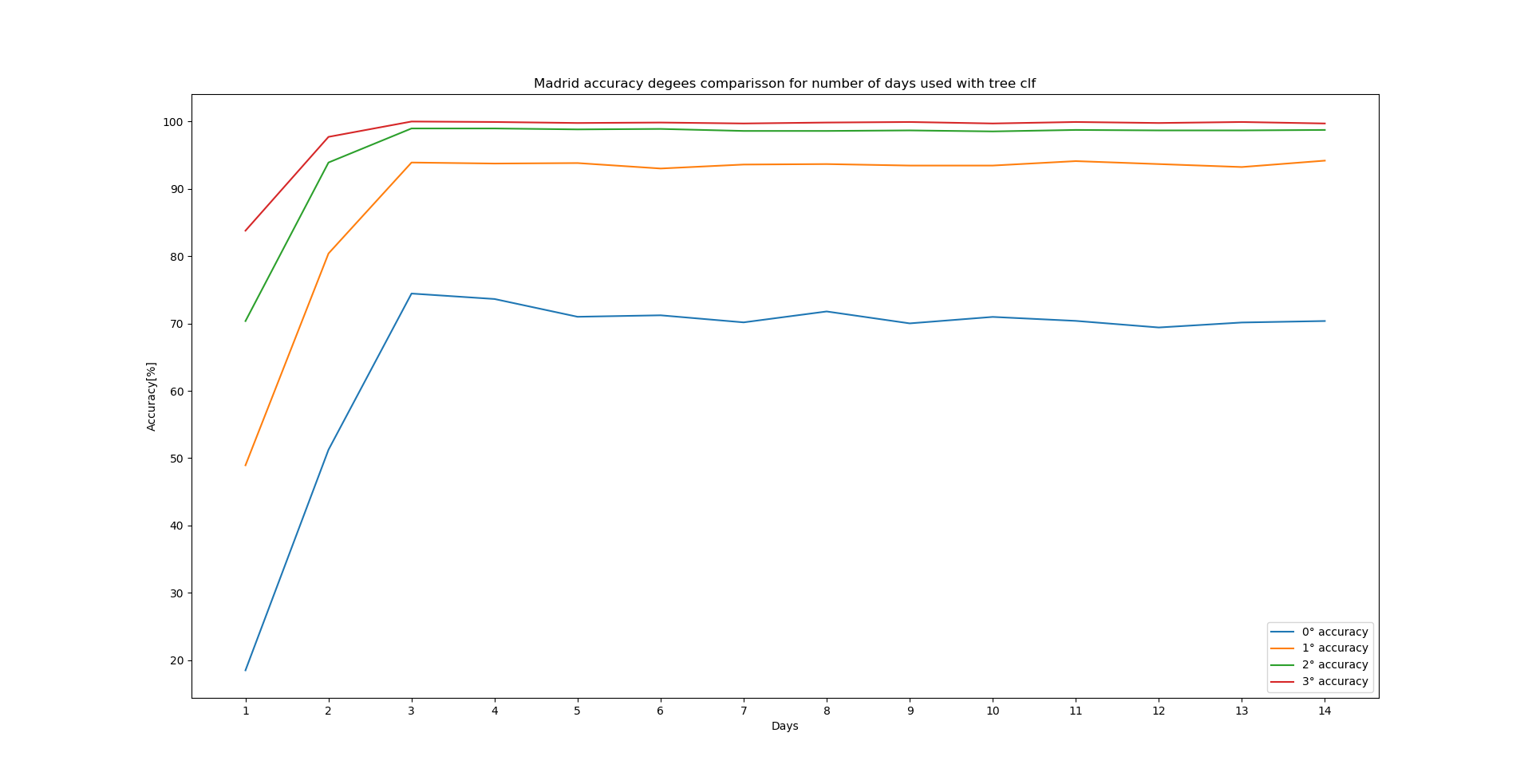
**תיאור הניסויים תוצאות ומסקנות**

**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.

לצורך ניסוי זה יצרנו 14 מסווגים מסוג "עץ החלטה" אשר אומנו על דוגמאות ממדידות בעיר מדריד, כאשר ההבדל בניהם הוא מספר הימים אחורנית מהם המידע של כל דוגמא מורכב.

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.

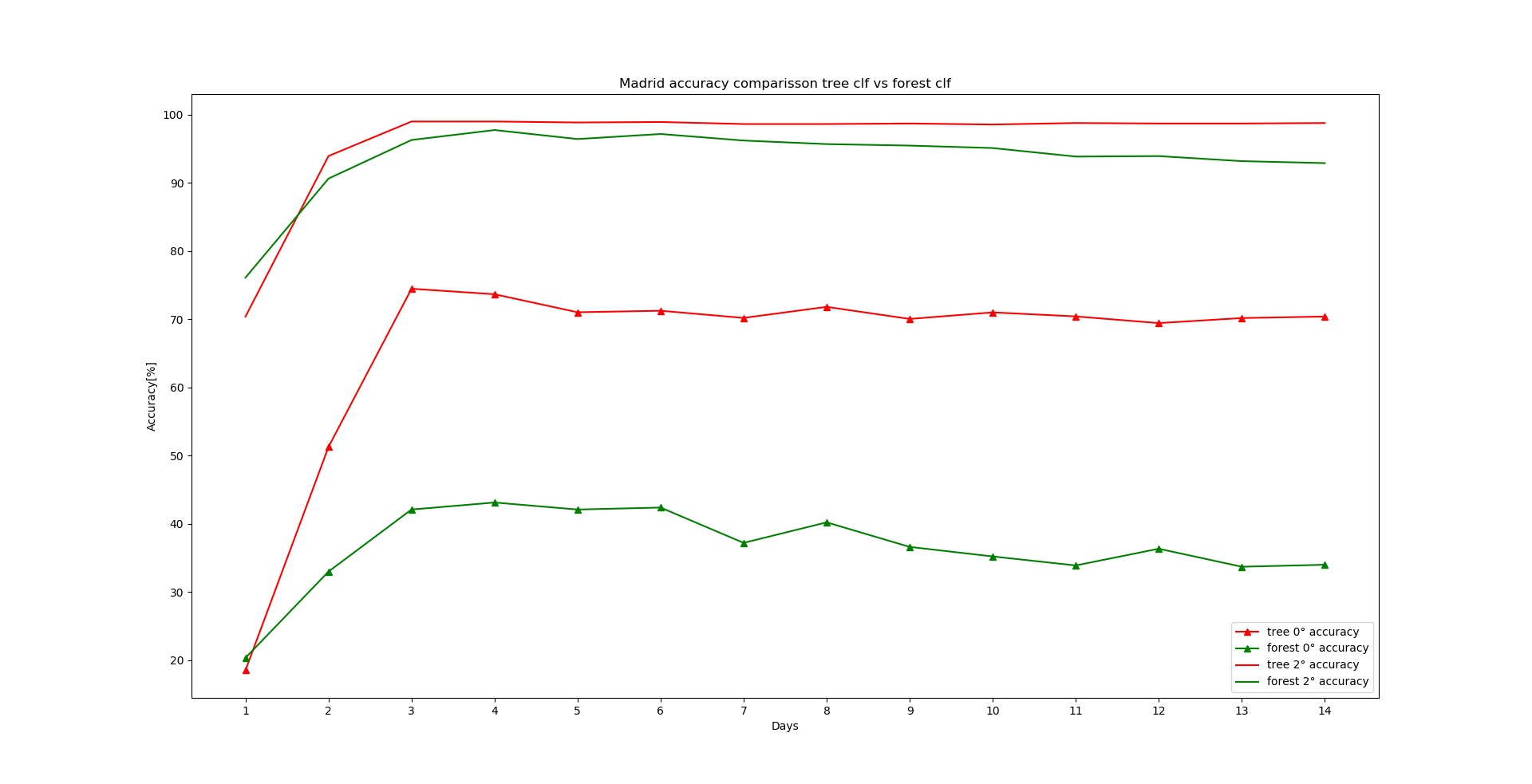
ואילו הסתכלות של יותר מ3 ימים אחורנית אינה מועילה ואף מוסיפה רעש שלעיתים פוגע קלות בביצועים.

**ניסוי שני:** השוואת מסווגים – עץ החלטה מול יער החלטה רנדומלי

תיאור הניסוי: 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 יצרנו 14 מסווגים מסוג "עץ החלטה" ו14 מסוג "יער החלטה רנדומלי" אשר אומנו על דוגמאות ממדידות בעיר מדריד, כאשר ההבדל בניהם, פרט לסוג המסווג, הוא מספר הימים אחורנית מהם המידע של כל דוגמא מורכב.

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



מסקנות: 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.

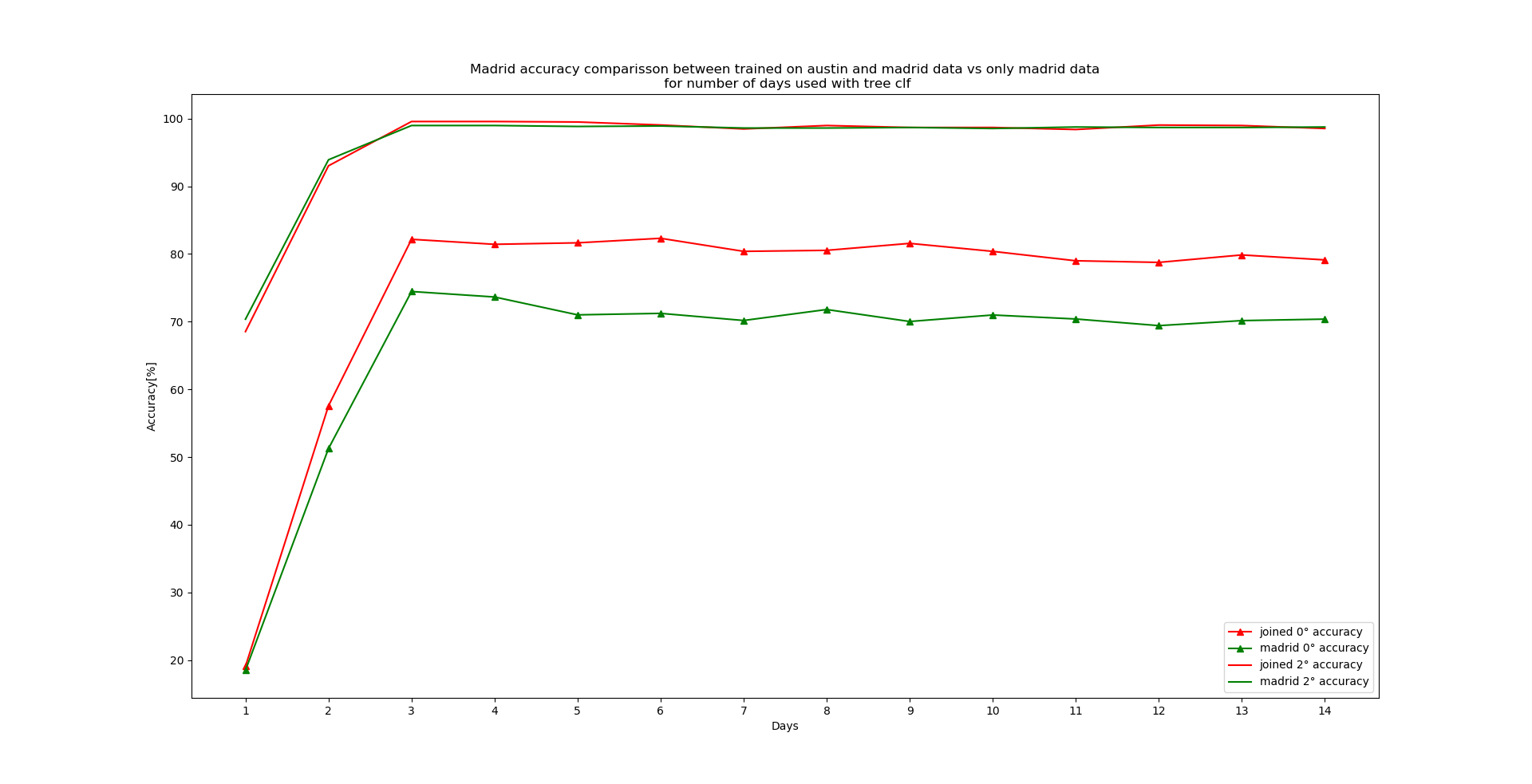
לעומת זאת, כאשר יש מידע רב ומגוון עצי ההחלטה יודעים להפריד בין המידע החשוב לפחות חשוב, לעומת יער החלטה רנדומלי שביצועיו נפגמים מריבוי תכונות לא חשובות.

**ניסוי 3:** התמודדות המסווג עם רעש

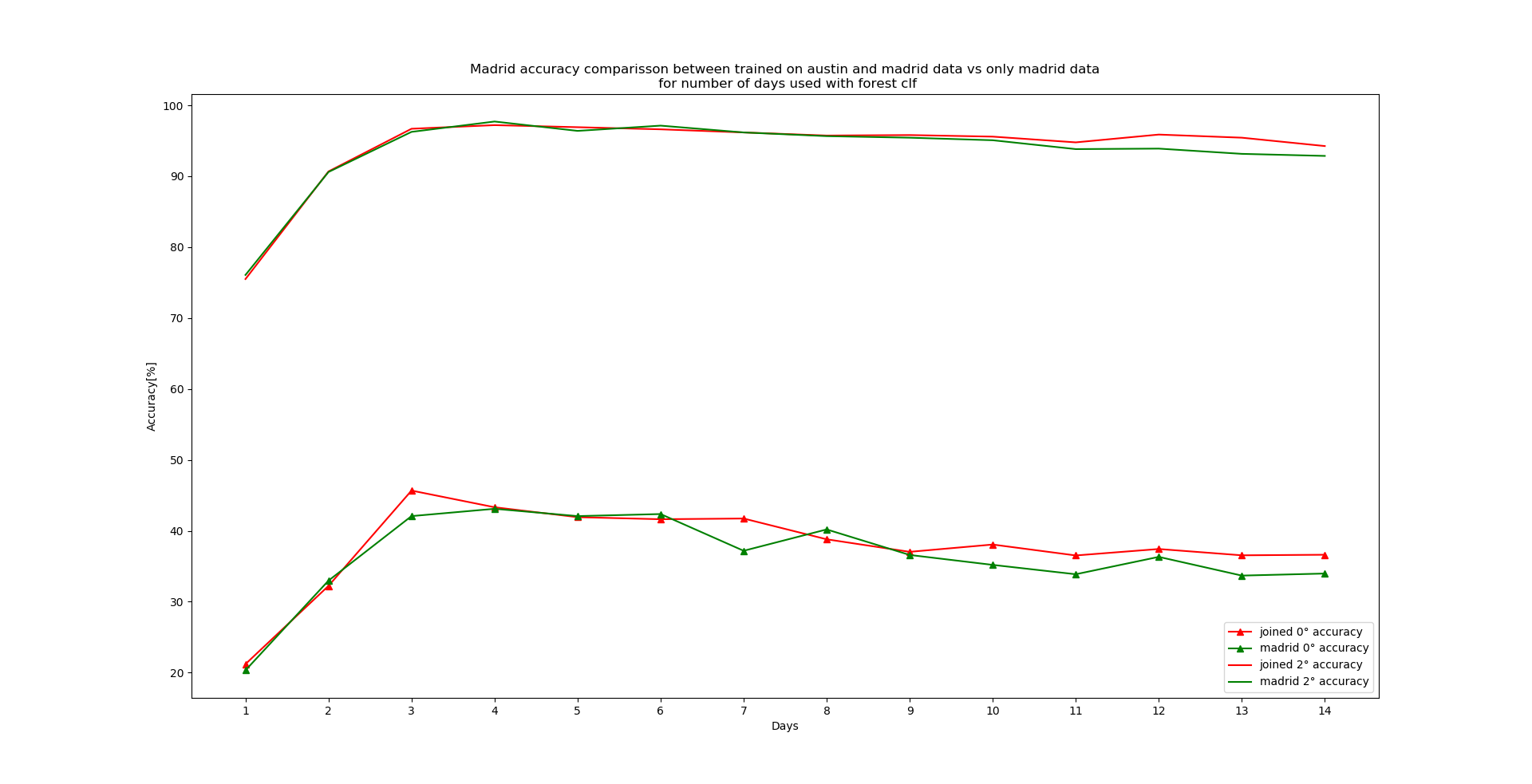
תיאור הניסוי: בניסוי זה אנו מעוניינים לבדוק את השפעת למידה מסט דוגמאות מורחב, הכולל עיר נוספת על ביצועי המסווג.

לצורך ניסוי זה יצרנו 14 מסווגים מסוג "עץ החלטה" ו14 מסוג "יער החלטה רנדומלי" אשר אומנו על דוגמאות ממדידות בעיר מדריד ומדידות בעיר אוסטין, כאשר ההבדל בניהם, פרט לסוג המסווג, הוא מספר הימים אחורנית מהם המידע של כל דוגמא מורכב.

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.

מסווג מסוג יער החלטה רנדומלי

בגרף מוצגים ביצועי מסווגי יער החלטה 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: עבור דיוק של 2 מעלות, התקבל כי התוצאות זהות לשני סוגי המסווגים, אלו שהתאמנו על מידע ממדריד בלבד ואלו שהתאמנו על המידע ממדריד ואוסטין ביחד.

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 ומויודעת.

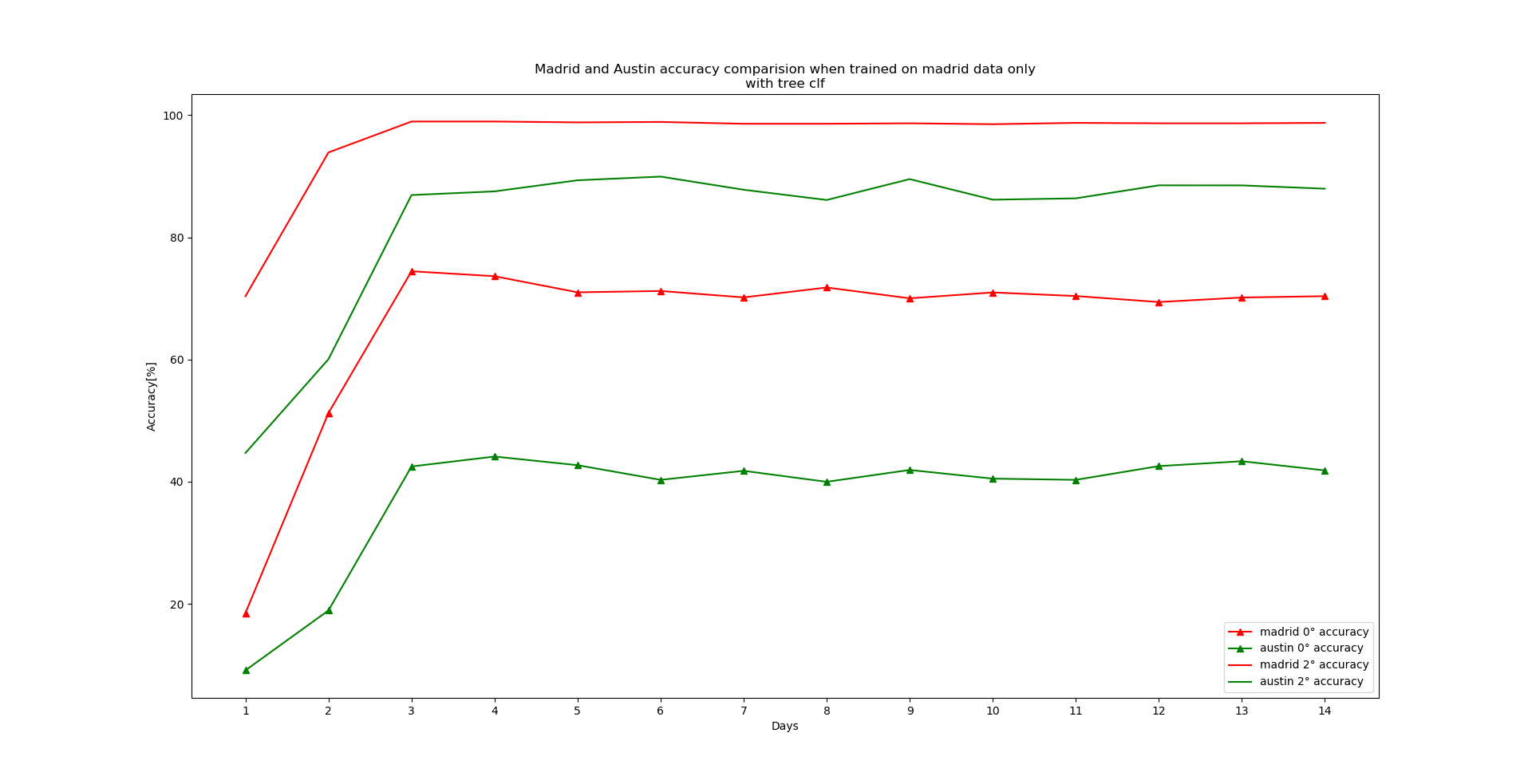
**ניסוי 4:** יכולת הכללת המסווג

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 יצרנו 14 מסווגים מסוג "עץ החלטה" ו14 מסוג "יער החלטה רנדומלי" אשר אומנו על דוגמאות ממדידות בעיר מדריד, כאשר ההבדל בניהם, פרט לסוג המסווג, הוא מספר הימים אחורנית מהם המידע של כל דוגמא מורכב.

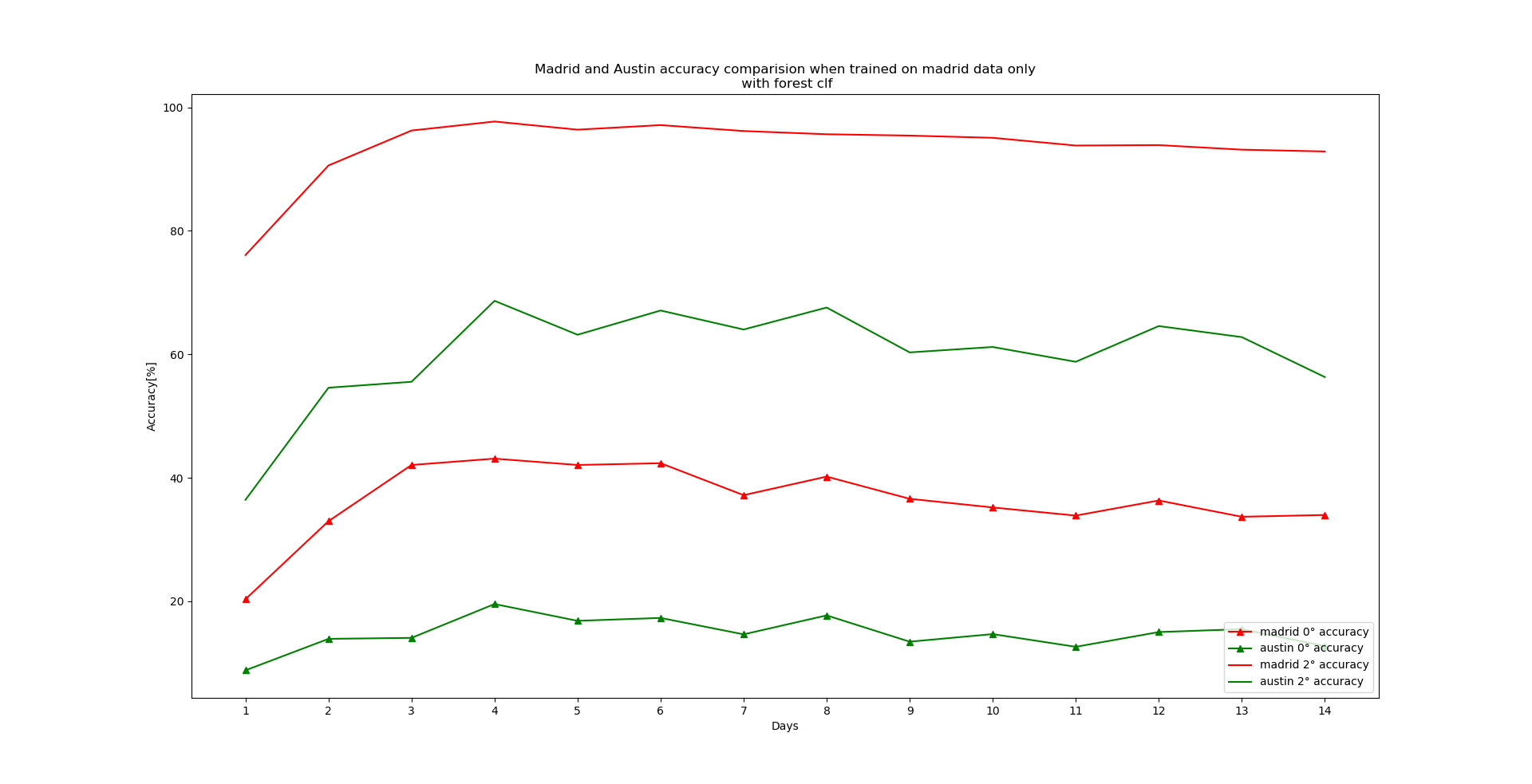
מסווג מסוג עץ החלטה

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.

**סיכום**

להוסיף: דיון בתוצאות, מה חסר במה שעשיתם, כיוונים להמשך המחקר וסיכום. זהו פרק חשוב ביותר. הקדישו לו את תשומת הלב הראויה

