**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)

**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, בלה בלה בלה בלה 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.

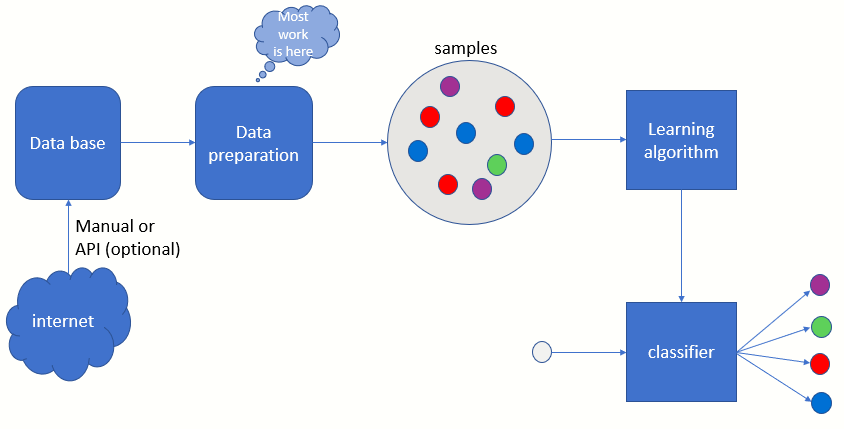
אימון המסווג:

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**

תוצאות ומסקנות. פרק זה יתאר כל ניסוי, יציג את התוצאות בטבלאות או גרפים וינתח אותן

**סיכום**

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

תיאור הבעיה

ישנן כמה בעיות שברצוננו לנסות לפתור, נציג את כולן ובמהלך התקדמות הפרוייקט ונחליט בעזרת המנחה לאיזה בעיות כדאי לגשת.

הבעיות שמעניינות אותנו הן ולא בהכרח רק:

* האם מחר יהיה יותר חם או יותר קר ביחס להיום
* האם ירד מחר גשם או לא
* חיזור רמת העננות מחר: בהיר\מעונן חלקית\מעונן\קודר
* האם תהייה מחר סופה או לא

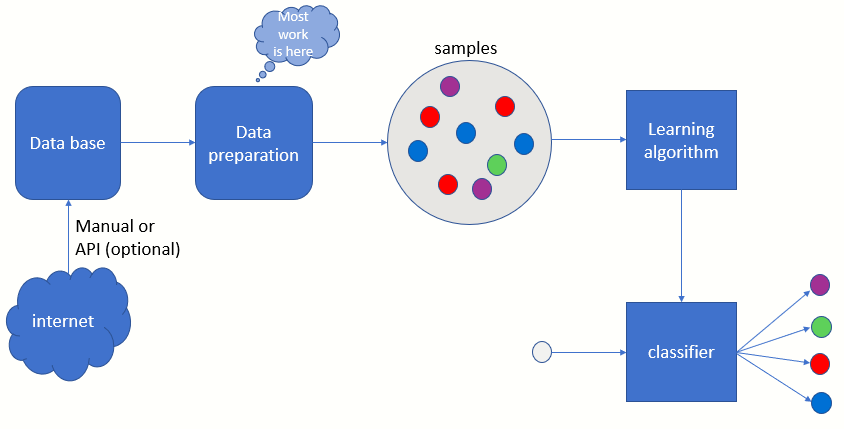
הצעת פתרון

הפתרון יתחלק לשלושה שלבים עיקריים:

* השגת היסטוריית מזג האוויר של ניו יורק מהאינטרנט
* ביצוע אגריגציה למידע כך שיתאים לשימוש באלגוריטמי למידה.
* שימוש בגרסא מורחבת של KNN ע"מ לבצע חיזוי.
* ביצוע ניסויים אוטומטיים ע"מ לשפר את דיוק האלגוריתם ע"י בדיקת פרמטרים.

המידע בו נשתמש יהיה היסטורית מזג האוויר בעיר ניו יורק, בנוסף קיימת אופציה להסתכלות על מזג אוויר בערים שכנות בשילוב עם הסתכלות על כיוון הרוח ע"מ לקבל חיזוי מדוייק יותר.

תיאור המערכת



\*קיימת אפשרות שנעשה אוטומציה שמקבלת עיר ומורידה את המידע הרלוונטי אליה.

החלק הראשון **והמשמעותי ביותר** של המערכת, הוא data preparation, בחלק זה ניקח את המידע שנשיג מהאינטרנט וניצור ממנו סט דוגמאות למידה ע"י:

* מילוי מידע חסר
* ביצוע אגריגציה
* ניקוי נעשים
* נרמול

החלק השני יהיה חלוקת הדוגמאות שקיבלנו ל3 קבוצות:

* קבוצות מבחן – כל הימים מתחילת 2018
* קבוצת הערכה – כל הימים בשנת 2017
* בקוצת אימון – הימים שקדמו ל2017

\*הגדלים הנ"ל אינם סופיים או מחייבים.

החלק השלישי יהיה שימוש באלגוריתם KNN הממוש כבר בסיפריות קיימות וקבלת מסווג לחיזוי.

ע"מ לשפר את דיוק האלגוריתם נבצע כיוון פרמטרים ע"י כתיבת אוטומציה שתבדוק את תוצאות אלגוריתם הלמידה לאחר מתן משקלים שונים לכל תכונה ובדיקת התוצאות לכל סט משקלים נתון על קבוצת ההערכה.

הערכת ביצועי המערכת

בכדי להעריך את ביצועי המערכת, נסתכל על מטריצת מטריצת הבלבול ונחשב את אחוז השגיעה ואחוז הדיוק. חשוב לציין, שבמקרה שלנו, אין סוג שגיאה ספציפי החמור יותר מאחר (לדוגמא להגיד שחולה בריא)

פרטים טכניים

שפת תכנות: פייתון 3

סביבת עבודה: pycharm

ספריות ומודולים: sklearn, pandas, numpy, scipy ועוד שנוסיף בהתאם להתקדמות הפרוייקט