

**פרויקט גמר**

למילוי חלקי של הדרישות לקבלת התואר **הנדסאי תוכנה**

**נושא הפרויקט:**

ניתוח וחקר הנתונים

**מגיש:**

תומר עובד-315112342

**העבודה בוצעה בהנחיית:**

שחר אוחנה

תשע"ט 2019-2018

אני הסטודנט תומר עובד. מס' ת.ז 315112342

החתום מטה מצהיר בזאת שכל עבודת הגמר/הפרויקט המוגש/ת בחוברת זו הינה פרי עבודתי בלבד.

על בסיס הנחייתו של המנחה ותוך הסתמכות על מקורות הידע והמידע האחרים המצויים בביבליוגרפיה המובאת בחוברת זאת.

אני מודע לאחריות שהנני מקבל על עצמי ע"י חתימתי על הצהרה זו שכול הנאמר בה הינו אמת ורק אמת.

חתימת מגיש החוברת: \_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_ אישור המנחה: \_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_ הנני מאשר הגשת החוברת להערכה: \_\_\_\_\_\_\_\_\_\_\_

תוכן עניינים

הצעת פרויקט..............................................................................................4

מבוא והקדמה..............................................................................................10

תהליכים רגשיים בכתיבת הפרויקט...................................................................11

פירוט והסברים על מודל הלמידה בפרויקט........................................................12-14

סדר הפעולות בסקריפט מכריעת הנתונים ועד להכנת במודל למדגם הבדיקה...... ..14-16

הסבר על האלגוריתם שמאפשר למודל ללמוד נתונים ולהבחן............................... 16-17

הסבר על הצורה הוויזואלית שהלקוח מקבל חזרה את המידע...................................18

הסכמות והמבני נתונים במסד הנתונים............................................................ 19-21

הסבר והצגת קבצי הפרויקט................................................ 22

בביבליוגרפיה......................................................................23

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| סמל ביה"ס: | 571281 |  | שם המכללה: | מכללת אורט סינגאלובסקי |

**הצעת נושא לפרוייקט להנדסאי / טכנאי הנדסת תוכנה [[1]](#footnote-1)1**

1. שם הפרויקט:Data Analysis
2. פרטי המנחה/ים: שם:שחר אוחנה

כתובת: ...................................... טלפון: ....................

1. פרטי המגיש/ים: שם : תומר עובד ת.ז. :315112342

כתובת:הגפן 346 בית אריה טל':0545404138

1. עבודה ביחיד / עבודה בזוג
2. אופי עבודת הפרויקט :תכנון ובניית אב טיפוס
3. מקום ביצוע הפרויקט :בית הספר
4. תיאור נושא הפרויקט:

אתר הנותן שרות למשתמשים בכל הנוגע לניתוח וחקר נתונים- big data

לכל משתמש באתר יהיה האפשרות לשלוח כל סוג נתונים אשר מאורגן בקובץ Excel ולקבל חיזוי מחושב על הנתונים

אשר שלח. ובנוסף יראה

המשתמש באופן גרפי את היחס בן הנתונים שלו לנתונים האחרים באותה

קבוצת נתונים . חישוב החיזוי וניתוח הנתונים נעשה באמצעות ספריות ה-machine learning בשפת התכנות python

1. סקירת מצב קיים, כולל פירוט הבעיות במצב הקיים

כיום בעידן הטכנולוגי, כמויות המידע שנצברות ומועברות הן בלתי נתפסות. ידע רב=כוח רב אז למה שלא ניקח את המידע הכה רב שמועבר בחיינו היומיומיים ננרמל אותו ( העברת המידע לחטיבת מידע אחת בלבד) ננתח אותו באמצעות פעולות מתמטיות ונסיק עליו מסקנות, מסקנות שיביאו תועלת בכל היבט ותחום בחיים :לצרכן, ליצרן ולמשתמש באתרי אינטרנט, לכל גוף שניתוח המידע מקדם אותו באופן כלשהוא.

1. פירוט הדרישות מהמערכת:

על המערכת למלא את הדרישות הבאות:

* אפשרות להרשמה וכניסה לאתר כמשתמש( (authentication
* הצגת תרשימי נתוני הקבצים שעליהן אותו מודל התאמן וכמו כן,הצגת תרשימי החיזוי שנעשו לאותו מודל זה.
* אפשרות למשתמשים בלבד ליצירת מודל חדש וניסיונות חיזוי על אותו מודל.
* שמירת כל הנתונים על מסד נתונים ושמירת המודלים על קבצי .sav
* הרצת תהליך בן (thread)אשר יריץ קובץ python ויבצע כל הקשור לניתוח הנתונים והmachine learning
* משתמשי האתר יוכלו להגיב על תרשימי המידע למחוק תרשימי מידע שהם יצרו בלבד (authorization) וכול שאר אפשריות המובנות מאליו במערכות בעלות משתמשים

10 מפרט טכני:

* CPU- Intel Core i7 2.70GHz
* RAM 8.00 GB
* Hard Disk 220 GB
* מערכת הפעלה windows 10 x64
* סביבת פיתוח visual studio code
* מהדר-node js compiler& python interpretation

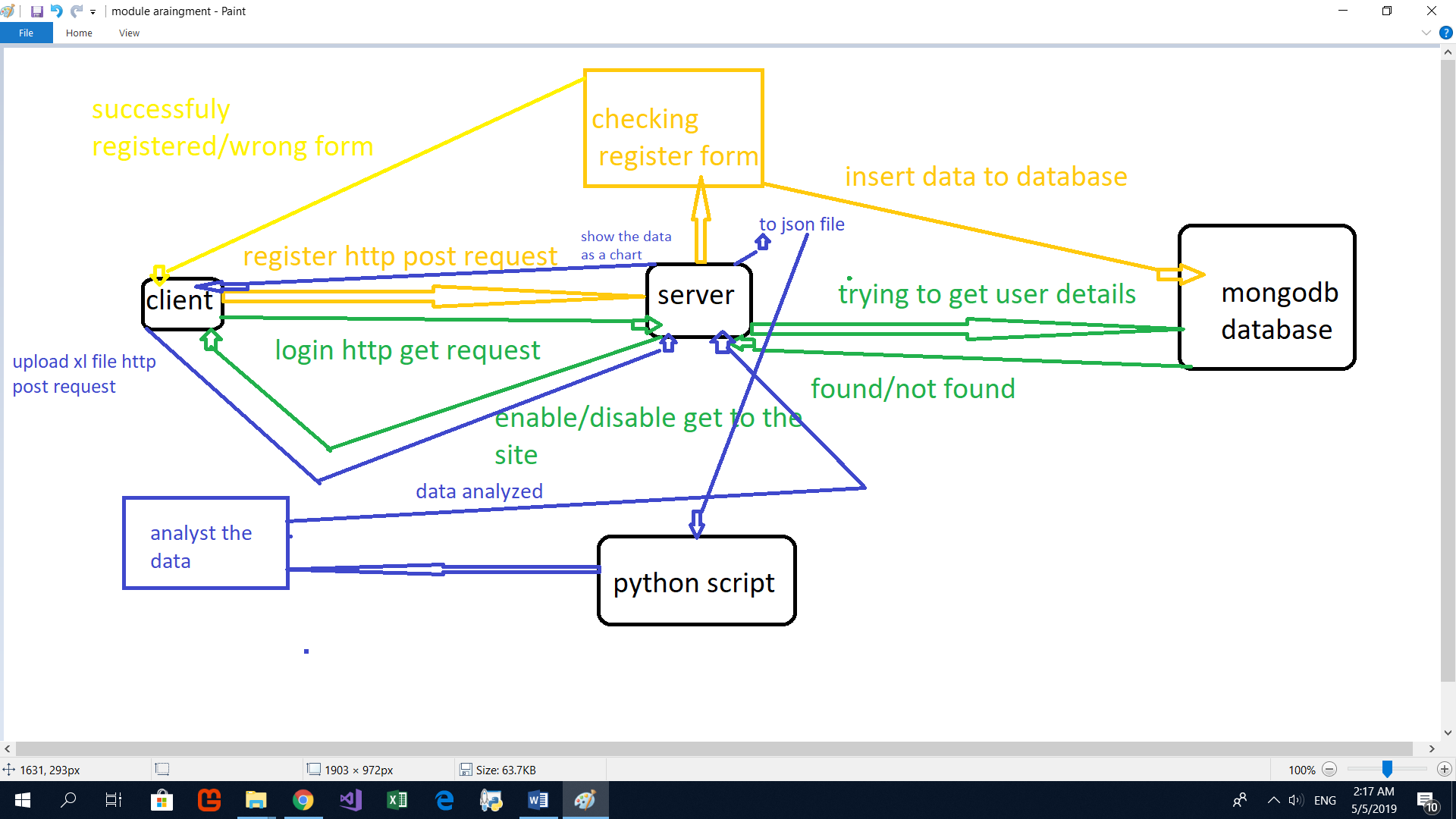
11.חלוקה משוערת למודלים

|  |
| --- |
| Client side-GUI  מראה למשתמש את כל המודלים ותרשימי המידע שנוצרו וכמובן שנותן לו את האפשרות  ליצור מודל משלו וחיזויים על אותו מודל  פעולות עיקריות  מתחבר לשרת, בוחר את סוג החיזוי איתו הוא מתכוון לעבוד(רגרסיה או סיווג(פירוט בהמשך)( המסד נתונים שולף את כל תרשימי המידע הקשורים לאותו סוג סיווג.  פעולה נוספת שלקוח יוכל לבצע זה העלאת קובץ excel עליו יתאמן המודל ואפשרות לקבל חיזוי חדש לאותו מודל זה וכמובן הצגת הנתונים באופן גרפי ויזואלי |

|  |
| --- |
| Server side  השרת באופן עקרוני וכללי אמור לשרת את הלקוח בכל פעולה אשר חפץ ובנוסף השרת יהווה גורם מתווך בין הלקוח לנתונים אשר חבויים במסד הנתונים  דוגמאות ספציפיות:  1.הרשמה לאתר  כאשר הלקוח ירצה להירשם לאתר, הלקוח ישלח http post request לשרת, השרת יבצע ולידציה על אותם נתונים שקיבל  ואם הוא יחליט שהנתונים מתאימים לדרישות יכניס את פרטיו למסד הנתונים ויכיר בו כמשתמש האתר  2.העלאת קבצי xl  אחרי שהלקוח יעלה קובץxl השרת ימיר אותו לקובץ json וישלח אותו כפרמטר לתהליך הבן שמבצע את קוד הpython שאחראי על ניתוח הנתונים של אותו קובץ והחזרת המידע המעובד לשרת שמציג אותו באופן ויזואלי ללקוח וכמובן שמירת כל המידע על מסד הנתונים  3.חיזוי תוצאה  לאחר שהמודל ילמד על די נתונים המשתמש יוכל לבחון את ביצועיו עם נתונים חדשים |

|  |
| --- |
| Database  מודול אשר אחראי על שליפת יצירה עריכה ומחיקה של נותנים מהמסד  המודל הזה אחראי על החזקת המידע הקשור למשתמשים |

זרימת הנתונים בן המודולים בפרויקט



12.מספר שעות מוקדשות לפרויקט 520

13.חלוקת עבודה בן מגישים-לא רלוונטי

14.לוח זמנים לביצוע הפרויקט

|  |  |  |
| --- | --- | --- |
| שלב | טווח תאריכים | |
| מ - | עד - |
| ניתוח המערכת | 1.1.2019 | 20.1.2019 |
| בניית מסד הנתונים | 21.1.2019 | 12.2.2019 |
| עיצוב המערכת | 14.2.20.19 | 28.2.2019 |
| פיתוח התוכנה/קידוד | 1.3.2019 | 12.4.2019 |
| ניפוי השגיאות | 13.4.2019 | 1.5.2019 |
| הרצות ניסיון | 1.5.2019 | 3.5.2019 |
| הגשת הפרוייקט |  |  |

15.טכנולוגיות שפות תכנות וספריות בשימוש הפרויקט

צד לקוח

* Visual studio code
* Html
* CSS
* Bootstrap4
* jQuery
* JavaScript

צד שרת

* Visual studio code
* Node js
* Express.js
* Npm
* MongoDb
* Python
* import numpy as np
* import matplotlib.pyplot as plt
* import pandas as pd
* import pickle
* from sklearn.compose import ColumnTransformer
* from sklearn.preprocessing import Normalizer, OneHotEncoder
* from sklearn.ensemble import RandomForestRegressor





הקדמה והרקע בבניית הפרויקט

נושא הפרויקט

נושא הפרויקט אותו אני מגיש הוא ניתוח וחקר הנתונים-:big data היכולת לקרוא נתונים בקנה מידה גדול מאוד מכל סוגי המידע, מכל תחום חיים וללמוד את היחס בן אותם הנתונים לבן מבחן התוצאה שנבע עקב אותם הנתונים. באופן מעשי- לקרוא מספר לא ידוע ולא תלוי של פרמטרים וערכם ואת התוצאה עליה הפרמטרים השפיעו וללמוד כיצד כל פרמטר משפיע על התוצאה ואת המשקל של אותו פרמטר ולאחר תהליך למידה על די מקרים להכיר באפשרות לחזות תוצאה עם מקרה חדש בעל אותם פרמטרים אך ערכים חדשים כמובן לאותם פרמטרים.

בשפה המקצועית תחום זה נקרא machine learning-בעברית, “מכונה לומדת", מכונה הכוונה לקוד שלומד התנהגות מסוימת של נתונים והשפעתם על תוצאה מסוימת ולאחר מכן היכולת להסיק ולחזות תוצאה עם ערכים חדשים לאותם הפרמטרים.

תהליך הלמידה שהקוד עובר הוא דומה לעיבוד המידע שבן אדם ומכך נגזר השם מכונה לומדת אך לא כמו בן אדם היחס בן כמות המידע שקוד מעבד ליחידת זמן אחת גדול בהרבה מאשר המוח האנושי יכול לעבד

אירועי "ערב" כתיבת הפרויקט

חצי שנה לפני הגשת הפרויקט מצבי ביחס לפרויקט לא היה משביע רצון. לא היה לי רעיון לא היה לי סביבת פיתוח מועדפת בה רציתי לעשות את הפרויקט ולא הכלים לבצע אותו,אך בדיוק באותו זמן קרו שתי אירועים במקביל אשר חיללו תפנית במצב:

1.הקורס התאורטי בmachine learning במכללה הגיע לסיומו,הרגע בו הבנתי שאני רוצה ללמוד יותר לעומק את התחום ולבסוף לממש את הידע בפרויקט גמר

2.מנהל המחלקה לתוכנה במכללה המליץ לי להכנס לעולם הweb עקב ביקוש איש web לתפקידים בצהל.

עקב שתי אירועים אלו החלטתי "להרביץ שתי ציפורים במכה אחת " ולעשות פרויקט בweb אשר משלב בתוכו machine learning. בקבלת החלטה זו היה חבוי אומץ רב והימור כבד מפני שאז לא היה לי שום ידע מקדים בweb לא כתבתי שום פרויקט בweb ולא ידעתי דבר על machine learning חוץ מהפירוש שלו לעברית אך ידעתי שכדי להגיע לגדולה צריך להאמין בעצמך לעבוד קשה ולא לפחד ממכשולים כי זה רק יחשל אותי

לאחר קבלת ההחלטה לעשות פרויקט גמר בממשק משתמש של אתר אינטרנט בנושא ניתוח נתונים בעזרת המכונה הלומדת קניתי שתי קורסים "און ליין" מהאקדמיה הווירטואלית "Udemy " 1.לימוד פיתוח בסביבת web עם

הטכנולוגיות החדשות ביותר, 2.machine learning-hands on with python

תחילה עשיתי כמובן את שתי הקורסים בערך ב3 חודשים ורק אז התחלתי לעבוד על הפרויקט עצמו. בהתחלה לא הפסיקה לעבור בתוכי המחשבה שאולי נכנסתי לנעליים שגדולות עליי בכמה מידות כי בכל זאת: ללמוד שתי קורסים שמכילים בממוצע 200 שעות הרצאה כל אחד ואחרי זה עוד לבנות פרויקט גמר להנדסאים שכולו מתבסס על הידע שהייתי אמור לרכוש מהקורסים האלו וכל זה תחת מסגרת הזמן שגודלה כ חמישה חודשים בלבד לא היה המצב הכי אידיאלי. עם זאת, ידעתי כאשר אני חורק שיניים ועובד קשה השמיים הם הגבול מבחינתי ועם חלוקת זמן מסודרת ומיטבית אצליח לעשות זאת וכך היה: בערך חודש וחצי לפני תאריך הגשת הפרויקט סיימתי את שתי הקורסים , התחלתי לקודד את הפרויקט בעצמי כאשר זוג אוזניות בתוך אוזניי וחלון של “StackOverFlow” (פורום באינטרנט שעוסק בהתמודדות עם באגים בתכנות ) פתוח במחשב

חלופות שהוצעו לפני הפרויקט

המחשבה הראשונה שלי הייתה לקחת את הפרויקט עשיתי שנה שעברה: משחק דו ממד בc# שמכיל בתוכו המון חוקיי מתמטיקה ופיזיקה אבל ללא שכבת תקשורת או שרת ולשדרג אותו להוסיף לו שרת שיאפשר משחק רב משתתפים לכתוב את הקוד שלי בצורה יותר מונחת בעצמים אבל זה היה הכי צפוי הכי מובן מאליו ואני במצבי לא רציתי לעשות פרויקט צפוי רציתי להכין פרויקט לא מובן מאליו שאת החומר שלו לא מלמדים במכללות הטכנולוגיות של אורט וכך עשיתי,כנגד כל הסיכויים

מניעים מאחורי כתיבת הפרויקט

הסיבות בגללם בחרתי בנושא זה לפרויקט:

הסיבה בגללה בחרתי לבצע את הפרויקט שלי בנושא של ניתוח נתונים בעזרת מכונה לומדת מורכבת מכמה סיבות

1.אני נמצא בתחום המחשבים והתכנות מזה 4 שנים.יצא לי לעשות פרויקטים בתקשורת,תכנות משחקי דו ממד ,אתרי אינטרנט פשוטים וכו’ אך לא יצע לי לנגוע בתחום הניתוח וחקר הנתונים. במקומי היום אני לא מתיימר לדעת מה קורא מאחורי הפונקציות בהם השתמשתי בקבצי הpython שלי במובן המתמטי אלא רק ידע תיאורטי אינטואיטיבי על איך האלגוריתמים עובדים ,אך היה לי חשוב להיכנס ברגל או שתיים לכל התחום הנפלא והאינסופי הזה שנקרא-machine learning א כי זה תחום בעליה הרי שאין היום תוכנה שלא משתמשת בספריות לימודי המכונה על מנת להסיק מידע על המשתמשים באותה תוכנה , וב כי אני לא יודע ,ואני אוהב לדעת ואני לא מפחד משמות גדולים כמו machine learning

2.בפרויקט הגמר של לימודי ההנדסאי היה לי חשוב לעסוק במשהו בעל חשיבות ביקוש וערך מוסף לא פרויקט רגיל ופשוט כדי שאנשים שעכשיו קורים את ספר זה ידעו שיש לי ידע ותוצר שכמות הנדסאי התוכנה שגם להם יש את זה קטן מאוד וכך זה יעלה את ערכי בסביבת ההנדסאים האחרים

מטרת הפרויקט :

מטרת הפרויקט בשבילי היה להתנסות וליצור יישום המורכב מתחומים שאני לא מכיר, שלא התנסיתי מעולם ועצם ההתנסות והיצירה מאותו תחום תניב לי פירות בהמשך הדרך.

על אף התפתחות וקידום ממשקי האפליקציות במכשירים החכמים יש להכיר ולהפנים יש את היתרון שיש לאפליקציית web על אפליקציית android, iOS או :windows אפליקציית web הינה אפליקציה אשר אינה תלויה במערכת הפעלה, היא יכולה לעבוד על מכשיר שמותקנת עליו מערכת הפעלה של android או מערכת הפעלה של apple או של Microsoft זה לא משנה ,כל עוד יש לאותו מכשיר דפדפן שיודע לגשת לכתובת של שרת מסוים ולהציג את המידע שהוא מקבל ממנו כדפי web (html).

מטרה נוספת של בניית הפרויקט בשבילי הייתה ליצור תוצר אשר ייתן מענה לאותם גופים, חברות או אנשים פרטיים אשר בידיהם נמצאים נתונים גדולי הנפח אך אין להם את יכולת הטכנית להפוך את הנתונים האלה למידע , אינדיקציה שאפשר לקחת ולעבוד אתה. כאן והפרויקט שלי נכנסים לתמונה: אותם גופים בעלי הנתונים מוזמנים לשלוח לאתר את הנתונים האלה בפורמט של xl, בנוסף הם מתבקשים להזין כמה פרטים טכניים פשוטים למען תפוקת עיבוד הנתונים וגם מה בדיוק הם רוצים לדעת על אותם הנתונים ששלחו

.

כתיבת הפרויקט-ההיבט המנטלי רגשי

חששות ומכשולים בעת כתיבת הפרויקט

עשיית הקורס בפיתוח אתרים גרר אתו את הצורך לעשות עם המרצה את האתר שעובדים עליו במשך כל הקורס. צד השרת אותו המרצה לימד ובו כתבנו את הצד שרת באתר שבנינו בקורס היהnode.js-java script בצד של השרת וזה גרם לסתירה בפרויקט :הצד שרת בפרויקט שלי ייכתב בnode.js אך כל החישובים והניתוח נתונים נעשה בpython איך אוכל להפריח סתירה זו? מחשבה זו הייתה לי כאבן בלב לאורך כל למדת הקורסים וידעתי בתוך תוכי שאהיה חייב להתגבר על סתירה זו. חשבתי באופן טיפשי להקים שרת נוסף על הlocal host ולהעביר את הקבצים כקובץ json

כשהגעתי לנקודת זמן בה אני צריך להתגבר על המכשול חיפשתי באינטרנט כיצד אוכל לקבל קובץ xl לשרת node.js להמיר אותו לקובץ json ולהעביר אותו כפרמטר לסקריפט בpython שאמור להתבצע אוטומטית לאחר העלאת הקובץ מצד הלקוח.

לאחר חיפושים רבים באינטרנט מצאתי ספרייה ב node.js המאפשרת להריץ ולקמפל קובץ כשורת קוד.

וכך עשיתי, אחרי שהמשתמש מעלה קובץ xl לאחת התיקיות בשרת השרת מבצע את אותו סקריפט ב python עם ה path של אותו הקובץ כפרמטר ולאחר מכן אותו סקריפט מחזיר לשרת את תוצאת הנתונים לאחר ניתוחן והשרת מחזיר ללקוח באופן ויזואלי וגרפי את אותם הנתונים.

תהליך הלמידה שעברתי בפרויקט והמסקנות שקיבלתי ממנו

כתיבת הפרויקט הייתה מלווה בהמון חוסרי מידע מבחינתי והמענה לחוסר המידע לא היה מובן מאליו מפני שלא נעזרתי בשום גורם אנושי חיצוני, אף אחד מחברי למכללה לא עשה את הפרויקט באותו נושא וסביבת העבודה ובגלל זה כאשר הגעתי לנקודה בה יש באג בקוד שלי או שאני לא יודע איך להמשיך מאותה נקודה הפתרון לכך לא היה נראה לעין. משאב מידע היחיד שבו הייתי תלוי ולא יכולתי בלעדיו הוא רשת האינטרנט. כיום ישנם כמויות גדולות של פורומים העוסקים בשאלות ופתרון בעיות בתחום התכנות וזאת הסיבה בגללה הצלחתי לבנות פרויקט בתחום ונושא שמעולם לא נגעתי ומעולם לא למדו אותי.

רוב עשיית הפרויקט שלי התבטא בחיפושים אחר פתרונות לבאגים שהיה לי בתוכנית ולא היה לי מושג למה הם קרו או איך אפשר לפתור אותם , אך כל פעם שהצלתי למצוא פתרון לבעיה שאז הייתה לי הגעתי, לאותה תובנה: בעזרת האינטרנט, עם מספיק זמן רצון וניסיונות הרצה תמיד אצליח להתגבר על המכשול שבאותו זמן היה לפניי.

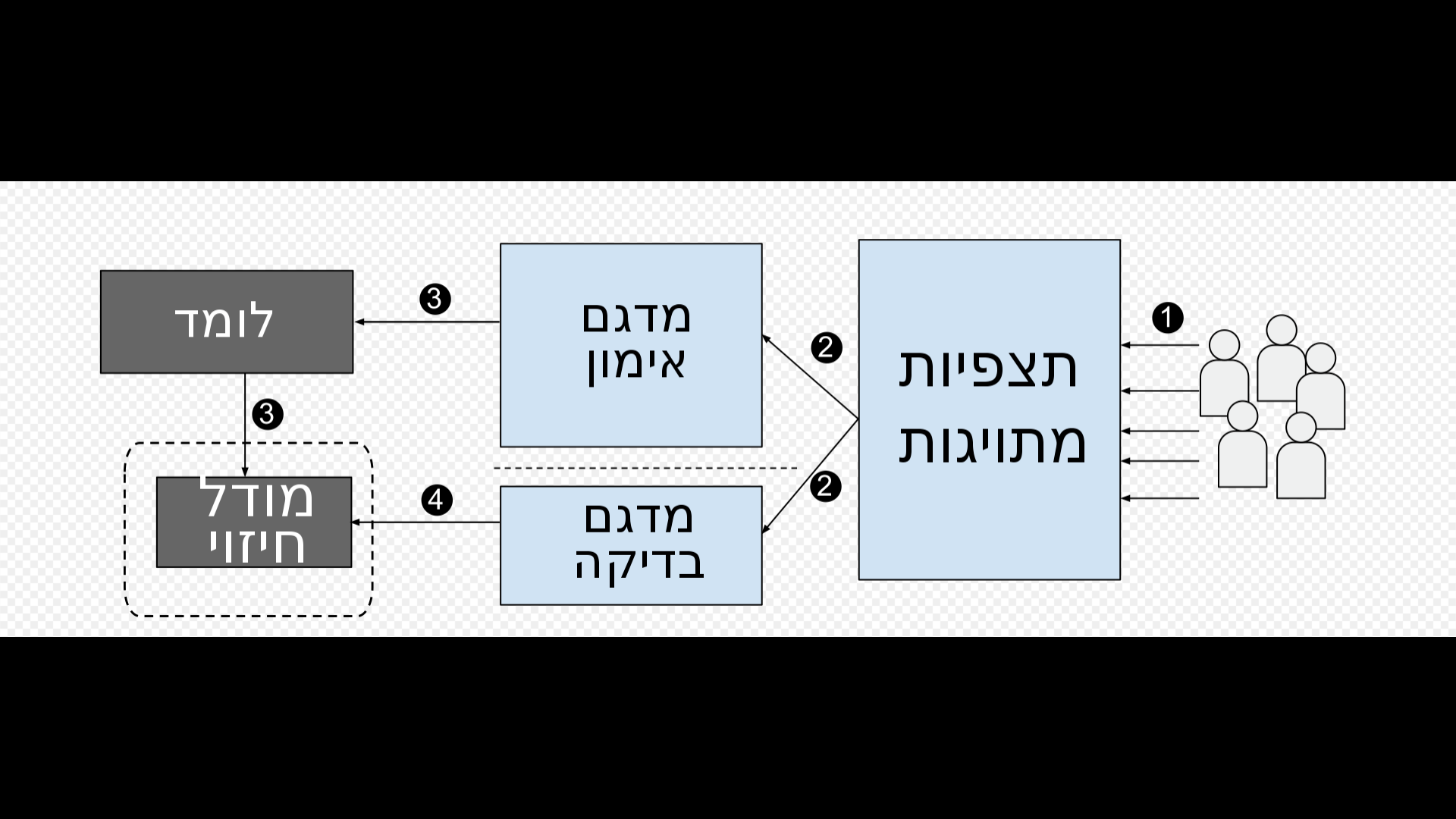
התמודדות עם הקשיים בפרויקט וההתגברות עליהם ב100% מהמקרים העלה לי את הביטחון העצמי והאמונה העצמית וגרם לי להבין שעם מספיק ומשמעת עצמית השמים הם הגבול.

פירוט ותיאור החלקים המרכיבים את הפרויקט

טכניקת לימוד המודל בפרויקט ותהליך יצירתו

הטכניקה שאפשרה לי לבנות את הפרויקט היא "למידה מונחת"  (Supervised learning) היא טכניקה המאפשרת לבנות "מכונות" שלומדות להכליל פתרונות על בסיס מאגר גדול של דוגמאות "פתורות". לדוגמה – הפרויקט שלי יודע לקבל קובץ xl שמכיל בתוכו שם של חברת סטארט-אפ וסכומי כסף שהושקעו בכל מיני תחומים (פרסום ניהול וכו'’) ולבסוף ידע המודל לחזות את הרווח של אותה חברת סטארט-אפ. האפשרות לחיזוי מותנת בכך שיאפשרו למכונה "להתאמן" על קובץ xl בו ישנה כמות גדולה של חברות סטארט-אפ ופירוט השקעות הכספים באותם התחומים והרווח של כל חברה וחברה ביחס להשקעת אותם הסכומים. בשפה פורמלית: למידה מונחית היא טכניקה ב[למידה חישובית](https://he.wikipedia.org/wiki/%D7%9C%D7%9E%D7%99%D7%93%D7%94_%D7%97%D7%99%D7%A9%D7%95%D7%91%D7%99%D7%AA) על מנת ללמוד [פונקציה](https://he.wikipedia.org/wiki/%D7%A4%D7%95%D7%A0%D7%A7%D7%A6%D7%99%D7%94) ("היפותזה") על בסיס סט אימון. סט האימון מכיל דגימות מתוך [התפלגות](https://he.wikipedia.org/wiki/%D7%94%D7%AA%D7%A4%D7%9C%D7%92%D7%95%D7%AA) משותפת של מרחב האלמנטים ומרחב התיוגים. מטרת הלמידה המונחית היא ללמוד על ההתפלגות המשותפת, ומתוך כך לנבא עבור אלמנט את התיוג שלו. בעיות שבהן עוסקת למידת מונחית וכמובן המודל בפרויקט שלי כוללים רגרסיה וסיווג .(פירוט על שתי המושגים בהמשך)

מערכת למידה מונחית

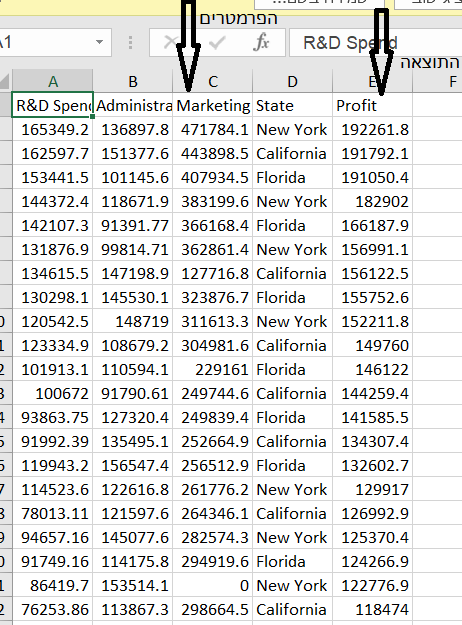


כמו שהסברתי בשלב הראשון נאספות התצפיות: המשתמש באתר תחילה שולח קובץ המהווה מדגם אימון, סט האימון המכיל דגימות באותו נושא אחר כך הלומד (אלגוריתם הלמידה) משתמש בדוגמאות המתויגות ממדגם האימון כדי ללמוד וליצור מודל חיזוי לאותו נושא . בשלב השני המשתמש יוכל לבדוק את מודל החיזוי באמצעות סט בדיקה ולפי תוצאת מדגם הבדיקה ניתן יהיה לבדוק את טיב המודל.

האופן בו השרת מצפה לקבל את המידע מהלקוח ודרישותיו

מנגנון ניתוח ועיבוד המידע באתר שלי דורש את המידע בפורמט של קבצי xl. על מנת שמנגנון זה יעבד את המידע כנצפה , קבצי ה-xl צריכים להיות מאורנים באופן הבא:

השורה הראשונה בקובץ מכילה את שמות הפרמטרים שלדעתו של המשתמש משפיעים על התוצאה והעמודה הימנית ביותר צריכה להיות המילה עמודת התוצאה. החל מהשורה השנייה ועד השורה האחרונה של הקובץ, כל שורה שמהוות בעצם מקרה צריכה להכיל את ערכי הפרמטרים לאותו מקרה/ שורה והעמודה הימנית באותה שורה מכילה את התוצאה של אותו המקרה לדוגמא בבעיית הרגרסיה שיש לנו כאן נתון קובץ של 50 חברות סטרטאפ שמפרט את הסכומים שהשקיעו בכל תחום(הפרמטרים שלנו) והרווח (התוצאה שלנו):



ה"פרמטרים" "והתוצאה" הם המגדירים את המושג רגרסיה :

בבעיות של רגרסיה מחלקים את הפרמטרים והתוצאה לשתי סוגים של משתנים-משתנים תלויים ומשתנים בלתי תלויים.

המשתנים הבלתי תלויים הם בעצם הפרמטרים, המשתנים שבעצם משפיעים על התוצאה ואינם תלויים אחד בשני. המשתנים התלויים בד"כ הם יחידים כלומר משתנה אחד והוא בעצם התוצאה. מכנים אותו כמשתנה תלוי כי ערכו בעצם תלוי ונקבע לפי ערכים של המשתנים הבלתי תלויים ,כמו בדוגמא שלהעליל המשתנים הבלתי תלויים אלו בעצם כמויות הכסף אשר השקיעו בכל תחום ,בכל פרמטר והמשתנה התלוי, הוא המשתנה שערכו נקבע לפי הערכים של המשתנים הבלתי תלויים, כלומר כמויות הכסף שהשקיעו בכל תחום (פרסום ,ניהול, שיווק) והוא הרווח -כמות הכסף שהסטארט-אפ הרוויח.

באופן הזה בעצם צריכים קבצי הxl אשר נשלחים מהלקוחות להיות מאורגנים: משתנים בלתי תלויים השם שלהם והערכים שלהם והחלק השני-התוצאה בכל אחד ואחד מהמקרים הללו.

מן המובן מאליו הלוקח כמובן צריך לידע את השרת בכמה פרמטרים בלתי תלויים מדובר על מנת שיוכל לחלק את המידע כראוי

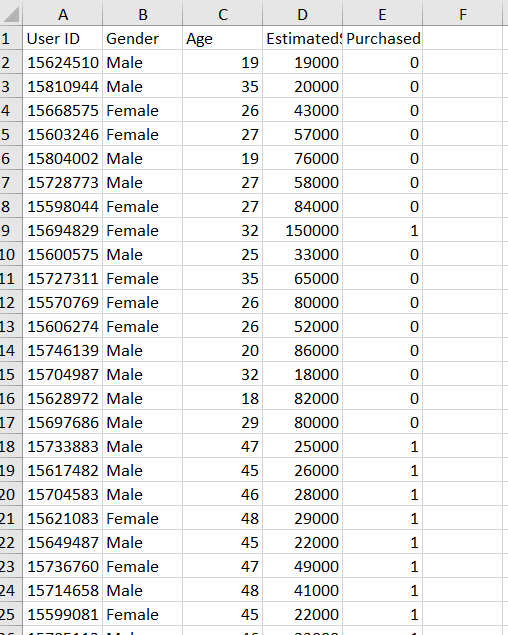
סיווג

**סיווג** הוא פעולה שמחלקת קבוצת עצמים לתת-קבוצות. למשל: ניתן לסווג את לקוחות הבנקים שמבקשים משכנתא לקבוצות למידה שונות: יחזירו או לא יחזירו **כלל סיווג**. לעיתים, כמו בדוגמה הראשונה, אין סיווג "נכון". במקרים אחרים (כמו בדוגמה השנייה) פעולת הסיווג אמורה לחקות חלוקה שאינה תלויה בפעולת הסיווג עצמה, ואז כלל הסיווג נקרא

באתר שלי ישנו מודל אשר מ קיבל קובץ עם פרטים על אנשים גיל תעודת זהות ומשכורת ותשובה האם הם יקנו את מה שפורסם להם ברדת החברתית בה אותם אנשים נמצאים

1-קנה

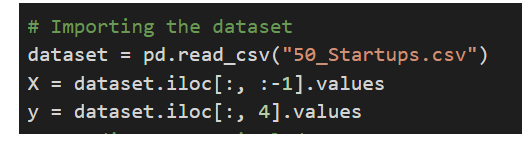
0-לא קנה



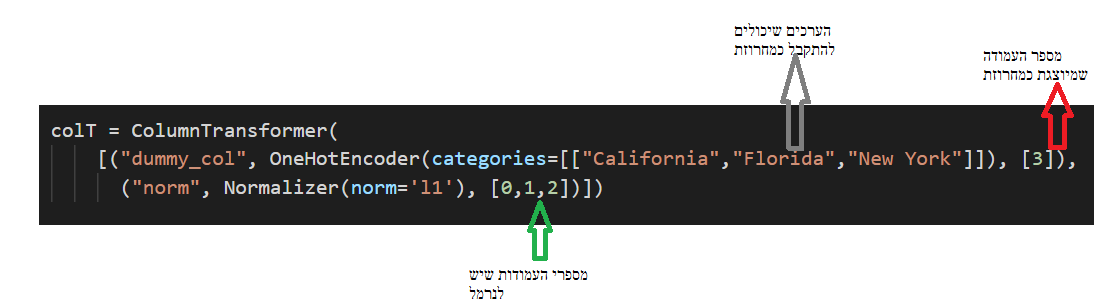
סדר הפעולות בסקריפט מכריעת הנתונים ועד להכנת במודל למדגם הבדיקה :

א. בניית המודל

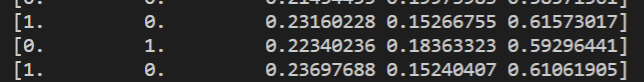
1.חלוקת סט נתוני האימון ולאחר מכן חלוקה למשתנים תלויים ובלתי תלויים X ו Y שאלו בעצם שתי וקטורים



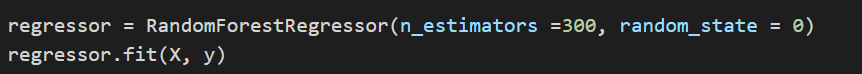
2.כינוס המידע לחטיבת מידע אחת(נרמול) הפיכת עמודות בהן הערכים מיוצגים כמחרוזות לערכים מספריים



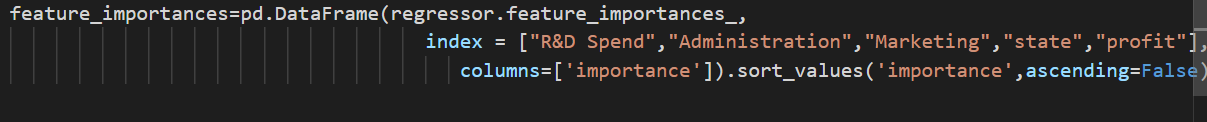
האופן שבו הנתונים מיוצגים לאחר מכן



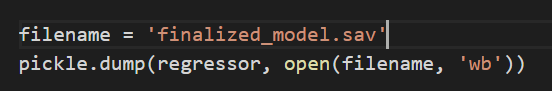
3.יצירת האובייקט שמכיל את האלגוריתם שיאמן את המודל והפעלתו



4. ניתוח החשיבות של כל פרמטר ופרמטר על התוצאה ומיונו בסדר עולה



5. שמירת המודל ויכולתו על קובץ



ב. בחינת יכולות המודל על חיזויים חדשים

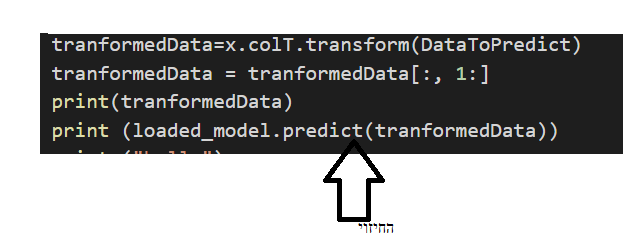
1.קריאת ערכי הפרמטרים



2טעינת המודל מן הקובץ



3.התאמת הנתונים באופן שהמודל מכיר וחיזוי התוצאה



הסבר עקרוני על האלגוריתם בעזרתו המודל מתאמן על סט הנתונים-RandomForestRegression/ classifier

צעד ראשון- לוקחים k נקודות/מקרים מסט הנתונים עליו המודל מתאמן

צעד שני-בונים עץ החלטה לאותם נקודות (אפרט מיד מה הוא עץ החלטה)

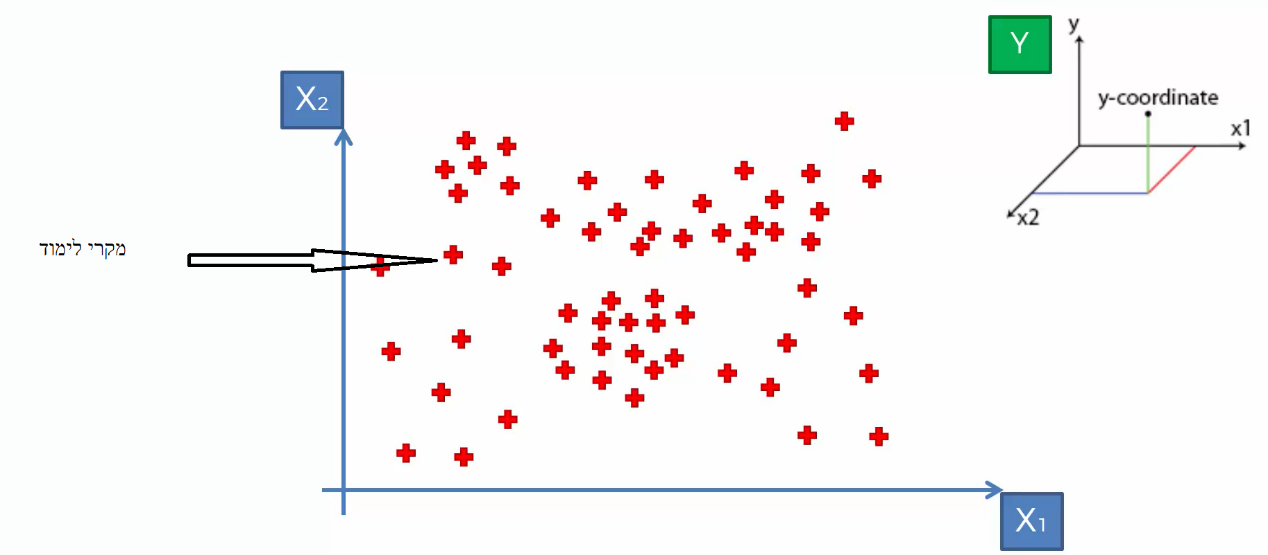
צעד שלישי-בוחרים מספר של עצים שאנחנו רוצים לבנות (forest) וחוזרים על צעדים אחת ושתיים

צעד רביעי (החיזוי)-לנקודה חדשה מכניסים אותה לכל עצי ההחלטה שקיימים לפי המספר שקבענו ומחשבים את ממוצע ערך הy שיתקבל מתוך כל אחד מהעצים האלו

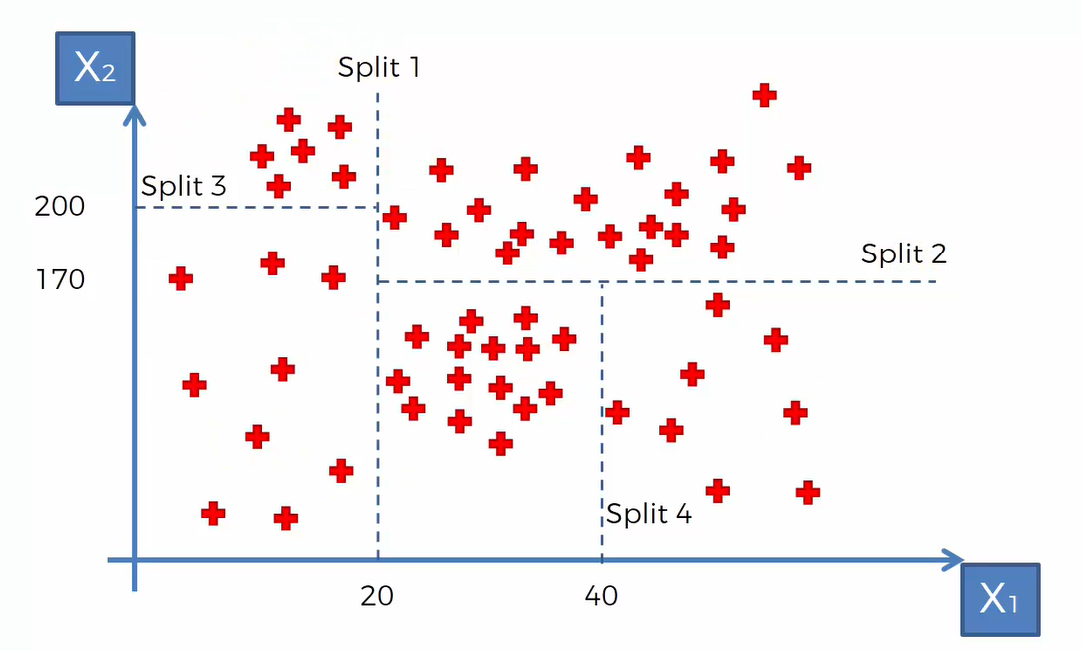
מהו עץ החלטה?

עץ החלטה המספק מיפוי בין תצפיות לערכים המתאימים עבורן. עץ החלטה יכול לשמש כמודל חיזוי, הממפה תצפיות על פריט ויוצר מסקנות על ערך היעד של הפריט. שמות תיאוריים יותר עבור עץ ההחלטות הם עצי סיווג או עצי רגרסיה

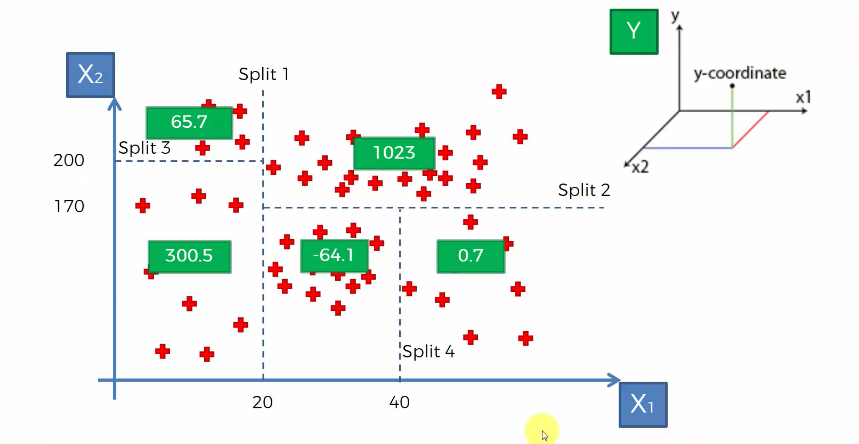
דוגמא לעץ החלטה על נתונים שמורכבים משתי משתנים בלתי תלויים ומשתנה תלוי, התוצאה



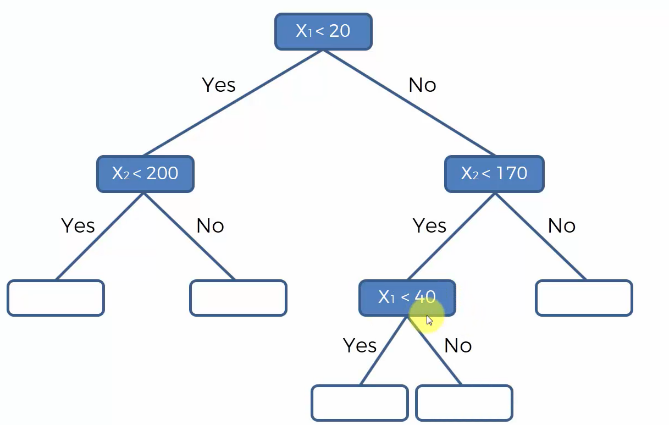
פיצול הנתונים לקבוצות ערכים מסוימות שנקבעו בגלל חוקים מתמטיים שאיני יודע ואין ביכולתי להסביר:



חישוב ממוצע שיעור הY לכל תחום באוסף הנתונים(שיעור הy ידוע לכל נתון ונתון כי אלו מקרי למידה לא מקרים שבוחנים את המודל אתם)



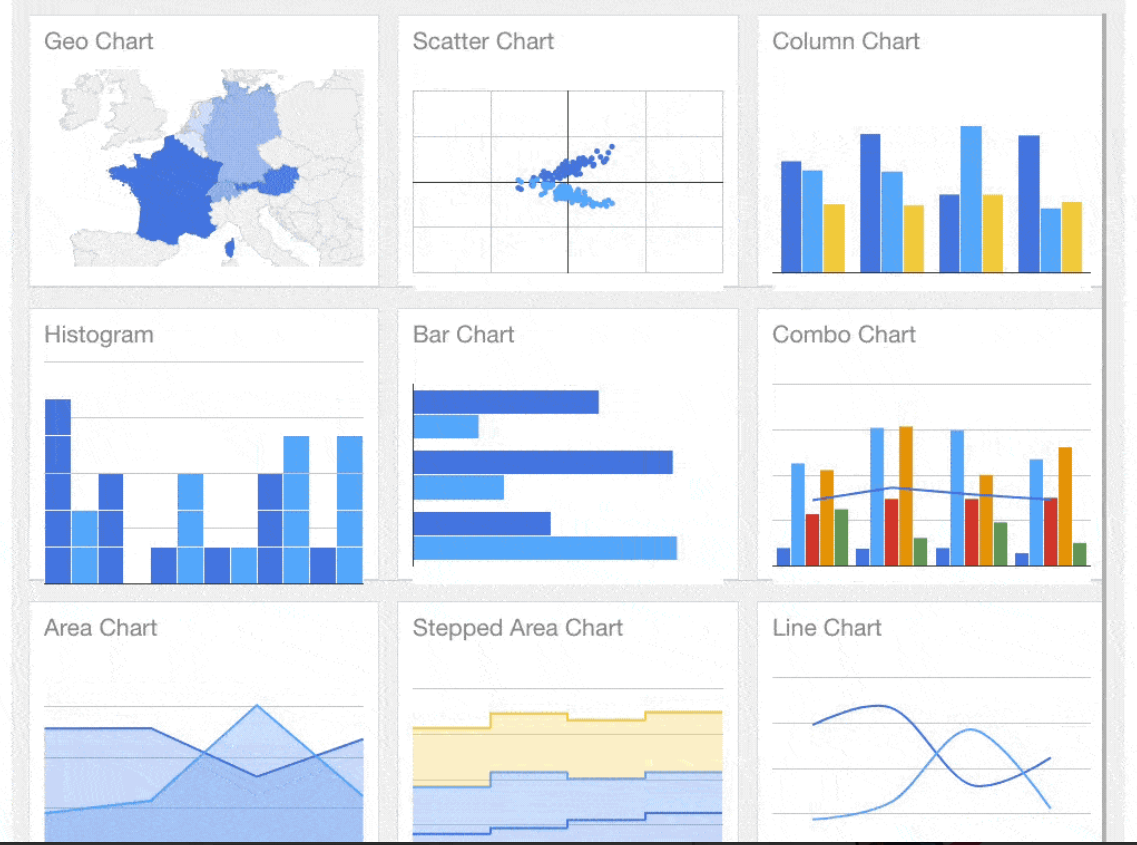
כאשר מגיעה נקודה חדשה ממפים ותוחמים אותה לפי העץ ההחלטה שהתקבל והשיעור הY של אותה נקודה כלומר, תוצאת החיזוי שלה ייקבע לפי שיעור הy הממוצע שחושב באותו התחום שהנקודה נמצאת בו



כך בעצם באופן עקרוני מתבצע האלגוריתם בו המודל משתמש כדי ללמוד לחזות ולנתח את הנתונים אשר הוא מקבל

הדרך שבה משתמש האתר רואה את כל כברת עבודת הניתוח הנתונים הללו בצורה ויזואלית

כאשר משתמש באתר שולח קובץ xl עליו הוא מתכנן לבנות את המודל הוא מקבל חזרה גרף עמודות שכל עמודה בה מציינת את גודל החשיבות של פרמטר מסוים, משתנה בלתי תלוי ביחס לאותו מודל. כמספר המשתנים הבלתי תלויים, כמספר העמודות בגרף.

המשתמש רואה את המודל כגרף עמודות ואת ניסיונות החיזוי כתתי גרף עמודות באותו הדף

יש להבין דבר חשוב:

שמשתמש שולח קובץ xl עליו המודל יתאמן וכאשר הוא שולח ניסיון מבחן/חיזוי גרף הנתונים שהמשתמש מקבל מייצג מידע שונה: שהמשתמש שולח נתוני מבחן הוא לא מקבל בחזרה את החשיבות של כל פרמטר ופרמטר כערך גרפי אלא הוא מקבל כמובן את תוצאת החיזוי כערך מספרי וגרף שמייצג את הערכים של ששלח כמידע מנורמל



הרשאות והיררכיות בפרויקט:

משתמשים אשר אינם רשומים באתר לא יכולים לעשות דבר חוץ מלצפות בגרפים וניסיונות חיזוי. משתמשים אשר הינם רשומים לאתר יכולים לבנות מודל, יכולים לבחון את המודל שיצרו ואף למחוק אותו.

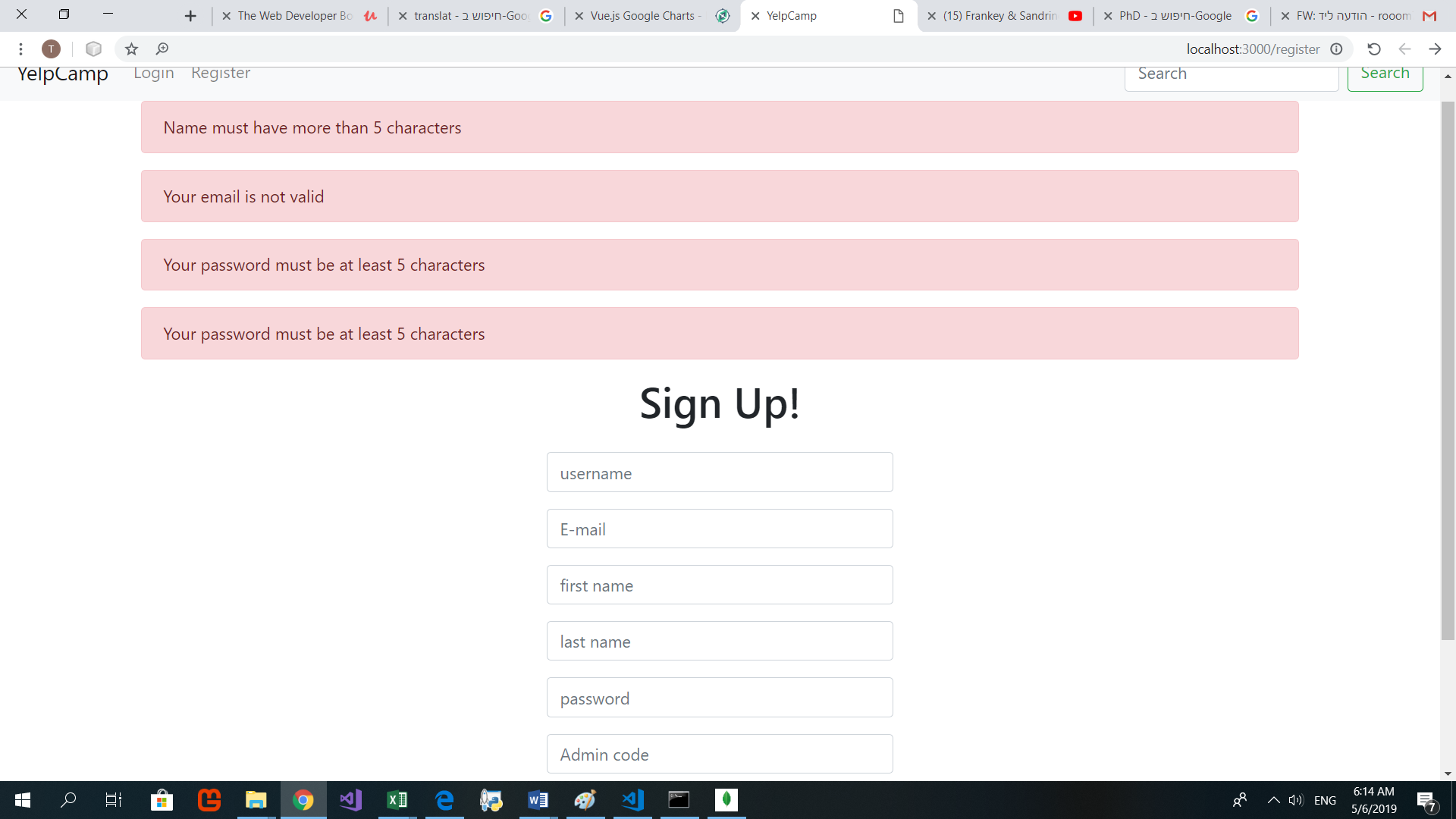
בנוסף, כל משתמשי האתר יכולים להגיב על מודלים וניסיונות חיזוי של אחרים אך כמובן מהמובן מאליו שלא למחוק תגובה או מודל כלשהו אשר אינם שלהם.

ישנו אדם אחד בכל המערכת אשר יכול לעשות כרצונו:הוא יכול למחוק מודל, למחוק ניסיון חיזוי, תגובה שלדעתו לא הייתה ראויה וכו'. הבן אדם הזה מכונה "admin " והמערכת מזהה אותו בעזרת הקשת קוד בעת הרישום של ה"admin" לאתר

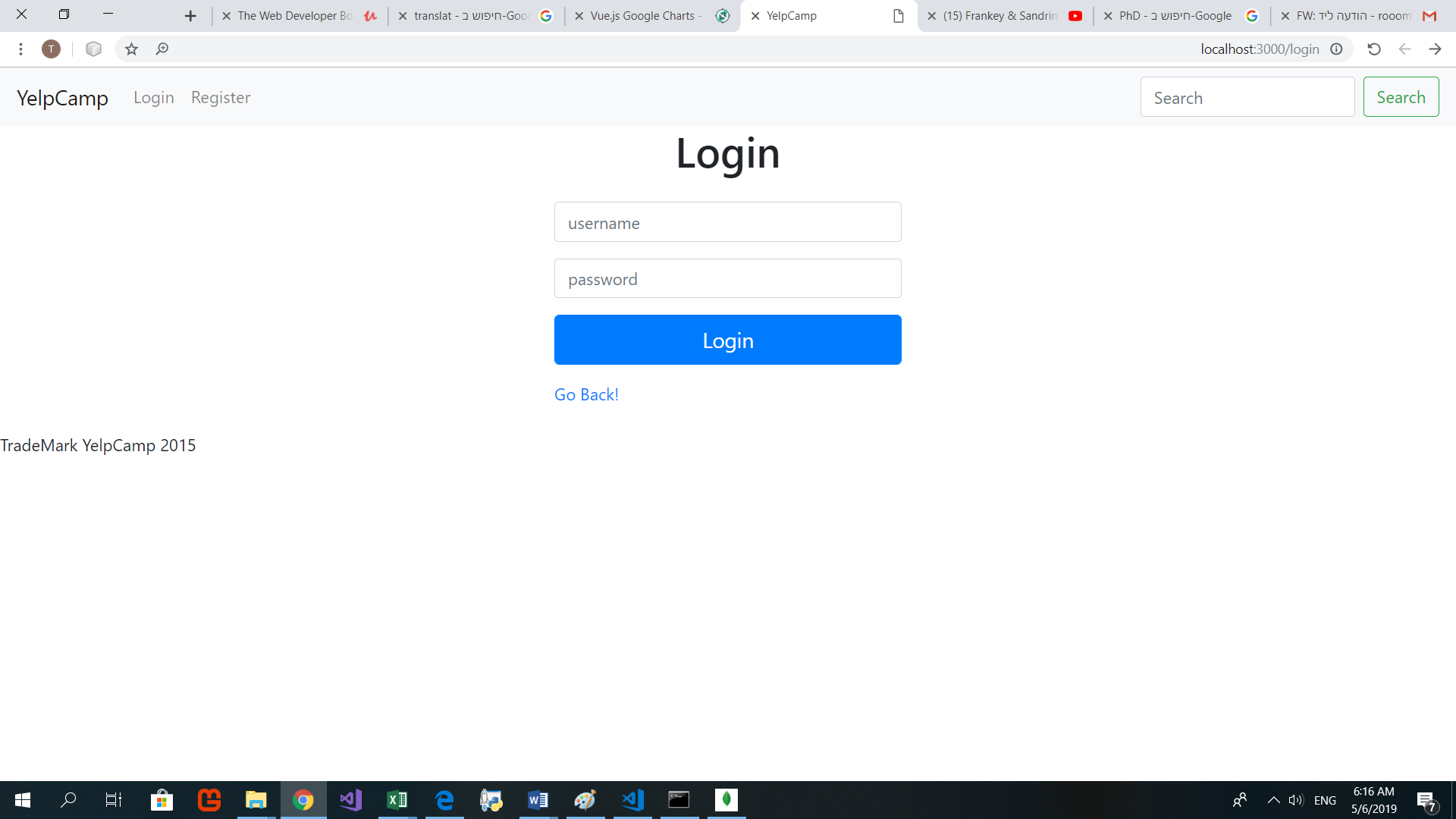
מעט אבטחה בפרויקט

שמשתמש נרשם לאתר,מטעמי אבטחה הסיסמא שלו מקודדת באמצעות "hash " בצורה עילאית ומכניסה את הסיסמא שלו למסד נתונים כצופן מורכב ומסובך שאין להבינו.

Register page:



Login page



הישויות ותכונותיהן במסד הנתונים

User

var UserSchema=new mongoose.Schema({

username:String,

password:String,

avatar:String,

firstName:String,

lastName:String,

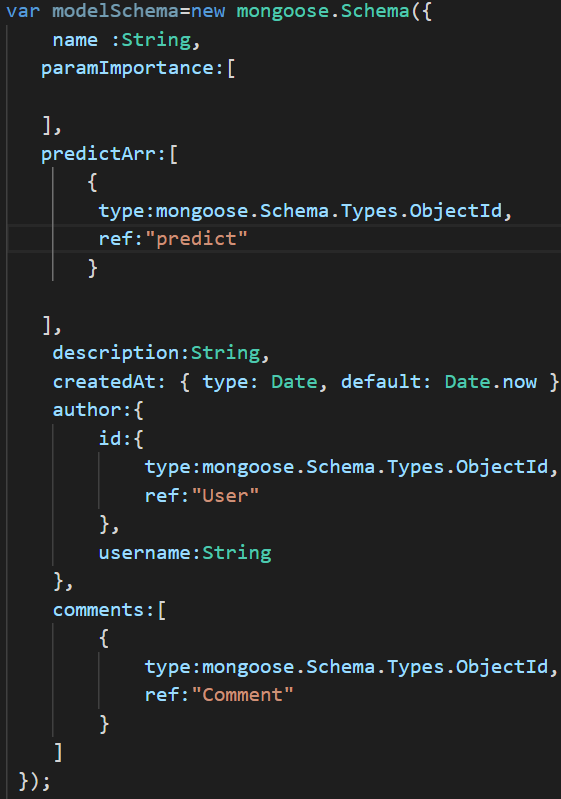
email:String,

isAdmin:{type:Boolean,default:false}

});

הישות משתמש אוגר בתוכו את התכונות אשר רואים כאן(שם משתמש סיסמא שם פרטי שם משפחה וערך שאומר האם הוא מנהל האתר או לא)

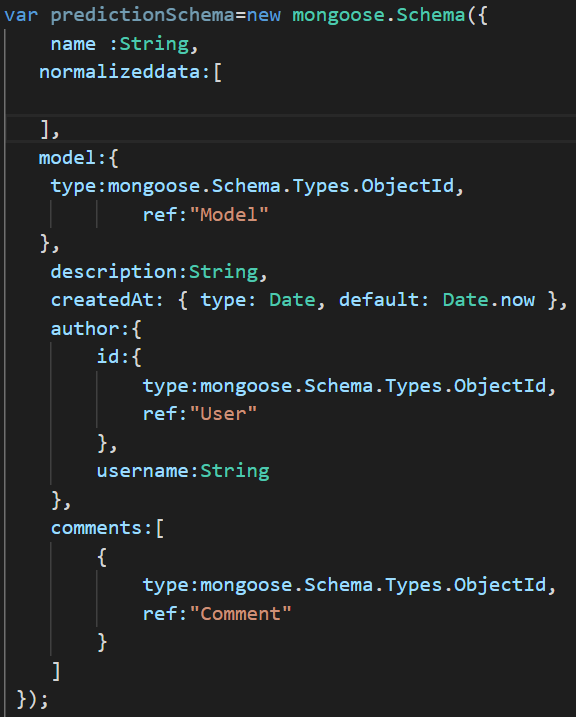
model



הישות model היא ישות מורכבת מפני שיש לה קשרים עם ישויות אחרות:

יש לה קשר של יחיד לרבים עם ישות משתמש: למשתמש אחד הרבה מודלים אך למודל מסוים משתמש אחד ,יש לה קשר של יחיד לרבים עם ישות חיזוי: למודל אחד הרבה חיזויים וקשר של יחיד לרבים עם ישות תגובה: לכל מודל הרבה תגובות אך תגובה מסוימת משויכת למודל אחד

Prediction



הישות prediction היא ישות מורכבת מפני שיש לה קשרים עם ישויות אחרות:

יש לה קשר של יחיד לרבים עם ישות משתמש: למשתמש אחד הרבה חיזויים ,יש לה קשר של יחיד לרבים עם ישות תגובה : לחיזוי אחד הרבה תגובות וקשר של יחיד לרבים עם ישות מודל : לכל מודל הרבה חיזויים

יישות תגובה

var commentSchema=mongoose.Schema({

text:String,

createdAt: { type: Date, default: Date.now },

author:{

id:{

type:mongoose.Schema.Types.ObjectId,

ref:"User"

},

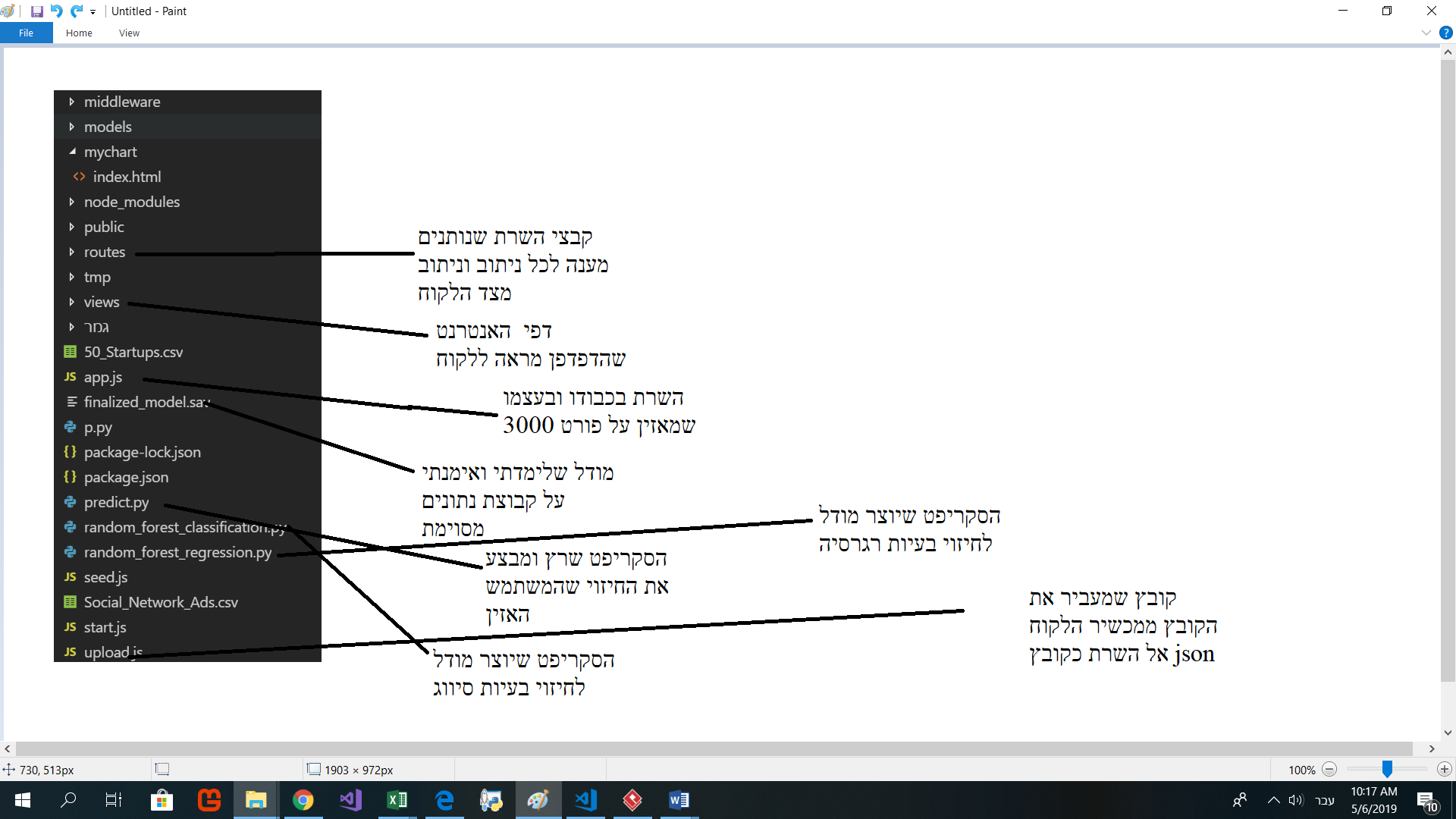
username:String

}

});

קשר של יחיד לרבים עם היישות משתמש :למשתמש אחד יש הרבה תגובות

הקבצים ותפקידם בפרויקט:



בביבליוגרפיה

[www.stackoverflow.com](http://www.stackoverflow.com)

[www.github.com](http://www.github.com)

[www.udemy.com](http://www.udemy.com)

[www.youtube.com](http://www.youtube.com)

1. [↑](#footnote-ref-1)