למידת מכונה 2020 – מטלה 1

על כל סטודנט לפתוח חשבון קאגל משלו, כל המטלות המעשיות יעשו באתר קאגל (*נספח 1*).

שימו לב, ההגשה ביחידים.

תחרות

התחרות עבור מטלה 1 נקראת **Titanic: Machine Learning from Disaster** (*נספח 2*).

הקישור לתחרות נמצא כאן: https://www.kaggle.com/c/titanic

בתחרות זו עליכם לבצע משימת סיווג.

עליכם לסווג את נוסעי הטיטניק בין 2 תוויות, אלה ששרדו את ההתנגשות ואלה שלא שרדו אותה.

לרשותכם מספר מועט של פיצ'רים, שאיתם אתם יכולים לשחק. אתם יכולים להוריד פיצ'רים או לחשב פיצ'רים חדשים שמתבססים עליהם.

תיאור הדאטה נמצא כאן: https://www.kaggle.com/c/titanic/data

קראו אותו היטב, הבינו מה המשמעות של כל עמודה וחישבו כיצד כל פיצ'ר עשוי להשפיע על התוצאה (מידת ההישרדות של הנוסע).

במקרה שיש לכם שאלות בנוגע לתחרות ולדאטה, אתם תמיד יכולים לכתוב ולקרוא שאלות של אחרים בפורום של התחרות: https://www.kaggle.com/c/titanic/discussion

אם אתם רוצים לראות מחברות של אחרים, לקבל השראה מהדרך בה הם פתרו את הבעיה, אתם יכולים להיכנס ל- https://www.kaggle.com/c/titanic/notebooks

שימו לב שהניקוד בתחרות מתבצע על ידי דיוק (Accuracy):

> **Metric**
>
> Your score is the percentage of passengers you correctly predict. This is known as accuracy.

דרישות הפתרון

על הפתרון שלכם להופיע במחברת אחת שתוגש בשני קבצים: קובץ ipynb שניתן להורדה מתפריט ה- File במחברת, וקובץ ה- html שניתן ליצירה מקובץ ה- ipynb על ידי ההדרכה פה.

יש לשים את 2 הקבצים בקובץ ZIP (שימו לב שזה צריך להיות ZIP ולא RAR) יחיד ולהעלות את קובץ ה- ZIP למטלה במודל.

יש לרשום בתחילת המחברת את השם ותעודת הזהות ולצרף קישור לדף המשתמש שלכם ב- Kaggle.

בשלב הראשון עליכם לחקור את הדאטה. מומלץ להשתמש בגרפים וטבלאות ולהסביר את ניתוח הדאטה שלכם בתאי Markdown. זה המקום לנתח את הפיצ'רים והשפעתם על הסיווג. זה גם המקום ליצור פיצ'רים חדשים אם אתם מוצאים לנכון לעשות זאת. תתייחסו למטלה הזאת כחידה, השלב הראשון בפתרון החידה הוא הבנת כל הנתונים והסקת מסקנות מיידיות. לאחר מכן, ניתן להתחיל לפתור את החידה.

עליכם לסווג את הדאטה על ידי שימוש ב- Logistic Regression או ב- MLP. הסבירו את בחירתכם ונמקו אותה (אם בחרתם ביותר מאחד, הסבירו מה ההבדלים ונתחו את התוצאות בהתאם).

חלקו את ה- Train Dataset ל- 2 קבוצות: Train ו- Validation. השתמשו בקבוצת ה- Validation כ- Test עבור הניסויים שלכם לפני ההגשה האמיתית.

שימו לב שאתם יכולים להגיש רשמית עד 10 הגשות ביום:

Submission Limits

You may submit a maximum of 10 entries per day.

כך, שכדי לא לבזבז אותן על ניסויים, מומלץ לבחון תיאוריות על ה- Validation לפני אימון מלא על כל ה- Train והגשת ה- Test להערכה.

נסו ערכים שונים של היפר-פרמטרים, נסו אימונים על תתי-קבוצות שונות של פיצ'רים ובדקו מה הם הפיצ'רים הטובים ביותר עבור החיזוי (יש לבדוק על ה- Validation לפני שבודקים על ה- Test).

הציגו גרפים של שגיאת (Loss) האימון ושל שגיאת הולידציה (לפי היפר פרמטרים שונים). ודיוק (Accuracy) האימון והולידציה (אתם יכולים גם להשוות בין האימון לולידציה, לנתח את ההבדלים ולהסיק מסקנות).

לבסוף, צרו את קובץ ההגשה והגישו אותו לתחרות. צלמו את דף ההגשות שלכם (אם יש לכם יותר מ- 10 הגשות, זה יספיק אם תצלמו רק את 10 ההגשות האחרונות) והדגישו את ההגשה הטובה ביותר שלכם. העלו את תמונת ההגשות למחברת. צלמו את המיקום שלכם ב- Leaderboard והעלו את גם התמונה הזאת למחברת.

סכמו את העבודה שלכם (הסיכום לא חייב להיות ארוך) והסבירו מה עשיתם, מה עבד טוב ומה עבד פחות טוב. הוסיפו בסוף המחברת רשימת מקורות למחברות אחרות שלקחתם השראה מהן ולאתרים או מדריכים שלמדתם מהם.

מבנה המחברת

1. שם, ת.ז., קישור למשתמש ה- Kaggle שלכם.
2. הסבר על התחרות, מה אתם מנסים לעשות ובאילו אמצעים אתם הולכים להשתמש.
3. חקירת הדאטה.
4. ניסויים על פיצ'רים שונים, מודלים שונים ובחירת היפר-פרמטרים (בדיקות יש לעשות על ה- validation).
5. הצגת גרפים וניתוח התוצאות.
6. תמונות ההגשות והמיקום ב- Leaderboard.
7. סיכום העבודה, מה עבד ומה לא עבד (אם יש דברים כאלה).
8. רשימת מקורות.

מבנה הציון

לכל אחד מהשלבים במבנה המחברת יש ערך בציון. גם למראה של המחברת יש ערך בציון. חשוב לשמור על מחברת יפה מסודרת וברורה. חשוב לשמור על קוד נקי ברור ופשוט. חשוב מאוד להבין שבסופו של דבר אתם תמיד תעבדו עם עוד אנשים, והם יצטרכו להבין את מה שכתבתם בצורה הקלה והנוחה ביותר. תחשבו שאתם מסבירים את מה שאתם עושים לתלמיד תיכון שמכיר פה ושם את המינוחים אך הוא לא מומחה גדול בנושא. הסבירו במילים פשוטות ושלבו קישורים במידה ואתם חושבים שההסבר במחברת לא מספיק. זכרו שתמונות וגרפים מובנים הרבה יותר ממילים, וביחד, תמונות גרפים ומילים, אפשר להסביר כמעט כל דבר לכל אחד.

קוד פשוט, מסודר, מוסבר ונקי – 10%

מחברת מסודרת, מוסברת, נקייה ונוחה לקריאה – 10%

השקעה, לימוד עצמי, עשייה מעבר למינימום הנדרש – 10%

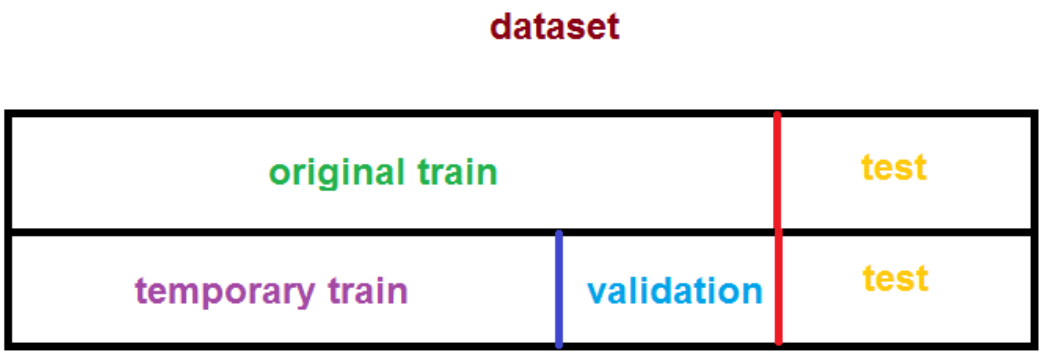
מימוש נכון של דרישות הפתרון ומבנה המחברת – 70%

יתכנו בונוסים על התרגיל הנוכחי, שיתקבלו עבור ביצועים שיפתיעו אותנו לטובה (לדוגמה, חקירת דאטה מושקעת מאוד), כמובן עד ניקוד מקסימלי של 100 בתרגיל הנוכחי.

הערות

שימו לב שכאשר רוצים לבצע ניסויים לפני הגשה, מחלקים את ה- original train ל- temporary train ו- validation.

החלוקה תראה בערך כך:



ה- Temporary Train משמש אותנו כ-Train וה- Validation משמש כ- Test שניתן לחשב בקלות את ערך השגיאה שלו, כדי לדעת אם אנחנו בכיוון או לא.

לאחר שבוחרים את כל הפיצ'רים וההיפר-פרמטרים, יש לאמן מחדש על ה-Original Train ואז ליצור קובץ הגשה על ה- Test.

אתם יכולים להתנסות ביחסי חלוקות שונים ל- Temporary Train ול- Validation, הסבירו בקצרה למה בחרתם ביחס המתאים (אין חוקים מוגדרים בנושא הזה, אתם מוזמנים לחפש באינטרנט הסברים על יחסי החלוקות שעובדים בדרך כלל).

מומלץ להשתמש בפונקציות קצרות וייעודיות ולא ליצור כפל קוד. מומלץ לתת שמות משמעותיים לפונקציות ולמשתנים ולהסביר לפני כל פונקציה מה היא עושה (אפשר גם בהערה).

יש להשתמש במודל Logistic Regression (על ידי האופטימייזר LogisticRegression או על ידי האסטימייטור SGDClassifier עם log loss או על ידי שניהם), או במודל MLP (על ידי MLPClassifier). במודלים אחרים תשתמשו בתרגילים הבאים.

חשוב להבין שבהרבה מקרים, אין נכון או לא נכון. ניסוי וטעייה זה חלק בלתי נפרד מלמידת מכונה. תבינו מה אתם עושים ותסבירו את המהלכים שלכם. יש מהלכים שיהיו מוטעים מבחינה מתודולוגית (לדוגמה, אימון על ה- Train ובדיקה, גם על ה- Train. זו טעות מכיוון שאם התאמנו על ה- Train, הגיוני שהטעות על ה- Train תהיה נמוכה, אך זה לא אומר הרבה על יכולת ההכללה של המודל ועל התוצאה שתהיה לו על ה- Test), ומהלכים שיהיו נכונים מבחינה מתודולוגית (לדוגמה, לחלק ל- Temporary Train ו- Validation ואז לבחון את יכולת ההכללה של המודל על ה- Validation).

חשוב להראות הבנה, לכתוב מה ניסיתם ועבד ומה ניסיתם ולא עבד, ולנסות להסביר את זה.

קישורים נחוצים

אתר קאגל – https://www.kaggle.com/

תחרות הטיטניק – https://www.kaggle.com/c/titanic

הדאטה של תחרות הטיטניק – https://www.kaggle.com/c/titanic/data

הפורום של תחרות הטיטניק – https://www.kaggle.com/c/titanic/discussion

המחברות של תחרות הטיטניק – https://www.kaggle.com/c/titanic/notebooks

אתר להמרת קבצי ipynb לקבצי html – https://htmtopdf.herokuapp.com/ipynbviewer/

אתר ללימוד Markdown – https://guides.github.com/features/mastering-markdown/

סיכום הדרישות הטכניות

דאטה: לבצע ניתוח דאטה למשימת קלסיפיקציה

מודל: לבחור אחד או יותר מ- Logistic Regression, MLP

ולידציה: לבצע חלוקה של ה- Train ל- Temporary Train ו- Validation

תוצאות: לבצע ניתוח תוצאות למשימת קלסיפיקציה (עם Accuracy)

נושאים שנלמדים במטלה זו

Data Investigation for Classification

Logistic Regression

MLP

Train, Validation, Test

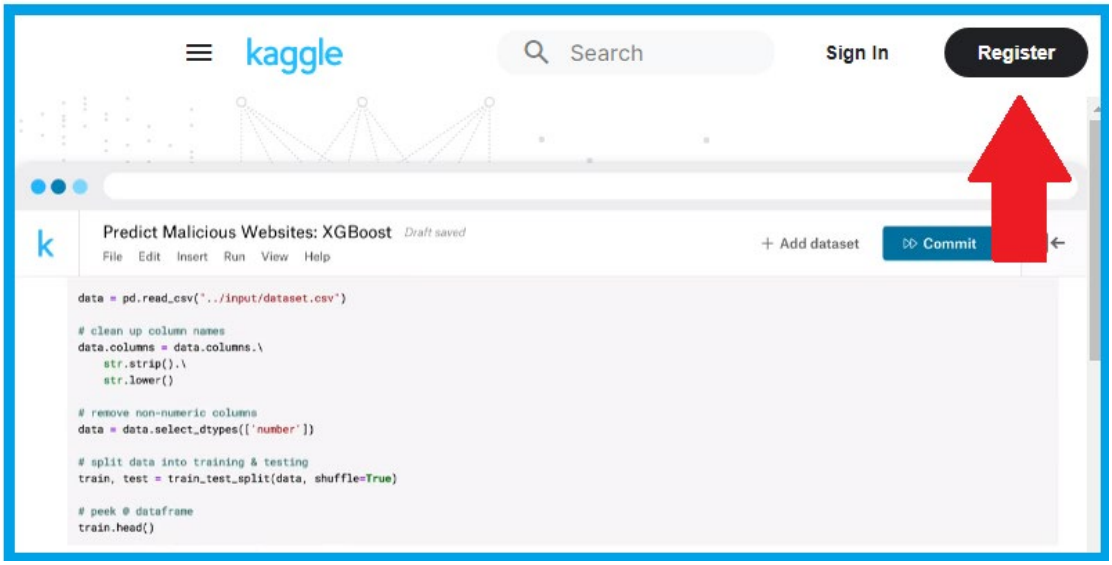
Accuracy

בהצלחה

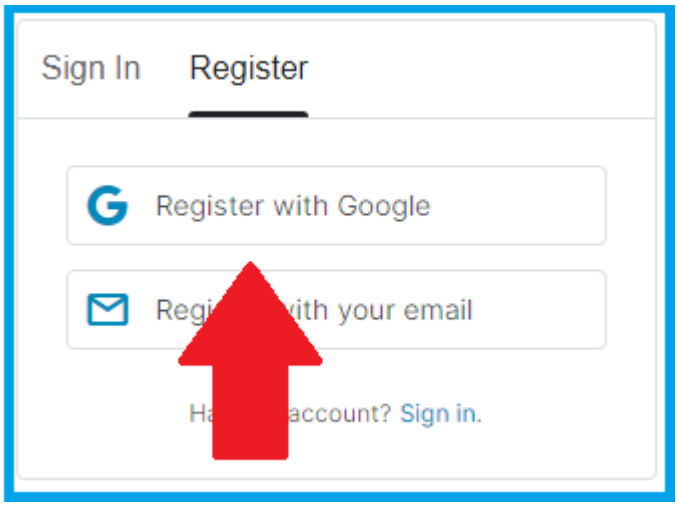
נספח 1 - הוראות פתיחת חשבון בקאגל

היכנסו לאתר https://www.kaggle.com/

לחצו על כפתור **Register**:



לחצו על **Register with Google**:

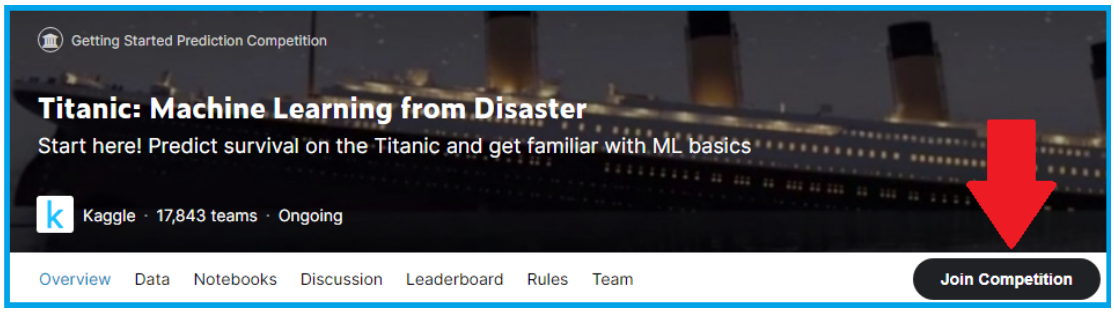


היכנסו עם משתמש Gmail וצרו את משתמש ה- Kaggle שלכם.

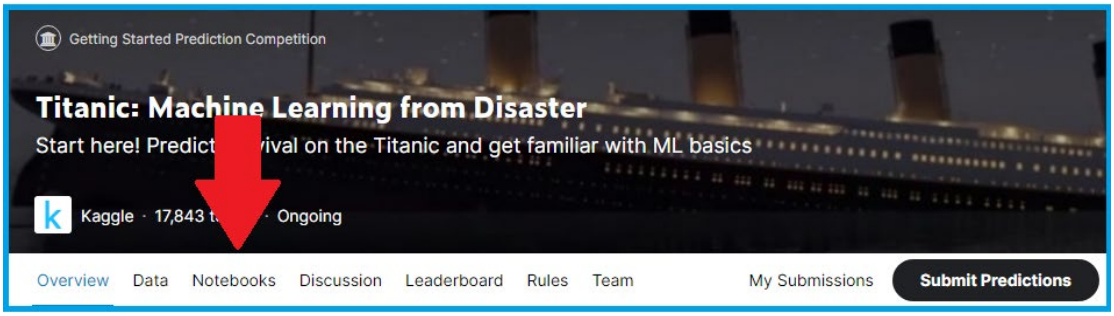
נספח 2 - הוראות הרשמה לתחרות ושימוש במחברת Kaggle

היכנסו לאתר התחרות https://www.kaggle.com/c/titanic

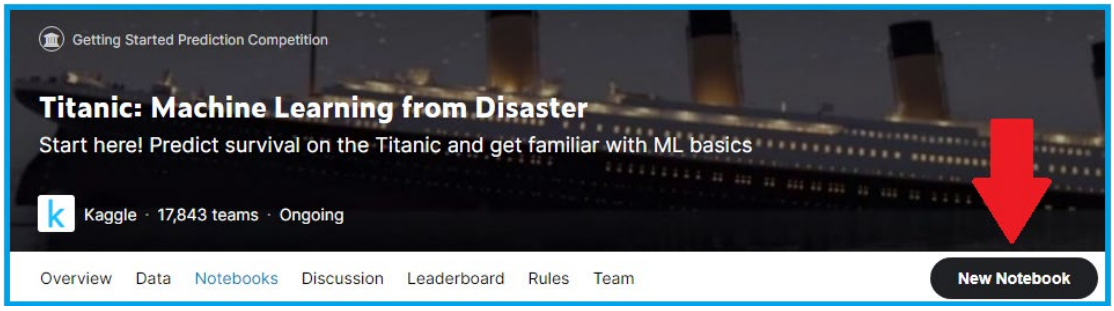
ולאחר מכן לחצו על **Join Competition**:



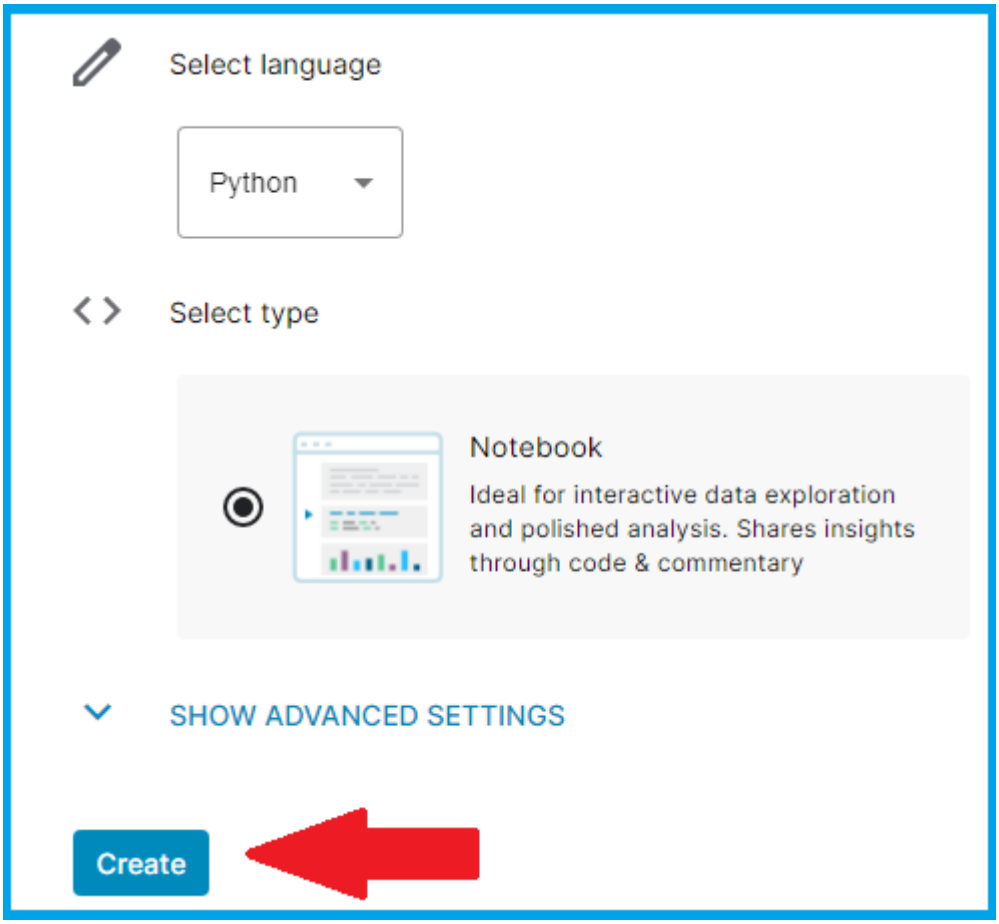
לאחר שנרשמתם לתחרות לחצו על **Notebooks**:



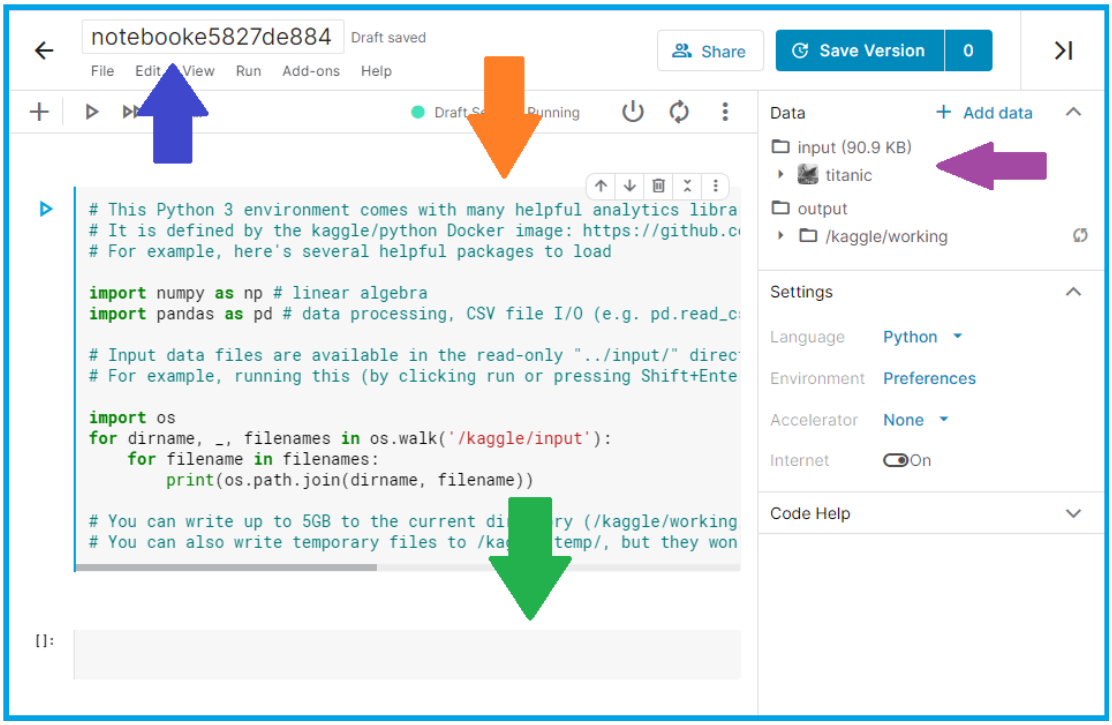
ליצירת מחברת חדשה עם הדאטה של התחרות, לחצו על **New Notebook**:



ודאו שהשפה היא Python וסוג הקובץ הוא Notebook. לאחר מכן, לחצו על **Create**:



תפתח לכם מחברת כזאת:



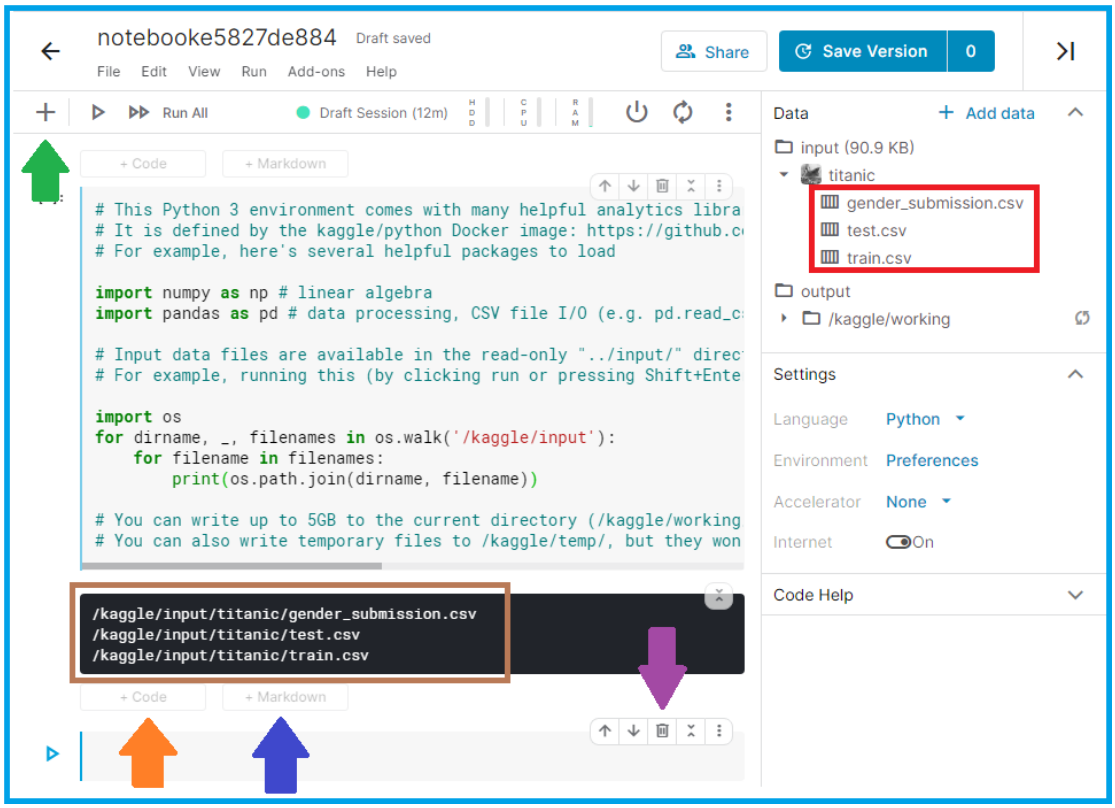
בחץ **סגול** ניתן לראות את הדאטה בתיקיית input. זה הדאטה של התחרות שבו תשתמשו כדי לבצע את החיזויים.

בחץ **כחול** ניתן לראות את שם המחברת, מומלץ לשנות אותו לשם משמעותי.

בחץ **כתום** ניתן לראות את התא הראשון (תמיד התא הראשון בכל מחברת חדשה בקאגל יהיה זהה). אם תריצו תא זה (Shift + Enter) תקבלו הדפסה של תוכן תיקיית input. התא הזה לא כל כך חשוב ואפשר למחוק אותו אם אין בו שימוש.

בחץ **ירוק** ניתן לראות את התא השני (התא הריק הראשון במחברת), בו ניתן לכתוב קוד.

לאחר הרצת (Shift + Enter) התא הראשון נקבל:



ניתן לראות שההדפסה (ריבוע **חום**) זהה לקבצים שנמצאים בתיקיית input (ריבוע **אדום**).

אם רוצים להוסיף תא קוד ניתן ללחוץ על הפלוס שמסומן בחץ **ירוק** או על הכפתור "+ Code" שמסומן בחץ **כתום** (יש כפתור כזה מתחת לכל תא).

אם רוצים להוסיף תא Markdown ניתן ללחוץ על כפתור "+ Markdown" שמסומן בחץ **כחול** (יש כפתור כזה מתחת לכל תא).

אם רוצים למחוק תא, ניתן ללחוץ על סמל הפח שמסומן בחץ **סגול**.