יש את הפרזנטציה ויש את הדוח כמו שהגשנו בתרגיל 1 ו2. כותבים אותו בR.

פרזנטציה

* צריך להציג את הדטה, להסביר מהו.
* להסביר מדוע בחרנו את הדטה.
* להוסיף מידע על הדטה עצמו כלומר לאיזה מינים אתה מתכנן לסוג.
* האם ניקית את הדאטה?
* מה אתה מקווה למצוא מהדטה בסוף הפוריט? להראות שוני בין המינים או השלבים השונים.
* צריך להשתמש באלגוריתם אחד על הדטה. כל האלגוריתמים מוכנים בקוד של רומי, הכל מוכן. לא צריך לשנות שורת קוד אחת אלא רק את ההסברים.
* לדבר על השלבים הבאים שאתה מתכנן, נצטרך להגיד איזה אלגוריתמים אנחנו רוצים לעשות. בפרזנטציה מציגים אחד אבל בפרויקט יש כמה.

R שפה מאוד קלה אבל התוכנה בלתי נסבלת. יש לנו עוד קוד על הדטה עצמו שהוא מוכן ואפשר להשתמש בו.

* לשים לש לזמן של 8 דקות
* להתאמן על שאלות
* לקרוא את סיכומי ההרצאות
* המצגת חייבת להיות באנגלית
* לא הרבה טקסט במצגת, כמה גרפים שאנחנו רוצים.
* להתמקד ב: למה בחרנו, מידע על הדטה, תוצאות ראשונית.
* לא להראות שום קוד במצגת , בדוח כן.
* צריך לראות ניתוח נתונים.

פרויקט  
עבור הדוח הסופי:

* הפחתת מימד
* 4 אלגוריתמים שנלמדו בקורס , עבור כל אלגוריתם להסביר למה נבחר. להתמקד בלמקסם את הביצועים של האלגוריתמים ולראות מי עובד על הנתונים.
* צריך לכתות EDA עבור המצגת
* הקדמה , סיכום, ממצאים ומסקנות
* תוצאות ופלוטים
* בסוף הדוח צריך להחליט מה הגישה הכי טובה
* צריך להגיש בPDF במודל
* לא יותר מ15 עמודים
* אם יש משהו חשוב לשים את זה בנספח
* צריך לשנות את המידות של הפלוטים
* את ההתפלגות של לפני שיפור אפשר לשים בנספח
* להגיש עד ה15.8

כדי להפחית מימד : PCA, UMAP, tSNE   
4 אלגוריתמים לבחור מאלו :KNN, עץ החלטה, SVM, Random forest, Kmeans

DATA

דטה מקורי: 4 קבצים של אדם, 2 קבצים בשלב מוקדם ו2 קבצים בשלב אחר(מוקדם ומאוחר). לכל אחד מהשלבים יש נקבה וזכר. הקבצים הם קבצי BED. הפיקים מייצגים אזורים פתוחים בדנא. יש גם דטה של עבר, יש 2 שלבים שכאשר לכל שלב יש נקבה וזכר. ך שאפשר להשוות לפי השלבים , לפי המין, לפי היצור. זה הדטה הגולמי.

דטה לאחר ניקוי:   
ראשית הורדנו חפיפות. היה פיקים שמוכלים בתוך פיקים אחרים ולכן אין הגיון לספור אותם פעמיים. בנוסף, ישנם פיקים שהם חופפים אזי איחדנו אותם. בנוסף, עשינו את כל הקבצים באותו פורמט כאשר יש לו ID, כרומוזום, התחלה, סיום, אורך.

גרפים:

* דיאגרמות וואן- כמה חפיפה יש בין גנומים שונים.
* דיאגמות פאי- כמה אחוזים הדטה מכיל מתוך כל הגנום.
* היסטוגרמות- של אורך פיק למשל. בבני אדם יש פחות פיקים אבל יותר ארוכים.

אם אנחנו רוצים לעשות אלגוריתם מסוים צריך לבחור אלגוריתם ופשוט להלביש עליו את הCSV.

משימות :

* לפתוח גיט- דניאל
* להעלות את הקוד של תרגיל 1 לגיט - שחר
* לשלוח קוד תרגיל 2 + קוד פרויקט גמר, את הקבצים- אור
* להפוך לCSV וניקיון ,לאחד קבצים בR ולהעלות לגיט- אור
* רקע והסבר על הדטה, למה בחרנו- אור
* EDA – שחר
* עיבוד מקדים+פלוטים(לפני ואחרי)-אור
* הפחתת מימדים+פלוטים- דניאל
* אלגוריתם אחד של למידת מכונה+פלוטים – דניאל
* עוד 3 אלגוריתמים של למידת מכונה- שחר
* הסברים לכל הפלוטים ומסקנות בסה"כ- אור
* לשלוח לדניאל גרפים ודברים למצגת
* מצגת לפרנטציה-דניאל
* לדבר על התוצאות
* להתאמן על המצגת
* לעצב דוח ,לדחוס אותו ל15 עמודים, להוריד אזהרות ולהטעין גרפים- אור

דד ליין:

לסיים את כל המשימות למצגת ולשלוח לדניאל- 25-26.06

מצגת ואלגוריתם 1 של הלמידת מכונה – 27.06

פגישת תוצאות ולהתאמן – 28.06

דוח סופי- 15.08