# תרגיל 3

## שאלה 1

### סעיף א׳

בחרתי את המאפיינים הבאים:

האם המילים הבאות מופיעות במשפט: הסתייגות, כבוד, רבותיי, תודה, חבר, כנסת, בבקשה, :.

מספר המופעים של: בעד, נגד, **ספרות**, פסיקים.

האם המשפט מסתיים ב: ?, ..

מספר הטוקנים במשפט.

### סעיף ה׳

תוצאות המסווג:

- Naive Bayes: 61.10%

- DecisionTree: 63.11%

- KNN: 58.16%

התוצאות שקיבלתי מפתיעות לטובה שכן מדובר רק ב-14 פיצ׳רים. במהלך הבדיקה והפיתוח שמתי לב כי פיצ׳רים המתייחסים למבנה המשפט לא עזרו שכן השונות במשפטים של ריבלין היא גדולה. הדבר נובע מכך שהוא החליף תפקידים – יו״ר, ח״כ, נשיא כאשר רוב הטקסטים של אדלשטיין הם מהתקופה בו הוא יו״ר - וכפי שהוצג בדו״ח הקודם, המשפטים של היו״ר קצרים ומכילים מילים ״פונקציונליות״. לאחר מכן שמתי לב כי ריבלין מרבה להשתמש במילים כגון ״רבותיי״ ו- ״בבקשה״ וכשהוספתי אותן הדיוק השתפר משמעותית.

## שאלה 2

### סעיף ב׳

- Naive Bayes: 52.04%

- DecisionTree: 59.23%

- KNN: 51.43%

ניתן לראות כי בחירה כללית מדי לסיווג איננה בהכרח טובה כאשר הקרופוס הוא בנושא מסויים.

# שאלה 3

### סעיף א׳

בטקסט יש 18950 מילים שונות עבור דגימה אקראית של קלטים בגודל המירבי עבור הקורפוס הכולל לשני הדוברים, לאחר ההגבלה ל- 10,000 משפטים לדובר, בטקסט יש 20497 מילים שונות (דגמתי אקראית משפטים ולכן המספר עלול להשתנות).

### סעיף ב׳

- Naive Bayes: 72.75%

- DecisionTree: 70.27%

- KNN: 64.77%

התוצאות טובות, כמצופה, שכן בחרנו את המילים מהקורפוס עצמו. אמנם השתמשנו ב-10Fold אך עדיין יש סכנה של Overfitting שכן לכל אחד מהדוברים יש סגנון דיבור ייחודי באופן יחסי. ניתן להסביר באופן יחסי את חוסר ההצלחה של KNN ביחס למסווגים האחרים הוא שאולי יש הרבה מילים משותפות ולכן המרחקים לא יהיו אינדיקציה טובה מספיק עבור סיווג הדובר.

## שאלה 4

### סעיף א׳

המילים שנבחרו הן:

2016, אדם, אומר, אותם, אני, אשר, את, אתה, בבקשה, בעד, דבר, האם, הוא, היה, הכנסת, התקבלה, ובכן, זה, חברי, חברת, יכול, כאשר, כבוד, כל, כמובן, כנוסח, לא, לבוא, להיות, לחבר, לחברת, לך, לשנת, מוכרח, מן, מס, מתכבד, נמנע, נמנעים, על, פה, צריך, רבה, רבותי, רוצה, שבו, שהוא, שהיא, שהצעת, תודה

לאחר בחירת המילים, נראה שחלק מהמילים שבחרתי בעצמי הופיעו, כגון, רבותי, בבקשה, נמנע וכו׳ כמו כן, המילים מדגימות כי הדוברים אמרו מילים פורמליות, כמצופה מיו״ר, ואולי בגלל זה האלגוריתמים לא עשו עבודה טובה יחסית בסיווג משפטים לדובר.

## סעיף ב׳

תוצאות הסיווג:

- Naive Bayes: 64.63%

- DecisionTree: 63.36%

- KNN: 61.77%

### סעיף ג׳

ההבדלים הינם משמעותיים לטובת סעיף 3. שימוש בכל הקורפוס הניב תוצאות טובות יותר מאשר 50 המילים הנפוצות ביותר, ההבדל נובע כנראה מ overfitting או מכך שהדוברים נשאו בתפקידים דומים ויש הרבה מן המשותף בסגנון הדיבור שנגזר מאופי תפקיד היו״ר.