**מבוא למערכות לומדות- תרגיל בית 1- דו״ח עבודה**

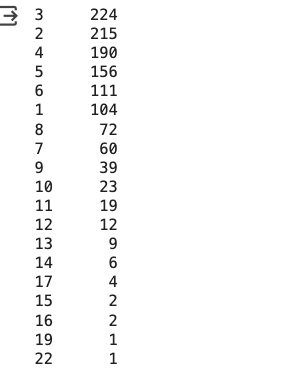
**מגישים:   
  
גל קסטן 316353176**

**חן פרי – 313283657**

(Q1) במסד הנתונים יש 1250 שורות ו26 עמודות.

(Q2)

Value counts of conversations\_per\_day:

****פלט:

בעולם האמיתי, המשתנה ״ מספר שיחות ליום״ מתייחס כנראה לממוצע מספר השיחות ביום שאדם פנים מול פנים. משתנה זה משמש כמעיין מדד לפעילותו החברתית של האדם ולחשיפה שלו למחלות מדבקות.

סוג המשתנה הוא אורדינאלי כיוון שמצד אחד המשתנה קטגורי (הוא דיסקרטי, ויש מספר סופי של ערכים שהוא יכול לקבל) אבל מצד שני קיים סדר טבעי בין הערכים שהמשתנה יכול לקבל (מדובר במספרים טבעיים עם יחס סדר) ויש משמעות כמותית לערכים גדולים/קטנים של כמות שיחות ביום.

(Q3)

|  |  |  |
| --- | --- | --- |
| Type | Description | Feature name |
| Other | Incremental id of the patient | **patient\_id** |
| Ordinal/ Continuous? | Age of the paitent | **age** |
| Categorical | Sex of the patient (Male/Female) | **sex** |
| Continuous | Weight of the patient in kilograms | **weight** |
| Categorical | The blood type of the patient(O/A/AB/B and +/-) | **blood\_type** |
| Other | Current location of the patient (latitude and longitude). | **current\_location** |
| Ordinal | The number of siblings the patient has | **num\_of\_siblings** |
| Ordinal | A Score describing the patient’s level of happiness | **happiness\_score** |
| Categorical | A categorical representation of the patient’s household income | **household\_income** |
| Ordinal | Average number of conversations the patient has everyday | **conversations\_per\_day** |
| Ordinal/ Continuous? | The sugar level measurements for the patient | **sugar\_levels** |
| Ordinal | The patient’s level of sport activity on scale from 0-5 | **sport\_activity** |
| Other | Textual description of any symptoms the patient may have reported | **symptoms** |
| Other | The date in which a PCR test was conducted | **pcr\_date** |
| Continuous | Measurement #1 from a PCR test | **PCR\_01** |
| Continuous | Measurement #2 from a PCR test | **PCR\_02** |
| Continuous | Measurement #3 from a PCR test | **PCR\_03** |
| Continuous | Measurement #4 from a PCR test | **PCR\_04** |
| Continuous | Measurement #5 from a PCR test | **PCR\_05** |
| Continuous | Measurement #6 from a PCR test | **PCR\_06** |
| Continuous | Measurement #7 from a PCR test | **PCR\_07** |
| Continuous | Measurement #8 from a PCR test | **PCR\_08** |
| Continuous | Measurement #9 from a PCR test | **PCR\_09** |
| Continuous | Measurement #10 from a PCR test | **PCR\_10** |

**(Q4)**

WE need to write this in Hebrew   
  
**(from the comments on reference)**   
It is important that we use the exact same split for our analysis so that we can reproduce the experiment and obtain the same models. Moreover, when we compare machine learning algorithms we would like them to be evaluated on the same subsets of the dataset.

**(FROM CHAT)**

Using the exact same split for all analyses is crucial because:

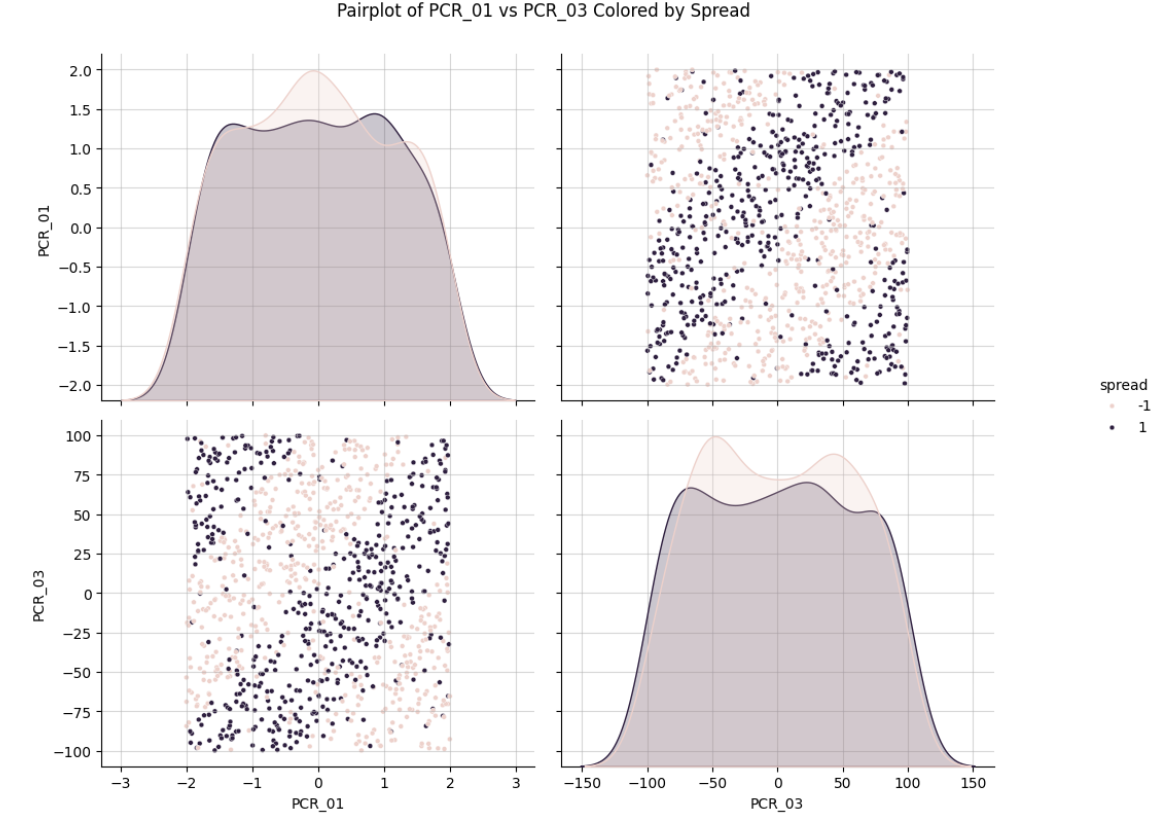
Reproducibility: It ensures that the results of your analysis can be reproduced by anyone else or in future experiments, which is a fundamental aspect of scientific research.

Fair Comparison: When comparing different models or configurations, using the same data split ensures that all models are evaluated under the same conditions, making the comparisons fair.

Avoid Data Leakage: Keeping a consistent split between training and testing data helps prevent data leakage. Data leakage occurs when information from the test set is inadvertently used to train the model, leading to overly optimistic performance estimates. By maintaining a strict separation and not using the test set for decision-making during model selection or tuning, you ensure that the model's performance evaluation is realistic and reflects its ability to generalize to unseen data.

Validation Consistency: When you eventually introduce a validation set for tuning hyperparameters or making decisions about the model, having a consistent training/test split ensures that the validation process is stable and reliable, further avoiding data leakage and ensuring that performance metrics are meaningful.

**(Q5)**



על בסיס התמונה, הפיצ׳רים pcr\_01, pcr\_02 עשויים להיות שימושיים לחיזוי spread ביחד מאחר וניתן לראות כי הדאטה כמעט פריד ל4 אזורים שונים (יש outliers). עם זאת, הדאטה לא פריד לינארית ולכן נצטרך מודל שיודע ללכוד קשרים מורכבים יותר.

כל פיצ׳ר בעצמו לא מספיק כדי לחזות את הדאטה- ההתפלגויות השוליות של שני המשתנים מראות שיש חפיפה בערכים של הקטגוריות של spread ואין כמעט אזורי הפרדה ולכן למשל לא למצוא ערך מפריד שיתן הפרדה לינארית בין תחום ערכי pcr (1 או 3) לבין סוג הspread.

**(Q6)**

**correlation between spread and PCR\_01**: 0.005909176667077393

**correlation between spread and PCR\_03:** -0.004310898790492347

מן הממצאים עולה כי הקורלציה בין pcr\_01 לpcr\_03 היא כמעט 0 כלומר הקשר הלינארי בין המשתנים חלש. ממצאים אלו תומכים במה שמצאנו קודם לכן, ראינו קודם כי כאשר מסתכלים על ההתפלגות השולית של כל אחד מן המשתנים לבד בהפרדה לפי קטגוריות spread, הייתה חפיפה בין הערכים, כלומר לא היתה הפרדה ברורה לעין בין טווח ערכים שpcr\_01 מקבל עבור spread=1 וטווח ערכים עבור spread = -1. באופן דומה גם עבור pcr\_03.

\*\*בבדיקת הקורלציה בודקים רק קשר בין כל משתנה pcr לspread אך לא את האפקט המשולב שלהם.

"Based on the pairplot, the features **PCR\_01** and **PCR\_03** might jointly offer predictive value for 'spread' since the data points are visually grouped into four distinct regions, albeit with some outliers present. This suggests that there may be a non-linear relationship where certain combinations of **PCR\_01** and **PCR\_03** values are associated with specific 'spread' categories. However, since the data does not exhibit linear separability, a model capable of capturing complex relationships, such as a kernel SVM or a neural network, would be more appropriate for this task.

Individually, each feature does not demonstrate a significant ability to predict 'spread'—the marginal distributions for both **PCR\_01** and **PCR\_03** show considerable overlap across the 'spread' categories, indicating no clear threshold or value range that separates 'spread' = 1 from 'spread' = -1.

The calculated correlation coefficients between 'spread' and each PCR feature are very close to zero (**PCR\_01**: 0.005909176667077393, **PCR\_03**: -0.004310898790492347), reinforcing the earlier observation that there is no strong linear relationship between these features and 'spread'. The lack of correlation supports the hypothesis that the relationship between the PCR features and 'spread' is non-linear, as linear correlation would not capture complex patterns such as clusters or non-linear arrangements seen in the pairplot.

Furthermore, the correlation between **PCR\_01** and **PCR\_03** being close to zero suggests that these features are largely independent of each other, which is beneficial for modeling as they may provide unique information to the predictive model.

In conclusion, while the individual linear correlations of **PCR\_01** and **PCR\_03** with 'spread' are negligible, their combined pattern as observed visually suggests the potential for predictive modeling when using non-linear methods. This analysis underscores the importance of considering multivariate effects and non-linear relationships in predictive modeling, and it points towards the need for sophisticated machine learning techniques that can leverage the complex structure in the data."