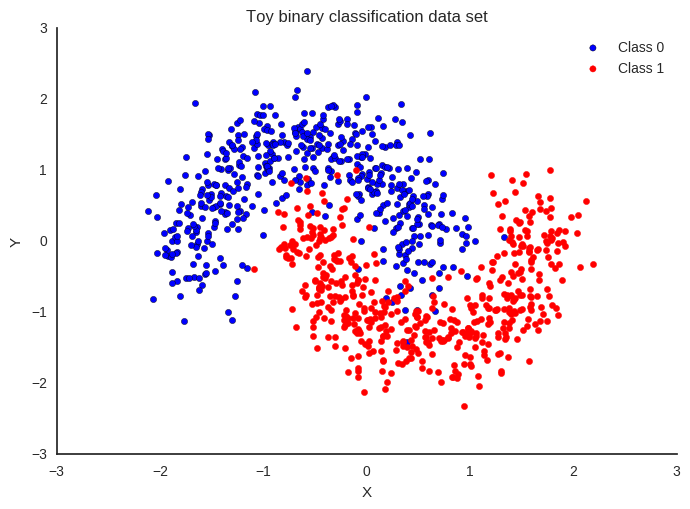


כריית ידע - פרויקט מסכם



מרצה: ד"ר אייל קולמן

עוזר הוראה: מר איתי מרגולין

מגישים:

|  |  |  |
| --- | --- | --- |
|  | שם | ת.ז. |
|  | נעם תור | 203302021 |
|  | תומר ינאי | 305187296 |
|  | יוגב מטלון | 201390408 |

תוכן עניינים

[תקציר 3](#_Toc485725263)

[מבוא 3](#_Toc485725264)

[מודל הסימולציה 4](#_Toc485725265)

[הנחות המודל 5](#_Toc485725266)

[תיאור המודל 5](#_Toc485725267)

[ניתוח המצב הקיים 7](#_Toc485725268)

[מדדים לבחינת המצב הקיים 7](#_Toc485725269)

[איכות השירות בפארק: 7](#_Toc485725270)

[מדדים תפעוליים: 7](#_Toc485725271)

[מדדי משרד הבריאות: 7](#_Toc485725272)

[סוג המערכת 7](#_Toc485725273)

[בחינת המדדים במודל הנוכחי 8](#_Toc485725274)

[הצעות לחלופה 10](#_Toc485725275)

[מבחן סטטיסטי השוואתי בין תוצאות מדדים 10](#_Toc485725276)

[מסקנות ודיון 11](#_Toc485725277)

[נספח א' – נתונים התחלתיים 12](#_Toc485725278)

[נספח ב' – נתונים אשר מקורם מחלק א' של הפרויקט 13](#_Toc485725279)

[נספח ג'- תיאור מפורט של חלקי המודל 14](#_Toc485725280)

[נספח ד' – חישובים נוספים 21](#_Toc485725281)

[חישוב מספר החזרות לסימולציה 21](#_Toc485725282)

[חישוב מדד סטטיסטי בין חלופות 21](#_Toc485725283)

# תקציר

TBD

# שלבי העבודה

שלב מספר 1 – Data Exploration

1. **בחירת נפוצים**– מבין תשעת הקלאסים הקיימים בדאטה בחרנו לסווג את המוצרים לשתי הקטגוריות הנפוצות ביותר: קטגוריה 2 ו-6. בעקבות בחירת קטגוריות אלו נותרנו עם 22,719 רשומות (מתוך 46,408).
2. **בדיקה האם התגיות מאוזנות** – האיזון חשוב מאוד כדי לוודא שהמודלים לא מסיקים מסקנות שגויות על סמך תגיות שהם לא מאוזנות (Data Label Distribution במחברת). ניתן לראות כי הלייבלים אכן מאוזנים (53% לקטגוריה 2 והשאר של 6).

**"מבט על" עבור הפיצ'רים** – ביצענו בדיקה עבור סוג (Type) כל פי'צר ובדיקה האם ישנם ערכים שאינם מספריים (או ערכי Null כמובן), תיאור סטטיסטי של כל אחד מהפיצ'רים (ממוצע, סטיית תקן, מינימום, מקסימום ורבעונים), ותרשים Box Plot לכל פי'צר. מניתוח ה-Box Plot ניתן לראות מעיכה דרמטית של הנתונים סביב ה-0 עבור כל הפיצ'רים.

1. **בדיקת התנהגות (התפלגות) הנתונים** – ביצענו בדיקות של התפלגויות הנתונים השונים (עבור כל feature) באמצעות גרפים העוזרים להבין את התנהגות הנתון בהשוואה לכל קטגוריה. בדיקות אלו מראות כי בכל אחד מהפיצ'רים הערך 0 הוא הנפוץ ביותר ובנוסף יכולות להעיד על קורלציה כפי שנפרט בהמשך. בהמשך לבדיקת ההתפלגויות, בדקנו עבור כל פי'צר את מידת ה"ספארסיות" שלו – מה אחוז המדויק של הערכים שאינם אפס בכל פיצ'ר.
2. **קורלציה** – בחלק זה בדקנו האם ישנה קורלציה בין הפיצ'רים השונים. קורלציה גבוהה תעיד כי עלינו לשקול הוצאת פיצ'רים, שכן אינן מסבירות מידע חדש. לאחר הבדיקה ראינו כי הפיצ'רים אינם קורלטיביים זה לזה.

שלב מספר 2 – Pre-Processing

1. **הסרת חריגים** – ראשית רצינו לענות על השאלה האם ישנם נתונים חריגים בדאטה אשר דורשים הסרה. בשלב הקודם ראינו כי כל העמודות הינן מסוג int, וכי אין ערכים שאינם מספרים או Null. מתרשים ה-Box Plot שעשינו בחלק הקודם, ראינו כי עבור כל הפי'צרים כמעט כל הערכים שהיו שונים מ-0 הוגדרו כ"חריגים". מן הסתם, אל לנו להתייחס לאלה כחריגים, משום שדווקא אלו עשויים להכיל את המידע הקריטי להכרעה בין 2 הקלאסים. מסיבה זו, לא נגדיר אותם כחריגים, וניעזר בהם בשלבים הבאים – ביצירת פיצ'רים חדשים, וביצוע נרמול. לאור האמור לעיל, לא נבצע הסרה מתוך הנתונים.
2. **נרמול הנתונים**

במקרה זה השתמשנו בנרמול מסוג MinMax עבור כל העמודות, לשם שימוש נכון במודלים בהמשך. הנרמול ממיר את הנתונים לאותה הסקלה – [0,1].

1. **השלמת מידע חסר**

כפי שניתן לראות בסיכום הנתונים, לא קיימים נתונים חסרים בדאטה. וכפי שציינו נתייחס ל-0 כנתון בעל משמעות ולא כמידע חסר.

1. **המרת טיפוסי נתונים**
2. עמודות is\_news – הערכים הומרו מטיפוס הנתונים String לטיפוס הנתונים Int.
3. עמודה News\_front\_Page –הערכים הומרו מטיפוס הנתונים String לטיפוס הנתונים Int.
4. עמודה alchemy\_categories – המרת טיפוס הנתונים String לטיפוס הנתונים Int (פאקטוריאלי).
5. alchemy\_category\_score – המרת טיפוס הנתונים מ-string ל-float.
6. **הסרת עמודות**

עמודה framebased הכילה כולה ערכי 0, לכן הסרנו אותה כי אין היא מוסיפה מידע חדש.

*שלב מספר 3 – Feature Selection*

*שלב מספר 4 – בניית מודל ראשוני והערכתו*

*שלב מספר 5 – בניית מודלים מתקדמים*

# סיכום