**פרויקט ברגרסיה ליניארית**

מטרת הפרויקט להקנות לסטודנטים כלים בסיסיים להתמודדות עם ניתוח סטטיסטי של מאגרי נתונים טבלאיים. הניתוח הסטטיסטי כולל הן ניתוח לא-פורמלי (תיאורי/גרפי/טבלאי) והן ניתוח פורמאלי בעזרת הכלים של רגרסיה ליניארית שנלמדים בקורס. במסגרת הקורס יבצעו הסטודנטים פרויקט של התנסות מעשית שיתייחס לשני ההיבטים הללו של הניתוח הסטטיסטי.

**הוראות הגשה חלק א':**

הגשת חלק א' תיעשה **בזוגות בלבד ועד לתאריך 18.12.2022 בשעה 23:59**.

ההגשה מתבצעת דרך המודל ותחולק לשתי תיקיות הגשה:

1. **תיבת הגשה חלק א':** בה תגישו **קובץ WORD יחיד** עם דו"ח חלק א' בפרויקט (שם הקובץ יהיה מספר הקבוצה בלבד). מבנה הדו"ח:
   1. העמוד הראשון יכלול את מספר הקבוצה ומספרי תעודת הזהות של חברי הקבוצה
   2. תוכן עניינים
   3. חלק א'
   4. נספחים
2. **תיבת הגשה חלק א' – קוד:** בה תגישו קובץ R יחיד עם הקוד של חלק א' בפרויקט

**מספיקה הגשה של אחד מחברי הקבוצה.** **הגשה שלא על פי ההנחיות תגרור הורדת נקודות.**

**שימו לב, אין להגיש את קובץ הנתונים שלכם, כלומר, אין לבצע אף שינוי בקובץ האקסל עצמו. במידה ויש צורך בשינויים בבסיס הנתונים, הם צריכים להתבצע בקוד בR בלבד.**

**יש ציפייה שכל חברי הצוות יהיו שותפים לעבודה על הפרויקט, על כל חלקיו. צוות הקורס שומר לעצמו את הזכות לתת ציון דיפרנציאלי לחברי הצוות, במקרים בהם תהיה שונות משמעותית בין חברי הצוות במידת ההשקעה והתרומה.**

**מספר דגשים:**

* הפרויקט בקורס כולל שני חלקים. חלק א' של הפרויקט סוכם ל100 נקודות והוא מהווה 30% מהציון הסופי של הפרויקט (בחלק ב', אתם נדרשים להישאר באותן הקבוצות).
* אורך הדו"ח של חלק א' לא יעלה על 15 עמודים (להוציא תוכן עניינים ונספחים). **כל עמוד נוסף מעבר למכסה שנקבעה יוריד 2 נקודות מציון הפרויקט**!
* **כל ניתוח המתבצע בתוכנה חייב להיות מגובה בצילום מסך של הפלט, בגוף הדוח או בנספחים.**
* המשמעות בלפרש את הנתונים/פלטים היא להסביר את התוצאות בהקשר של הנתונים ולא לכתוב במילים את מה שרואים בפלט/גרף, אלא לחזור אל הנתונים ולנסות להסביר מה בנתונים גרם לתוצאות הללו.
* יושם דגש בבדיקה על יצירתיות ומקוריות החשיבה בהקשר של ניתוח הנתונים שימוש בגרפים יצירתיים וכו'.

**חלק א' – בחירת בסיס נתונים וניתוח סטטיסטי של הנתונים (30%)**

1. בחירת מאגר נתונים:

תחת לשונית פרויקט הועלתה תיקיה בשם "Data sets" ובה קבצי נתונים עבור הפרויקט. לאחר סגירת הרישום לקבוצות, כל קבוצה תקבל במייל את שם ה-Data set שעליו יבצעו את הפרויקט. הבחירה תעשה באופן אקראי ע"י צוות הקורס. בכל קובץ excel, מתואר בקצרה את המשמעות של כל אחד מהעמודות בטבלה. **בנוסף, עבור כל Data set, מצוין מי הוא המשתנה המוסבר בהערה בסוגריים**.

* על כל אחד מה Data setsהמוצעים **יעבדו לכל היותר 9 זוגות**.
* חל איסור מוחלט על שיתוף פעולה בין צוותים העובדים על אותו נושא. **חשד מבוסס לשיתוף פעולה יוביל לתלונה חמורה לוועדת משמעת.**
* **בהמשך, בחלק ב' של הפרויקט, יינתנו נקודות בונוס לצוותים שיראו תוצאות טובות ביחס לקבוצות האחרות עם אותו הData set.**

1. יצירת טבלת משתנים:

יש לצרף בסעיף זה טבלה עם הסבר על כל אחד מהמשתנים: (10 נק').

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| סוג המשתנה -מוסבר/מסביר | סימון במודל | יחידת מידה | סוג המשתנה – רציף / קטגוריאלי | הסבר קצר על המשתנה |

1. תיאור המשתנים:

נמקו מדוע כל אחד מהמשתנים המסבירים עשוי לתרום ולהוסיף מידע לגבי מגמות השינוי במשתנה המוסבר (10 נק').

1. תיאור קשרים בין משתנים:

העזרו בפונקציית plotעל המשתנים המסבירים הרציפים ועל המשתנה המוסבר שלכם כפי שלמדתם בתרגול המעבדה. נתחו את הקשרים העיקריים בין המשתנים המסבירים השונים ובנוסף בין המשתנים המסבירים למשתנה המוסבר. הסבירו במפורש את המנגנון שמקשר סיבה ותוצאה עבור הקשרים הללו ופרטו האם מדובר בקשר מדגמי/ סיבתי. האם זיהיתם קשרים (קורלציה) בין מסבירים כלשהם? האם לדעתכם יש סיבה לקשרים אלו? או להפך, האם אתם מניחים קשר סיבתי אך לא רואים את הקורלציה? מדוע? (10 נק').

1. ניתוח תיאורי של המשתנים:

הציגו ניתוח תיאורי של כל אחד מהמשתנים הרציפים במאגר: ממוצע, חציון, סטיית תקן, תחום בין-רבעוני, אסימטריה (skewness).

משתנים קטגוריאליים ללא חשיבות לסדר - יש להציג את החלוקה הפנימית בין הקטגוריות של המשתנה הקטגוריאלי, השכיחויות לפי כל קטגוריה.

פרשו את תוצאות הניתוח (10 נק').

1. ניתוח חריגים:

בדקו האם ישנן רשומות או נתונים חריגים. נמקו מדוע לדעתכם נתונים אלו יצאו חריגים וכיצד לדעתכם צריך לטפל בכל אחד מהמקרים החריגים. במידה ובחרתם להוריד חריגים, בצעו זאת. (10 נק').

1. פונקציית צפיפות והתפלגות מצטברת:

הציגו את פונקציית צפיפות ההסתברות (היסטוגרמה) ואת פונקציית ההתפלגות המצטברת עבור לפחות שלושה משתנים מסבירים. פרשו את התוצאות, תוך התייחסות לשיאים, זנבות, וכיו"ב (15 נק').

1. ייצוג קשרים בעזרת תרשימים:

חפשו והציגו קשרים "מעניינים" בין משתנים - צפויים ולא-צפויים. הסבירו את המשמעות של קשרים אלה. הציגו לפחות חמישה תרשימים שונים בין זוגות שונים של משתנים לבחירתכם (לפחות שניים מהתרשימים יכללו משתנה קטגוריאלי). נדרש להיעזר בתרשימי פיזור (scatterplot), תרשים פסיפס (mosaic) וכו'. פרשו את התוצאות, תוך התייחסות לתלות בין המשתנים, למגמות, להשפעות הדדיות וכו'. יושם דגש בבדיקה של סעיף זה בהקשר של ניתוח הנתונים ושימוש בגרפים יצירתיים (15 נק').

1. טבלאות שכיחות:

הציגו לפחות שתי טבלאות שכיחות חד-ממדיות ושתי טבלאות שכיחות דו-ממדיות. במידת הצורך כדאי לשקול איחוד ערכים אפשריים של המשתנה לקבוצות. הסבירו את משמעות הממצאים (10 נק').

* רצוי להוסיף ניתוחים תיאוריים/טבלאיים/גרפיים נוספים שיסייעו להבנת מאפייני מאגר הנתונים שבחרתם. 10 הנקודות הנותרות (על מנת להשלים לציון 100) מתבססות על איכות הדו"ח של חלק א'.

שאלות בנושא הפרויקט יש להעלות לפורום הרלוונטי.

**בהצלחה**!