**פרויקט ברגרסיה ליניארית**

**הוראות הגשה חלק ב':**

הגשת חלק ב' תיעשה **בזוגות בלבד (אותן זוגות כמו בחלק א') ועד ליום לפני המבחן בשעה 23:59**.

ההגשה מתבצעת דרך המודל ותחולק לשתי תיקיות הגשה:

1. **תיבת הגשה חלק ב':** בה תגישו **קובץ WORD יחיד** עם דו"ח חלק ב' 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 ב'
   4. נספחים
2. **תיבת הגשה חלק ב' – קוד:** בה תגישו קובץ R יחיד עם הקוד של חלק ב' בפרויקט

**מספיקה הגשה של אחד מחברי הקבוצה.** **אין צורך להגיש שנית את חלק א' של הפרויקט.** **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 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:**

* חלק ב' של הפרויקט סוכם ל100 נקודות והוא מהווה 70% מהציון הסופי של הפרויקט.
* אורך הדו"ח חלק ב' לא יעלה על 15 עמודים (להוציא שער, תוכן עניינים כללי, תוכן עניינים עם רשימות טבלאות והגרפים 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.**
* **כל ערך מספרי בו אתם משתמשים חייב להיות מגובה ע"י פלט מהתוכנה המראה שזהו אכן הערך שתואם את הנתונים שלכם.**
* במהלך חלק ב' של הפרויקט יתכן ותבצעו ניסיונות רבים בקוד, משחקים עם טרנספורמציות ונתונים. מומלץ מאוד ואף הכרחי לציין בדוח אך רק את השלבים החשובים והתורמים לשיפור המודל ולא לתאר את כל תהליך המחשבה במלואו. הדוח נחשב איכותי בין היתר כשישנו רצף מובן והגיוני של המתואר בו.
* יש לשים לב לאופן הגשת הדוח - אבחנה בין טבלאות ופלטים שיינתנו בגוף הדוח או בנספחים; כותרות מסודרות לגרפים ולטבלאות (כותרת לטבלה מעליה; כותרת לגרף מתחתיו); תוכן עניינים ורשימת טבלאות, גרפים ונספחים (בתחילת הדוח); התייחסות מפורשת לכל טבלה או גרף בגוף הדוח; הסבר מפורט של פלטים שהתקבלו; הקפדה על ראשי פרקים.
* 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.
* מטרת הפרויקט היא ביצוע תהליך ניתוח נתונים סטטיסטי מתחילתו ועד סופו. תהליך ניתוח נתונים מורכב משני רבדים - הניתוחים הסטטיסטיים הפורמליים, והקישור של הניתוח לנתונים - על כן נדרשת בקיאות בנושא הנבחר, ולפעמים צריך לבדוק קצת מעבר מה גורם למספרים לצאת כפי שהם יוצאים.
* יושם דגש בבדיקה על יצירתיות ומקוריות החשיבה בהקשר של ניתוח הנתונים שימוש בגרפים יצירתיים וכו'. **בנוסף, יינתנו עד 5 נקודות בונוס לקבוצות אשר יראו תוצאות טובות יותר ביחס לשאר הקבוצות עם אותו הData set בהתאם לפרמטרים: מדד טיב התאמה מתוקנן, AIC, מקוריות המודל (מבחינת אינטראקציות וטרנספורמציות) וסבירות המודל (פשוט, ונותן תמונה טובה על הקשרים בין המשתנים השונים).**

**חלק ב' – ניתוח פורמאלי של מאגר הנתונים (70%)**

1. תקציר מנהלים:

יש לכתוב תקציר מנהלים המסכם את חלק ב' של הפרויקט (מטרות, צעדים שננקטו, מסקנות ותוצאות) (10 נק').

יש לצרף לסעיף זה את טבלת המשתנים מחלק א' סעיף 2.

1. עיבוד מקדים:
   1. הסרה של משתנים:

בשלב זה, עבור משתנים שלדעתכם לא קשורים למשתנה המוסבר, או לא מוסיפים עליו מידע שיקלו האם יש להסירם מהמודל ופרטו מדוע אתם חושבים כך. התדיינו על שני משתנים לפחות (10 נק').

* 1. התאמת משתנים:

בדקו והחליטו האם יש צורך בהגדרה מחדש של משתנים או בהתאמתם לבסיס הנתונים שלכם. עבור משתנה קטגוריאלי, לעיתים יש צורך לבצע איחוד קטגוריות. עבור משתנה רציף לעיתים יש צורך בהפיכתו למשתנה קטגוריאלי (דיסקרטיזציה), או למשל הגדרת משתנה חדש שהוא הפרש או סכום בין משתנים קיימים. חשבו על 3 התאמות לפחות והחליטו אם יש לבצע אותן. (15 נק')

* 1. הגדרת משתנה דמה:

בחרו משתנה קטגוריאלי אחד והגדירו בדו"ח את משתני הדמה שנגזרים ממנו. (5 נק').

* 1. הגדירו והוסיפו משתני אינטראקציה:

בחרו מדגם של משתני אינטראקציה מסוימים (לפחות שלושה) כאשר עליכם להסביר האם ומדוע, לפי השערתכם, משתנים אלו יכולים לתרום לחיזוי המשתנה המוסבר. הגדירו את משתני האינטראקציה כפי שלמדתם, גם בקוד העבודה וגם בדוח. מצופה להראות תהליך של חשיבה והבנה מדוע יש לאינטראקציה סיבה להוות השפעה ומדוע לא.

לבסוף, התאימו את המודל שמכיל את המשתנים המעובדים, כולל את משתני הדמה והאינטראקציה שהגדרתם. הציגו את מדדי הביצוע שלו. (20 נק').

1. התאמת המודל ובדיקת הנחות המודל:
   1. בחירת משתני המודל: (15 נק')

בשלב זה בחרו קריטריון להשוואה בין מודלים (AIC, BIC, R2adj), והריצו שני אלגוריתמים לבחירת משתנים המתבססים על אותו קריטריון. האם שני האלגוריתמים הביאו לאותו המודל? דונו בתוצאות.

\*\* כיוון שסעיף 4 ירד אנו ממליצים לכם לנסות להחזיר משתנים שהורדתם בשלב מוקדם יותר של הפרויקט ולבדוק האם כבר עכשיו החזרתם תשפר את תוצאות ביצועי המודל.

* 1. בדיקת ויזואלית של הנחות המודל: (15 נק')

צרו תרשים שגיאות מנורמלות מול תחזיות כפי שלמדתם בתרגול. בצעו בחינה ויזואלית של השערות הלינאריות ושוויון השונויות והסבירו האם לדעתכם מתקיימות ולמה. צרו תרשים QQplot ובחנו את השערת הנורמאליות.

* 1. בדיקת פורמלית של הנחות המודל: (10 נק')

בצעו מבחני השערות (עליהם דיברנו בתרגול) לבדיקת הנחת הנורמאליות והנחת הלינאריות. דונו בתוצאות.

**סעיף 4 ירד- לא להגשה**

1. שיפור המודל: (20 נק')

בדקו האם יש צורך בהוספה של משתני טרנספורמציה או בהחלפת משתנה קיים בטרנספורמציה של אותו המשתנה. נמקו את בחירתכם.

בשלב זה ניתן לשקול בחינה מחדש של משתנים אשר הושמטו מהמודל, במידה והינכם חושבים כי הם תורמים להסברת המשתנה המוסבר.

יש להגיע עד לבניית מודל רגרסיה ע"י אחד מהאלגוריתמים השונים לבחירת המודל ע"פ השיטות שנלמדו בהרצאות.

**זכרו - המטרה העיקרית שלכם בחלק ב' היא לבנות את המודל הכי טוב שאתם מסוגלים, מודל שעוזר להסביר את המשתנה המוסבר בצורה הטובה ביותר.**

שאלות יש להעלות לפורום הרלוונטי בלבד.

**בהצלחה**!