**נושא הפרויקט:** **מודל רב-ממדי להערכת סיכון אשראי: מסגרת לחיזוי תוצאות, הלוואות ומדדים פיננסיים**

# הצעה לביצוע פרויקט גמר – פרויקט מחקרי

## מבוא

סיכון אשראי הוא אחד הגורמים המרכזיים שמוסדות פיננסיים מתמודדים עימם בעת מתן הלוואות. הבנת דפוסי החזרי הלוואות וזיהוי גורמי הסיכון העיקריים מסייעים בקבלת החלטות פיננסיות מושכלות. התקדמות בתחומי למידת המכונה והאנליטיקה מאפשרת כיום לפתח מודלים רב-משתניים לחיזוי תוצאות הלוואות תוך התייחסות למספר משתנים תלויים בו-זמנית. מחקר זה יתמקד בבניית מודל רב-ממדי להערכת סיכון אשראי, תוך שימוש בנתוני הלוואות ממשיים שמקורם בפלטפורמת Kaggle.

תחום ניהול סיכוני האשראי מהווה חלק בלתי נפרד מהמערכת הפיננסית המודרנית, ונמצא במוקד המחקר הפיננסי והיישומי זה עשורים. מוסדות פיננסיים, כדוגמת בנקים, חברות כרטיסי אשראי ופלטפורמות הלוואות דיגיטליות, פועלים בשוק שבו סיכון האשראי משפיע על רווחיותם ויציבותם. ניהול סיכונים זה החל להתפתח בראשית המאה ה-20 עם השימוש הראשוני במודלים סטטיסטיים כמו ניתוח רגרסיה וציוני אשראי (Credit Scores).

עם השנים, כניסת מערכות מידע מתקדמות ונתוני עתק (Big Data) שיפרו משמעותית את יכולות הניתוח והחיזוי. טכניקות מסורתיות, כמו דירוג FICO, סיפקו בסיס למדידה אחידה של סיכון אשראי. כיום, הטכנולוגיות המודרניות מתמקדות באינטגרציה של למידת מכונה ואנליטיקה מתקדמת ליצירת מודלים מותאמים אישית. התחום רלוונטי במיוחד להנדסת תעשייה וניהול, שכן הוא דורש שילוב בין כלים סטטיסטיים מתקדמים, תהליכי קבלת החלטות ואופטימיזציה.

הנדסת תעשייה וניהול עוסקת בין היתר באופטימיזציה של תהליכים מורכבים וקבלת החלטות מבוססת נתונים. תכנון מודל סיכון אשראי דורש הבנה מעמיקה של מתודולוגיות כמו כריית נתונים, למידת מכונה, ניהול נתוני עתק (Big Data Management) וניהול סיכונים. בנוסף, התחום מקדם פיתוח פתרונות טכנולוגיים התומכים בקבלת החלטות במערכות מורכבות ובעלות אי-ודאות גבוהה.

## הבעיה המחקרית

העולם הפיננסי מתמודד עם אתגרים הולכים וגוברים בניהול סיכוני אשראי, שהם גורם קריטי להצלחת מוסדות פיננסיים. גישות מסורתיות להערכת סיכון אשראי מתמקדות בניתוח חד-ממדי, בו כל משתנה תלוי נחזה בנפרד, מה שמגביל את יכולת המודלים ללכוד קשרים מורכבים בין המשתנים. ניתוח נפרד עלול להוביל לתובנות חלקיות ואף להחמצת קשרים משמעותיים. עם התקדמות טכנולוגיות למידת המכונה והזמינות של נתוני עתק, נוצרה הזדמנות לפתח מודלים רב-משתניים המסוגלים לחזות בו-זמנית מספר משתנים תלויים, תוך התחשבות במורכבות היחסים ביניהם.

שאלות המחקר:

מהם המשתנים הבלתי תלויים המשפיעים באופן משמעותי על סטטוס ההלוואה (שולמה במלואה/חדלות פירעון)?

1. כיצד ניתן לבנות מודל רב-משתני שיחזה בו-זמנית את סטטוס ההלוואה, שיעור הריבית ודירוג האשראי?
2. עד כמה ביצועי המודל הרב-משתני טובים יותר לעומת מודלים חד-משתניים מסורתיים?
3. מהן ההשלכות המעשיות של ממצאי המודל למוסדות פיננסיים בקבלת החלטות?

השערות מחקר:

1. משתנים כמו דירוג FICO, יחס חוב להכנסה (dti) ושיעור ניצול האשראי (revol\_util) משפיעים באופן משמעותי על סיכון האשראי.
2. מודל רב-משתני המבוסס על למידת מכונה יפיק תחזיות מדויקות יותר בהשוואה למודלים מסורתיים.(צריך לעשות השוואה)
3. שימוש בגישה רב-משתנית תאפשר זיהוי קשרים סמויים בין משתנים תלויים, כגון ריבית וסטטוס ההלוואה.
4. ניתוח המודל יחשוף תובנות ייחודיות על דפוסי החזר הלוואות ויתמוך בשיפור מדיניות ניהול האשראי.

## מטרות הפרויקט

1. פיתוח מודל חיזוי רב-משתני: לבחון את האפשרות לחזות בו-זמנית משתנים תלויים הקשורים לסיכון אשראי, כגון סטטוס הלוואה, ריבית ודירוג אשראי.
2. זיהוי גורמי מפתח: לנתח משתנים בלתי תלויים (כגון נתוני לווה, היסטוריית אשראי והתחייבויות פיננסיות) המשפיעים על סיכון אשראי.
3. הערכת ביצועי המודל: להשוות את ביצועי המודל הרב-ממדי למודלים מסורתיים ולהצביע על יתרונותיו.
4. תרומה יישומית: לספק תובנות יישומיות למוסדות פיננסיים לשיפור תהליכי מתן הלוואות.

## מתודולוגית העבודה

1. איסוף נתונים:

* שימוש בנתוני LendingClub מ-Kaggle.
* הנתונים כוללים משתנים כמו סטטוס ההלוואה, דירוג FICO, הכנסה שנתית, יחס חוב להכנסה וריבית.

1. עיבוד נתונים:

* ניקוי נתונים: טיפול בערכים חסרים ושגיאות.
* המרת משתנים קטגוריים לקידוד מתאים.
* נרמול משתנים רציפים כמו הכנסה ורמת חוב.

1. בנייה ואימון המודל:

* אימון מודל למידת מכונה בגישה רב-משתנית (Multi-Task Learning).
* שימוש באלגוריתמים לדוגמא Random Forest, XGBoost ורשתות נוירונים.

1. הערכת המודל:

* מדדי ביצועים עבור כל משתנה תלוי: דיוק (Accuracy), שורש ממוצע ריבועי (RMSE), ו-F1 Score.

1. ניתוח הממצאים:

* זיהוי המשתנים הבלתי תלויים המשפיעים ביותר.
* השוואת ביצועי המודל הרב-משתני למודלים חד - משתניים.

## מקורות

אתרי אינטרנט/מסדי נתונים

<https://www.kaggle.com/datasets/jorgesalasf/credit-risk-dataset>

<https://www.kaggle.com/datasets?search=loan.csv>