כלי עזר- הוראות להקמת סביבת עבודה : <http://sundog-education.com/spark-python/>

(ניתן להקים סביבת עבודה גם עם כלים אחרים)

1. לפניך קובץ המכיל נתוני אשראי של לקוחות (קובץ שנלקח מרשת האינטרנט)
2. כתוב תוכנית בשפת Python תוך שימוש ב- Spark הנותנת מענה לשאלות הבאות:
   1. הצג רשימת הערכים היחודיים של עמודת credit\_history
   2. הצג את כל בקשות האשראי שמעל סכום של 10,000(credit\_amount) לגברים מובטלים מעל גיל 30
   3. הצג את ממוצע סכומי האשראי מקובצים לפי מטרה (purpose) - התוצאות יוצגו ממוינות מגבוה לנמוך
   4. בונוס- הצג כמה לקוחות מועסקים יותר מ- 4 שנים (employment)
3. עליך לרשום מודל חיזוי האם לקוח הוא טוב או רע למתן אשראי – משתנה החיזוי class

קוים מנחים:

* יש להשתמש במודל logistic regression (בונוס – הרצה על מודלים נוספים)
* יש לבחור את ה- features הכי משפיעים שיהיו בשימוש המודל
* יש לאמן את המודל על 80% הנתונים ולבחון אותם על ה- 20% הנותרים

(train data\test data)

בסיום הרצת המודל הצג מדדי איכות של המודל ( + הסבר) וכמות ההצלחות בחיזוי על אוכלוסיית הבדיקה