Comparison of the Support Vector Classifier algorithm with the Decision Tree algorithm for Credit Card Fraud Detection with the Goal of Improving Accuracy

S. Tharun Sainatha Reddy, P. Sriramya

[**קישור ל GitHub**](https://github.com/nirma100/Credit-Card-Fraud-Detection-Project-Exploring-Machine-Learning-Methods)

**מגיש:** ניר מעיין

**קורס:** כריית נתונים

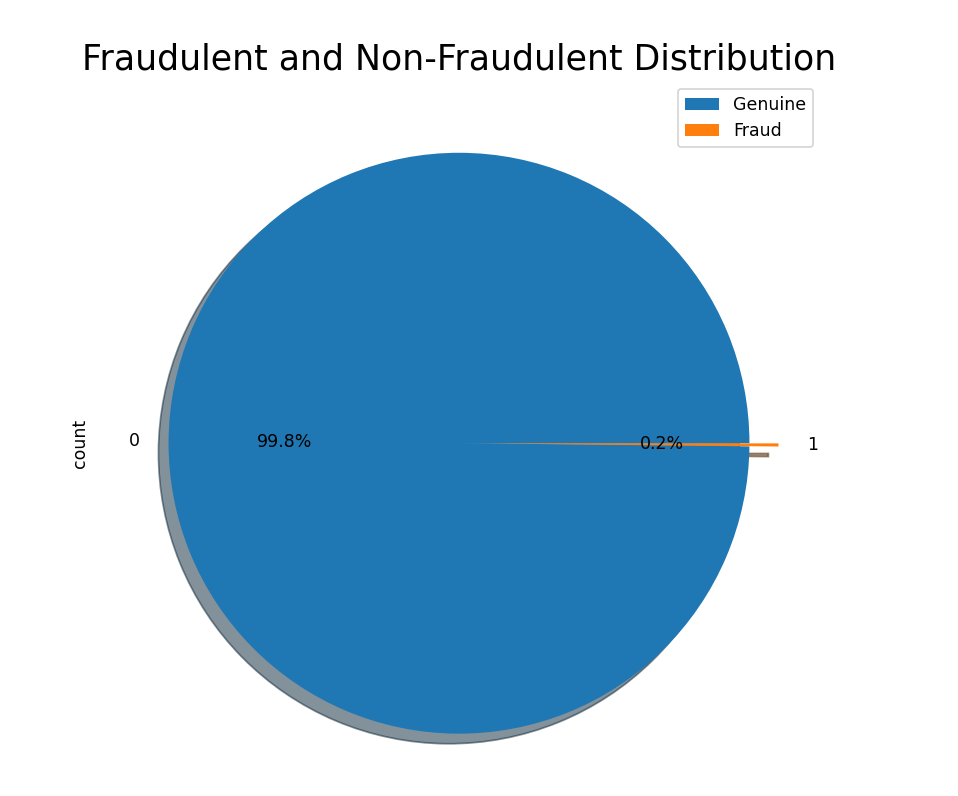
**מרצה:** ד"ר דימה אלברג

**תקציר:**

מחקר זה בחן את יעילותם של שני אלגוריתמים ללמידת מכונה, Decision Tree (DT) ו-Support Vector Classifier (SVC), בזיהוי הונאות בכרטיסי אשראי. הונאות בכרטיסי אשראי הן בעיה עולמית חמורה, והן עלולות לגרום לנזק כלכלי משמעותי.

המחברים השתמשו במאגר נתונים בעל כ- 56 אלף רשומות ו-31 עמודות (המאגר מתעדכן כעת בעל כ- 284 אלף רשומות). מתוך מאגר המידע בוצעו 10 איטרציות גודל המדגם השתנה בין האיטרציות (לכל היותר 1000 תצפיות במדגם) כאשר קבוצה 1 ל DT וקבוצה 2 ל SVC ובכל קבוצה 21 תצפיות של הונאה.

לתהליך לימוד המכונה נעשה שימוש ב-80% מהנתונים להכשרת האלגוריתמים, וב-20% הנותרים לבדיקת הדיוק שלהם.

התוצאות הראו כי אלגוריתם ה-SVC השיג דיוק גבוה יותר מאשר אלגוריתם ה-DT. אלגוריתם ה-SVC השיג דיוק של 98.59%, בעוד ש-DT השיג דיוק של 94.86%. ההבדל בדיוק בין שני האלגוריתמים הוא משמעותי, וכי הוא מצביע על כך שה-SVC הוא כלי יעיל יותר לזיהוי הונאות בכרטיסי אשראי.

המחברים סבורים כי אלגוריתם ה-SVC הוא כלי יעיל לזיהוי הונאות בכרטיסי אשראי. הם מציעים כי מחקרים עתידיים צריכים לחקור שיטות לשיפור דיוקם של שני האלגוריתמים. לדוגמה, ניתן לחקור שיטות לשיפור איכות הנתונים המאומנים, או שיטות לשיפור יכולתם של האלגוריתמים להתמודד עם נתונים לא סטנדרטיים.

גרף 1: יחס בין סיווג 0 ו- 1 במשתנה המטרה

\*יש לציין כי אם נסתכל בגרף 1 נראה כי קיים מיעוט מובהק במשתנה המוסבר ( 492 תצפיות הונאה מתוך כ 284 אלף תצפיות) והבחירה דווקא במחיקת רוב המידע לשם פתירת הבעיה לא בהכרח היה הפתרון האידיאלי ועל כך אבחר להעמיק בסוף הפרויקט.

**מבוא וסקירת ספרותית:**

במאמר זה סקירת הספרות היא כחלק מהמבוא אשר מספק סקירה על מטרות המחקר והסיבות לבחירת המתודולוגיה.

ההקדמה במאמר תחילה מציגה את נושא המחקר שהינו "הונאות כרטיסי אשראי". הכותבים מסבירים כי הבעיה שתלווה את המחקר הינה הונאות בכרטיסי אשראי אשר הינן בעיה חמורה הגורמת לנזקים כלכליים משמעותיים וכי כיום הונאה בכרטיסי אשראי היא הנפוצה והנפוצה ביותר מכל ההונאות (Dal Pozzolo et al. 2018), עוד ציינו כי לפי המחקר של   
Li et al. (2021) אשר חקר את דיוק הרגרסיה הלוגיסטית ומצא כי SVC מציג מודל בעל דיוק גבוה יותר. לאחר מכן, הכותבים מתמקדים במטרת המחקר, שהיא להשוות את יעילותם של שני אלגוריתמים ללמידת מכונה, Decision Tree ו-Support Vector Classifier לפי המלצת המחקר של Lucas Yvan (2019) אשר חקר את הדיוק של מודלי DT ו- SVC בזיהוי הונאות בכרטיסי אשראי.

בתחום הזיהוי ההונאה, המחקר בשילוב למידת מכונה הם נושאים חזקים ונמצאים בפיתוח ובמחקר כבר מספר שנים רב. תחום זה מהווה אתגרים עיקריים לחברות אשראי ולבנקים עקב עליות רמות ההונאה בסביבה הפיננסית ולכן עקב פרסומו בשנת 2023 מאמר זה לא מציג מסקנות חדשניות בנושא. בנוסף במאמר לא קיימת סקירה תיאורטית על כל מודל והבחירה בהן התבצעה על סמך המלצה ממחקר אחר(Lucas 2019) ועל כן אבחר לפרט עליהם בקצרה:

אלגוריתם Decision Tree הוא אלגוריתם למידת מכונה פשוט יחסית, המבוסס על חלוקת הנתונים לקטגוריות על בסיס תכונות מפתח. האלגוריתם בונה עץ דמוי תרשים זרימה, כאשר כל צומת בעץ מייצג החלטה, וכל ענף מייצג תוצאה אפשרית של ההחלטה לכן אם ביצע פעולה בזמן מסוים האם זה מעיד על אמינות פעולת אשראי זו.

אלגוריתם Support Vector Classifier הוא אלגוריתם למידת מכונה מורכב יותר, המבוסס על מציאת קווים או מישורים המפרידים בין נתונים מקבוצות שונות אשר במקרה זה יאפשרו לעשות קלסיפיקציה של הונאות ופעולות תמימות.

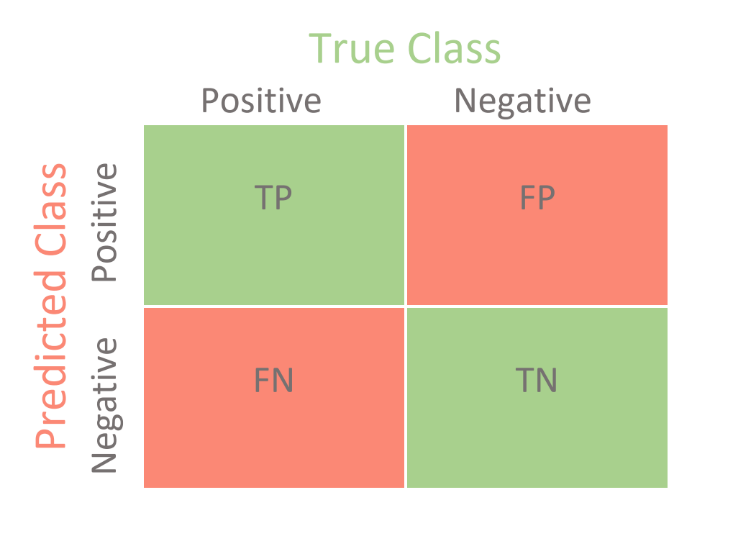
**השערות ושיטות המחקר:**

ההשערה העיקרית במחקר זה הינה שמודל SVC יהיה מודל טוב יותר בעל רמת דיוק גבוהה יותר בזיהוי הונאות אשראי. אך עם זאת טענו שגם DT יהווה מודל טוב בעל אחוזי דיוק גבוהים.

שיטות המחקר שהוצגו במאמר היו יצירת קוד של כל מודל לניתוח ולבסוף הערכת המודל לפי ערכים כמו Recall, Precision וכדומה אשר הוזכרו קודם במבוא. זאת על מנת לקבל דעה מושכלת, איזה מודל הוא הטוב ביותר.

שיטות המחקר קשורות למודלים של למידת מכונה ופחות לכריית נתונים כמו מערכת נוירונים אשר בעזרתה ניתן להבין התנהגויות וקשרים מורכבים (אנחנו נרצה לבדוק דיוק של מודל זה, ע"י תיקון המיעוט נוכל לקחת דאטה גדול יותר ובכך להשתמש במודלי כריית נתונים כמו מערכת נוירונים אשר גם אותה נבחר לממש).

שיטות המחקר המוצגות במאמר אינן חדשות, הם מפרטות את הדרך הפרקטית להערכת כל מודל ואף מציעות למחקר המשך לחקור כיצד ניתן עוד לשפר את המודלים כדי להגיע לתוצאות אמינות יותר.

לבסוף מסבירים את האופן הפרקטי של הערכת נתוני כל מודל אותו בחרו המחברים בעזרת המשוואות המשוואות הבאות (עם כי חסר הרחבה תיאורטית לכן אבחר להרחיב על משמעות של כל נתון:   
1) **Confusion** **table**

Confusion table הוא טבלה שמראה כמה נתונים סווגו כנכונים באמת, כנכונים בטעות, כטועים באמת וכטועים בטעות.

TP: True Positive -- TN: True Negative FP: False Positive -- FN: False Negative

איור 2: Confusion Table

2) **Precision**

Precision הוא מדד שמראה כמה מהנתונים שנחשבו כנכונים באמת נכונים.

Precision = True Positives / (True Positives + False Positives)

3)**Recall**

Recall הוא מדד שמראה כמה מהנתונים האמיתיים סווגו כנכונים.

Recall = True Positives / (True Positives + False Negatives)

4) **Accuracy**

Accuracy הוא מדד שמראה כמה מהנתונים סווגו כראוי.

Accuracy = (True Positives + True Negatives) / (Total)

5) **F-measure** –   
A black text on a white background

Description automatically generated

F-measure מחשב את הממוצע הרמוני של Precision ו-Recall. הממוצע הרמוני נותן משקל גבוה יותר לנתונים הנדירים יותר, כגון False Positives ו-False Negatives.

* **זיהוי הונאות בכרטיסי אשראי:** בתחום זה, חשוב לזהות את כל ההונאות, גם אם זה אומר לסווג כמה עסקאות אמיתיות כעסקאות הונאה. במקרה זה, F-measure גבוה יותר מ-Precision.
* **אבחון מחלות:** בתחום זה, חשוב לזהות את כל המחלות, אך גם חשוב לא לסווג אנשים בריאים כחולים. במקרה זה, F-measure גבוה יותר מ-Recall.

A white paper with black text

Description automatically generatedA white paper with black text

Description automatically generatedובנוסף מרחיבים באופן הפרקטי כיצד יתבצע כל מודל בעזרת הפסאודו קוד הבא:

תרשים 2:פסאודו קוד לאלגוריתם DT

תרשים 1:פסאודו קוד לאלגוריתם SVC

\*פסאודו קוד הוא תיאור מצומצם ולא רשמי לאלגוריתם של תוכנית מחשב. פסאודו קוד משתמש במוסכמות של שפות תכנות, אך מיועד לקריאה של בני אדם ולא לקריאה על ידי מחשב.

**אלגוריתמים ומודלים:**

מחקר זה אינו כולל מודלים של כריית נתונים אלא למידת מכונה בלבד והמידע בו השתמשו החוקרים הינו פומבי וזמין ב Kaggle .

**Decision Tree**

אלגוריתם Decision Tree הוא אלגוריתם למידת מכונה גנרטיבי המשתמש במבנה עץ כדי לייצג החלטות. המודל משתמש בסדרה של שאלות כדי להגיע למסקנה, כאשר כל שאלה מתבססת על תכונה אחת של הנתונים.

* אלגוריתם Decision Tree מתחיל עם קבוצת נתונים שבה לכל נתון יש מספר תכונות. המודל מתחיל בצומת אחד בעץ, ואז שואל שאלה אחת על כל תכונה. התשובה לשאלה קובעת את הענף שאליו ילך המודל.

**יתרונות:**

* + קל להבין ולהשתמש בו
  + לא דורש נורמליזציה של נתונים
  + יכול לזהות תבניות מורכבות

**חסרונות:**

* + אימון המודל עלול להיות ארוך
  + עלול להיות רגיש לנתונים לא תקינים, ונוטה להיות לא יציב
* אלגוריתם Support Vector Classifier הוא אלגוריתם למידת מכונה פורמלי המבוסס על מציאת קווים או מישורים המפריד בין נתונים מקבוצות שונות.

לדוגמה, באלגוריתם Support Vector Classifier ניתן להשתמש כדי לזהות תבניות של הונאות אשראי על ידי מציאת קו המפריד בין נתונים של הונאות ונתונים של פעולות תקינות.

אלגוריתם Support Vector Classifier מוצא את הקו או המישור המפריד בין נתונים מקבוצות שונות בצורה כזו שהמרחק בין הקו או המישור לנתונים מכל קבוצה הוא מקסימלי.

הנתונים שנמצאים על קו או מישור המפריד נקראים Support Vectors. הנתונים אלו חשובים עבור המודל מכיוון שהם קובעים את המיקום של קו או המישור.

**יתרונות:**

* + מותאם לזיהוי בין קבוצות כאשר ישנו הבדל מובהק ביניהם
  + מספק ביצועים טובים בבסיסי מידע מרובי תכונות

**חסרונות:**

* + מורכב יותר להבין אותו ולהשתמש בו
  + לא מתאים למאגרי מידע גדולים

**דיון ותוצאות המחקר:**

תוצאות המחקר המרכזיות הינן:

* SVC הוא אלגוריתם מדויק יותר לזיהוי הונאות בכרטיסי אשראי מאשר DT.
* הדיוק של SVC הוא 98.59%, בעוד שדיוק של DT הוא 94.86%.
* A table with numbers and text

  Description automatically generatedעם זאת, השגיאה הממוצעת של SVC היא מעט גבוהה יותר מזו של DT.

ניתן לראות (טבלה 1) את האיטרציות השונות ואת המדגם והדיוק באותה איטרציה לפי כל מודל (ניתן לשים לב איך הדיוק משתנה עם שינוי המדגם)

טבלה 1: תוצאות הדיוק של כל אלגוריתם מהמאמר

A table with numbers and a few words

Description automatically generated with medium confidenceלאחר מכן החוקרים בודקים בעזרת מבחן t את מובהקות ההבדל בדיוק של כל מודל, וברמת מובהקות של 5% נמצא כי אכן SVC מדויק יותר:

טבלה 2: כמות התצפיות, ממוצע וסטיית תיקון של דיוק כל אלגוריתם

**A table of test results

Description automatically generatedטבלת מבחן t דו זנבי לבדיקת ההבדל בין שני המודלים ומבחן Levene לשוויון שונויות:** ראשית ניתן לראות שה- P-value~=0 ולכן אין הבדל בין שונויות הקבוצות לאחר מכן במבחן T נראה כי גם כאן ה P-value~=0 **לכן נראה כי יש הבדל מובהק וכי ממוצע SVC גבוה יותר בר"מ 5%**

טבלה 3: תוצאות מבחני T ו Levene

ובנוסף הגרף הבא מתאר מראה את הדיוק של כל מודל וניתן לראות כי הדיוק של SVC גבוה יותר:

A graph of a bar

Description automatically generated with medium confidence

גרף 2: ממוצעי כל קבוצה +- שני סטיות תקן

לבסוף בשלב הדיון טוענים כי יש מקום למחקר המשך, הכולל:

1. שימוש בטכניקות אופטימיזציה כדי להפחית את השגיאה הממוצעת של SVC.
2. שימוש באלגוריתם של כריית נתונים על מנת ליצור מערכת שמזהה תבניות מורכבות יותר.
3. שילוב של אלגוריתמים שונים כדי להגביר את דיוק הפלט

**פרק המשך למחקר: ביקורת והצעות לשיפור**

**הצעה 1 לשיפור: תיקון ביצוע ""UnderSampling ע"י איזון כמות תצפיות ההונאות והרגילות**

כאשר ניסיתי לשחזר את המודל נראה כי ישנו חוסר יציבות בשל המיעוט בתצפיות ההונאה. נשים לב כי בשל שיטת החישוב של הדיוק ע"י Confusion\_Table והמיעוט המובהק במשתנה המטרה נוצר עיוות בדיוק כך שאם נשים לב לדוגמא ב SVC כאשר N=600 , **3 מתוך 4 הונאות לא זוהו ועדיין הדיוק 96.67%**

|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
| Model | TP | FP | FN | TN | Accuracy | comment |
| Desicion Tree | 1 | 1 | 3 | 115 | 96.67% | N=600 |
| 1 | 0 | 3 | 126 | 97.69% | N=650 |
| 4 | 1 | 0 | 135 | 99.29% | N=700 |
| 4 | 0 | 0 | 146 | 100.00% | N=750 |
| 3 | 1 | 1 | 155 | 98.75% | N=800 |
| 3 | 1 | 1 | 165 | 98.82% | N=850 |
| 2 | 0 | 2 | 176 | 98.89% | N=900 |
| 3 | 0 | 1 | 186 | 99.47% | N=950 |
| 2 | 1 | 2 | 195 | 98.50% | N=1000 |
| 4 | 0 | 0 | 206 | 100.00% | N=1050 |
| Support Vector Classifier | 1 | 1 | 3 | 115 | 96.67% | N=600 |
| 1 | 0 | 3 | 126 | 97.69% | N=650 |
| 4 | 1 | 0 | 135 | 99.29% | N=700 |
| 4 | 2 | 0 | 144 | 98.67% | N=750 |
| 4 | 1 | 0 | 155 | 99.38% | N=800 |
| 3 | 0 | 1 | 166 | 99.41% | N=850 |
| 2 | 1 | 2 | 175 | 98.33% | N=900 |
| 3 | 0 | 1 | 186 | 99.47% | N=950 |
| 2 | 1 | 2 | 195 | 98.50% | N=1000 |
| 4 | 0 | 0 | 206 | 100.00% | N=1050 |

טבלה 4: תוצאות שחזור הניסוי – דיוק, גודל מדגם, סוג המודל ו- ConfusionTable

A screenshot of a table

Description automatically generatedוכאשר נבדק שוב מבחן T ולוין התוצאות היו הפוכות מתוצאות המאמר:

לפי טבלה 6 נראה כי אין הבדל בממוצעי הקבוצות אך ישנו הבדל בשונות הקבוצות בר"מ 5%

טבלה 5: תוצאות שחזור הניסוי – ממוצע וסטיית תקן דיוק של כל מודל



טבלה 6: תוצאות שחזור הניסוי – מבחן T ו-Levene

A blue and white squares with black text

Description automatically generatedA table with numbers and a yellow line

Description automatically generatedולהלן השיפור, בהתאם לטכניקת ה- "UnderSampling" אותה נקטו החוקרים, ננסה לתקן ולאזן את המתודולוגיה שתתאים עדיין ל Confusion\_Table , בכך שכמות ההונאות תהיה שווה לכמות הפעולות האמיתיות אך לא גדולה מכמות ההונאות הכוללת במסמך (492 הונאות):

טבלה 7: אחרי שיפור 1- ממוצע וסטיית תקן דיוק של כל מודל

טבלה 9: אחרי שיפור 1- מבחן T ו-Levene

טבלה 8: אחרי שיפור 1- דיוק, גודל מדגם, סוג המודל ו- ConfusionTable

כעת ניתן לראות שהמודל יותר מאוזן ומכאן הדיוק שהוא משקף אמין יותר, נרצה לציין כי על אף שהתוצאה הגבוהה ביותר אכן הייתה של מודל SVC לא נמצא הבדל מובהק בין ממוצעי הקבוצות אלא רק בשונות. **לכן לא ניתן להניח ש SVC מודל טוב יותר לזיהוי הונאות ברמת מובהקות של 5%.**

**הצעה 2 לשיפור: שיפור אפשריים למודלים הנוכחיים (לאחר ייצוב המודל)**

2.1) ידוע כי כמות העלים ב – DT משנה את ה MAE (ממוצע השגיאות המוחלטות) וע"י הצבת כמות עלים שונה ובדיקת ה MAE האופטימלי (הנמוך ביותר) A graph of a number of scores

Description automatically generated with medium confidenceעלולה לשפר משמעותית את תוצאות המודל בהתאם לעיקרון זהנבחר להשתמש RandomForestClassifier אשר הינו מודל משופר של DT הפועל על מנת לאזן בין ה- OverFitting ל .UnderFitting2.2) עוד נשים לב שבעזרת MI אשר מציגה מידע משותף בין המשתנה המוסבר למוסברים ניתן להזין את המודל בעמודות אלו ובכך לנסות לשפר את דיוק שני המודלים.

לפי גרף 3 ניתן לראות כי עמודות

V11,V12,V10,V14,V16 ו- V17 מכילות הכי הרבה מידע משותף עם המשתנה המוסבר ולכן אבחר את ששת עמודות אלו לאימון המודלים.

גרף 3: מידע משותף בין משתנה מוסבר למסבירים

**A screenshot of a computer

Description automatically generatedA screen shot of a computer

Description automatically generatedלהלן התוצאות לאחר השיפורים:**

**A blue and black text

Description automatically generated**

טבלה 12: מבחן T ו-Levene אחרי שיפור 2.1 ו-2.2- מבחן T ו-Levene

טבלה 10: אחרי שיפור 2.1 ו-2.2- ממוצע וסטיית תקן הדיוק של כל מודל

טבלה 11: אחרי שיפור 2.1 ו-2.2- דיוק, גודל מדגם וסוג המודל

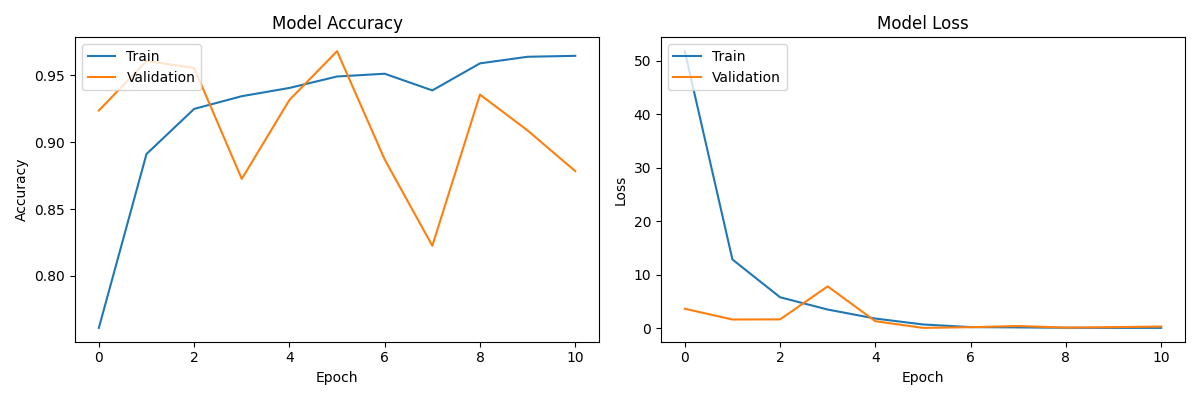
ניתן לראות בטבלה 12 שאין הבדל מובהק בממוצעי הקבוצות אך אין שוויון שונויות בר"מ 5%. אך נציין כי ממוצעי הדיוק אכן עלו במעט לאחר השיפורים (DT – 92.2% לעומת 91.5%, SVC- 92.52% לעומת 91.19%).

3) **מערכת נוירונים בשילוב Data-Augmentation ו- SMOTE:**

בבניית מערכת הנוירונים אשר הינה מערכת לכריית שיעילותה גבוהה מלימוד מכונה כשמדובר בממסדי נתונים גדולים והבנת התנהגויות מורכבות.על מנת להשתמש במודל כריית נתונים נעבור מגישת "UnderSampling" אותה נקטנו בלימוד מכונה ונעבור לגישת "OverSampling" ע"י ייצור תצפיות דמה ובכך נאזן את יחס המיעוט-רוב. קיימת טכניקת **Data-Augmentation** המשרתת עיקרון זה לדוגמא בנושא ComputerVision ע"י הסתכלות מזוויות שונות ניתן להרחיב את מאגר המידע משמעותית ובהקשר לממסד הנתונים שלנו המורכב מערכים נומריים, נראה כי אלגוריתם **SMOTE** Synthetic Minority) (Oversampling Technique משרת עיקרון זה ע"י ייצור תצפיות הונאה מלאכותיות. בנוסף על מנת להחמיר נשתמש גם בחבילת compute\_class\_weight שנועדה לתת משקל לסוגי תצפיות ומיועדת למדגמים לא מאוזנים כמו שלנו, ולבסוף ניעזר ב- early\_stop כדי לנסות למנוע OverFitting.



טבלה 13: מערכת נוירונים - דיוק, גודל מדגם, סוג המודל ו- ConfusionTable



גרף 4: מימין תרשים הפסד כתלות במספר ה- epochs משמאל דיוק המודל כתלות במספר ה- epochs.

בכחול אלו נתוני המידע שעליו אומן המודל ובכתום המידע שנשמר לאימות המודל

בכחלו

נראה כי אכן המודל לכריית נתונים מראה תוצאות מרשימות יותר ובעל דיוק של 95.78% בזיהוי הונאות, אך נשים לב כי המיעוט עדיין משפיע משמעותית וכי המודל עדיין מתקשה לזהות מידע חדש (לפי גרף 4 נראה כי ציון מידע ה-validation כ-87%)

**References:**

**קישור למאמר:** [**https://sifisheriessciences.com/journal/index.php/journal/article/view/463/446**](https://sifisheriessciences.com/journal/index.php/journal/article/view/463/446)

**קישור למאגר המידע:** [**https://www.kaggle.com/datasets/mlg-ulb/creditcardfraud**](https://www.kaggle.com/datasets/mlg-ulb/creditcardfraud)