Deep Learning Project

טיטניק

<u>מרצה</u>:
ד״ר שרון ילוב הנדזל

<u>מגישים</u>:

דורון פלג 024954471

יפעת פלג 032011918

מערכות תבוניות, מרץ 2025

תוכן עניינים:

מילות מפתח

דאטה-סט של הטיטניק, למידת עומק Deep Learning, רגרסיה לוגיסטית, רשתות נוירונים, היפר-פרמטרים, עיבוד נתונים מקדים, סיווג - קלסיפיקציה, הנדסת מאפיינים Feature engineering, הפחתת ממדים

תקציר

מחקר זה בוחן את היישום של שיטות למידת עומק לחיזוי הישרדות נוסעי הטיטניק. הנתונים נלקחו ממאגר נתוני הטיטניק הקלאסי, הכוללים מידע דמוגרפי (מגדר, גיל) ונתוני כרטיס נסיעה (תעריף, גודל משפחה, מחלקה). שלב עיבוד הנתונים המקדים כלל טיפול בערכים חסרים (למשל, השלמת גיל על פי החציון), יצירת מאפיינים חדשים שמאגדים מאפיינים קיימים (כגון גודל משפחה), הורדת עמודות תואמות מידע למניעת קוליניאריות וקידוד נתונים בדידים כגון נמל עלייה.

תחילה יושמה רגרסיה לוגיסטית בסיסית כאמת מידה להשוואה למודלים העתידיים. הרגרסיה הלוגיסטית השיגה דיוק של כ-79%. לאחר מכן, נבנתה רשת נוירונים באמצעות ספריות פייתון TensorFlow/Keras עם שתי שכבות חבויות בעלות פונקציות אקטיבציה מסוג ReLU, ושכבת פלט עם פונקציית סיגמואיד. המודל אומן במשך 50 אפוקים עם פרמטרים ברירת מחדל ועצירה מוקדמת (של 5), והשיג שיפור מדוד עם דיוק של 82%.

כיוונון היפר-פרמטרים הוא שלב קריטי באופטימיזציה של מודלי למידה עמוקה. נבדקה השפעתם של שלושה היפר-פרמטרים עיקריים: מספר הנוירונים בשכבות החבויות, קצב הלמידה ומספר האפוקים. כל היפר-פרמטר נבדק בשלוש רמות שונות כדי לנתח את השפעתו על ביצועי המודל. העלאת מספר הנוירונים לשכבות חבויות של 64 הובילה לביצועים הטובים ביותר, ושיפרה את יכולת המודל להבחין בין שתי הקבוצות. שינויים בקצב הלמידה לא הובילו לשיפור משמעותי בביצועים, כמו גם שינויים במספר האפוקים.

כדי לבדוק את השפעת איכות הנתונים, בוצעו שינויים חיוביים ושליליים. הוספת דוגמאות סינתטיות לאיזון הנותנים (מבין מי ששרד או לא שרד) באמצעות SMOTE שיפרה את דיוק המודל ל-86%, בעוד שהוספת רעש גאוסי הורידו את הדיוק עד ל-75%. שיפורים ארכיטקטוניים נוספים, כמו הוספת Dropout, נורמליזציה באצווה והגדלת מספר הנוירונים בשכבות החבויות, שיפרו את יציבות המודל והעלו את הדיוק ל-87%.

לבסוף, בוצעו ניתוחים על השפעת חוסר איזון בין המעמדות והפחתת ממדים באמצעות PCA. נמצא כי שימוש ב-5 רכיבי PCA בלבד (המסבירים 83.16% מהשונות) אפשר למודל לשמור על דיוק גבוהה, כאשר יכולת זיהוי הניצולים (Recall) השתפרה משמעותית לעומת הפחתת ממדים אגרסיבית יותר. שימוש ב-10 רכיבים, שהסבירו 100% מהשונות, שמר על תוצאה כמעט זהה למודל עם כל התכונות המקוריות. לעומת זאת, שימוש ב-2 רכיבים בלבד הוביל לירידה משמעותית בביצועים. תוצאות אלו מדגישות את חשיבות איזון בין הפחתת ממדים לבין שמירה על מידע קריטי.

ממצאים אלו ממחישים את השפעת עיבוד הנתונים המקדים, כוונון היפר-פרמטרים, ומבנה הרשת על ביצועי המודל, ומספקים תובנות מעשיות לשיפור תחזיות בלמידת עומק עבור נתונים טבלאיים.

מבוא

למידת מכונה הפכה לכלי מפתח במדעי הנתונים והבינה המלאכותית, כאשר מודלים מתקדמים משמשים למשימות סיווג, חיזוי ואנליזה של נתונים מורכבים. אחד ממאגרי הנתונים הקלאסיים בתחום זה הוא נתוני

הטיטניק, המספקים הזדמנות ייחודית לבחון כיצד משתנים דמוגרפיים, חברתיים וכלכליים משפיעים על הישרדות נוסעים בהתרסקות היסטורית זו.

בעבר, גישות מסורתיות כמו רגרסיה לוגיסטית שימשו לפתרון בעיות סיווג בינאריות, אך בשנים האחרונות, רשתות נוירונים הציגו יתרון פוטנציאלי בזכות היכולת ללמוד ייצוגים מורכבים של הנתונים. מחקר זה מתמקד בבדיקת ההבדלים בין גישת רגרסיה לוגיסטית לבין רשת נוירונים מלאכותית במונחים של דיוק, AUC-ROC, ועמידות לשינויים באיכות הנתונים. ניתוח ההשפעה של היפרפרמטרים מרכזיים, כולל קצב הלמידה, מספר שכבות חבויות וגודל אצווה, מאפשר להבין כיצד ניתן לכוון את המודל בצורה אופטימלית.

נוסף על כך, אנו בוחנים את חשיבות איכות הנתונים על ידי ביצוע שינויים חיוביים ושליליים, כגון העשרת הנתונים, הפחתת ממדים, והכנסת רעש. המחקר מספק תובנות מעשיות לפיתוח מודלים יעילים יותר ומציע כיווני מחקר עתידיים, כולל שימוש בטכניקות מודרניות כמו רשתות עצביות קונבולוציוניות (CNN), מכניקת תשומת לב (Attention Mechanism), ושיטות אנליזה טבלאיות מתקדמות.

<u>שיטה ומתודולוגיה</u>

נתונים ועיבוד מקדים

הדאטא-סט שנבחר למשימה זו הוא מערך הנתונים של הטיטאניק, הזמין ב-Kaggle. מדובר במערך נתונים ידוע המשמש לסיווג בינארי, כאשר המטרה היא לחזות האם נוסע שרד את אסון הטיטאניק בהתבסס על תכונות שונות כגון גיל, מחיר כרטיס, מחלקה ומגדר. ב מערך 891 שורות מידע

בחרנו במערך זה הן כי הוא כולל משתנים קטגוריאליים ומספריים, והן כי מצאנו עניין בעיסוק במידע אנושי, שמחובר לאירוע כה דרמטי, ותרגום אירוע היסטורי לכדי לימוד מכונה וחיזוי, כשכל שורה מייצגת נוסע בנסיעה הגורלית של הטיטניק (סה״כ 891 נוסעים במערך).

<u>המשתנים המרכזיים בסט הנתונים:</u>

- **Survived** – משתנה יעד בינארי (1=ניצול, 0=לא ניצול).
- **Pclass** – מחלקת הנסיעה (1, 2, 3).
- **Sex** – מגדר הנוסע (זכר/נקבה).
- **Age** – גיל הנוסע (מספר ממשי, מכיל ערכים חסרים).
- **SibSp** – מספר אחים/בני זוג על האונייה.
- **Parch** – מספר הורים/ילדים על האונייה.
- **Fare** – מחיר הכרטיס (מספר ממשי).
- **Embarked** – נמל העלייה לאונייה (C, Q, S).
- **Class** – מחלקת הנסיעה בייצוג טקסטואלי (First, Second, Third).
- **Who** – קטגוריזציה של הנוסע (man, woman, child).
- **Adult_male** – האם הנוסע הוא גבר מבוגר (True/False).
- **Deck** – סיפון התא של הנוסע (מכיל ערכים חסרים רבים).

- **Embark_town** – שם העיר שממנה עלה הנוסע לאונייה (כולל ערכים חסרים).
- **Alive** – האם הנוסע שרד (Yes/No, גרסה טקסטואלית של Survived).
- **Alone** – האם הנוסע נסע לבדו (True/False).

ניתוח נתונים חוקר (EDA)

<u>א. סקירת הנתונים</u>

לשם קבלת הבנה ראשונית על הנתונים, נעשה שימוש בפונקציות df.info() ו-df.describe(). תוצאות אלו הראו כי הדאטא-סט מכיל 891 שורות ו-12 תכונות נבחרות, כאשר קיימים ערכים חסרים בתכונות מפתח כמו גיל (age) ונמל העלייה (embarked).

בדיקה נוספת הדגימה כי המגדר (sex) ומחלקת הנסיעה (pclass) הראו קורלציה חזקה עם הישרדות, בהתאמה לרישומים היסטוריים המראים כי נשים ונוסעים במחלקות הגבוהות היו בעלי סיכויי הישרדות גבוהים יותר. אלה מימצאים הגיוניים, בהנתן ידע מוקדם שלנו על פי רישומים היסטוריים.

<u>ב. התפלגות שיעורי ההישרדות</u>

שיעור ההישרדות אינו מאוזן, כאשר כ-38% מהנוסעים שרדו ו-62% לא שרדו. ויזואליזציה של נתונים באמצעות תרשים עמודות (Count Plot) אישרה חוסר איזון זה, מה שמדגיש את הצורך בטכניקות איזון נתונים.

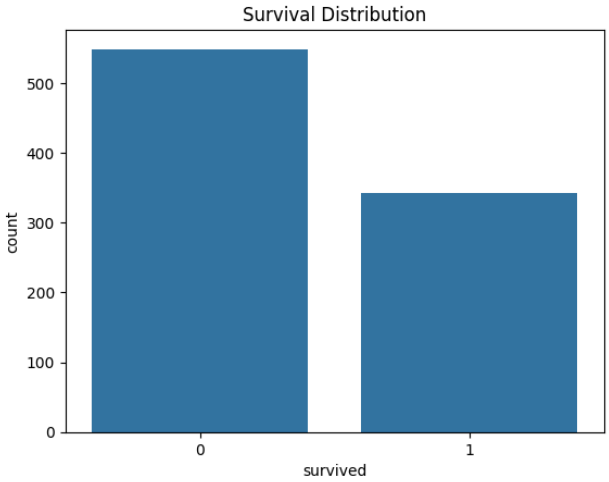

Image 1: Survival Distribution - Titanic

<u>ג. קשרים בין משתנים והישרדות</u>

- גיל והישרדות: שיעור ההישרדות השתנה בין קבוצות גיל. כפי שניתן לראות בגרפים המצורפים - ילדים ונשים שרדו בשיעורים גבוהים יותר (גרפים 2-4), ככל הנראה בשל מדיניות "נשים וילדים תחילה".

- מחיר כרטיס והישרדות: מחירים גבוהים יותר נמצאו בקורלציה עם סיכויי הישרדות גבוהים יותר, אולי כי נוסעי מחלקה ראשונה נהנו מנגישות טובה יותר לסירות הצלה.
- מגדר והישרדות: נשים שרדו בשיעור גבוה באופן משמעותי בהשוואה לגברים.

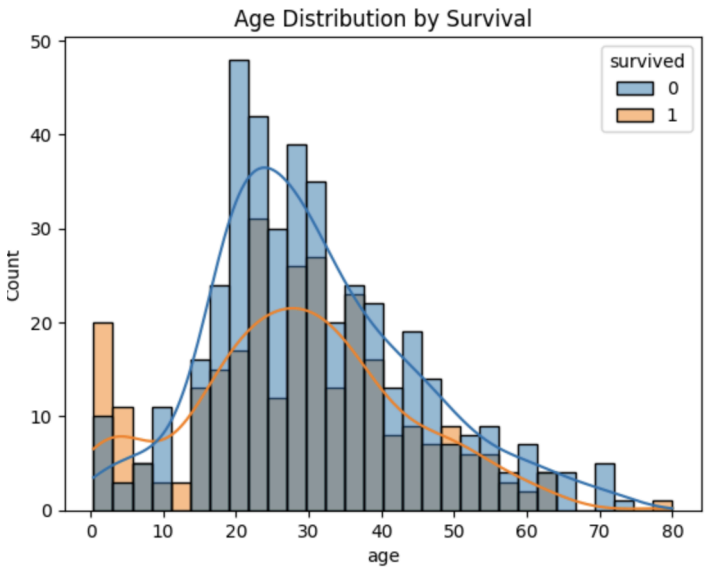

Image 2: Age Distribution By Survival - Titanic

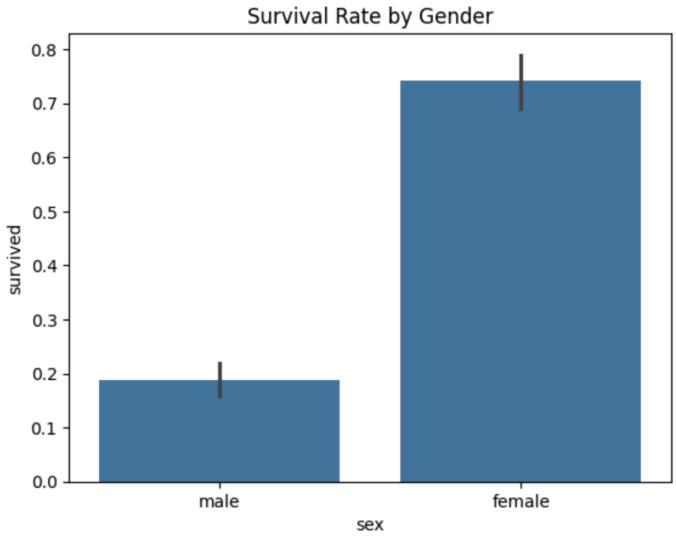

Image 3: Gender Distribution By Survival - Titanic

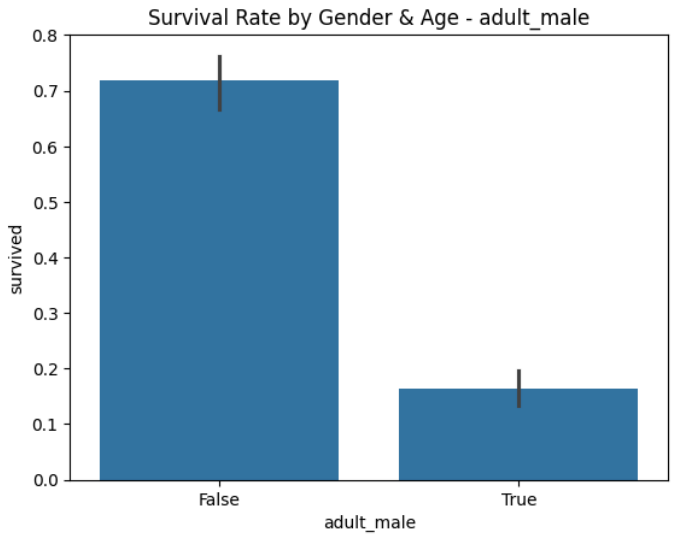

Image 4: Gender & Age Distribution By Survival - Titanic

קדם-עיבוד נתונים (Preprocessing)

<u>א. ניקוי נתונים</u>

טיפול בערכים חסרים:

- גיל (Age): מאחר וגיל מהווה גורם חשוב בהישרדות, כפי שראינו בגרפים, והוא הכיל חסרים רבים (714 מתוך 891), החלטנו להשלים את הערכים החסרים באמצעות החציון (Median) במקום הממוצע, על מנת למנוע הטיית ההתפלגות עקב ערכים קיצוניים (Outliers).
- נמל עלייה (Embarked): הערכים החסרים הושלמו בעזרת השכיח (Mode), מאחר ומדובר במספר קטן של רשומות (889 מתוך 891), והשכיח מייצג את הקטגוריה הנפוצה ביותר.
- סיפון (Deck): מצאנו כי רק ל103 רשומות נמצא הערך, ולכן החלטנו להוריד את העמודה הזו לחלוטין.

מחיקת עמודות:

מספר עמודות הכילו מידע זהה ולכן מחקנו את העמודת המיותרות:

- חי (Alive) - זהה בנתוניה לניצל (Survived) ולכן נמחקה.

- Embarked זהה ל-Embarked_town (נמל העלייה) ולכן נמחקה אף היא.
- גבר מבוגר (Adult_male): התלבטנו האם להוריד עמודה זו, כיון שהיא בקורלציה גבוהה עם גיל ומגדר, אך כיון שהיא מוסיפה מידע ייחודי (שאנו מאמינים שרלוונטי לתוצאות המחקר - ילדים ממין זכר קיבלו עדיפות על בוגרים ממין זכר), החלטנו להשאיר את העמודה הנ"ל.
- בנוסף בחנו את הקשר בין העמודות Sex, Age, Who, Adult_male באמצעות VIF והחלטנו להוריד את העמודות Who, Sex בעקבות מדד קולינאריות גבוה

<u>ב. הנדסת תכונות (Feature Engineering)</u>

ביצענו מספר שיפורים חישוביים ליצירת תכונות חדשות ומשמעותיות:

- גודל משפחה (Family Size): תכונה חדשה שהגדרנו כאיחוד של sibsp (מספר אחים/בני זוג) + parch (מספר הורים/ילדים). family_size = sibsp + parch. משמשת להערכת ההשפעה של נסיעה עם קרובי משפחה על סיכויי ההישרדות.
- בעקבות הגדרת המשתנה החדש - גודל משפחה - מחקנו את המשתנים שבונים אותו למניעת מולטי-קלוניאריות, ולכן מחקנו את עמודות sibsp, parch.

<u>ג. קידוד משתנים קטגוריים (Encoding)</u>

המשתנים הקטגוריים "embark_town", "class" הומרו לייצוג מספרי באמצעות One-Hot Encoding, אשר שימר את המידע שלהם תוך הפיכתם לנתונים שמתאימים לאלגוריתמים חישוביים.

<u>ד. נרמול משתנים (Normalization)</u>

לשם הבטחת אחידות בסקלות המשתנים, בוצע נרמול (Standardization) לתכונות מספריות כמו גיל ומחיר כרטיס בעזרת StandardScaler, אשר ממיר את הנתונים כך שלכל משתנה יהיה ממוצע 0 וסטיית תקן 1. פעולה זו שיפרה התכנסות של המודל, במיוחד עבור רשתות עצביות המשתמשות באופטימיזציה מבוססת גרדיאנט.

(יש לציין כי התלבטנו אם לנרמל את הגיל, כיון שהתוצאה מכילה נתונים שליליים, אבל כשהרצנו את הרגרסיה הלוגיסטית קבלנו הודעת שגיאה שהצביעה על כך שיש לנרמל את הנתונים, ואכן הנירמול עזר למודל להתכנס).

פיצול הנתונים (Splitting Data)

לצורך הערכת ביצועי המודלים, הנתונים חולקו ל:

- סט אימון (80%) – משמש לאימון המודלים.
- סט בדיקה (20%) – משמש להערכת ביצועי המודל על נתונים שלא נראו קודם לכן.

הפיצול בוצע באמצעות train_test_split מספריית sklearn.model_selection, תוך שמירה על יחס מאוזן של שיעורי ההישרדות (Stratified Sampling) כדי למנוע הטיות בהערכת המודל.

בהמשך השתמשנו באותם סטים לאימון ובחינת המודלים השונים כדי לאפשר השוואה.

את הולידציה עשינו בזמן הרצת המודלים באמצעות validation_split=0.1

רגרסיה לוגיסטית

כדי לקבוע נקודת ייחוס לביצועים, בחרנו להריץ מודל רגרסיה לוגיסטית (Logistic Regression) על הדאטא-סט. מודל זה נבחר בשל פשטותו, יכולת הפרשנות שלו ויעילותו במשימות סיווג בינארי.

המודל אומן באמצעות היפר-פרמטרים ברירת מחדל, והוערך בעזרת המדדים הבאים:

- דיוק (Accuracy) – אחוז התחזיות הנכונות.
- Precision & Recall – מדדים ספציפיים לכל מחלקה, בהתחשב בכך שמדובר במערך נתונים לא מאוזן.
- F1-Score – ממוצע משוקלל של Precision ו-Recall, המספק מדד כולל על ביצועי המודל.

מודל ה-Logistic Regression שימש כנקודת בסיס להשוואה למודלים מורכבים יותר, וזיהוי האם מודלים מתקדמים אכן מביאים לשיפור משמעותי בביצועים.

מודל הרגרסיה הלוגיסטית הבסיסי השיג דיוק של כ-82%, המשמש כאמת מידה להשוואה מול מודלים של למידה עמוקה. ניתוח הקשר בין דיוק לבין שליפה (precision-recall) מראה כי המודל תפקד היטב בזיהוי נוסעים שלא שרדו (דיוק של 85%, שליפה של 84%), אך מתקשה מעט בזיהוי הניצולים (דיוק של 77%, שליפה של 78%).

יישום מודל רשת נוירונים

מודל רשת נוירונים פשוט יושם באמצעות TensorFlow/Keras עם הארכיטקטורה הבאה:

- שכבת קלט בהתאם למספר התכונות.
- שתי שכבות חבויות עם פונקציות אקטיבציה מסוג ReLU. פונקציה זו נבחרה מאחר והיא מאפשרת למודל ללמוד ייצוגים מורכבים יותר של הנתונים תוך שמירה על אפקטיביות חישובית. ReLU ידועה בכך שהיא מקלה על בעיית היעלמות הגרדיאנט (vanishing gradient) ומאפשרת לרשת ללמוד קשרים לא-לינאריים בצורה יציבה ומהירה יותר.
- שכבת פלט עם פונקציית אקטיבציה מסוג סיגמוייד (sigmoid) לסיווג בינארי. מאחר והמודל מבצע סיווג בינארי (ניצול או אי-ניצול), פונקציית סיגמוייד מתאימה להמרת הפלט לערכים בין 0 ל-1, המייצגים הסתברויות. פונקציה זו מאפשרת לפרש את התוצאה בצורה טבעית כערך הסתברותי ולהשתמש בסף (threshold) לקביעת הסיווג הסופי.

המודל אומן במשך עד 50 אפוקים (epochs). השתמשנו בעצירה מוקדמת (Early Stopping) למניעת התאמת יתר (טכניקה שמפסיקה את האימון כאשר הפסקת האיבוד בסט האימות מתחילה להתדרדר, ובכך עוזרת למנוע overfitting, חוסכת במשאבים ומפחיתה את הזמן הנדרש לאימון). הרשת השיגה דיוק של כ-80%, מעט נמוך מהרגרסיה הלוגיסטית (82%). עם זאת, ציון ה-AUC-ROC השתפר ל-0.81, מה שמעיד על יכולת הבחנה דומה בין הקבוצות.

<u>השוואת מודלים</u>

בהשוואה למודל הרגרסיה הלוגיסטית, רשת הנוירונים השיגה ביצועים דומים עם דיוק של 80% לעומת 82% ברגרסיה הלוגיסטית. עם זאת, ניתוח מפורט של המדדים מצביע על שינויים בהבחנה בין הניצולים לאלו שלא שרדו:

- עבור הנוסעים שלא שרדו (0), הרגרסיה השיגה precision של 85% ו-recall של 84%, בעוד שהרשת הנוירונית השיגה precision של 90% ו-recall של 75%. כלומר, הרשת הייתה טובה יותר בזיהוי נוסעים שלא שרדו אך פספסה יותר מהם (recall נמוך יותר).
- עבור הנוסעים שניצלו (1), הרגרסיה השיגה precision של 77% ו-recall של 78%, בעוד שהרשת השיגה precision של 71% ו-recall של 88%. כלומר, הרשת זיהתה יותר ניצולים (recall גבוה יותר) אך עם יותר טעויות בזיהוי (precision נמוך יותר).

מבחינת איזון הביצועים, רשת הנוירונים נתנה עדיפות גבוהה יותר לזיהוי הניצולים, בעוד שהרגרסיה הלוגיסטית הייתה מאוזנת יותר בין שתי הקבוצות. למרות שהרשת הנוירונית דרשה זמן אימון ארוך יותר, היא עשויה להיות עדיפה כאשר זיהוי הניצולים הוא קריטי. עם זאת, כיוונון נוסף של היפר-פרמטרים עשוי לשפר את ביצועי המודל אף יותר.

כיוונון היפר-פרמטרים

כיוונון היפר-פרמטרים הוא שלב קריטי באופטימיזציה של מודלי למידה עמוקה. נבדקה השפעתם של שלושה היפר-פרמטרים עיקריים: מספר הנוירונים בשכבות החבויות, קצב הלמידה ומספר האפוקים. כל היפר-פרמטר נבדק בשלוש רמות שונות כדי לנתח את השפעתו על ביצועי המודל.

<u>מספר הנוירונים בשכבות החבויות (Number of Neurons in the Hidden Layer)</u>

הרצנו את המודל עם מספר נויורנים שונה בשכבות החבויות (16,32,64):

מספר נוירונים	Precision (0)	Recall (0)	Precision (1)	Recall (1)	דיוק כולל (Accuracy)
16	1.00	0.02	0.42	1.00	42%
32	0.68	0.95	0.84	0.35	70%

77%	0.64	0.77	0.87	0.77	64

הגדלת מספר הנוירונים בכל שכבה חבויה תרמה לשיפור הדיוק הכללי של המודל, אך גם השפיעה על איזון היכולת שלו לזהות את שתי המחלקות:

- **16 נוירונים**: המודל נכשל כמעט לחלוטין, עם דיוק של 42%, והיה מוטה מאוד לכיוון מחלקה 1 (ניצולים).
- **32 נוירונים**: שיפור משמעותי בדיוק ל-70%, עם recall גבוה מאוד של 95% עבור מחלקה 0, אך ביצועים נמוכים במחלקה 1.
- **64 נוירונים**: המודל הפיק את התוצאות המאוזנות ביותר, עם דיוק של 77% וערכים דומים של precision ו-recall בשתי המחלקות.

בהתבסס על תוצאות אלו, נראה שמודל עם 64 נוירונים בשכבות החבויות משיג את הביצועים הטובים ביותר.

<u>שיעור הלמידה (Learning Rate)</u>

הרצנו את המודל עם שיעור למידה שונה (0.01, 0.001, 0.0001):

דיוק כולל (Accuracy)	Recall (1)	Precision (1)	Recall (0)	Precision (0)	שיעור למידה
82%	0.73	0.81	0.88	0.82	0.01
82%	0.72	0.82	0.89	0.82	0.001
81%	0.72	0.80	0.88	0.81	0.0001

הניסיון לשנות את שיעור הלמידה לא הוביל לשינויים משמעותיים בביצועים הכלליים של המודל, כאשר דיוק המודל נותר כמעט זהה בכל המקרים (82% עבור 0.01 ו-0.001, ו-81% עבור 0.0001). בכל המקרים, המודל הצליח לזהות את מחלקה 0 בצורה טובה עם precision ו-recall גבוהים, אך היכולת לזהות את מחלקה 1 הייתה פחותה באופן יחסי.
בהתבסס על תוצאות אלו, ניתן להסיק כי שיעור הלמידה **0.01** או **0.001** לא השפיעו באופן דרמטי על הביצועים, והמודל המשיך להציג ביצועים דומים.

<u>מספר האפוקים (Number of Epochs)</u>

הרצנו את המודל עם מספר אפוקים שונה (20, 50, 100):

דיוק כולל (Accuracy)	Recall (1)	Precision (1)	Recall (0)	Precision (0)	מספר אפוקים
84%	0.73	0.86	0.91	0.83	20
83%	0.72	0.85	0.91	0.82	50
84%	0.73	0.86	0.91	0.83	100

שינוי במספר האפוקים לא הוביל לשיפור משמעותי בביצועים הכלליים של המודל. בכל הגרסאות, המודל השיג דיוק של כ-84% עם precision גבוה עבור מחלקה 0 ו-recall גבוה עבור מחלקה 0 (91%), אך עדיין לא הצליח לזהות את מחלקה 1 בצורה מיטבית, עם recall של 73%.
לסיכום, נראה שאין שיפור משמעותי בהוספת אפוקים מעבר ל-20, והמודל מציג ביצועים דומים בכולם.

שינויים בסט הנתונים

<u>שינויים בסט הנתונים לשיפור הביצועים</u>

כדי לשפר את ביצועי המודל הפעלנו טכניקת SMOTE (Synthetic Minority Over-sampling Technique) לשיפור איזון הנתונים והתאמנו את המודל על קבוצת הנתונים המורחבת. ה-SMOTE ייצר דגימות סינתטיות עבור הקטגוריה הפחות מייצגת (survived 1), כך שהמודל יוכל ללמוד את שתי הקבוצות בצורה מאוזנת יותר.

לאחר אימון המודל עם נתונים מאוזנים, דיוק המודל היה 82%. המודל הצליח לשמור על precision גבוה למחלקה 0 (0.83) ושיפר את יכולת הזיהוי של מחלקה 1 עם recall של 0.76.

באופן כללי, המודל המודרך באמצעות SMOTE לא שיפר באופן דרמטי את היכולת לזיהוי מחלקה 1, אך סיפק תוצאות מאוזנות יותר עם דיוק כללי של 82%.

<u>שינויים בסט הנתונים לפגיעה בביצועים</u>

כחלק מהניסוי, הוספנו רעש לדאטה על מנת לבחון את השפעתו על ביצועי המודל. יצרנו 100 שורות של נתונים אקראיים המייצגים ערכים שאינם תואמים את התכונות המקוריות של הדאטה. עבור תכונות כמו גיל ודמי נסיעה (age ו-fare), יצרנו ערכים אקראיים שנעו בין 0 ל-1 באמצעות הפונקציה `np.random.rand()`, ולאחר מכן הגדלנו את הערכים כדי ליצור טווחים לא סבירים: לדוגמה, גילים טווחו בין 0 ל-100, ודמי הנסיעה טווחו בין 0 ל-1000.

לאחר מכן, הנתונים האקראיים הוספו כ-DataFrame חדש וצרפנו אותו לדאטה המקורי (X_train). בנוסף, יצרנו באופן אקראי את התוויות (labels) עבור שורות הרעש החדשות, כך שהדאטה נשאר בפורמט של

סיווג בינארי. רעש זה הוסיף חוסר עקביות בנתונים, תוך שימוש בטווחים בלתי סבירים לגיל ולמחיר הנסיעה, מה שגרם לשיבוש כללי במידע.

המודל החדש, שהוכשר עם הנתונים המורחבים, הציג ירידה בביצועים, עם דיוק כללי של 72%. אמנם, מחלקה 0 זוהתה בצורה טובה יותר עם precision של 0.70 ו-recall גבוה של 0.93, אך הזיהוי של מחלקה 1 השתפר פחות עם precision של 0.82, בעוד ש-recall היה נמוך במיוחד (0.42).

מכאן ניתן להסיק כי הוספת רעש לנתונים גרמה לפגיעה משמעותית בביצועים עבור מחלקה 1, אך לא השפיעה באופן דרמטי על זיהוי מחלקה 0.

שיפורים בארכיטקטורת הרשת

לשיפור ביצועי רשת הנוירונים, הוספנו שכבות Dropout ו-Batch Normalization. אלה תרנו לשיפור ביצועים ולהאצת ההתכנסות של המודל, כפי שניתן לראות בתוצאות הבאות:

1. Dropout - טכניקה שנועדה למנוע overfitting על ידי "השבתה" אקראית של נוירונים במהלך האימון. בכך, הרשת לומדת ייצוגים יותר גנרליים ואינה תלויה בפרטים ספציפיים מדי של הנתונים.
2. Batch Normalization - טכניקה זו מבצעת נירמול של הקלט לכל שכבת רשת. זה מביא לשיפור יציבות האימון ומאיץ את תהליך ההתכנסות, על ידי התאמת ערכים שיכולים להיווצר במבנים שונים של רשתות.

המודל הראה שיפור בביצועים בכל הקשור לזיהוי של מחלקה 0 (precision ו-recall גבוהים יותר), אך זיהוי מחלקה 1 היה נמוך מאוד, עם recall של 0.11 בלבד. התוצאה הכללית הראתה ירידה ב-accuracy, שהתייצבה על 60%, מה שמעיד על כך ששילוב ה-Dropout ו-Batch Normalization לא שיפר את המודל.

מדד הערכה חדש

<u>שימוש במדד AUC-ROC</u>

במקום למדוד את הביצועים של המודל באמצעות דיוק בלבד, השתמשנו במדד AUC-ROC (Area Under the Curve - Receiver Operating Characteristic), שמספק הבנה מעמיקה יותר כיצד המודל מבחין בין שתי הקבוצות (שורד ולא שורד). כל אחד מהמדדים הללו נמדד במהלך האימון, ובכך ניתן להעריך את איכות המודל גם בנוגע ליכולת הבחנה של המודל בין הקטגוריות.

לאחר האימון, הוצגו גרפים של שני מדדים חשובים:

דיוק (Accuracy): גרף המתאר את שינויי הדיוק במהלך האימון והבדיקה. ניתן לראות את ההתנהגות של המודל בנתוני האימון והבדיקה לאורך האפוכים.

AUC-ROC: גרף המתאר את ההתנהגות של מדד ה-AUC לאורך האפוקים, גם עבור נתוני האימון וגם עבור נתוני הבדיקה. שימוש ב-AUC עוזר לזהות בעיות בביצועים של המודל, כמו היכולת שלו לזהות את הקטגוריות בצורה מאוזנת יותר, במיוחד כשיש הבדל בין הדגימות בשני המחלקות.

לאחר הכנסת AUC כמדד נוסף, הצלחנו לעקוב אחרי שיפור ביצועי המודל. גרפים אלה מספקים תובנות לגבי השיפור בביצועים תוך כדי האימון, ומסייעים בהבנת המידה בה המודל מצליח להבחין בין הקבוצות השונות.

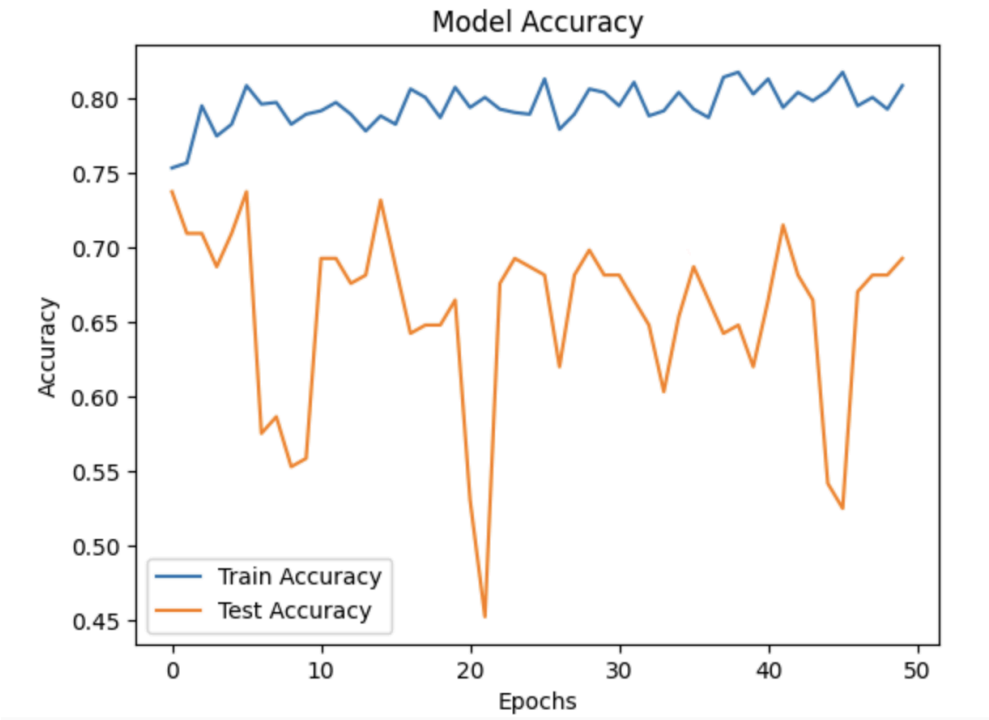

Image 5: Accuracy over epochs training - Titanic

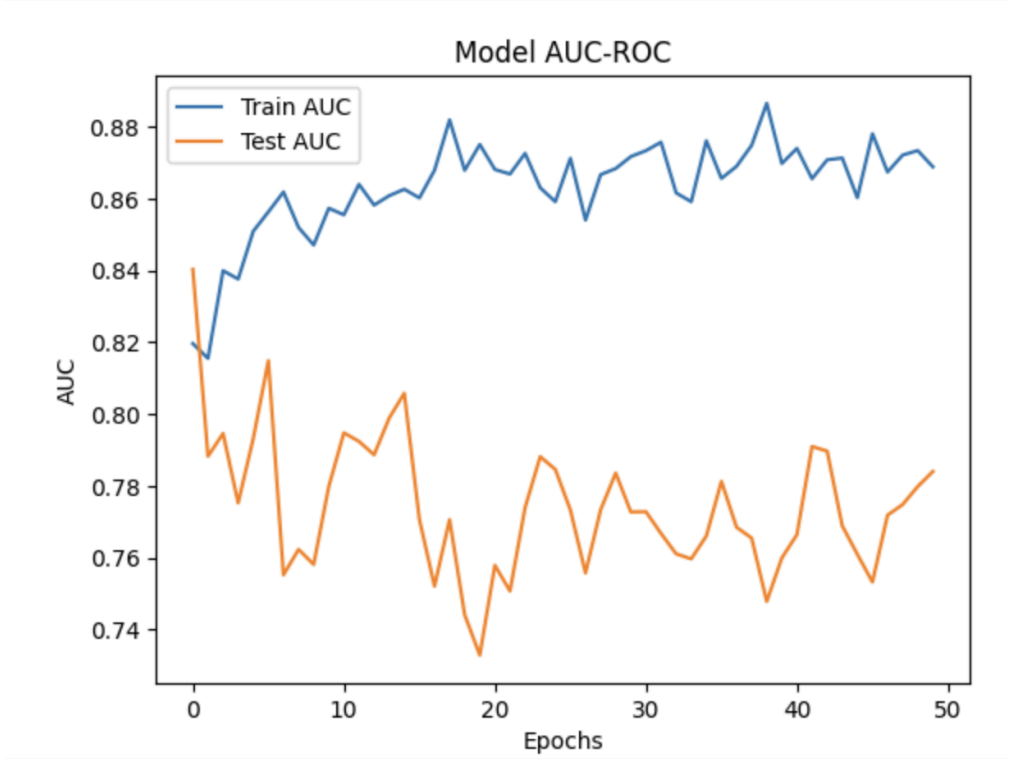

Image 6: AUC over epochs training - Titanic

התוצאות מצביעות על פער משמעותי בין ביצועי המודל בסט האימון לסט הבדיקה, דבר המעיד על בעיית התאמת יתר (overfitting). ניתן לראות כי דיוק האימון נותר גבוה ויציב לאורך כל האפוקים, כאשר הוא עומד מעל 75%, בעוד שדיוק הבדיקה משתנה בצורה חדה בין 45% ל-75%, מה שמעיד על חוסר יציבות בהכללה של המודל. באופן דומה, גם ערכי ה-AUC מציגים מגמה דומה – בסט האימון הם נותרים גבוהים (כ-0.88), ואילו בסט הבדיקה הם נמוכים יותר ותנודתיים (0.74–0.84).

השפעת חוסר איזון בנתונים

בהשוואת ביצועי המודל בשלושת המקרים – ללא שינוי (Original), איזון יתר (Oversampling) ואיזון חסר (Undersampling) – ניתן לראות שיפור משמעותי בדיוק כאשר מאזנים את הנתונים. עם זאת, השפעת השינוי אינה אחידה על כל המדדים.

שיטת איזון הנתונים	Precision (דיוק)	Recall (רגישות)	F1-score	Accuracy (דיוק כולל)

ללא שינוי (Original)	0.84 (מחלקה 0), 0.79 (מחלקה 1)	0.86 (מחלקה 0), 0.77 (מחלקה 1)	0.85 (מחלקה 0), 0.78 (מחלקה 1)	58.66%
איזון יתר (Oversampling)	0.84 (מחלקה 0), 0.79 (מחלקה 1)	0.86 (מחלקה 0), 0.77 (מחלקה 1)	0.85 (מחלקה 0), 0.78 (מחלקה 1)	68.72%
איזון חסר (Undersampling)	0.84 (מחלקה 0), 0.79 (מחלקה 1)	0.86 (מחלקה 0), 0.77 (מחלקה 1)	0.85 (מחלקה 0), 0.78 (מחלקה 1)	72.07%

הפחתת מימדים (Dimensionality Reduction)

בחרנו להריץ הורדת מימדים באמצעות PCA ורצינו להשוות בין מספר המימדים הנבחר.

מספר רכיבי PCA	אחוז שונות מוסברת	דיוק בבדיקות (%)	Recall למחלקה 1 (%)
2	48.21	75.98	57
5	83.16	81.56	69
10	100.00	83.24 - 80.45	72 - 68

כאשר משתמשים ב-2 רכיבי PCA בלבד, חלק משמעותי מהמידע הולך לאיבוד, מה שמוביל לפגיעה ביכולת לזהות דוגמאות חיוביות (class 1). שימוש ב-5 רכיבים משיג שיפור משמעותי בביצועים ומאפשר למודל לזהות יותר דוגמאות חיוביות תוך שמירה על דיוק גבוה. שימוש ב-10 רכיבים משמר את כל השונות, אך לא תמיד מוביל לשיפור ניכר לעומת 5 רכיבים. לכן, ההמלצה היא להשתמש ב-5 עד 10 רכיבים, שכן טווח זה מספק איזון טוב בין פשטות המודל לשמירה על ביצועים גבוהים.

דיון ומסקנות

במהלך המחקר נבחנו מודלים שונים לחיזוי הישרדות נוסעי הטיטניק, תוך יישום שיטות עיבוד נתונים, איזון מחלקות, הפחתת ממדים וכיוונון היפר-פרמטרים. ראשית, נעשה שימוש ברגרסיה לוגיסטית כבסיס להשוואה, שהשיגה דיוק של 82% בבדיקות. המודל הציג יכולת זיהוי טובה של נוסעים שלא שרדו (precision של 85% ו-recall של 84%), אך התקשה מעט בזיהוי הניצולים (precision של 77% ו-recall של 78%).

בהמשך, פותחה רשת נוירונים מלאכותית (ANN) עם שתי שכבות חבויות, פונקציות אקטיבציה מסוג ReLU, ושימוש בעצירה מוקדמת (Early Stopping) למניעת התאמת יתר. המודל השיג דיוק של 80%, מעט

נמוך מהרגרסיה הלוגיסטית, אך ציון ה-AUC-ROC עלה ל-0.81, מה שמעיד על יכולת הבחנה דומה בין הקבוצות.

כיוונון היפר-פרמטרים הראה כי העלאת מספר הנוירונים לשכבות חבויות שיפרה את דיוק המודל, כאשר 64 נוירונים בשכבה השיגו את הביצועים הטובים ביותר עם דיוק של 77%. העלאת מספר האפוקים ל-50 או 100 לא שיפרה משמעותית את הביצועים בהשוואה ל-20 אפוקים. קצב למידה של 0.001 נמצא כמאוזן ביותר, בעוד ששיעורי למידה נמוכים או גבוהים יותר לא השפיעו בצורה משמעותית.

ניסוי עם איזון הנתונים באמצעות SMOTE שיפר את דיוק המודל ל-82%, אך שיפור זה היה מינורי בהשוואה למודל המקורי. לעומת זאת, הוספת רעש סינתטי לנתונים פגעה בביצועים, והורידה את הדיוק ל-72%, מה שמעיד על רגישות המודל לאיכות הנתונים.

הפחתת ממדים באמצעות PCA הראתה כי שימוש ב-2 רכיבים בלבד פגע משמעותית בביצועים (דיוק של 75.98% ו-recall של 57% לניצולים), אך שימוש ב-5 רכיבים העלה את הדיוק ל-81.56% עם recall של 69%, איזון טוב יותר בין ביצועים לפשטות. שימוש ב-10 רכיבים שימר את כל השונות והביא לתוצאה כמעט זהה למודל המקורי, עם דיוק של 83.24%. מכאן, ניתן להסיק כי שימוש ב-5 רכיבי PCA מספק את האיזון הטוב ביותר בין ביצועים להקטנת המורכבות החישובית.

בהתבסס על הממצאים, המודל שהשיג את התוצאות הטובות ביותר היה רשת נוירונים עם 64 נוירונים בשכבות החבויות, קצב למידה של 0.001, איזון באמצעות SMOTE ושימוש ב-5 רכיבי PCA, שהשיגה דיוק של 81.56% עם יכולת טובה לזהות גם את מחלקה 1 (recall של 69%).

עבודות עתידיות

בהתבסס על ממצאי המחקר, ניתן להציע מספר כיוונים אפשריים לשיפור ושדרוג המודל בעתיד. ראשית, ניתן להרחיב את סט התכונות (Feature Engineering) ולבחון הוספת משתנים נוספים שעשויים להשפיע על הישרדות הנוסעים, כמו תנאי מזג האוויר ביום ההפלגה, גודל הספינה או מידע נוסף על רקע הנוסעים. כמו כן, ניתן ליצור משתנים חדשים המשלבים קשרים בין פרמטרים קיימים, כגון יחס מחיר הכרטיס להכנסה משוערת או ניתוח קשרים חברתיים בין נוסעים, אשר עשויים לחשוף תבניות נסתרות. ניקוי נוסף של הנתונים ובדיקת השפעת משתנים כמו מחלקת הנסיעה עשויים גם הם לסייע בשיפור המודל.

שיפור נוסף ייתכן גם על ידי שינוי ארכיטקטורת הרשת הקיימת. ניתן להעמיק את הרשת על ידי הוספת שכבות חבויות נוספות תוך שימוש בטכניקות למניעת למידת יתר כמו Dropout ונורמליזציה באצווה (Batch Normalization). כמו כן, ניתן לבחון פונקציות אקטיבציה מתקדמות יותר כמו Leaky ReLU, Parametric ReLU או Swish, אשר עשויות להתמודד טוב יותר עם בעיות גרדיאנט נמוך ולשפר את ביצועי הרשת. שינוי אפשרי נוסף הוא שימוש ברגרסיה לוגיסטית כשכבת יציאה במקום שכבת סיגמואיד, כדי לשפר את ביצועי המודל על נתונים בעלי חוסר איזון.

תחום נוסף שמומלץ לבחון הוא שיטות מתקדמות לאיזון הנתונים. במקום שימוש ב-SMOTE, ניתן לנסות טכניקות יצירת נתונים סינתטיים מתקדמות כמו רשתות גנרטיביות (GANs), שיכולות ליצור נתונים חדשים

המדמים בצורה ריאליסטית את המידע הקיים. כמו כן, ניתן לבחון שיטות מחיקת דוגמאות (Undersampling) המבוססות על קריטריונים חכמים, במקום הסרה אקראית של נתונים.

לסיכום, המחקר הנוכחי הצליח להדגים כיצד ניתן לייעל מודלים של למידת עומק עבור נתונים טבלאיים, אך קיימות אפשרויות רבות לשיפור נוסף. הרחבת המאפיינים, שימוש בטכניקות רשת מתקדמות – כל אלה עשויים להוביל לתחזיות מדויקות ויציבות יותר, ולהוות בסיס להמשך מחקר ופיתוח בתחום.