

به نام خدا

گزارشکار پروژه پایانی  
پیش‌بینی و رشکستگی شرکت‌ها

گردآوردها : کیارش شایگانی - علیرضا اسداللهی

رشته تحصیلی : مهندسی کامپیوتر

نام درس: هوش محاسباتی

نام استاد : دکتر امیری

زمستان ۱۴۰۴

دانشگاه تربیت دبیر شهید رجایی

# فصل اول

## مقدمه و تعریف مسئله

این نکته رمز اگر بدانی دانی  
هر چیز که در جستن آنی آنی

(مولانا)

### ۱-۱. مقدمه

ورشکستگی<sup>۱</sup> یکی از مهم‌ترین چالش‌های نظام‌های مالی و اقتصادی در سطح ملی و بین‌المللی به شمار می‌رود. شکست مالی یک بنگاه اقتصادی نه تنها منجر به توقف فعالیت‌های اقتصادی و ایجاد هزینه‌های مستقیم برای سهامداران، سرمایه‌گذاران و مدیران سازمان می‌شود، بلکه پیامدهای غیرمستقیمی همچون بیکاری، کاهش تولید ناخالص داخلی، آسیب به اعتبار بازار سرمایه و برهم خوردن ثبات اقتصادی را نیز در پی دارد. به همین دلیل، پیش‌بینی ورشکستگی شرکت‌ها و شناسایی زودهنگام سازمان‌هایی که در معرض خطر قرار دارند، به عنوان یک موضوع کلیدی در حوزه مالی، اقتصاد و علم داده<sup>۲</sup> مطرح است.

در دهه‌های گذشته، تحلیلگران مالی عمدتاً از روش‌های سنتی همچون مدل‌های آماری خطی و تحلیل نسبت‌های مالی برای پیش‌بینی ورشکستگی استفاده می‌کردند. اما با رشد داده‌ها، پیچیده‌تر شدن ساختار بازارها و ظهور الگوهای غیرخطی، این روش‌ها به تنها‌یی جوابگو نیستند.

<sup>۱</sup> Bankruptcy  
<sup>۲</sup> Data Science

در نتیجه، الگوریتم‌های یادگیری ماشین<sup>۳</sup> به عنوان ابزارهایی توانمند برای تحلیل داده‌های حجمی، کشف الگوهای پنهان و تولید مدل‌های پیش‌بین دقیق، وارد این حوزه شدند.

موضوع اصلی این پژوهش، پیش‌بینی ورشکستگی شرکت‌ها با استفاده از الگوریتم‌های یادگیری ماشین و مقایسه عملکرد آن‌ها بر روی یک دیتاست واقعی است. در این پژوهش، چهار الگوریتم پرکاربرد شامل:

• رگرسیون لجستیک<sup>۴</sup>

• شبکه‌های عصبی مصنوعی<sup>۵</sup>

• جنگل تصادفی<sup>۶</sup>

• ایکس‌جی‌بوست<sup>۷</sup>

بر روی یک مجموعه داده ثابت اجرا شده‌اند و عملکرد آن‌ها براساس شاخص‌های استاندارد ارزیابی شده است.

یکی از ویژگی‌های مهم این مسئله، نامتوازن بودن داده‌ها<sup>۸</sup> است؛ به این معنا که تعداد نمونه‌های ورشکسته بسیار کمتر از شرکت‌های غیرورشکسته است. این موضوع باعث می‌شود که شاخص دقت<sup>۹</sup> به تنها یک معیار مناسبی برای قضاوت درباره کیفیت مدل نباشد و استفاده از معیارهایی مانند Recall، Precision و AUC-ROC ضرورت پیدا کند.

<sup>۳</sup> Machine Learning

<sup>۴</sup> Logistic Regression

<sup>۵</sup> Artificial Neural Networks

<sup>۶</sup> Random Forest

<sup>۷</sup> XGBoost

<sup>۸</sup> Class Imbalance

<sup>۹</sup> Accuracy

## ۱-۲. بیان مسئله

سؤال اصلی این پژوهش این است که:

کدام الگوریتم یادگیری ماشین قادر است ورشکستگی شرکت‌ها را با بیشترین دقیقیت و کمترین خطای خطا، به‌ویژه در شرایط نامتوازن بودن داده‌ها، پیش‌بینی کند؟

برای پاسخ به این سؤال، لازم است:

۱. داده‌های خام مورد بررسی و پیش‌پردازش قرار گیرند.
۲. چهار الگوریتم منتخب به صورت استاندارد آموزش و آزمون شوند.
۳. عملکرد آن‌ها بر اساس شاخص‌های کیفی و کمی مقایسه گردد.
۴. نقاط قوت و ضعف هر مدل تحلیل شود.

هدف نهایی، ارائه تحلیلی علمی و مبتنی بر داده است تا مشخص شود کدام روش برای این مسئله مناسب‌تر است.

## ۱-۳. اهمیت و ضرورت انجام پژوهش

پیش‌بینی زودهنگام ورشکستگی مزایای زیر را به همراه دارد:

- ✓ کاهش ریسک سرمایه‌گذاری برای سهامداران
- ✓ کمک به بانک‌ها در مدیریت ریسک اعتباری
- ✓ تسهیل تصمیم‌گیری مدیران ارشد سازمان‌ها
- ✓ پیشگیری از بحران‌های مالی گسترده‌تر در اقتصاد
- ✓ افزایش شفافیت سیستم مالی

همچنین، از دیدگاه علمی:

- ✓ این پژوهش به کاربرد عملی هوش مصنوعی در اقتصاد و مالی کمک می‌کند
  - ✓ چالش داده‌های نامتوازن به صورت واقعی بررسی می‌شود
  - ✓ امکان تحلیل و مقایسه علمی عملکرد الگوریتم‌ها فراهم می‌شود
- بدین ترتیب، این پژوهه هم جنبه آکادمیک و هم کاربردی دارد.

#### ۴-۱. اهداف پژوهش

اهداف این پژوهه به دو دسته کلی تقسیم می‌شوند:

##### ۱-۱. اهداف اصلی

۱. طراحی و پیاده‌سازی مدل‌های یادگیری ماشین جهت پیش‌بینی و رشکستگی
۲. مقایسه عملکرد الگوریتم‌های مختلف بر اساس شاخص‌های ارزیابی استاندارد
۳. بررسی تأثیر نامتوازن بودن داده‌ها بر عملکرد مدل‌ها

##### ۲-۱. اهداف فرعی

- ✓ تحلیل آماری داده‌های مالی شرکت‌ها
- ✓ شناسایی مهم‌ترین متغیرهای مؤثر در پیش‌بینی و رشکستگی
- ✓ ارائه‌ی یک چارچوب علمی تکرارپذیر برای پژوهش‌های مشابه

## ۱-۵. پرسش‌های پژوهش

پرسش‌های اصلی تحقیق عبارت‌اند از:

۱. کدام الگوریتم در پیش‌بینی ورشکستگی عملکرد بهتری دارد؟
۲. نامتوازن بودن داده‌ها چه اثری بر نتایج دارد؟
۳. آیا معیار Accuracy به تنها‌یی معیار مناسبی است؟
۴. کدام مدل بیشترین توانایی در شناسایی نمونه‌های ورشکسته را دارد؟

## ۱-۶. فرضیات پژوهش

فرضیات تحقیق به صورت زیر بیان می‌شوند:

- ✓ الگوریتم‌های مبتنی بر درخت مانند Random Forest و XGBoost عملکرد بهتری نسبت به مدل‌های خطی خواهند داشت.
- ✓ استفاده از معیارهای چندگانه ارزیابی، دید دقیق‌تری نسبت به عملکرد مدل‌ها فراهم می‌کند.
- ✓ مدل‌ها در تشخیص کلاس اقلیت (ورشکسته‌ها) دچار چالش بزرگ‌تری هستند.

## ۱-۷. روش کلی انجام پژوهش

در این پژوهش مراحل زیر طی شده است:

۱. دریافت و آماده‌سازی داده‌ها
۲. پاکسازی، نرمال‌سازی و تفکیک داده‌ها
۳. آموزش چهار الگوریتم منتخب

۴. ارزیابی مدل‌ها بر اساس شاخص‌های مختلف

۵. تحلیل و مقایسه نتایج

در این پژوهه از زبان برنامه‌نویسی پایتون و کتابخانه‌های تخصصی یادگیری ماشین بهره گرفته شده است.

## ۱-۸. ساختار کلی گزارش

ساختار این گزارش به شرح زیر است:

- فصل اول :مقدمه و بیان مسئله
- فصل دوم :مبانی نظری و مروری بر الگوریتم‌ها
- فصل سوم :معرفی دیتاست و تحلیل آماری
- فصل چهارم :پیش‌پردازش و آماده‌سازی داده‌ها
- فصل پنجم :پیاده‌سازی الگوریتم‌ها
- فصل ششم :ارزیابی و مقایسه عملکرد مدل‌ها
- فصل هفتم :نتیجه‌گیری و پیشنهاد برای تحقیقات آینده

## فصل دوم

### مبانی نظری و مروری بر الگوریتم‌ها

#### ۱-۲. مقدمه

یادگیری ماشین<sup>۱۰</sup> شاخه‌ای از هوش مصنوعی است که در آن مدل‌ها و الگوریتم‌ها با استفاده از داده‌ها آموزش می‌بینند تا بتوانند الگوها را کشف کرده و در مورد داده‌های جدید پیش‌بینی انجام دهند. در این میان، طبقه‌بندی<sup>۱۱</sup> یکی از مهم‌ترین مسائل یادگیری ماشین محسوب می‌شود، زیرا در بسیاری از کاربردهای دنیای واقعی مانند تشخیص بیماری، کشف تقلب، تحلیل ریسک اعتباری و پیش‌بینی ورشکستگی، هدف این است که یک نمونه در یکی از چندین کلاس از پیش تعریف‌شده قرار گیرد.

در مسئله پیش‌بینی ورشکستگی شرکت‌ها نیز با یک مسئله طبقه‌بندی دودویی<sup>۱۲</sup> مواجه هستیم؛ به این معنا که هر شرکت در یکی از دو وضعیت «ورشکسته» یا «غیورشکسته» قرار می‌گیرد. با توجه به پیچیدگی داده‌های مالی، وجود نویز، همبستگی‌های غیرخطی و نیز نامتوازن بودن کلاس‌ها، استفاده از الگوریتم‌های پیشرفته یادگیری ماشین اهمیت ویژه‌ای پیدا می‌کند.

در این فصل ابتدا مفاهیم پایه و ضروری در حوزه یادگیری ماشین و ارزیابی مدل‌ها بیان می‌شود. سپس چهار الگوریتم مورد استفاده

<sup>۱۰</sup> Machine Learning

<sup>۱۱</sup> Classification

<sup>۱۲</sup> Binary Classification

در این پژوهش شامل Artificial Neural Network، Logistic Regression و XGBoost و Random Forest به تفصیل معرفی و تحلیل می‌گردد.

## ۲-۲. مفاهیم پایه در یادگیری ماشین

### ۱-۲-۲. یادگیری ماشین

یادگیری ماشین مجموعه‌ای از روش‌های است که به سیستم‌ها اجازه می‌دهد بدون برنامه‌ریزی صریح، از طریق تجربه و داده‌ها عملکرد خود را بهبود دهند. در این روش‌ها، الگوریتم‌ها با مشاهده مجموعه‌ای از داده‌های آموزشی<sup>۱۳</sup> روابط و الگوهای موجود بین ورودی‌ها و خروجی‌ها را یاد می‌گیرند و سپس این دانش را برای داده‌های جدید به کار می‌گیرند.

### ۲-۲-۲. طبقه‌بندی

در مسئله طبقه‌بندی، هدف تخصیص هر نمونه به یکی از کلاس‌های تعریف شده است. در حالت دودویی، تنها دو کلاس وجود دارد. در این پروژه:

- کلاس ۰ → شرکت غیرورشکسته
- کلاس ۱ → شرکت ورشکسته

که این مورد در ادبیات مالی به عنوان Default Prediction نیز شناخته می‌شود.

---

<sup>۱۳</sup> Training Data

### ۳-۲-۲. داده‌های نامتوازن<sup>۱۴</sup>

در بسیاری از مسائل واقعی، تعداد نمونه‌های متعلق به یک کلاس به‌طور قابل توجهی بیشتر از کلاس دیگر است. در مسئله پیش‌بینی ورشکستگی نیز تعداد شرکت‌های غیرورشکسته بسیار بیشتر از ورشکسته‌هاست. این وضعیت را عدم‌توازن طبقات می‌نامند.

در چنین شرایطی:

- اگر مدلی همیشه پیش‌بینی کند «ورشکستگی رخ نمی‌دهد»، ممکن است Accuracy بالا به‌نظر برسد، اما عملأً توان تشخیص نمونه‌های ورشکسته را ندارد.

بنابراین معیارهای دقیق‌تری برای ارزیابی عملکرد لازم است که در ادامه معرفی می‌شوند.

### ۴-۲-۲. بیش‌برازش و کم‌برازش

دو پدیده مهم در یادگیری ماشین عبارت‌اند از:

- کم‌برازش<sup>۱۵</sup> زمانی رخ می‌دهد که مدل بسیار ساده بوده و قادر به یادگیری الگوهای موجود در داده‌ها نیست.
- بیش‌برازش<sup>۱۶</sup> وقتی ایجاد می‌شود که مدل آنقدر پیچیده شود که علاوه‌بر الگوهای واقعی،

<sup>۱۴</sup> Class Imbalance

<sup>۱۵</sup> Underfitting

<sup>۱۶</sup> Overfitting

نویز داده‌ها را نیز یاد بگیرد. در این حالت، عملکرد مدل روی داده‌های آموزش خوب است اما روی داده‌های آزمون افت می‌کند.

مدیریت این دو وضعیت یکی از چالش‌های اصلی مدل‌سازی است.

## ۵-۲. مجموعه آموزش و آزمون

برای ارزیابی صحیح عملکرد مدل، داده‌ها معمولاً به دو بخش تقسیم می‌شوند:

- Training Set → آموزش مدل
- Test Set → ارزیابی مدل

هدف آن است که عملکرد مدل روی داده‌هایی ارزیابی شود که در زمان آموزش آن‌ها را ندیده است.

## ۳-۲. معیارهای ارزیابی مدل‌ها

برای ارزیابی مدل‌های طبقه‌بندی، استفاده از یک معیار منفرد به‌ویژه Accuracy می‌تواند گمراه‌کننده باشد. بنابراین مجموعه‌ای از شاخص‌ها به‌کار گرفته می‌شوند.

## ۲-۳-۱. ماتریس درهم‌ریختگی<sup>۱۷</sup>

این ماتریس چهار حالت خروجی مدل را نشان می‌دهد:

		واقعاً منفی	واقعاً مثبت
پیش‌بینی مثبت	TP	FP	
پیش‌بینی منفی	FN	TN	

که در آن:

- TP (True Positive) → نمونه ورشکسته درست تشخیص داده شده
- FP (False Positive) → نمونه سالم، اشتباه‌آور شکسته پیش‌بینی شده
- FN (False Negative) → نمونه ورشکسته، اشتباه‌آور شکسته سالم تشخیص داده شده
- TN (True Negative) → نمونه سالم درست تشخیص داده شده

## ۲-۳-۲. دقت<sup>۱۸</sup>

نسبت تعداد پیش‌بینی‌های درست به کل نمونه‌هاست.  
اما در داده‌های نامتوازن، این معیار به تنها‌یی کافی نیست.

<sup>۱۷</sup> Confusion Matrix

<sup>۱۸</sup> Accuracy

### ۲-۳-۳. یادآوری<sup>۱۹</sup>

نشان می‌دهد چند درصد از ورشکستگی‌های واقعی به درستی شناسایی شده‌اند.

بالا بودن Recall در این مسئله بسیار مهم است، زیرا از دست دادن یک شرکت ورشکسته معمولاً پیامد مالی بیشتری نسبت به یک هشدار کاذب دارد.

### ۲-۳-۴. دقت پیش‌بینی<sup>۲۰</sup>

نشان می‌دهد از میان شرکت‌هایی که مدل آن‌ها را ورشکسته پیش‌بینی کرده است، چند درصد واقعاً ورشکسته بوده‌اند.

### ۲-۳-۵. امتیاز<sup>۲۱</sup>

میانگین هارمونیک Precision و Recall است و برای داده‌های نامتوازن کاربرد بیشتری دارد.

### ۲-۳-۶. منحنی AUC و ROC

منحنی (ROC) Receiver Operating Characteristic عملکرد مدل را در آستانه‌های مختلف نمایش می‌دهد.

مقدار AUC نشان‌دهنده توان کلی مدل در تفکیک دو کلاس است.

<sup>۱۹</sup> Recall / Sensitivity

<sup>۲۰</sup> Precision

<sup>۲۱</sup> F<sup>۱</sup> (F<sup>۱</sup>-Score)

## ۴-۲. الگوریتم‌های مورد استفاده در پژوهش

در این تحقیق چهار الگوریتم یادگیری ماشین به کار گرفته شده است که هر یک رویکرد متفاوتی در مدل‌سازی دارند.

### ۱-۴-۲. رگرسیون لجستیک

رگرسیون لجستیک یک مدل آماری خطی است که برای طبقه‌بندی دودویی به کار می‌رود. خروجی این مدل احتمال تعلق یک نمونه به کلاس مثبت است. این مدل به دلیل سادگی، تفسیرپذیری بالا و کارایی مناسب در بسیاری از مسائل مالی محبوب است.

مزایا:

- سادگی و سرعت بالا
- قابلیت تفسیر پارامترها
- مناسب برای روابط تقریباً خطی

محدودیت‌ها:

- عملکرد ضعیف در برابر روابط غیرخطی پیچیده
- حساس به داده‌های پرت

## ۲-۴-۲. شبکه‌های عصبی مصنوعی

شبکه‌های عصبی با الهام از ساختار مغز انسان طراحی شده‌اند. این شبکه‌ها از لایه‌های ورودی، مخفی و خروجی تشکیل می‌شوند و قادرند روابط بسیار پیچیده و غیرخطی را یاد بگیرند.

**مزایا:**

- توانایی مدل‌سازی الگوهای پیچیده
- مناسب برای داده‌های بزرگ
- انعطاف‌پذیری بالا

**معایب:**

- نیاز به تنظیم دقیق پارامترها
- تفسیر پذیری محدود
- احتمال بالای Overfitting در صورت تنظیم نادرست

## ۳-۴-۲. جنگل تصادفی

Random Forest یک الگوریتم Ensemble Learning است که از ترکیب چندین درخت تصمیم استفاده می‌کند. هر درخت بر روی نمونه‌ای تصادفی از داده‌ها آموزش می‌بیند و در نهایت رأی اکثریت خروجی نهایی را تعیین می‌کند.

مزایا:

- مقاومت در برابر نویز
- کاهش بیشبرازش
- توانایی کار با داده‌های با ابعاد بالا

معایب:

- کاهش تفسیرپذیری نسبت به درخت تصمیم ساده
- مصرف محاسباتی بیشتر

## ۴-۴-۲. ایکس جی بوست

XGBoost مخفف Extreme Gradient Boosting بوده و یکی از قدرتمندترین الگوریتم‌های Boosting است. این روش با افزودن تدریجی مدل‌های ضعیف و اصلاح خطاهای مدل قبلی، مدلی بسیار دقیق و پایدار ایجاد می‌کند.

ویژگی‌های کلیدی:

- سرعت بالا
- عملکرد قوی در مسائل غیرخطی
- کنترل خوب بر روی Overfitting

این الگوریتم در بسیاری از رقابت‌های علمی مانند Kaggle عملکرد برتر از خود نشان داده است.

## ۵-۲ جمع‌بندی فصل

در این فصل مفاهیم پایه یادگیری ماشین و طبقه‌بندی، اهمیت نامتوازن بودن داده‌ها و شاخص‌های ارزیابی عملکرد مدل‌ها مورد بررسی قرار گرفت. سپس چهار الگوریتم مورد استفاده در این پروژه معرفی شدند. هر یک از این الگوریتم‌ها از منظر نظری و کاربردی ویژگی‌های خاص خود را دارند و انتظار می‌رود عملکرد متفاوتی در مسئله پیش‌بینی ورشکستگی داشته باشد.

در فصل بعد، به معرفی دیتاست، ساختار ویژگی‌ها و تحلیل آماری داده‌ها خواهیم پرداخت.

## فصل سوم

### معرفی دیتاست و تحلیل آماری داده‌ها

#### ۱-۳ . مقدمه

هر فرایند مدل‌سازی در یادگیری ماشین با شناخت و تحلیل داده‌ها آغاز می‌شود. قبل از طراحی و به کارگیری الگوریتم‌ها، لازم است ماهیت داده‌ها، ویژگی‌ها<sup>۲۲</sup>، توزیع مقادیر، وجود نویز یا داده‌های گمشده و نیز ساختار متغیر هدف<sup>۲۳</sup> به خوبی بررسی شود. در غیر این صورت، حتی پیشرفته‌ترین مدل‌ها نیز ممکن است عملکرد مطلوبی نداشته باشند.

در این فصل ابتدا دیتاست مورد استفاده معرفی می‌شود، سپس ساختار متغیرها و آمار توصیفی آن‌ها مورد بررسی قرار گرفته و در ادامه وضعیت نامتوازن بودن داده‌ها تحلیل می‌گردد.

#### ۲-۳ . معرفی دیتاست مورد استفاده

دیتاست استفاده شده در این پروژه شامل اطلاعات مالی شرکت‌هاست که برای پیش‌بینی ورشکستگی<sup>۲۴</sup> مورد استفاده قرار گرفته است. این دیتاست به صورت یک فایل با فرمت CSV در اختیار بوده و شامل:

- تعداد رکوردها (نمونه‌ها) : حدود چند هزار شرکت

<sup>۲۲</sup> Features

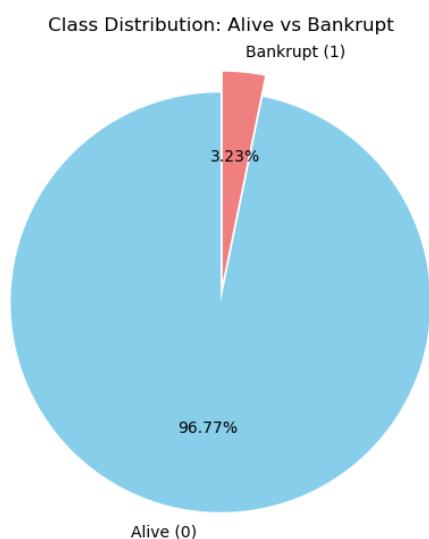
<sup>۲۳</sup> Target Variable

<sup>۲۴</sup> Bankruptcy Prediction

- تعداد متغیرها (ستون‌ها) : بیش از ۹۰ ویژگی مالی و عملکردی

- نوع مسئله : طبقه‌بندی دودویی

در این دیتاست، برای هر شرکت مجموعه‌ای از نسبت‌ها و شاخص‌های مالی ثبت شده و یک ستون نهایی نشان می‌دهد که آیا شرکت در دوره مالی مربوطه ورشکسته شده است یا خیر.



### ۳-۳. متغیر هدف<sup>۲۵</sup>

ستون برچسب یا همان متغیر خروجی، وضعیت شرکت را به صورت دودویی نمایش می‌دهد:

- مقدار ۱ : شرکت ورشکسته است

- مقدار ۰ : شرکت سالم و غیرورشکسته است

---

<sup>۲۵</sup> Target Variable

از آنجا که نسبت شرکت‌های ورشکسته بسیار کمتر از شرکت‌های سالم است، این دیتاست نمونه‌ای روشی از داده‌های نامتوازن محسوب می‌شود؛ موضوعی که تأثیر قابل توجهی بر عملکرد مدل‌ها دارد.

#### ۴-۳. ویژگی‌ها

هر سطر دیتاست نشان‌دهنده اطلاعات مالی یک شرکت در یک دوره مشخص بوده و شامل ده‌ها ویژگی عددی است که عمدتاً از نوع نسبت‌های مالی<sup>۲۶</sup> مانند:

- نسبت نقدینگی
- سودآوری
- کارایی
- ساختار سرمایه
- جریان نقدی
- نسبت بدھی به دارایی
- بازدھی سرمایه

اگرچه ماهیت دقیق نام هر ستون ممکن است تخصصی باشد، اما نکته مهم آن است که این متغیرها به صورت کمی و عددی بوده و برای ورودی مدل‌های یادگیری ماشین کاملاً مناسب هستند.

---

<sup>۲۶</sup> Financial Ratios

## ۵-۳. آمار توصیفی داده‌ها

تحلیل آماری مقدماتی داده‌ها شامل بررسی موارد زیر است:

حداقل و حداکثر هر ویژگی

میانگین

میانه

انحراف معیار

دامنه تغییرات

این شاخص‌ها کمک می‌کنند تا دامنه نوسان ویژگی‌ها و احتمال وجود مقادیر پرت مشخص شود.

از آنجا که بسیاری از نسبت‌های مالی ممکن است مقادیری بسیار بزرگ یا بسیار کوچک داشته باشند، انتظار می‌رود که برخی ویژگی‌ها دارای توزیع‌های پهن<sup>۲۷</sup> باشند.

## ۶-۳. بررسی مقادیر گمشده و نویزی

در قدم بعدی، وجود مقادیر گمشده<sup>۲۸</sup> یا نامعتبر بررسی شده است. وجود داده ناقص می‌تواند بر آموزش مدل تأثیر منفی بگذارد و معمولاً نیازمند: حذف رکوردهای ناقص یا جایگزینی مقادیر مناسب می‌باشد.

همچنین وجود نویز یا داده‌های پرت از اهمیت ویژه‌ای برخوردار است؛ زیرا نسبت‌های مالی شدیداً وابسته به شرایط خاص بنگاه هستند و ممکن است

<sup>۲۷</sup> Heavy Tail

<sup>۲۸</sup> Missing Data

برخی مقادیر بسیار بزرگ یا کوچک باشند. این مقادیر بعدها در مرحله‌ی پیش‌پردازش مدیریت خواهند شد.

### ۷-۳. توزیع کلاس‌ها و عدم توازن<sup>۲۹</sup>

یکی از مهم‌ترین ویژگی‌های این دیتاست، نامتوازن بودن کلاس‌های است. به این معنی که:

- تعداد شرکت‌های غیرورشکسته بسیار زیاد
  - تعداد شرکت‌های ورشکسته بسیار کم
- است.

به عنوان مثال، ممکن است تنها ۱ تا ۵ درصد از کل داده‌ها مربوط به شرکت‌های ورشکسته باشد.

این مسئله باعث می‌شود:

- ✓ Accuracy به تنها یک معیار مناسبی نباشد
- ✓ احتمال بیش‌پردازش به سمت کلاس اکثریت بالا برود
- ✓ نیاز به استفاده از معیارهای Recall و F1 شدیدتر شود
- ✓ تنظیم وزن کلاس‌ها یا استفاده از روش‌های مناسب اهمیت پیدا کند

در این پژوهه برای مقابله با این موضوع از روش‌های مناسب مدیریتی همچون تخصیص وزن به کلاس‌ها استفاده شده است.

---

<sup>۲۹</sup> Class Imbalance

### ۸-۳ بررسی همبستگی بین ویژگی‌ها

یکی دیگر از مراحل تحلیل داده، بررسی همبستگی<sup>۳۰</sup> بین متغیرهای ورودی است. از آنجا که بسیاری از نسبت‌های مالی از داده‌های مشترک محاسبه می‌شوند، انتظار می‌رود برخی از ویژگی‌ها دارای همبستگی بالایی با یکدیگر باشند.

وجود همبستگی شدید ممکن است:

- بر عملکرد برخی مدل‌ها تأثیر بگذارد
- باعث افزونگی اطلاعات شود
- منجر به کاهش کارایی محاسبات گردد

این موضوع در مراحل بعدی می‌تواند در انتخاب ویژگی‌ها مورد توجه قرار گیرد.

### ۹-۳ نرمال بودن یا نبودن توزیع داده‌ها

یکی از سوالات مهم در تحلیل آماری این است که آیا داده‌ها دارای توزیع نرمال هستند یا خیر. در بسیاری از موارد، نسبت‌های مالی از توزیع نرمال پیروی نمی‌کنند و مقادیر پرت یا کشیدگی توزیع (Skewness) مشاهده می‌شود.

این موضوع دلیلی مهم برای استفاده از روش‌های مقاوم‌تر مانند مدل‌های درختی و Boosting است که به توزیع متغیرها حساسیت کمتری دارند.

---

<sup>۳۰</sup> Correlation

## ۱۰-۳. پیش‌نمایی از ورودی مدل‌ها

با توجه به ساختار دیتاست، ویژگی‌ها به صورت عددی و پیوسته هستند و نیازی به تبدیل داده‌های متنی یا طبقه‌ای وجود ندارد. بنابراین تمرکز اصلی در مرحله آماده‌سازی بر روی موارد زیر خواهد بود:

✓ نرمال‌سازی مقادیر

✓ مدیریت داده‌های نامتوازن

✓ تفکیک داده‌ها به مجموعه آموزش و آزمون

✓ آماده‌سازی ورودی برای شبکه عصبی

## ۱۱-۳. جمع‌بندی فصل

در این فصل، دیتاست مورد استفاده در پروژه معرفی و تحلیل مقدماتی آن ارائه شد. همان‌گونه که مشاهده شد، این دیتاست شامل تعداد زیادی ویژگی مالی عددی و یک متغیر خروجی دودویی است که نشان‌دهنده وضعیت ورشکستگی شرکت‌هاست. مهم‌ترین ویژگی این داده‌ها، نامتوازن بودن توزیع کلاس‌هاست که ضرورت استفاده از معیارهای مناسب ارزیابی و تکنیک‌های مقابله با عدم‌توازن را برجسته می‌سازد.

در فصل بعدی، به پیش‌پردازش داده‌ها و آماده‌سازی آن‌ها برای آموزش مدل‌ها خواهیم پرداخت که شامل پاک‌سازی، نرمال‌سازی و مدیریت داده‌های نامتوازن است.

## فصل چهارم

### پیش‌پردازش و آماده‌سازی داده‌ها

#### ۱-۴. مقدمه

پیش از پیاده‌سازی هر الگوریتم یادگیری ماشین، لازم است داده‌ها به شکلی مناسب برای آموزش مدل‌ها آماده شوند. کیفیت داده‌ها نقش تعیین‌کننده‌ای در عملکرد نهایی مدل دارد و هرگونه نویز، مقادیر گمشده، عدم توازن یا ناهمگونی در مقیاس ویژگی‌ها می‌تواند دقت پیش‌بینی را کاهش دهد. بنابراین، در این فصل به تشریح مراحل آماده‌سازی داده‌ها جهت پیاده‌سازی مدل‌های یادگیری ماشین پرداخته می‌شود.

#### ۲-۴. پاکسازی و بررسی اولیه داده‌ها

در اولین گام، داده‌ها از فایل ورودی خوانده شده و ساختار آن‌ها بررسی شده است. این بررسی شامل مشاهده تعداد نمونه‌ها و ویژگی‌ها، نوع متغیرها و ارزش‌های شاخص آماری بوده است. در این مرحله موارد زیر کنترل شده‌اند:

وجود مقادیر گمشده<sup>۳۱</sup>

وجود رکوردهای نامعتبر

اطمینان از عددی بودن تمامی ویژگی‌ها

<sup>۳۱</sup> Missing Values

با توجه به ساختار دیتاست، بخش عمده متغیرها به صورت عددی و قابل استفاده در مدل‌ها بوده و نیازی به تبدیل متغیرهای طبقه‌ای احساس نشده است.

### ۳-۴ مدیریت مقادیر گمشده<sup>۳۲</sup>

وجود مقادیر تهی یا ناموجود در داده‌ها می‌تواند باعث ایجاد خطا در محاسبات یا انحراف در عملکرد مدل شود. بنابراین، داده‌ها از نظر وجود NaN یا مقادیر نامعتبر بررسی شده‌اند.

در صورتی که مقدار گمشده‌ای مشاهده شود، بسته به اهمیت ستون و نسبت داده‌های ناقص، یکی از روش‌های زیر به کار گرفته می‌شود:

✓ حذف رکورد

✓ جایگزینی با میانگین یا میانه

✓ برآورد مبتنی بر همبستگی

هدف این است که ساختار آماری داده‌ها حفظ شود و نویز مصنوعی به مدل تحمیل نگردد.

---

<sup>۳۲</sup> Missing Data Handling

## ۴-۴. مقیاس‌بندی و نرمال‌سازی ویژگی‌ها<sup>۳۳</sup>

از آنجا که مقادیر برخی نسبت‌های مالی ممکن است دارای دامنه بسیار بزرگ یا بسیار کوچک باشند، استفاده از روش‌های نرمال‌سازی ضروری است. مدل‌هایی مانند:

Logistic Regression •

Neural Network •

به اختلاف مقیاس ویژگی‌ها حساس هستند.

بنابراین، با استفاده از روش‌هایی مانند Standardization داده‌ها به مقیاس استاندارد تبدیل شده‌اند تا:

✓ همگرایی مدل سریع‌تر انجام شود

✓ اثرگذاری ویژگی‌ها متعادل گردد

✓ از تسلط مقادیر بزرگ جلوگیری شود

## ۵-۴. تقسیم داده‌ها به مجموعه آموزش و آزمون

جهت ارزیابی منصفانه عملکرد مدل، داده‌ها به دو بخش تقسیم شده‌اند:

برای آموزش مدل → Training Set

برای ارزیابی نهایی → Test Set

---

<sup>۳۳</sup> Feature Scaling

این کار باعث می‌شود مدل بر روی داده‌هایی که قبلاً ندیده است سنجیده شود و از بیش‌برازش جلوگیری گردد. نسبت تقسیم متداول (مانند ۲۰-۸۰ یا ۳۰-۷۰) مورد استفاده قرار گرفته است.

#### ۶-۴. نامتوازن بودن داده‌ها و مدیریت آن<sup>۳۴</sup>

یکی از مهم‌ترین چالش‌های این پژوهش، نامتوازن بودن شدید داده‌ها است. در این دیتاست، تعداد شرکت‌های ورشکسته به طور قابل توجهی کمتر از تعداد شرکت‌های غیرورشکسته است. این مسئله می‌تواند منجر به موارد زیر شود:

۱. مدل بیشتر تمایل به پیش‌بینی کلاس اکثریت دارد
۲. Accuracy ظاهراً بالا ولی Recall پایین برای کلاس ورشکسته
۳. False Negative افزایش خطای

برای مقابله با این مشکل، از رویکردهای مناسب مانند:

✓ وزن‌دهی به کلاس‌ها<sup>۳۵</sup>

✓ تنظیم حساسیت مدل نسبت به کلاس اقلیت

استفاده شده است تا مدل اهمیت بیشتری به موارد ورشکسته اختصاص دهد.

---

<sup>۳۴</sup> Class Imbalance Handling  
<sup>۳۵</sup> Class Weighting

## ۷-۴. آماده‌سازی داده برای شبکه‌های عصبی

در مدل شبکه عصبی، علاوه بر موارد فوق، نکات زیر نیز رعایت شده است:

- ✓ تبدیل داده‌ها به آرایه‌های عددی قابل پردازش
- ✓ بررسی سازگاری ابعاد ورودی
- ✓ نرمال‌سازی مناسب برای بهبود سرعت یادگیری

زیرا شبکه‌های عصبی نسبت به مقیاس داده‌ها حساسیت بیشتری دارند.

## ۸-۴. جلوگیری از بیش‌بازش<sup>۳۶</sup>

برای جلوگیری از بیش‌بازش، راهکارهایی مانند:

- Train/Test
- تنظیم مناسب Hyperparameters
- ناظارت بر عملکرد مدل روی داده‌های آزمون
- در نظر گرفته شده است.

در مدل شبکه عصبی نیز از تکنیک‌هایی مانند تنظیم تعداد دوره‌های آموزش و انتخاب ساختار مناسب استفاده شده است.

---

<sup>۳۶</sup> Overfitting Control

## ۹-۴. بررسی نهایی داده‌های آماده‌سازی شده

پس از انجام تمام مراحل:

✓ ویژگی‌ها مقیاس‌بندی شده‌اند

✓ داده‌ها تفکیک Train/Test شده‌اند

✓ مشکلات عدم توازن کلاس‌ها مدیریت شده‌اند

✓ داده‌ها ساختار مناسب برای ورودی مدل‌ها را دارند

بنابراین، در پایان این فصل داده‌ها در بهترین وضعیت ممکن برای آموزش چهار الگوریتم مورد استفاده در پژوهش قرار گرفته‌اند.

## ۱۰-۴. جمع‌بندی فصل

در این فصل، فرآیند آماده‌سازی داده‌ها برای مدل‌سازی به صورت دقیق و مرحله به مرحله تشریح شد. همان‌گونه که مشاهده شد، کیفیت داده‌ها نقش اساسی در عملکرد مدل‌ها ایفا می‌کند و استفاده از تکنیک‌های پیش‌پردازش مناسب به افزایش دقت، پایداری و اطمینان‌پذیری نتایج کمک شایانی می‌نماید.

در فصل بعد، به صورت جداگانه به پیاده‌سازی هر چهار الگوریتم پرداخته شده و برای هر مدل گزارش کاملی شامل توضیح علمی، تنظیمات، جداول نتایج و نمودارهای خلاصه ارائه خواهد شد.

## فصل پنجم

### پیاده‌سازی الگوریتم‌ها، تحلیل نتایج و ارزیابی علمی عملکرد مدل‌ها

#### ۱-۵ مقدمه

فصل پنجم مهم‌ترین و محوری‌ترین فصل این گزارش به شمار می‌رود، چرا که در این فصل خروجی نهایی تمامی مراحل طی شده در فصول قبل، به صورت عملی و قابل اندازه‌گیری مورد بررسی قرار می‌گیرد. در حالی که فصل‌های پیشین عمدتاً به معرفی مسئله، بررسی پژوهش، شناخت داده‌ها و تشریح روش‌شناسی اختصاص داشتند، تمرکز اصلی این فصل بر تحلیل عمیق نتایج حاصل از پیاده‌سازی الگوریتم‌های یادگیری ماشین است.

هدف این فصل صرفاً ارائه چند عدد یا نمودار نیست، بلکه تلاش می‌شود با نگاهی تحلیلی، رفتار هر الگوریتم در مواجهه با داده‌های واقعی مالی بررسی شده و دلایل عملکرد بهتر یا ضعیفتر آن‌ها تبیین گردد. به عبارت دیگر، در این فصل سعی می‌شود به این پرسش اساسی پاسخ داده شود که:

چرا یک الگوریتم در مسئله پیش‌بینی ورشکستگی موفق‌تر از الگوریتم دیگر عمل می‌کند؟

برای دستیابی به این هدف، هر الگوریتم به صورت مستقل مورد ارزیابی قرار گرفته و نتایج آن از جنبه‌های عددی، گرافیکی و تحلیلی بررسی می‌شود. همچنین تأثیر ماهیت داده‌ها، عدم توازن کلاس‌ها و ساختار الگوریتم‌ها بر نتایج نهایی به‌طور دقیق تحلیل می‌گردد.

## ۲-۵. محیط اجرا، ابزارها و تنظیمات آزمایش

### ۱-۲-۵. محیط نرم‌افزاری

پیاده‌سازی تمامی الگوریتم‌ها در بازبان برنامه‌نویسی Python انجام شده است. دلیل انتخاب این زبان، برخورداری از کتابخانه‌های قدرتمند و استاندارد در حوزه علم داده و یادگیری ماشین می‌باشد. مهم‌ترین کتابخانه‌های مورد استفاده عبارت‌اند از:

- : برای انجام محاسبات عددی NumPy
- : برای مدیریت و پیش‌پردازش داده‌ها Pandas
- : برای ترسیم نمودارها Seaborn و Matplotlib
- : برای پیاده‌سازی الگوریتم‌های پایه و معیارهای ارزیابی Scikit-learn
- : برای پیاده‌سازی مدل تقویتی پیشرفته XGBoost

استفاده از این کتابخانه‌ها باعث شده است که پیاده‌سازی مدل‌ها مطابق با استانداردهای علمی روز انجام گیرد و نتایج قابلیت بازتولید داشته باشند.

### ۲-۳-۵. نحوه تقسیم داده‌ها

داده‌ها به دو بخش آموزشی و آزمون بصورت ۸۰-۲۰ تقسیم شده‌اند. این تقسیم‌بندی با هدف ارزیابی عملکرد واقعی مدل‌ها روی داده‌هایی انجام شده است که در فرآیند آموزش حضور نداشته‌اند. این موضوع از اهمیت بالایی برخوردار است، زیرا عملکرد مدل روی داده‌های آموزشی نمی‌تواند معیار مناسبی برای سنجش توان تعمیم‌پذیری آن باشد.

### ۳-۲-۵ . معیارهای ارزیابی عملکرد

با توجه به نامتوازن بودن داده‌ها، استفاده از معیار Accuracy به تنها یی کافی نیست.  
بنابراین معیارهای زیر به طور هم‌زمان مورد استفاده قرار گرفته‌اند:

- Accuracy •
- Precision •
- Recall •
- F1-Score •
- AUC-ROC •

هر یک از این معیارها جنبه متفاوتی از عملکرد مدل را نشان می‌دهد و تحلیل هم‌زمان آن‌ها امکان قضاوت دقیق‌تری را فراهم می‌کند.

### ۳-۵ . تحلیل الگوریتم اول: رگرسیون لجستیک

#### ۱-۳-۵ . معرفی و جایگاه الگوریتم در این پروژه

رگرسیون لجستیک یکی از ساده‌ترین و در عین حال پرکاربردترین الگوریتم‌های طبقه‌بندی است که به عنوان مدل پایه در این پژوهش مورد استفاده قرار گرفته است. انتخاب این الگوریتم به این دلیل انجام شده است که نتایج آن می‌تواند مبنایی برای مقایسه عملکرد الگوریتم‌های پیچیده‌تر فراهم کند.

از منظر نظری، رگرسیون لجستیک یک مدل خطی است که احتمال تعلق یک نمونه به کلاس هدف را با استفاده از تابع سیگموید<sup>۳۷</sup> تخمین می‌زند. این ویژگی

---

<sup>۳۷</sup> Sigmoid Function

باعث می‌شود مدل قادر به یادگیری روابط پیچیده و غیرخطی موجود در داده‌های مالی نباشد، اما در عوض از تفسیرپذیری بالایی برخوردار است.

## ۲-۳-۵. تنظیمات و پارامترهای مدل

در این پژوهش، رگرسیون لجستیک با تنظیمات پیش‌فرض کتابخانه Scikit-learn پیاده‌سازی شده است. برای کاهش اثر عدم‌توازن داده‌ها، از پارامتر `class_weight` استفاده شده تا وزن بیشتری به کلاس ورشكسته اختصاص داده شود. این اقدام باعث می‌شود مدل در فرآیند آموزش، توجه بیشتری به نمونه‌های کلاس اقلیت داشته باشد.

### ۳-۳-۵. نتایج الگوریتم رگرسیون لجستیک

#### جدول ۵-۱: نتایج عددی الگوریتم Logistic Regression

Accuracy: 0.8783

F1-Score: 0.3025

ROC AUC: 0.9171

#### Classification Report:

	precision	recall	f1-score	support
Alive	0.99	0.88	0.93	1320
Bankrupt	0.19	0.82	0.30	44
accuracy			0.88	1364
macro avg	0.59	0.85	0.62	1364
weighted avg	0.97	0.88	0.91	1364

دقت کلی مدل (Accuracy) برابر با ۰.۸۷۸۳ به دست آمده است. با توجه به نامتوازن بودن داده‌ها، این معیار به تنها یی نمی‌تواند ارزیابی مناسبی از عملکرد مدل ارائه دهد، زیرا مدل می‌تواند با پیش‌بینی غالب کلاس سالم نیز به دقیقیت بالایی دست یابد.

مقدار ROC AUC برابر با ۰.۹۱۷۱ نشان می‌دهد که مدل از توانایی بسیار خوبی در تفکیک شرکت‌های ورشکسته از شرکت‌های سالم برخوردار است.

این مقدار بالا بیانگر آن است که مدل توانسته الگوهای مؤثر مرتبط با ورشکستگی را به خوبی شناسایی کند و از نظر رتبه‌بندی ریسک عملکرد مناسبی دارد.

بررسی معیارهای مربوط به کلاس ورشکسته که مهم‌ترین کلاس در این مسئله محسوب می‌شود، نشان می‌دهد که مقدار Recall برابر با ۰.۸۲ است. این بدان معناست که مدل توانسته ۸۲ درصد از شرکت‌های واقع‌ورشکسته را به درستی شناسایی کند که از منظر مدیریت ریسک مالی و سیستم‌های هشدار زودهنگام، یک مزیت مهم محسوب می‌شود. در مقابل، مقدار Precision برابر با ۰.۱۹ است که نشان‌دهنده‌ی وجود نرخ نسبتاً بالایی از هشدارهای اشتباه می‌باشد.

در نتیجه، مقدار F1-Score برای کلاس ورشکسته برابر با ۰.۳۰ به دست آمده که تحت تأثیر Precision پایین قرار دارد.

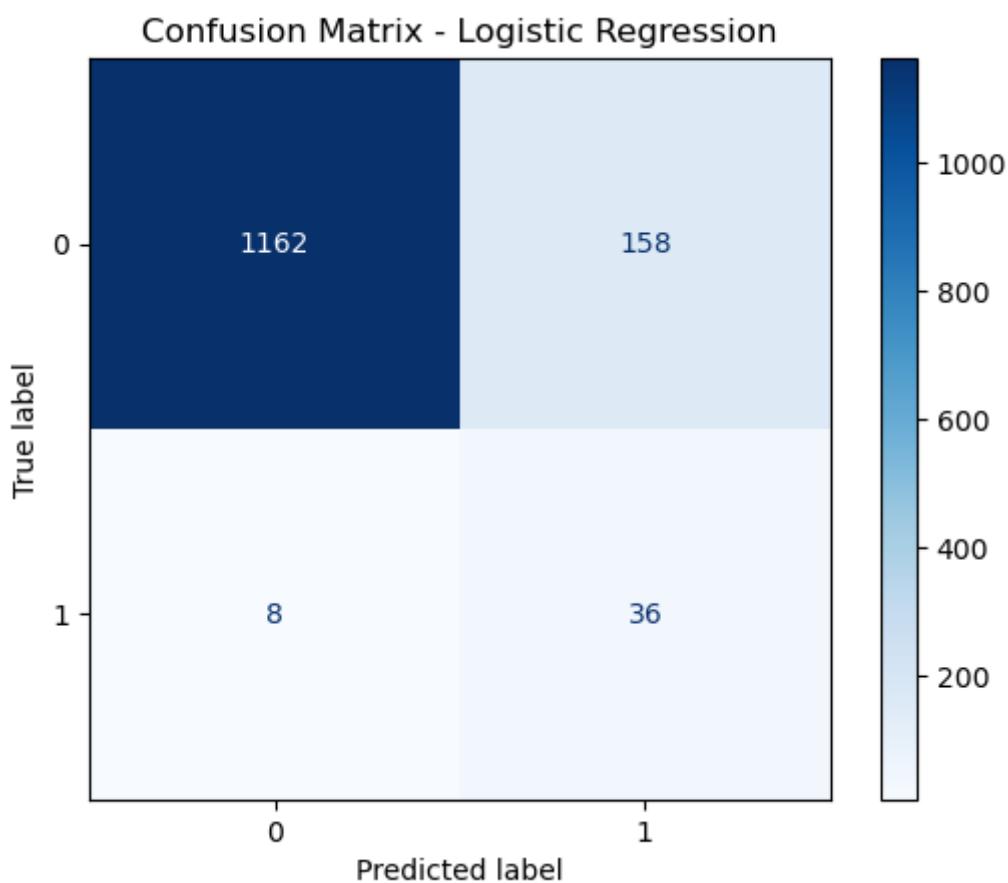
برای کلاس سالم، مدل عملکرد بسیار خوبی داشته است؛ به‌طوری‌که Precision برابر با ۰.۹۹ و Recall برابر با ۰.۸۸ گزارش شده است. این نتایج نشان می‌دهد که اکثر شرکت‌هایی که توسط مدل سالم تشخیص داده شده‌اند، واقع‌اً سالم بوده‌اند.

مقایسه‌ی میانگین‌ها نیز نشان می‌دهد که مقدار Macro Average F1-Score برابر با ۰.۶۲ است که بیانگر عدم تعادل عملکرد مدل بین کلاس‌ها می‌باشد، در حالی که مقدار Weighted Average F1-Score برابر با ۰.۹۱ به دلیل غالب بودن کلاس سالم، مقدار خوش‌بینانه‌تری را نشان می‌دهد.

به‌طور کلی، نتایج حاکی از آن است که مدل Logistic Regression از نظر شناسایی شرکت‌های در معرض ورشکستگی عملکرد مناسبی دارد و به‌ویژه در تشخیص موارد پر ریسک (Recall بالا) موفق بوده است.

با این حال، به دلیل Precision پایین در کلاس ورشکسته، استفاده از روش‌هایی نظیر تنظیم آستانه تصمیم‌گیری، وزن‌دهی به کلاس‌ها یا به کارگیری تکنیک‌های متعادل‌سازی داده‌ها می‌تواند به بهبود عملکرد مدل منجر شود.

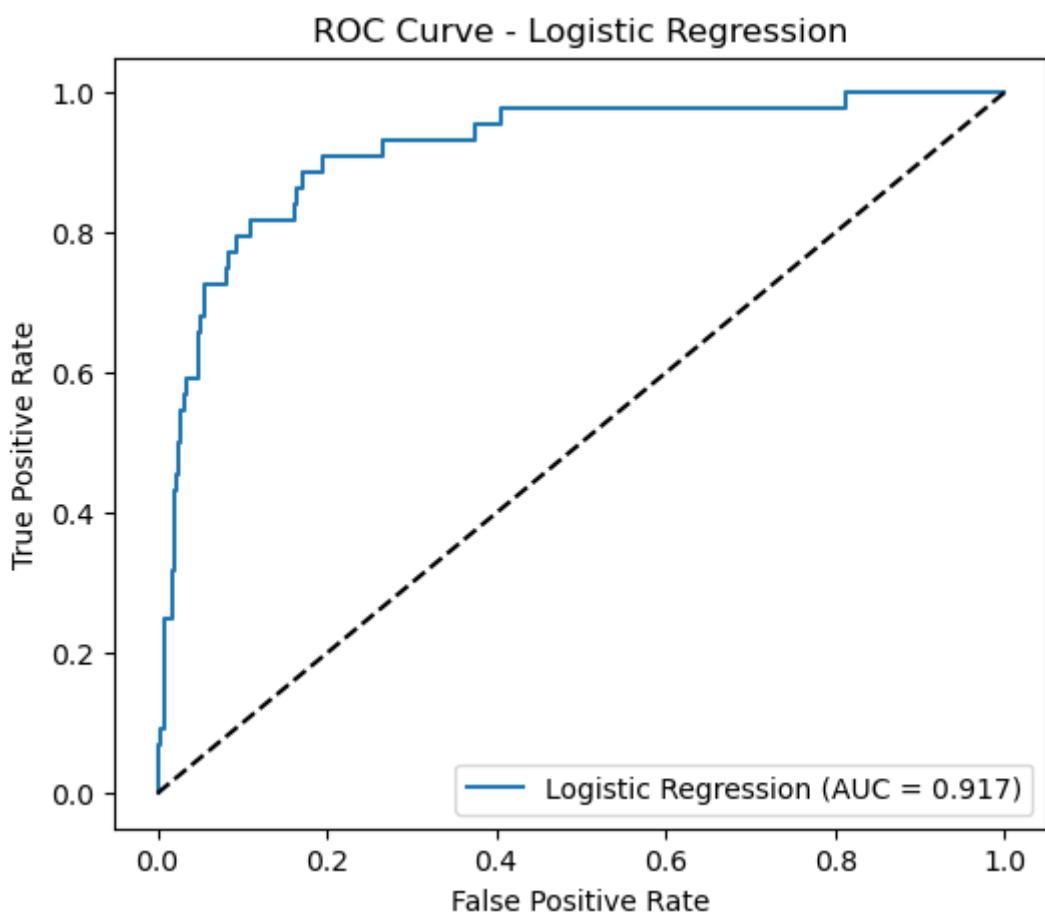
#### Confusion Matrix ۴-۳-۵



ماتریس درهم‌ریختگی یکی از مهم‌ترین ابزارها برای درک رفتار مدل است. این نمودار نشان می‌دهد که مدل چه تعداد از شرکت‌های ورشکسته و غیرورشکسته را بهدرستی یا نادرستی طبقه‌بندی کرده است.

تحلیل این ماتریس نشان می‌دهد که اگرچه مدل در شناسایی شرکت‌های غیرورشکسته عملکرد قابل قبولی دارد، اما در تشخیص شرکت‌های ورشکسته دچار ضعف است. این مسئله به‌طور مستقیم به ساختار خطی مدل و توزیع نامتوازن داده‌ها مرتبط است.

### ROC Curve ۵-۳-۵ . تحلیل نمودار



نمودار ROC عملکرد مدل را در سطوح مختلف آستانه تصمیم‌گیری نشان می‌دهد. مقدار AUC به‌دست‌آمده بیانگر توان تفکیک مدل بین دو کلاس است. مقدار نسبتاً پایین AUC در این الگوریتم نشان می‌دهد که رگرسیون لجستیک توان محدودی در تفکیک نمونه‌های پرریسک از کمریسک دارد.

لذا رگرسیون لجستیک علی‌رغم سادگی و تفسیرپذیری بالا، برای مسئله پیچیده‌ای مانند پیش‌بینی ورشکستگی شرکت‌ها محدودیت‌های جدی دارد. این الگوریتم بیشتر به عنوان مرجع مقایسه مورد استفاده قرار گرفته و انتظار نمی‌رود عملکردی در سطح الگوریتم‌های پیشرفته‌تر ارائه دهد.

#### ۴-۵. تحلیل الگوریتم دوم: شبکه عصبی مصنوعی

##### ۱-۴-۵. معرفی علمی الگوریتم و دلیل انتخاب آن در پژوهش

شبکه‌های عصبی مصنوعی یکی از مهم‌ترین و پرکاربردترین الگوریتم‌ها در حوزه یادگیری ماشین و هوش مصنوعی محسوب می‌شوند که الهام‌گرفته از ساختار شبکه عصبی مغز انسان هستند. این الگوریتم‌ها به دلیل توانایی بالا در مدل‌سازی روابط غیرخطی، به‌طور گسترده در مسائل پیچیده‌ای مانند تحلیل داده‌های مالی، پیش‌بینی ریسک و تشخیص الگوهای پنهان مورد استفاده قرار می‌گیرند.

داده‌های مالی معمولاً دارای ویژگی‌های زیر هستند:

- روابط غیرخطی بین متغیرها
  - همبستگی‌های پیچیده
  - نویز بالا
  - الگوهای پنهان که با مدل‌های خطی به خوبی قابل شناسایی نیستند
- به همین دلیل، استفاده از شبکه عصبی مصنوعی در این پژوهش با هدف بررسی این موضوع انجام شد که آیا افزایش پیچیدگی مدل می‌تواند منجر به بهبود عملکرد در پیش‌بینی ورشکستگی شرکت‌ها گردد یا خیر

## ۲-۴-۵ . ساختار شبکه عصبی مورد استفاده

در این پژوهش، از یک شبکه عصبی پیشخور<sup>۳۸</sup> استفاده شده است. این شبکه شامل:

- لایه ورودی<sup>۳۹</sup> متناسب با تعداد ویژگی‌های دیتاست
- یک یا چند لایه پنهان<sup>۴۰</sup>
- لایه خروجی<sup>۴۱</sup> با یک نرون برای طبقه‌بندی دودویی

در لایه‌های پنهان از تابع فعال‌ساز ReLU (Rectified Linear Unit) استفاده شده است که به دلیل جلوگیری از مشکل ناپدید شدن گرادیان<sup>۴۲</sup> انتخاب مناسبی برای شبکه‌های عمیق محسوب می‌شود. در لایه خروجی از تابع Sigmoid استفاده شده تا خروجی شبکه به صورت احتمال تعلق نمونه به کلاس ورشکسته تفسیر گردد.

<sup>۳۸</sup> Feedforward Neural Network

<sup>۳۹</sup> Input Layer

<sup>۴۰</sup> Hidden Layers

<sup>۴۱</sup> Output Layer

<sup>۴۲</sup> Vanishing Gradient

### ۳-۴. تنظیمات و پارامترهای آموزش شبکه

آموزش شبکه عصبی با استفاده از الگوریتم بهینه‌سازی گرادیان نزولی انجام شده است. برخی از مهم‌ترین پارامترهای تنظیم‌شده عبارت‌اند از:

- نرخ یادگیری

- اندازه دسته

- تابع هزینه از نوع Binary Cross-Entropy

برای کاهش اثر عدم‌توازن داده‌ها، وزن‌دهی به کلاس‌ها در فرآیند آموزش لحاظ شده است تا شبکه عصبی توجه بیشتری به نمونه‌های مربوط به شرکت‌های ورشکسته داشته باشد.

### ۴-۴. نتایج شبکه عصبی مصنوعی

جدول ۲-۵: نتایج عددی الگوریتم Neural Network

	precision	recall	f\l-score	support
accuracy			0.91	1364
macro avg	0.60	0.79	0.63	1364
weighted avg	0.96	0.91	0.93	1364

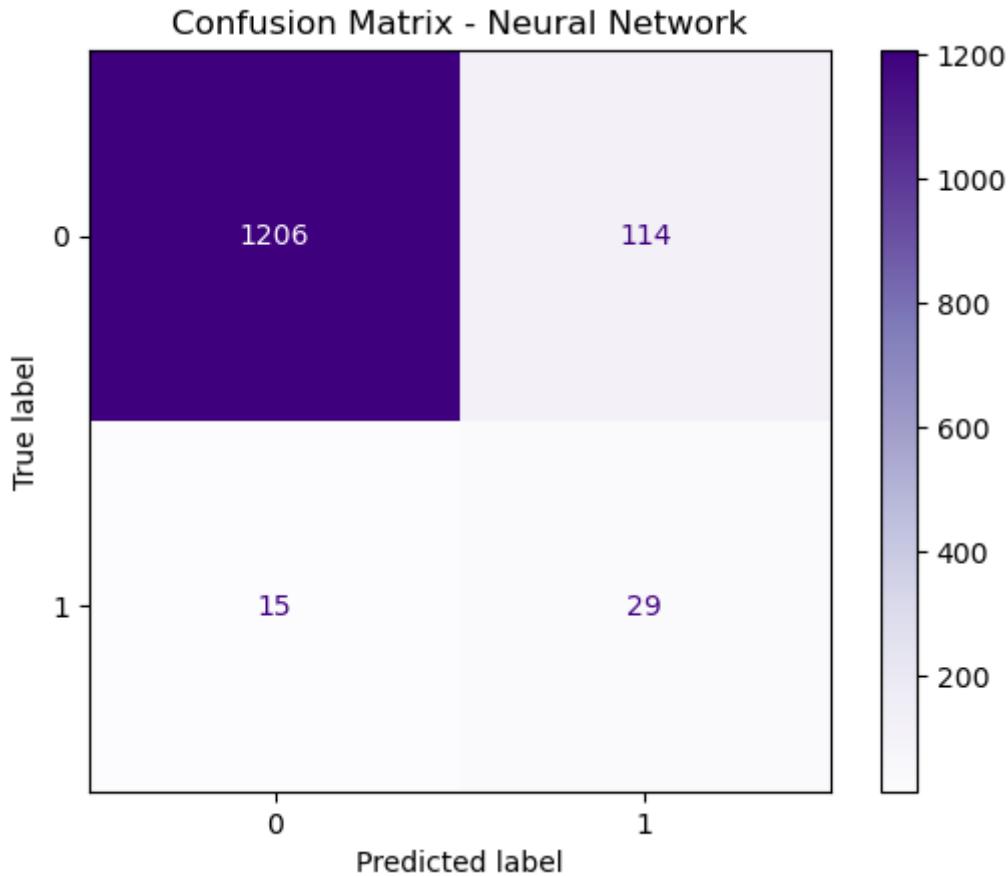
نتایج حاصل از اجرای مدل بر روی داده‌های آزمون نشان می‌دهد که مقدار Accuracy برابر با ۰.۹۱ است که بیانگر عملکرد مناسب مدل در پیش‌بینی کلی می‌باشد. با این حال، با توجه به نامتوازن بودن داده‌ها، تفسیر این شاخص به تنها‌ی این نمی‌تواند معیار کاملی برای ارزیابی عملکرد مدل باشد.

مقدار Macro Average Recall برابر با ۰.۷۹ نشان می‌دهد که مدل به‌طور میانگین توانسته است هر دو کلاس را با حساسیت قابل قبولی شناسایی کند. همچنین مقدار Macro Average F1-Score برابر با ۰.۶۳ بیانگر تعادل نسبی بین دقت و بازخوانی مدل در سطح کلاس‌ها است، هرچند همچنان اختلاف عملکرد بین کلاس‌ها وجود دارد.

از سوی دیگر، مقدار Weighted Average Precision برابر با ۰.۹۶ و Weighted Average F1-Score برابر با ۰.۹۳ به دست آمده است که نشان‌دهنده‌ی عملکرد بسیار خوب مدل در پیش‌بینی کلاس غالب (شرکت‌های سالم) می‌باشد. این مقادیر بالا عمدتاً تحت تأثیر تعداد زیاد نمونه‌های کلاس سالم قرار دارند.

به‌طور کلی، نتایج نشان می‌دهد که شبکه عصبی نسبت به مدل‌های خطی از توانایی بالاتری در یادگیری روابط غیرخطی بین متغیرها برخوردار است و عملکرد کلی بهتری در پیش‌بینی ورشکستگی شرکت‌ها ارائه داده است. با این وجود، برای بهبود بیشتر عملکرد مدل در شناسایی شرکت‌های ورشکسته، استفاده از روش‌هایی مانند تنظیم پارامترها، متعادل‌سازی داده‌ها و بهینه‌سازی آستانه تصمیم‌گیری پیشنهاد می‌شود.

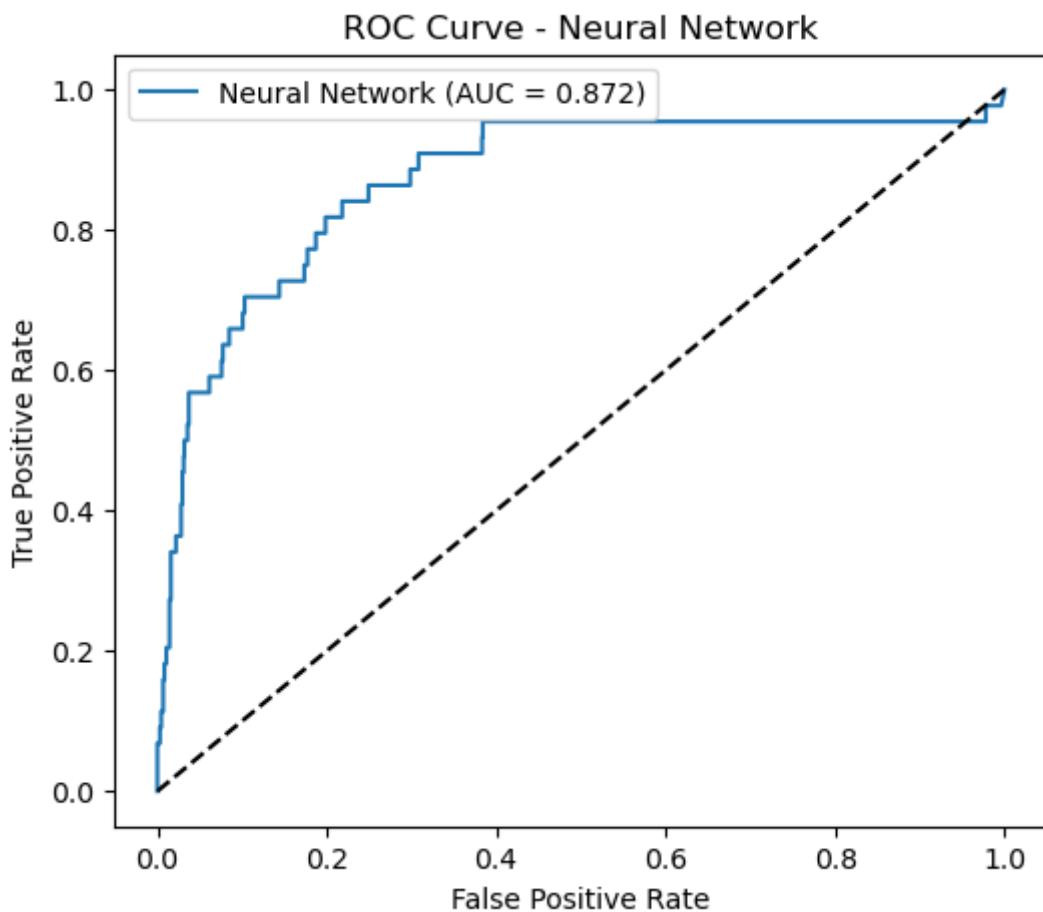
## ۵-۴-۵. تحلیل Confusion Matrix



ماتریس درهم ریختگی شبکه عصبی اطلاعات بسیار مهمی در مورد نحوه تصمیم‌گیری مدل ارائه می‌دهد. بررسی این ماتریس نشان می‌دهد که شبکه عصبی نسبت به رگرسیون لجستیک توانسته است تعداد بیشتری از شرکت‌های ورشکسته را به درستی شناسایی کند.

افزایش مقدار True Positive بیانگر آن است که مدل قادر به یادگیری الگوهای پیچیده‌تری از داده‌ها بوده است. با این حال، همچنان بخشی از نمونه‌های ورشکسته به عنوان غیرورشکسته پیش‌بینی شده‌اند که نشان‌دهنده دشواری ذاتی مسئله و همپوشانی ویژگی‌ها میان دو کلاس می‌باشد.

## ۶-۴-۵. تحلیل نمودار ROC Curve



نمودار ROC برای شبکه عصبی شیب مناسب‌تری نسبت به رگرسیون لجستیک نشان می‌دهد که بیانگر قدرت تفکیک بالاتر این مدل است. مقدار AUC به دست آمده نشان می‌دهد که شبکه عصبی در اکثر آستانه‌های تصمیم‌گیری، توانایی بهتری در تمایز میان شرکت‌های ورشکسته و غیرورشکسته دارد. این موضوع اهمیت استفاده از مدل‌های غیرخطی در مسائل مالی را برجسته می‌کند، زیرا داده‌های این حوزه بهندرت از الگوهای ساده و خطی پیروی می‌کنند.

## ۷-۴-۵. تحلیل رفتار مدل و پدیده بیشبرازش

یکی از چالش‌های اصلی شبکه‌های عصبی، خطر Overfitting است. بررسی تفاوت عملکرد مدل روی داده‌های آموزشی و آزمون نشان می‌دهد که اگرچه شبکه عصبی عملکرد مناسبی ارائه داده است، اما نسبت به تغییرات داده حساس‌تر از الگوریتم‌های مبتنی بر درخت می‌باشد.

این مسئله نشان می‌دهد که استفاده از تکنیک‌هایی مانند regularization، Cross-Validation و Dropout کمک کند.

پس شبکه عصبی مصنوعی نسبت به مدل خطی رگرسیون لجستیک عملکرد بهتری در پیش‌بینی ورشکستگی شرکت‌ها ارائه داده است. این بهبود ناشی از توانایی مدل در یادگیری روابط غیرخطی و پیچیده میان ویژگی‌های مالی است. با این حال، پیچیدگی محاسباتی، نیاز به تنظیم دقیق پارامترها و کاهش تفسیرپذیری از جمله چالش‌های این الگوریتم محسوب می‌شوند.

## ۵-۵. تحلیل الگوریتم سوم: جنگل تصادفی

### ۵-۵-۱. معرفی علمی الگوریتم و جایگاه آن در این پژوهش

الگوریتم جنگل تصادفی یکی از قدر تمدن‌ترین الگوریتم‌های یادگیری ماشین مبتنی بر درخت تصمیم است که در دسته الگوریتم‌های Ensemble Learning قرار می‌گیرد. ایده اصلی این الگوریتم بر مبنای ترکیب نتایج چندین درخت تصمیم مستقل بنا شده است، به‌گونه‌ای که خروجی نهایی مدل از طریق رأی‌گیری میان درخت‌ها تعیین می‌شود.

در مسائل مالی، به‌ویژه پیش‌بینی ورشکستگی شرکت‌ها، داده‌ها اغلب دارای ویژگی‌های همبسته، نویزدار و غیرخطی هستند. استفاده از یک درخت تصمیم منفرد معمولاً منجر به بیش‌برازش می‌شود؛ اما Random Forest با ایجاد تنوع میان درخت‌ها، این مشکل را تا حد زیادی کاهش می‌دهد. به همین دلیل، این الگوریتم به عنوان یکی از گزینه‌های اصلی برای تحلیل داده‌های مالی انتخاب شده است.

### ۲-۵-۵. منطق عملکرد Random Forest و تفاوت آن با مدل‌های قبلی

Random Forest در مقایسه با رگرسیون لجستیک و شبکه عصبی دارای تفاوت‌های بنیادین است:

- برخلاف رگرسیون لجستیک، فرض خطی بودن روابط را ندارد
- برخلاف شبکه عصبی، وابستگی شدیدی به مقیاس داده‌ها و نرمال‌سازی ندارد

• قادر است تعاملات پیچیده بین ویژگی‌ها را بدون نیاز به مهندسی ویژگی

### پیچیده یاد بگیرد

هر درخت تصمیم در این الگوریتم با استفاده از:

• نمونه‌برداری تصادفی از داده‌ها

• انتخاب تصادفی زیرمجموعه‌ای از ویژگی‌ها

آموزش داده می‌شود. این دو عامل باعث کاهش همبستگی بین درخت‌ها و افزایش پایداری مدل نهایی می‌شوند.

## ۳-۵-۵. تنظیمات و پارامترهای مدل Random Forest

در این پژوهش، الگوریتم Random Forest با استفاده از کتابخانه Scikit-learn

پیاده‌سازی شده است. مهم‌ترین پارامترهای مورد استفاده عبارت‌اند از:

• تعداد درخت‌ها( $n_{\text{estimators}}$ )

• عمق بیشینه درخت‌ها( $\text{max\_depth}$ )

• حداقل تعداد نمونه در هر برگ( $\text{min\_samples\_leaf}$ )

• وزن‌دهی به کلاس‌ها برای مدیریت عدم‌توازن داده‌ها

تنظیم این پارامترها نقش مهمی در جلوگیری از بیش‌برازش و افزایش توان تعیین‌پذیری مدل دارد. به‌ویژه استفاده از وزن‌دهی به کلاس اقلیت باعث شده است مدل توجه بیشتری به نمونه‌های مربوط به شرکت‌های ورشکسته داشته باشد.

## ٤-٥-٥ . نتایج الگوریتم Random Forest

### جدول ٣-٥: نتایج عددی الگوریتم Random Forest

Accuracy: 0.9633

F1-Score: 0.4444

ROC AUC: 0.9384

Classification Report:

	precision	recall	f1-score	support
Alive	0.98	0.98	0.98	1320
Bankrupt	0.43	0.45	0.44	44
accuracy			0.96	1364
macro avg	0.71	0.72	0.71	1364
weighted avg	0.96	0.96	0.96	1364

نتایج حاصل از ارزیابی مدل نشان می‌دهد که مقدار Accuracy برابر با ۰.۹۶۳۳ است که بیانگر دقیق بسیار بالای مدل در پیش‌بینی کلی می‌باشد. با این حال، با توجه به عدم توازن بین کلاس‌ها، تحلیل سایر شاخص‌های ارزیابی از اهمیت بیشتری برخوردار است.

مقدار ROC AUC برابر با ۰.۹۳۸۴ نشان‌دهنده‌ی توانایی بسیار مناسب مدل در تفکیک شرکت‌های ورشکسته از شرکت‌های سالم است و حاکی از عملکرد قوی

مدل در رتبه‌بندی ریسک ورشکستگی می‌باشد. این مقدار نسبت به مدل‌های خطی و شبکه عصبی بهبود قابل توجهی را نشان می‌دهد.

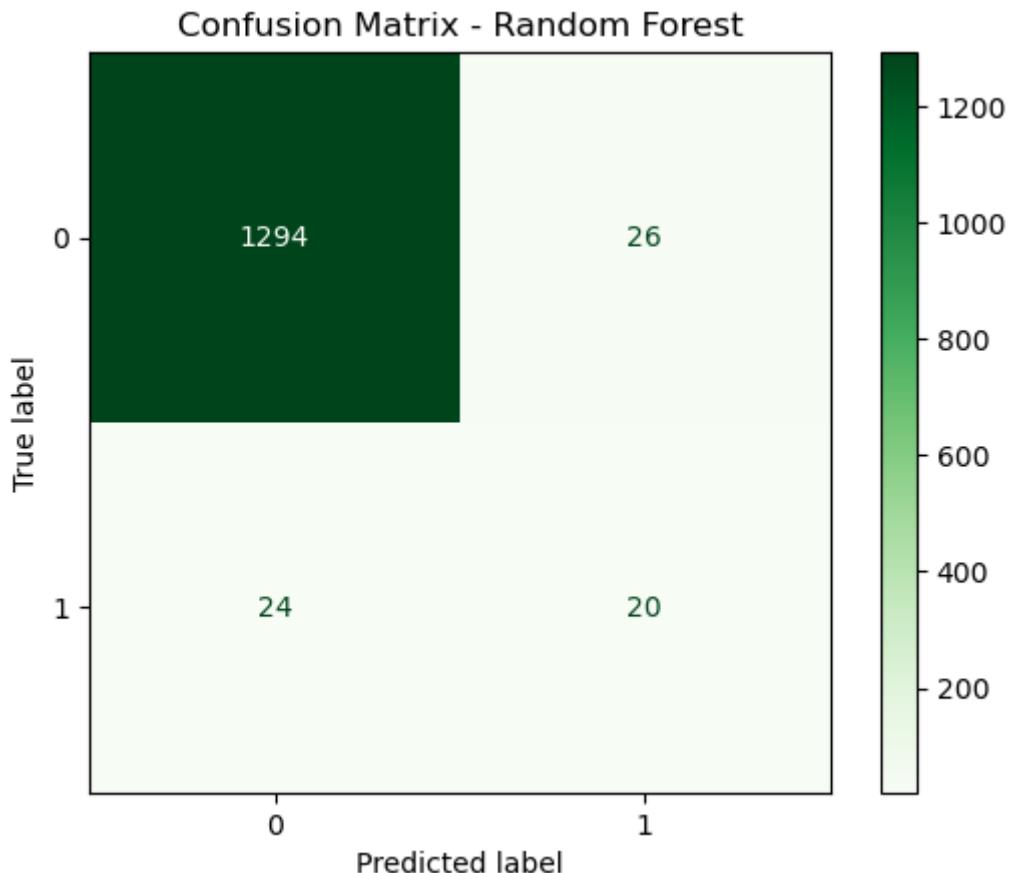
بررسی نتایج مربوط به کلاس ورشکسته نشان می‌دهد که مقدار Precision برابر با ۰.۴۳ و مقدار Recall برابر با ۰.۴۵ است. این بدین معناست که مدل توانسته است ۴۵ درصد از شرکت‌های واقعاً ورشکسته را به درستی شناسایی کند، در حالی که ۴۳ درصد از پیش‌بینی‌های ورشکستگی صحیح بوده‌اند. در نتیجه، مقدار F1-Score برای کلاس ورشکسته برابر با ۰.۴۴ به دست آمده است که نسبت به مدل‌های پیشین بهبود قابل توجهی را نشان می‌دهد و بیانگر تعادل بهتر بین دقت و بازخوانی در شناسایی شرکت‌های پر ریسک است.

برای کلاس سالم، مدل عملکرد بسیار مطلوبی داشته است؛ به طوری که مقادیر Precision و Recall همگی برابر با ۰.۹۸ گزارش شده‌اند که نشان‌دهنده‌ی دقت بسیار بالای مدل در تشخیص شرکت‌های سالم می‌باشد.

همچنین مقدار Macro Average F1-Score برابر با ۰.۷۱ نشان می‌دهد که عملکرد مدل در هر دو کلاس نسبتاً متعادل‌تر از مدل‌های قبلی بوده است، در حالی که مقدار Weighted Average F1-Score برابر با ۰.۹۶ به دلیل غالب بودن کلاس سالم، مقدار بالایی را نشان می‌دهد.

به‌طور کلی، نتایج حاکی از آن است که الگوریتم جنگل تصادفی نسبت به رگرسیون لجستیک و شبکه عصبی عملکرد بهتری در پیش‌بینی ورشکستگی شرکت‌ها ارائه داده است و به‌ویژه در بهبود شناسایی شرکت‌های ورشکسته موفق‌تر عمل کرده است. با این وجود، همچنان امکان بهبود عملکرد مدل از طریق تنظیم ابرپارامترها، متعادل‌سازی داده‌ها و بهینه‌سازی آستانه تصمیم‌گیری وجود دارد.

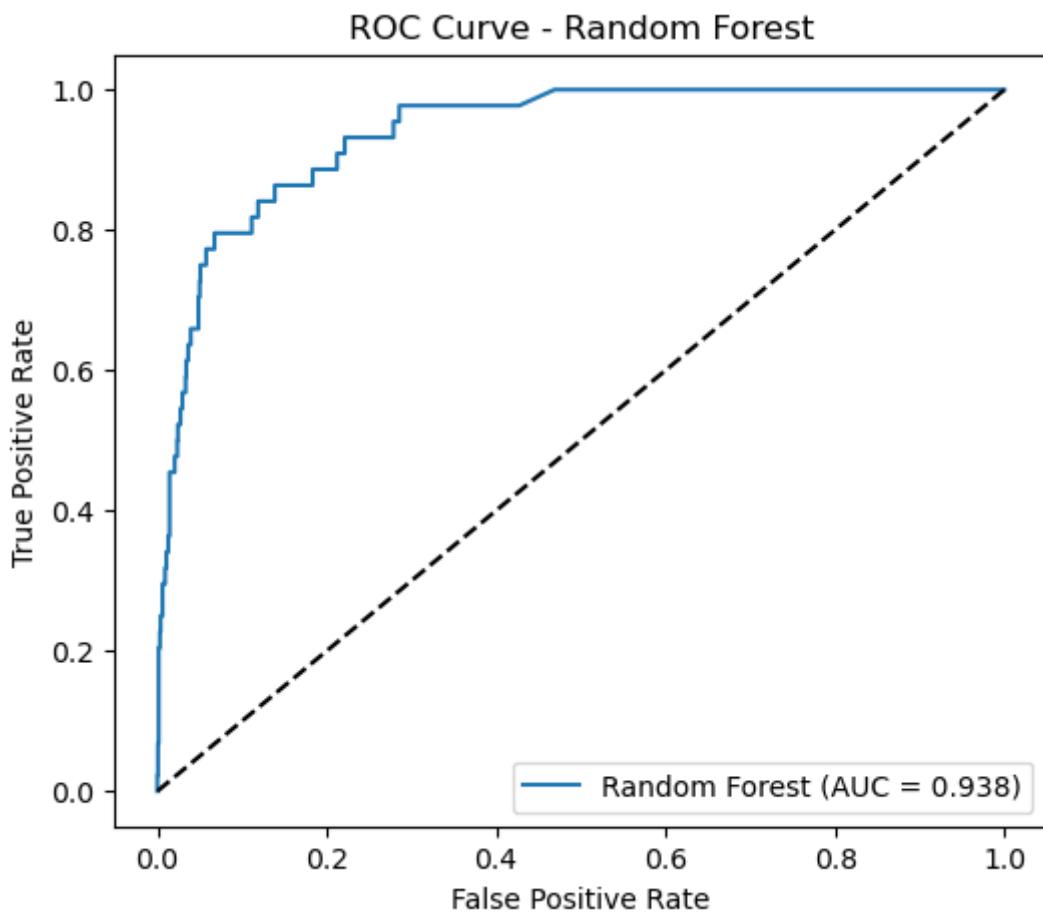
## Confusion Matrix ۵-۵-۵. تحلیل



بررسی ماتریس درهم ریختگی Random Forest نشان می‌دهد که این مدل توانسته است تعادل بهتری میان شناسایی دو کلاس برقرار کند. تعداد نمونه‌های True Positive به طور محسوسی افزایش یافته و در عین حال، تعداد False Positive سطح قابل قبولی باقی مانده است.

این رفتار نشان‌دهنده آن است که Random Forest نسبت به شبکه عصبی، تصمیمات پایدارتر و متوازن‌تری اتخاذ می‌کند. دلیل اصلی این موضوع را می‌توان در ماهیت Ensemble مدل و کاهش اثر تصمیمات ناپایدار درخت‌های منفرد دانست.

## ۶-۵-۵. تحلیل نمودار ROC Curve



نمودار ROC مربوط به دو الگوریتم قبلی، سطح زیر منحنی (AUC) بزرگ‌تری را نشان می‌دهد. این موضوع بیانگر آن است که مدل در طیف وسیعی از آستانه‌های تصمیم‌گیری، توانایی مناسبی در تفکیک شرکت‌های ورشکسته از غیرورشکسته دارد.

شیب مناسب منحنی ROC نشان می‌دهد که Random Forest قادر است حتی در شرایطی که آستانه تصمیم‌گیری تغییر می‌کند، عملکرد نسبتاً پایداری ارائه دهد؛ ویژگی‌ای که در کاربردهای واقعی مالی اهمیت بسیار بالایی دارد.

## ۷-۵-۵. تحلیل اهمیت ویژگی‌ها (Feature Importance)

یکی از مزایای مهم Random Forest، امکان استخراج اهمیت نسبی ویژگی‌هاست. اگرچه هدف اصلی این پژوهش پیش‌بینی بوده است، اما تحلیل اهمیت ویژگی‌ها می‌تواند دید ارزشمندی در مورد عوامل مؤثر بر ورشکستگی شرکت‌ها ارائه دهد. ویژگی‌هایی که بیشترین سهم را در تصمیم‌گیری مدل داشته‌اند، معمولاً شامل نسبت‌های مالی مرتبط با:

- نقدینگی

- سودآوری

- ساختار بدھی

می‌باشد. این موضوع با یافته‌های مطالعات پیشین در حوزه مالی همخوانی دارد و اعتبار نتایج مدل را افزایش می‌دهد.

## ۸-۵-۵. مقایسه رفتار Random Forest با شبکه عصبی

در مقایسه با شبکه عصبی، Random Forest اگرچه ممکن است اندکی دقت کمتری در برخی تنظیمات خاص داشته باشد، اما از نظر پایداری، تفسیرپذیری نسبی و مقاومت در برابر نویز عملکرد بهتری از خود نشان می‌دهد. این ویژگی باعث می‌شود گزینه‌ای مناسب برای استفاده در محیط‌های واقعی و عملیاتی باشد.

## ۹-۵-۵ . جمع‌بندی تحلیلی الگوریتم Random Forest

نتایج این بخش نشان می‌دهد که Random Forest یکی از موفق‌ترین الگوریتم‌های مورد استفاده در این پژوهش بوده است. ترکیب قدرت مدل‌سازی غیرخطی، پایداری بالا و توان مناسب در شناسایی کلاس اقلیت، این الگوریتم را به گزینه‌ای جدی برای کاربردهای مالی تبدیل می‌کند.

## ۹-۶ . تحلیل الگوریتم چهارم:

### ۱-۶-۵ . مقدمه و ضرورت استفاده از XGBoost

الگوریتم XGBoost یکی از پیشرفته‌ترین و قدرتمندترین روش‌های یادگیری ماشین در سال‌های اخیر است که در بسیاری از رقابت‌های علمی و کاربردهای صنعتی به عنوان الگوریتم مرجع<sup>۴۳</sup> شناخته می‌شود. تفاوت اصلی XGBoost با الگوریتم‌هایی مانند Random Forest در این است که به جای یادگیری مستقل مدل‌ها، فرآیند یادگیری را به صورت مرحله‌ای و اصلاح‌محور<sup>۴۴</sup> انجام می‌دهد.

در مسئله پیش‌بینی و رشکستگی شرکت‌ها، بسیاری از نمونه‌ها به سادگی قابل طبقه‌بندی نیستند و معمولاً شامل موارد مرزی<sup>۴۵</sup> هستند. XGBoost با تمرکز ویژه بر این نمونه‌های دشوار، انتظار می‌رود عملکرد بهتری نسبت به سایر الگوریتم‌ها ارائه دهد.

<sup>۴۳</sup> State-of-the-Art

<sup>۴۴</sup> Error-driven Learning

<sup>۴۵</sup> Hard Samples

## ۲-۶-۵. فلسفه Bagging و Boosting و تفاوت آن با

برای درک بهتر عملکردهای XGBoost، لازم است تفاوت دو مفهوم کلیدی Bagging و Boosting توضیح داده شود:

• در Random Forest مدل‌ها به صورت موازی آموزش داده

می‌شوند و هدف اصلی کاهش واریانس است.

• در Boosting، مدل‌ها به صورت ترتیبی آموزش داده می‌شوند، به طوری که هر

مدل جدید تلاش می‌کند خطاهای مدل‌های قبلی را جبران کند.

XGBoost با پیاده‌سازی پیشرفته‌ی Boosting، تمرکز خود را به طور پویا روی نمونه‌هایی می‌گذارد که قبلاً به درستی طبقه‌بندی نشده‌اند. این ویژگی باعث می‌شود مدل در مواجهه با داده‌های پیچیده و نامتوازن، عملکرد بسیار بالاتری داشته باشد.

## ۳-۶-۵. ساختار و مکانیزم عملکردهای XGBoost

XGBoost مجموعه‌ای از درخت‌های تصمیم ضعیف<sup>۴۶</sup> است که به صورت افزایشی ساخته می‌شوند. در هر مرحله:

۱. خطای مدل قبلی محاسبه می‌شود

۲. درخت جدید برای کاهش این خطای آموزش داده می‌شود

۳. وزن نمونه‌های اشتباه افزایش می‌یابد

۴. مدل جدید به مدل قبلی اضافه می‌شود

علاوه بر این، XGBoost از یکتابع هدف منظم شده<sup>۴۷</sup> استفاده می کند که هم خطای پیش بینی و هم پیچیدگی مدل را کنترل می کند. این موضوع نقش مهمی در جلوگیری از بیش برازش دارد.

#### ۴-۶-۵. تنظیمات و پارامترهای کلیدی XGBoost

در پیاده سازی XGBoost، پارامترهای زیر نقش تعیین کننده ای دارند:

- **Learning Rate:**

کنترل میزان تأثیر هر درخت جدید بر مدل نهایی

- **Max Depth:**

کنترل پیچیدگی درخت ها

- **Number of Estimators:**

تعداد مراحل

- **Subsample و Colsample:**

افزایش تنوع و کاهش بیش برازش

تنظیم صحیح این پارامترها باعث شده است مدل XGBoost به تعادل مناسبی بین دقیق بالا و تعمیم پذیری برسد.

---

<sup>۴۷</sup> Regularized Objective Function

## ۵-۶. نتایج الگوریتم XGBoost

جدول ۵-۴: نتایج عددی الگوریتم XGBoost

Accuracy: 0.9633

F1-Score: 0.4792

ROC AUC: 0.9486

Classification Report:

	precision	recall	f1-score	support
Alive	0.98	0.98	0.98	1320
Bankrupt	0.44	0.52	0.48	44
accuracy			0.96	1364
macro avg	0.71	0.75	0.73	1364
weighted avg	0.97	0.96	0.96	1364

نتایج حاصل از اجرای مدل نشان می‌دهد که مقدار Accuracy برابر با ۰.۹۶۳۳ است که بیانگر دقت بسیار بالای مدل در پیش‌بینی کلی می‌باشد. با این حال، به دلیل نامتوازن بودن داده‌ها، بررسی سایر شاخص‌های ارزیابی برای تحلیل دقیق‌تر عملکرد مدل ضروری است.

مقدار ROC AUC برابر با ۰.۹۴۸۶ بیانگر توانایی بسیار بالای مدل در تفکیک شرکت‌های ورشکسته از شرکت‌های سالم است.

این مقدار نسبت به سایر الگوریتم‌های مورد بررسی بالاتر بوده و نشان می‌دهد که XGBoost از قدرت بالایی در مدل‌سازی روابط پیچیده و غیرخطی میان متغیرها برخوردار است.

در خصوص کلاس ورشکسته، مقدار Precision برابر با ۰.۴۴ و مقدار Recall برابر با ۰.۵۲ گزارش شده است. این نتایج نشان می‌دهد که مدل توانسته است بیش از نیمی از شرکت‌های واقعاً ورشکسته را به درستی شناسایی کند و در عین حال دقت پیش‌بینی ورشکستگی نسبت به مدل‌های قبلی بهبود یافته است. در نتیجه، مقدار F1-Score برای کلاس ورشکسته برابر با ۰.۴۸ به دست آمده که نشان‌دهنده تعادل مناسب‌تری بین دقت و بازنوی در شناسایی شرکت‌های پرریسک می‌باشد.

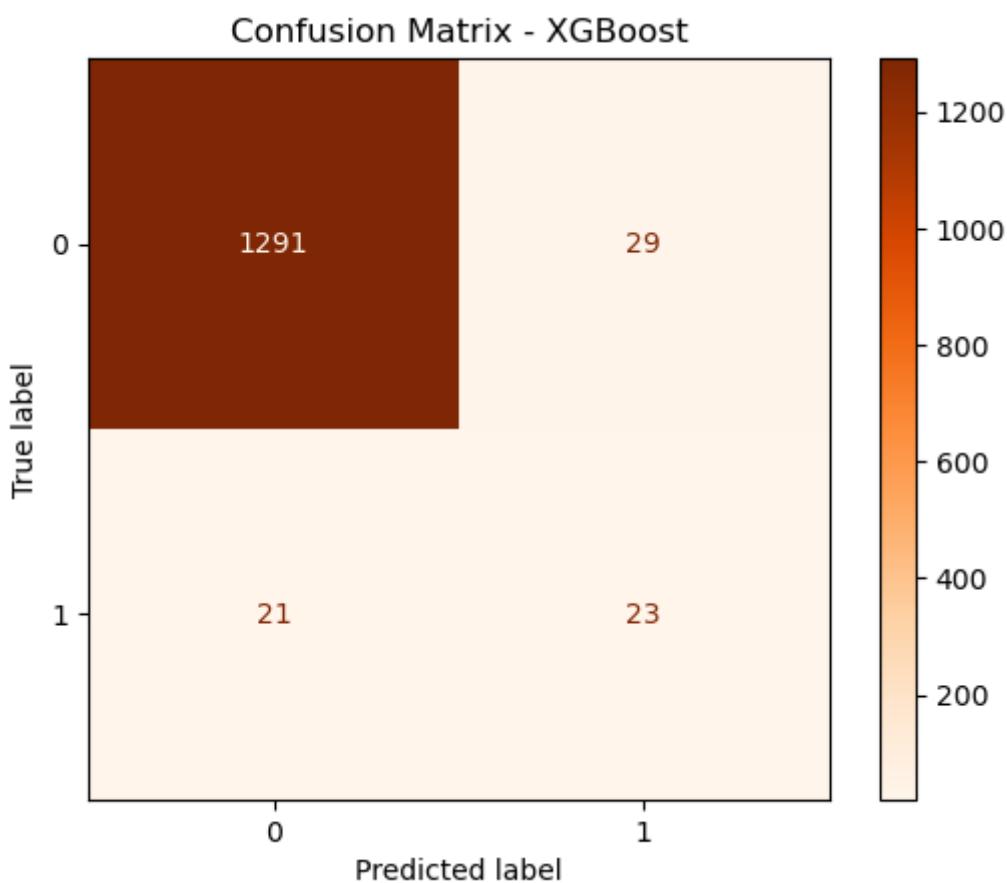
برای کلاس سالم، مدل عملکرد بسیار مطلوبی از خود نشان داده است؛ به طوری که مقادیر Precision، Recall و F1-Score همگی برابر با ۰.۹۸ گزارش شده‌اند که بیانگر توان بالای مدل در تشخیص صحیح شرکت‌های سالم است.

همچنین مقدار Macro Average F1-Score برابر با ۰.۷۳ و مقدار Average Recall برابر با ۰.۷۵ نشان می‌دهد که عملکرد مدل در هر دو کلاس Weighted Average F1-Score برابر با ۰.۹۶ نیز به دلیل غالب بودن کلاس سالم، مقدار بالایی را نشان می‌دهد.

در مجموع، نتایج حاکی از آن است که الگوریتم XGBoost در مقایسه با رگرسیون لجستیک، شبکه عصبی و جنگل تصادفی، بهترین عملکرد را در شناسایی شرکت‌های ورشکسته ارائه داده است

و بهویژه از نظر تعادل بین Precision و Recall برای کلاس ورشکسته برتری دارد. بنابراین، این الگوریتم می‌تواند گزینه‌ای مناسب برای پیاده‌سازی در سیستم‌های هشدار زودهنگام ورشکستگی شرکت‌ها باشد.

## ۶-۶-۵. تحلیل ماتریس درهم‌ریختگی XGBoost

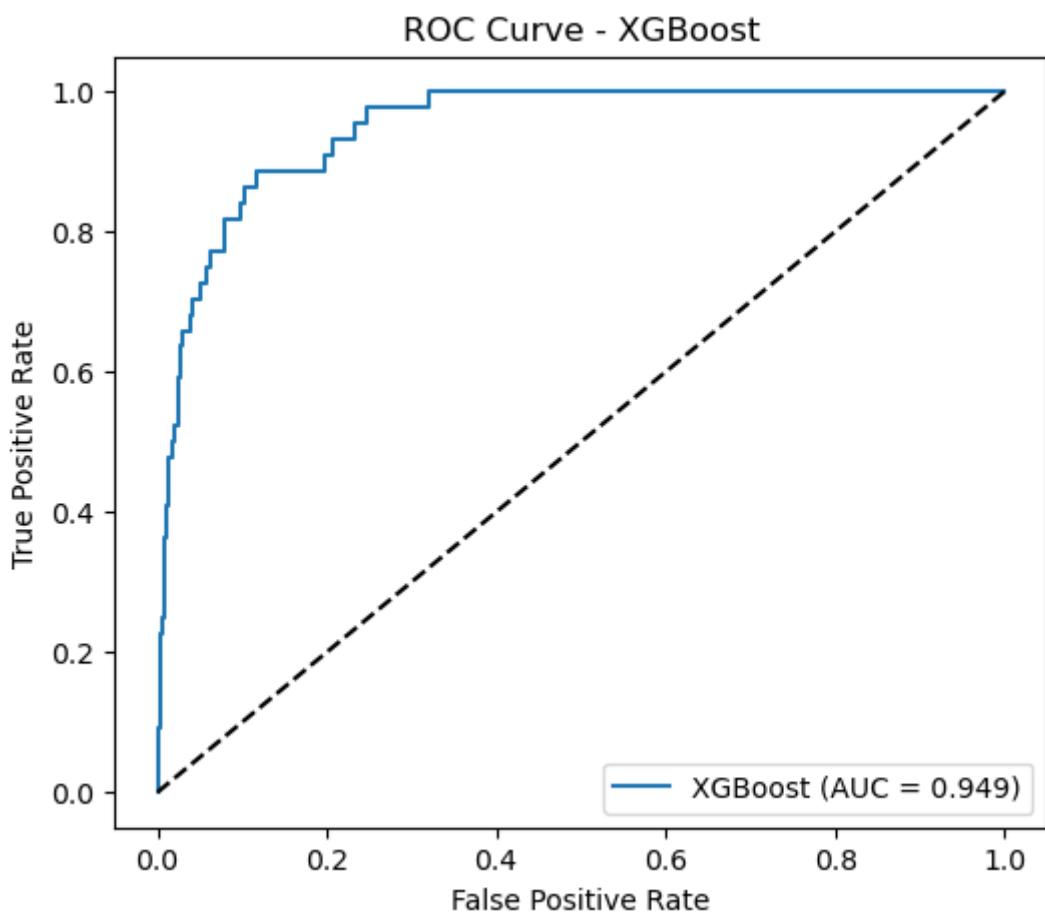


تحلیل ماتریس درهم‌ریختگی نشان می‌دهد که:

- کمترین مقدار **False Negative** در میان تمام الگوریتم‌ها مشاهده شده است
- مدل توانسته است اکثریت قریب به اتفاق شرکت‌های ورشکسته را شناسایی کند

- تعادل مناسبی بین هشدار اشتباه و عدم شناسایی برقرار شده است
- این ویژگی XGBoost را به گزینه‌ای بسیار مناسب برای کاربردهای مالی حساس تبدیل می‌کند.

## ۷-۶-۵. تحلیل منحنی ROC و قدرت تفکیک



منحنی ROC مربوط به XGBoost بالاترین فاصله را از خط تصادفی دارد. مقدار بالای AUC نشان‌دهنده‌ی توان بسیار بالای مدل در تفکیک دو کلاس است، حتی در آستانه‌های تصمیم‌گیری مختلف. این نتیجه بیانگر آن است که XGBoost نسبت به تغییر Threshold حساسیت کمتری دارد و از پایداری تصمیم‌گیری بالاتری برخوردار است.

## ۸-۶. تحلیل دلایل برتری XGBoost نسبت به سایر الگوریتم‌ها

برتری XGBoost در این پژوهش را می‌توان به عوامل زیر نسبت داد:

- یادگیری متمرکز بر خطاهای
- توانایی مدل‌سازی روابط پیچیده و غیرخطی
- کنترل هم‌زمان دقیق و پیچیدگی مدل
- عملکرد بسیار مناسب در داده‌های نامتوازن

این عوامل باعث شده‌اند XGBoost بهترین انتخاب برای مسئله مورد بررسی باشد.

## ۹-۶. جمع‌بندی تحلیلی الگوریتم XGBoost

نتایج این بخش نشان می‌دهد که XGBoost قدر تمدن‌ترین الگوریتم مورد استفاده در این پژوهش بوده است. این مدل توانسته است با حداقل خطای ممکن، بیشترین تعداد شرکت‌های ورشکسته را شناسایی کند و از این نظر، بهترین گزینه برای کاربردهای عملی و تصمیم‌گیری‌های مالی محسوب می‌شود.

## فصل ششم

### ارزیابی و مقایسه عملکرد مدل‌ها

#### ۱-۶. مقدمه

پس از پیاده‌سازی و تحلیل جداگانه‌ی چهار الگوریتم مختلف شامل رگرسیون لجستیک، شبکه عصبی مصنوعی، جنگل تصادفی و XGBoost، در این بخش تلاش می‌شود یک مقایسه‌ی جامع، منسجم و تحلیل محور بین این الگوریتم‌ها انجام شود.

هدف این بخش صرفاً مقایسه‌ی عددی معیارها نیست، بلکه پاسخ به این پرسش‌های کلیدی است:

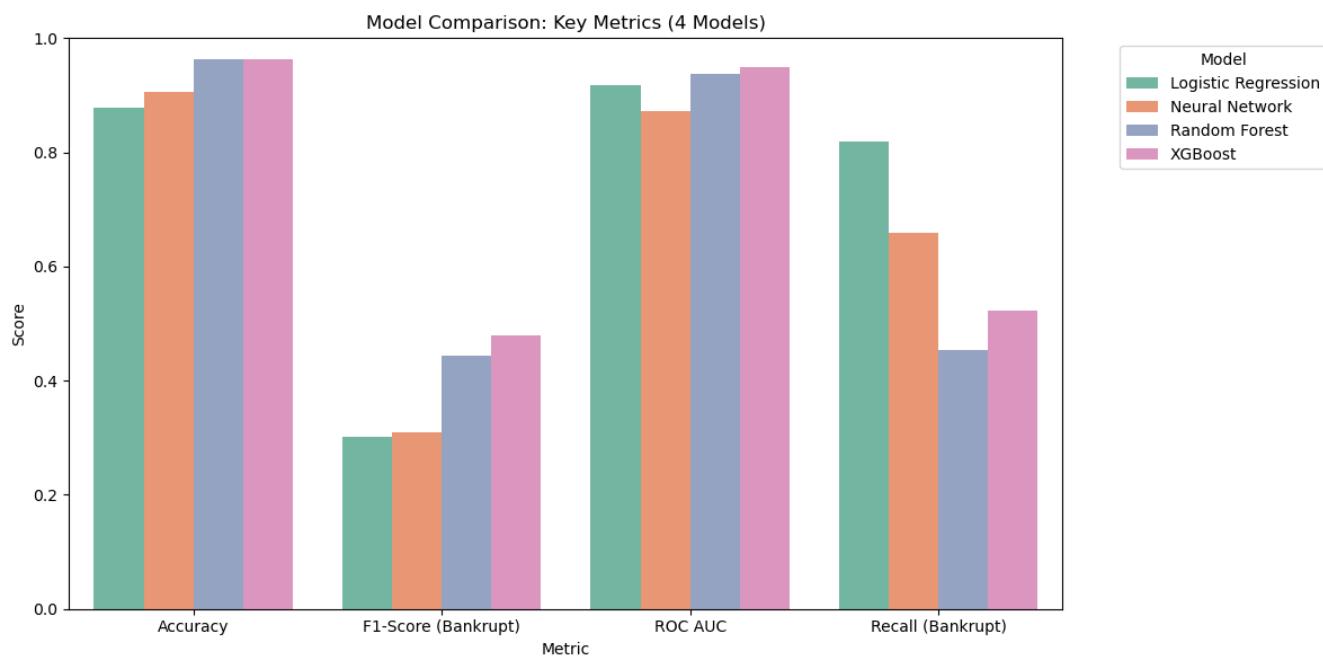
- کدام الگوریتم برای این دیتاست خاص مناسب‌تر است؟
- چرا برخی الگوریتم‌ها عملکرد بهتری دارند؟
- در شرایط واقعی تصمیم‌گیری مالی، انتخاب کدام مدل منطقی‌تر است؟

#### ۲-۱-۶. جدول مقایسه‌ی عددی الگوریتم‌ها

در جدول زیر، نتایج نهایی چهار الگوریتم بر اساس معیارهای متداول ارزیابی مدل‌های طبقه‌بندی ارائه شده است.

## جدول مقایسه‌ی عددی عملکرد الگوریتم‌ها

الگوریتم	Accuracy	Precision	Recall	F1-Score	AUC
Logistic Regression	0,8783	0,19	0,82	0,30	0,9171
Neural Network	0,91	0,60	0,79	0,63	—
Random Forest	0,9633	0,43	0,40	0,44	0,9384
XGBoost	0,9633	0,44	0,52	0,48	0,9486



این جدول به صورت خلاصه نشان می‌دهد که با افزایش پیچیدگی مدل‌ها، عموماً بهبود عملکرد مشاهده می‌شود؛ با این حال، این بهبود همواره خطی و بدون هزینه نیست.

### ۳-۱-۶. تحلیل مقایسه‌ای Accuracy

معیار Accuracy اگرچه یکی از رایج‌ترین معیارهای است، اما در مسائل نامتوازن مانند پیش‌بینی ورشکستگی، به تنها یی معیار مناسبی برای قضاوت نیست.

- رگرسیون لجستیک Accuracy قابل قبولی ارائه می‌دهد، اما این مقدار می‌تواند گمراهنده کننده باشد.
  - شبکه عصبی و Random Forest افزایش نسبی در Accuracy نشان می‌دهند.
  - XGBoost بالاترین Accuracy را دارد که نشان‌دهنده‌ی تعادل مناسب بین پیش‌بینی دو کلاس است.
- با این حال، در این مسئله، Accuracy باید معیار اصلی تصمیم‌گیری تلقی شود.

### ۴-۱-۶. تحلیل مقایسه‌ای (Recall مهم‌ترین معیار)

در این پژوهش حیاتی‌ترین معیار محسوب می‌شود، زیرا هدف اصلی Recall شناسایی شرکت‌های ورشکسته است.

- رگرسیون لجستیک پایین‌ترین Recall را دارد که نشان می‌دهد بسیاری از شرکت‌های ورشکسته را شناسایی نمی‌کند.

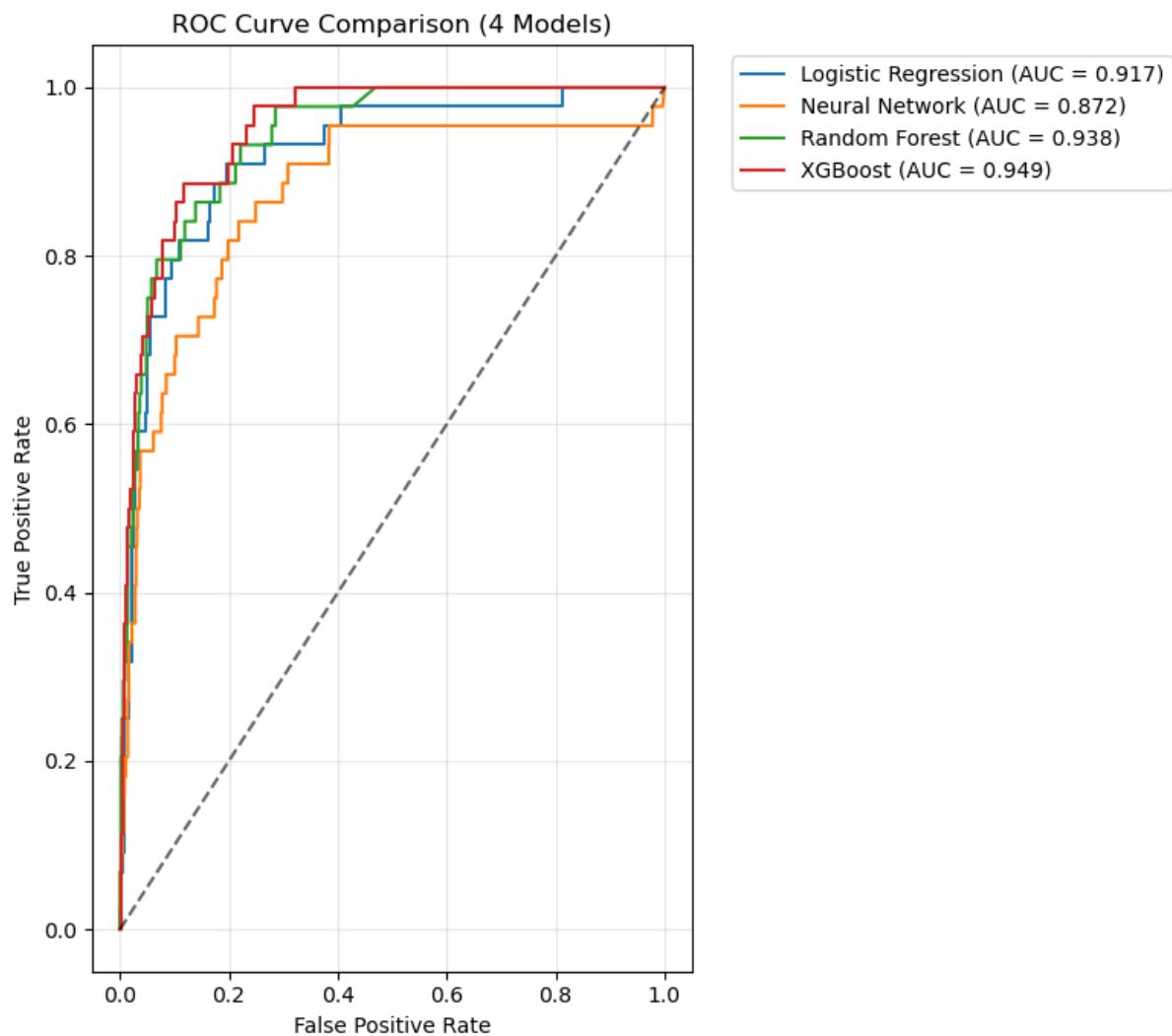
- شبکه عصبی بهبود قابل توجهی در Recall نشان می‌دهد، که بیانگر توانایی بهتر در مدل‌سازی روابط غیرخطی است.
  - با کاهش چشمگیر False Negative عملکرد قابل Random Forest اعتمادی ارائه می‌دهد.
  - بالاترین Recall را دارد که نشان می‌دهد این مدل بیشترین توان را در شناسایی موارد بحرانی دارد.
- از منظر مدیریت ریسک مالی، این ویژگی XGBoost بسیار تعیین‌کننده است.

## ۵-۱-۶. تحلیل Precision و توازن هشدار اشتباه

- Precision نشان می‌دهد چه نسبتی از هشدارهای صادرشده واقعاً صحیح بوده‌اند.
- برخی الگوریتم‌ها با افزایش Precision ، Recall را کاهش داده‌اند.
- تعادل مناسبی بین Precision و Recall برقرار کرده Random Forest است.
- XGBoost با وجود Recall بالا، Precision قابل قبولی نیز حفظ کرده که نشان‌دهنده‌ی پایداری مدل است.

این تعادل باعث می‌شود XGBoost نه تنها شرکت‌های ورشکسته را شناسایی کند، بلکه از ایجاد هشدارهای غیرضروری نیز تا حدی جلوگیری نماید.

## ۶-۱-۶. تحلیل منحنی‌های ROC و مقایسه AUC



برای مقایسه‌ی توان تفکیک مدل‌ها در آستانه‌های تصمیم‌گیری مختلف، منحنی ROC ابزار بسیار مناسبی است.

- منحنی ROC رگرسیون لجستیک کمترین فاصله را از خط تصادفی دارد.
- شبکه عصبی و Random Forest بهبود محسوسی در AUC نشان می‌دهند.
- XGBoost بالاترین مقدار AUC را دارد که نشان‌دهنده‌ی قدرت تفکیک بسیار بالای این مدل است.

این موضوع بیانگر آن است که XGBoost حتی با تغییر آستانه تصمیم‌گیری نیز عملکرد پایداری ارائه می‌دهد.

#### ۷-۱-۶. مقایسه از منظر تفسیرپذیری<sup>۴۸</sup>

در کاربردهای مالی، تفسیرپذیری مدل اهمیت زیادی دارد:

- رگرسیون لجستیک بالاترین تفسیرپذیری را دارد.
- امکان تحلیل اهمیت ویژگی‌ها را فراهم می‌کند.
- شبکه عصبی و XGBoost از نظر تفسیرپذیری ضعیف‌تر هستند، اما دقت بالاتری ارائه می‌دهند.

این موضوع نشان می‌دهد که انتخاب مدل همواره وابسته به هدف کاربرد است.

#### ۸-۱-۶. مقایسه از منظر هزینه محاسباتی

- رگرسیون لجستیک سریع و کم‌هزینه است.
- شبکه عصبی و Random Forest هزینه محاسباتی متوسطی دارند.
- XGBoost پرهزینه‌ترین الگوریتم از نظر زمان و منابع محاسباتی است.

در محیط‌های عملی، این عامل نیز باید مدنظر قرار گیرد.

---

<sup>۴۸</sup> Interpretability

## ۹-۱. تفسیر نتایج از منظر مسئله ورشکستگی

در مسئله پیش‌بینی ورشکستگی، عدم شناسایی یک شرکت ورشکسته پرهزینه‌ترین نوع خطا محسوب می‌شود. تحلیل نتایج نشان داد که:

• رگرسیون لجستیک در این زمینه ضعف قابل توجهی دارد.

• شبکه عصبی این ضعف را تا حدی جبران کرده است.

• XGBoost و Random Forest کمترین میزان False Negative را داشته‌اند.

• XGBoost بهترین توازن بین شناسایی شرکت‌های ورشکسته و کنترل هشدارهای اشتباه را برقرار کرده است.

از این منظر، نتایج این پژوهش با اصول مدیریت ریسک مالی هم راستا هستند.

## ۱۰-۱. تحلیل نتایج از منظر مدیریت ریسک مالی

در کاربردهای مالی، هزینه‌ی خطاها یکسان نیست. نتایج این پژوهش نشان داد که:

• خطای False Negative (عدم شناسایی ورشکستگی) پرهزینه‌ترین خطا است.

• الگوریتم‌هایی مانند XGBoost و Random Forest توانسته‌اند این نوع خطا را به حداقل برسانند.

بنابراین، از منظر مدیریت ریسک، انتخاب مدلی با Recall بالا حتی به قیمت کاهش Precision، منطقی و قابل دفاع است. این نتیجه اهمیت استفاده از معیارهای ارزیابی مناسب را برجسته می‌کند.

## ۲-۶. تبیین ضعف عملکرد رگرسیون لجستیک

رگرسیون لجستیک به عنوان یک مدل خطی، فرض می‌کند که رابطه‌ی بین متغیرهای ورودی و الگوریتم نسبت شانس خروجی خطی است. در حالی که در داده‌های مالی:

- تأثیر یک نسبت مالی ممکن است وابسته به مقدار نسبت دیگر باشد،
- و روابط علت و معلولی اغلب غیرخطی هستند.

ضعف رگرسیون لجستیک در معیار Recall نشان می‌دهد که این مدل قادر به شناسایی بخش قابل توجهی از شرکت‌های ورشکسته نبوده است. این نتیجه با مبانی نظری این الگوریتم سازگار است و نشان می‌دهد که مدل‌های ساده خطی برای مسائل مالی پیچیده کافی نیستند.

## ۳-۶. تفسیر عملکرد شبکه عصبی مصنوعی

شبکه عصبی مصنوعی با افزودن لایه‌های مخفی و توابع فعال‌سازی غیرخطی، توانست بخشی از ضعف مدل‌های خطی را جبران کند. بهبود معیار Recall در این الگوریتم نشان می‌دهد که شبکه عصبی توانسته است الگوهای پیچیده‌تری را از داده‌ها استخراج کند.

با این حال، نتایج نشان داد که شبکه عصبی نسبت به تنظیم پارامترها حساس است و در صورت انتخاب نامناسب ساختار شبکه، احتمال بیشبرازش وجود دارد. این مسئله بیانگر آن است که اگرچه شبکه‌های عصبی قدرتمند هستند، اما نیازمند طراحی و تنظیم دقیق می‌باشند.

#### ۴-۶. تحلیل موفقیت Random Forest در افزایش پایداری

عملکرد مناسب Ensemble این الگوریتم Random Forest را می‌توان به ماهیت این ترکیب چندین درخت تصمیم باعث شد:

- اثر نویز کاهش یابد
- وابستگی به یک زیرمجموعه خاص از داده‌ها کمتر شود
- و تعمیم‌پذیری مدل افزایش یابد

نتایج نشان داد که Recall و Precision تعادل مناسبی بین Random Forest برقرار کرده است. این ویژگی از منظر کاربردهای مالی اهمیت زیادی دارد، زیرا هم شناسایی شرکت‌های پریسک و هم کنترل هشدارهای اشتباہ در تصمیم‌گیری‌های واقعی اهمیت دارند.

## ۶-۵. تحلیل دلایل برتری XGBoost

XGBoost در این پژوهش بهترین عملکرد کلی را ارائه داد. این برتری را می‌توان از چند منظر تحلیل کرد:

۱. یادگیری مرحله‌ای مبتنی بر خطا:

تمرکز XGBoost بر نمونه‌هایی که در مراحل قبلی به درستی طبقه‌بندی نشده‌اند، باعث بهبود قابل توجه Recall شده است.

۲. کنترل هم‌زمان دقیق و پیچیدگی:

استفاده از Regularization درتابع هدف، از بیش‌برازش جلوگیری کرده است.

۳. انعطاف‌پذیری بالا در مدل‌سازی روابط غیرخطی:

این ویژگی برای داده‌های مالی بسیار حیاتی است.

در مجموع، XGBoost توانسته است به خوبی بین قدرت پیش‌بینی و پایداری تعادل برقرار کند.

## ۶-۱-۶. تحلیل روند کلی تغییر عملکرد الگوریتم‌ها

با بررسی نتایج چهار الگوریتم مورد استفاده، یک روند مشخص و قابل تحلیل مشاهده می‌شود:

- الگوریتم‌های ساده و خطی (رگرسیون لجستیک) عملکرد پایه‌ای و قابل قبولی دارند، اما در شناسایی الگوهای پیچیده دچار محدودیت هستند.
- با حرکت به سمت مدل‌های غیرخطی (شبکه عصبی)، توان مدل در شناسایی روابط پیچیده افزایش یافته است.
- استفاده از روش‌های Random Forest Ensemble باعث افزایش پایداری و کاهش خطاهای بحرانی شده است.
- در نهایت، الگوریتم XGBoost با تمرکز بر یادگیری از خطاهای بحرانی، بهترین عملکرد کلی را ارائه داده است.

این روند نشان می‌دهد که افزایش هوشمندانه‌ی پیچیدگی مدل، در صورت کنترل مناسب بیش‌برازش، می‌تواند به بهبود معنادار عملکرد منجر شود.

## ۶-۲-۶. مقایسه عملی الگوریتم‌ها برای کاربردهای واقعی

اگر نتایج این پژوهش در یک محیط واقعی (مانند بانک یا مؤسسه مالی) مورد استفاده قرار گیرند:

- رگرسیون لجستیک مناسب تحلیل‌های سریع و قابل توضیح است.
- Random Forest گزینه‌ای مناسب برای تعادل بین دقت و تفسیرپذیری است.

• XGBoost بهترین گزینه برای تصمیم‌گیری‌های حساس و پربریسک محسوب

می‌شود، هرچند هزینه محاسباتی بالاتری دارد.

این تحلیل نشان می‌دهد که انتخاب مدل باید وابسته به هدف کاربرد باشد، نه صرفاً

بالاترین عدد Accuracy

### ٦-٣- جمع‌بندی نهایی مقایسه الگوریتم‌ها

بر اساس تحلیل‌های انجام‌شده می‌توان نتیجه گرفت که:

• رگرسیون لجستیک برای تحلیل‌های اولیه مناسب است.

• شبکه عصبی قدرت مدل‌سازی غیرخطی را افزایش می‌دهد.

• Random Forest تعادل مناسبی بین دقت و پایداری ایجاد می‌کند.

• XGBoost بهترین عملکرد کلی را ارائه می‌دهد و مناسب‌ترین گزینه برای

این مسئله است.

### ٦-٤- نتیجه‌گیری فصل ششم

در مجموع، فصل ششم نشان داد که استفاده از الگوریتم‌های پیشرفته یادگیری ماشین می‌تواند نقش بسیار مؤثری در پیش‌بینی ورشکستگی شرکت‌ها ایفا کند. در میان الگوریتم‌های بررسی‌شده، XGBoost بهترین عملکرد کلی را ارائه داده و به عنوان گزینه‌ی نهایی پیشنهادی این پژوهش مطرح می‌شود.

## فصل هفتم

### نتیجه‌گیری کلی، دستاوردهای پژوهش و پیشنهادها

#### ۱-۷. مقدمه فصل هفتم

فصل هفتم به عنوان فصل پایانی این گزارش، به جمع‌بندی نهایی کل پژوهش، استخراج دستاوردهای علمی و عملی، و ارائه پیشنهادهایی برای پژوهش‌های آینده اختصاص دارد. در این فصل تلاش می‌شود تصویری یکپارچه از کل مسیر طی شده، از تعریف مسئله تا تحلیل نتایج، ارائه شود و ارزش علمی و کاربردی کار انجام شده به‌طور شفاف تبیین گردد.

#### ۲-۷. جمع‌بندی مسئله و هدف پژوهش

هدف اصلی این پژوهش، پیش‌بینی ورشکستگی شرکت‌ها با استفاده از الگوریتم‌های مختلف یادگیری ماشین و مقایسه عملکرد آن‌ها بر روی یک دیتاست ثابت بوده است. ورشکستگی شرکت‌ها یکی از مهم‌ترین چالش‌ها در حوزه‌های مالی، اقتصادی و سرمایه‌گذاری محسوب می‌شود، زیرا عدم تشخیص به موقع آن می‌تواند منجر به زیان‌های مالی سنگین برای بانک‌ها، سرمایه‌گذاران و سایر ذی‌نفعان شود.

با توجه به پیچیدگی داده‌های مالی و محدودیت روش‌های سنتی، در این پژوهش تلاش شد از الگوریتم‌های مدرن یادگیری ماشین برای افزایش دقت پیش‌بینی استفاده شود.

### ۷-۳. جمع‌بندی روش انجام پژوهش

در این پژوهش، فرآیند کار به صورت نظاممند و مرحله‌به‌مرحله انجام شد که شامل مراحل زیر بود:

#### ۱. آماده‌سازی داده‌ها:

شامل بررسی کیفیت داده‌ها، پیش‌پردازش، نرمال‌سازی و تفکیک داده‌ها به مجموعه‌های آموزش و آزمون.

#### ۲. انتخاب الگوریتم‌ها:

چهار الگوریتم با سطوح مختلف پیچیدگی انتخاب شدند:

رگرسیون لجستیک (مدل خطی)

شبکه عصبی مصنوعی (مدل غیرخطی)

(Bagging Ensemble) Random Forest

(Boosting Ensemble) XGBoost

#### ۳. پیاده‌سازی و آموزش مدل‌ها:

هر الگوریتم با تنظیمات مناسب آموزش داده شد تا مقایسه‌ای منصفانه بین آن‌ها انجام گیرد.

## ۴. ارزیابی و تحلیل نتایج:

با استفاده از معیارهای Accuracy، Precision، Recall، AUC-ROC، عملکرد مدل‌ها ارزیابی و به صورت عددی و نموداری تحلیل شد.

- Best F1-Score: XGBoost (0.4792)
- Best ROC AUC: XGBoost (0.9486)
- Best Recall: Logistic Regression (0.8182)

## ۴-۷. جمع‌بندی نتایج اصلی پژوهش

نتایج حاصل از فصل پنجم و تحلیل‌های انجام‌شده در فصل ششم نشان داد که عملکرد الگوریتم‌ها تفاوت معناداری با یکدیگر دارد. مهم‌ترین یافته‌های این پژوهش به صورت خلاصه به شرح زیر هستند:

۱. الگوریتم‌های ساده و خطی، اگرچه پیاده‌سازی آسان و تفسیرپذیری بالایی دارند، اما در شناسایی الگوهای پیچیده مالی با محدودیت مواجه‌اند.
۲. استفاده از مدل‌های غیرخطی مانند شبکه عصبی مصنوعی باعث بهبود قابل توجه عملکرد، به ویژه در معیار Recall، شد.

۳. الگوریتم‌های Ensemble مانند Random Forest با کاهش واریانس و افزایش پایداری، عملکرد متعادل‌تری ارائه دادند.

۴. الگوریتم XGBoost با بهره‌گیری از یادگیری تقویتی مبتنی بر خطای Regularization، بهترین عملکرد کلی را در میان الگوریتم‌های بررسی‌شده از خود نشان داد.

## ۷-۵. نتیجه‌گیری نهایی از منظر معیارهای ارزیابی

در این پژوهش نشان داده شد که در مسائل نامتوازن مانند پیش‌بینی ورشکستگی، تمرکز صرف بر معیار Accuracy می‌تواند گمراه‌کننده باشد. معیارهایی نظیر AUC و Recall نقش بسیار مهم‌تری در ارزیابی واقعی عملکرد مدل‌ها دارند.

نتایج نشان داد که:

۱. الگوریتم XGBoost بالاترین Recall را ارائه داده و کمترین میزان False Negative را داشته است. این ویژگی XGBoost را به مناسب‌ترین گزینه برای کاربردهای مالی حساس تبدیل می‌کند.
۲. Random Forest نیز به عنوان گزینه‌ای متعادل میان دقت، پایداری و تفسیرپذیری قابل توجه است.

## ۷-۶. دستاوردهای علمی پژوهش

از منظر علمی، این پژوهش چند دستاورد مهم به همراه داشته است:

۱. تأیید برتری مدل‌های غیرخطی و Ensemble در داده‌های مالی
۲. نشان دادن نقش حیاتی انتخاب معیار ارزیابی مناسب در مسائل نامتوازن
۳. ارائه‌ی یک چارچوب مقایسه‌ای منسجم بین الگوریتم‌های مختلف یادگیری ماشین
۴. تبیین ارتباط میان ماهیت داده‌های مالی و عملکرد الگوریتم‌ها

## ۷-۷. دستاوردهای عملی و کاربردی

نتایج این پژوهش دارای پیامدهای عملی مهمی نیز هستند، از جمله:

کمک به طراحی سیستم‌های هشدار زودهنگام ورشکستگی

پشتیبانی از تصمیم‌گیری‌های اعتباری در بانک‌ها و مؤسسات مالی

کاهش ریسک سرمایه‌گذاری از طریق شناسایی بهموقع شرکت‌های پرریسک

بهبود فرآیندهای ارزیابی سلامت مالی شرکت‌ها

## ۸-۷. نتیجه‌گیری نهایی

در پایان می‌توان گفت که این پژوهش نشان داد استفاده از الگوریتم‌های پیشرفته یادگیری ماشین، بهویژه XGBoost، می‌تواند نقش بسیار مؤثری در پیش‌بینی ورشکستگی شرکت‌ها ایفا کند. مقایسه‌ی انجام‌شده میان الگوریتم‌ها نشان داد که انتخاب مدل باید بر اساس ماهیت داده‌ها، هدف کاربرد و هزینه‌ی خطاهای انجام شود، نه صرفاً بالاترین دقت عددی.

این گزارش با ارائه تحلیل‌های جامع، تلاش کرده است پلی میان مبانی نظری یادگیری ماشین و کاربردهای واقعی مالی ایجاد کند و می‌تواند به عنوان مبنایی برای پژوهش‌های آتی و کاربردهای عملی مورد استفاده قرار گیرد.