LUISS

Dipartimento di Impresa e Management
Corso di laurea in Marketing

Cattedra: Machine Learning in Marketing

Applicazione di modelli di intelligenza artificiale generativa nell'analisi di scenari complessi.

Un approccio basato sull'arricchimento della conoscenza storica disponibile attraverso la generazione e il filtraggio di dati sintetici.

LUIGI LAURA — RELATORE

FRANCESCO SALATE SANTONE — CORRELATORE

VECCHIONE DOMENICO, 788261 — CANDIDATO

Anno Accademico 2024/2025

Indice dei contenuti

Abstract

L'intelligenza artificiale generativa rappresenta un'opportunità innovativa per migliorare la precisione e l'efficienza della valutazione del rischio di credito, un processo critico per le istituzioni finanziarie. Questo progetto di tesi studia l'uso di modelli generativi per creare scenari creditizi realistici e diversificati basati su dati storici e attuali. Tali scenari permettono di analizzare un ampio spettro di esiti, migliorando la comprensione del rischio e la resilienza dei portafogli. Dopo la selezione di un dataset rappresentativo, un'analisi esplorativa dei dati e la scelta dell'architettura più adatta, il modello verrà addestrato per apprendere le distribuzioni e le dipendenze tra le variabili. Una volta validata la qualità dei dati generati, verrà implementato un filtro per produrre scenari specifici, che saranno valutati attraverso modelli di rischio esistenti per analizzare il comportamento di fenomeni di interesse. L'obiettivo finale è ottenere una valutazione del rischio più accurata, robusta e interpretabile, con applicazioni nella simulazione di eventi estremi, analisi personalizzate e verifica dell'affidabilità dei modelli attuali.

1 Introduzione e obiettivi

L'intelligenza artificiale generativa rappresenta una frontiera innovativa che sta rivoluzionando numerosi settori, e tra questi offre nuove e promettenti opportunità per migliorare la precisione e l'efficienza del processo di valutazione del rischio di credito, un'attività cruciale per le banche, le istituzioni finanziarie e altri enti che operano nel campo del credito, dove la capacità di prevedere con accuratezza la probabilità di insolvenza di un debitore può fare la differenza tra profitti consistenti e perdite significative. Questo processo, tradizionalmente basato su modelli statistici e dati storici, si trova spesso a dover affrontare la complessità di un mondo finanziario in continua evoluzione, caratterizzato da variabili economiche mutevoli, comportamenti dei consumatori imprevedibili e eventi rari ma di grande impatto, come crisi economiche o pandemie, che possono mettere a dura prova la resilienza dei portafogli creditizi.

L'introduzione dell'IA generativa in questo contesto (Kasztelnik, 2024) consente di superare alcune delle limitazioni dei metodi convenzionali, ampliando la capacità di analisi attraverso la creazione di scenari creditizi ipotetici ma diversificati e ragionevolmente realistici, che permettono di esplorare un ventaglio molto più ampio di possibili esiti rispetto a quanto offerto dai soli dati storici disponibili.

Questi scenari non sono semplici proiezioni lineari del passato, ma rappresentano simulazioni sofisticate che tengono conto delle interdipendenze tra molteplici variabili – come il reddito, il debito, la storia creditizia, l'età, l'occupazione e fattori macroeconomici – e delle loro possibili evoluzioni in contesti diversi, offrendo così una visione più completa e dinamica del rischio. In questo modo, si raffinano le possibilità di comprendere e valutare il rischio di credito, non solo in termini di probabilità di default, ma anche rispetto alla capacità di un portafoglio di resistere a shock inattesi o di rispondere a situazioni specifiche, come l'ingresso di un nuovo segmento di clientela o l'impatto di una politica monetaria restrittiva.

L'obiettivo ultimo è quello di trasformare un processo che, pur sofisticato, può risultare rigido o limitato dalla quantità e qualità dei dati storici, in uno strumento più flessibile, predittivo e adattabile, capace di supportare decisioni finanziarie strategiche con un grado di precisione e robustezza mai raggiunto prima.

L'idea di base che guida questa applicazione dell'IA generativa è quella di sviluppare un modello capace di generare sinteticamente scenari di rischio creditizio realistici e diversificati, utilizzando come punto di partenza un insieme di dati storici e attuali che riflettano il comportamento passato dei debitori e le condizioni economiche in cui tali comportamenti si sono manifestati, integrandoli con tecniche avanzate di apprendimento automatico per estrapolare e creare nuove situazioni plausibili.

Questo modello non si limita a replicare i dati osservati, ma apprende le distribuzioni statistiche sottostanti e le relazioni tra le variabili, generando scenari che possono includere combinazioni di eventi non presenti nel dataset originale, come una recessione improvvisa accompagnata da un aumento dei tassi di interesse o una crescita economica rapida che influisce sulla capacità di rimborso dei prestiti. In pratica, il processo di generazione ha lo scopo di creare quantità di dati di sintesi realistici e significativamente descrittivi di un fenomeno di interesse, per poter consentire attraverso lo studio delle distribuzioni statistiche risultanti, l'osservazione di situazioni non adeguatamente rappresentate nei dati storici originari.

I dati rappresentativi di uno specifico scenario possono essere generati in maniera tale da sovrastimare il fenomeno oggetto di studio e consentire una migliore osservazione dei fattori di maggiore interesse ad esso correlati come una sorta di microscopio statistico che permetta di fare un'operazione di zoom su specifici elementi della distribuzione (ad es. per capire cosa potrebbe succedere in determinate situazioni sui ricavi). Una volta generati questi scenari, essi possono essere analizzati utilizzando uno dei modelli di valutazione del rischio di credito già esistenti – come i modelli di regressione logistica, gli score di credito tipo FICO o i sistemi basati su alberi decisionali – per ottenere una stima più precisa e robusta del rischio associato a ciascun caso simulato.

Questo approccio consente non solo di migliorare la capacità predittiva del processo, ma anche di rendere le valutazioni più trasparenti e motivabili, un aspetto cruciale in un settore altamente regolamentato dove le decisioni devono essere giustificate sia agli stakeholder interni che alle autorità di vigilanza, come la Banca Centrale Europea o l'Autorità Bancaria Europea. Ad esempio, una banca potrebbe utilizzare il modello generativo per simulare come un portafoglio di mutui reagirebbe a un aumento della disoccupazione del 5%, valutando non solo la probabilità media di default, ma anche

gli scenari peggiori (tail risk) che potrebbero emergere in una coda della distribuzione, fornendo così una stima più completa del rischio rispetto a un'analisi basata solo sui dati passati. Inoltre, la possibilità di generare scenari diversificati permette di testare la sensibilità dei modelli di rischio a variazioni specifiche delle condizioni economiche o demografiche, offrendo una visione più granulare e personalizzata che può essere adattata a diversi tipi di clientela, come piccole imprese, famiglie o grandi aziende.

Le applicazioni pratiche di questo approccio sono molteplici e spaziano dalla simulazione di eventi estremi alla personalizzazione delle analisi, fino alla spiegazione delle decisioni prese dai modelli di rischio, rispondendo così a diverse esigenze delle istituzioni finanziarie.

Una delle possibilità più interessanti è la simulazione di eventi estremi, come una crisi finanziaria globale o un crollo del mercato immobiliare, per valutare la resilienza dei portafogli creditizi in condizioni di stress, un'attività che le banche sono già tenute a svolgere nell'ambito degli stress test regolamentari, ma che con l'IA generativa può essere condotta con una varietà e una profondità di scenari molto maggiori rispetto ai metodi tradizionali, che spesso si limitano a poche combinazioni predefinite di variabili macroeconomiche. Ad esempio, un modello generativo potrebbe simulare migliaia di scenari in cui il PIL diminuisce del 3%, i tassi di interesse salgono al 5% e la disoccupazione raggiunge il 10%, variando leggermente questi parametri per coprire un'ampia gamma di possibilità e osservando come tali variazioni influenzano il tasso di default di un portafoglio di prestiti al consumo, permettendo alla banca di identificare i punti di vulnerabilità e di pianificare strategie di mitigazione, come l'aumento delle riserve di capitale o la riduzione dell'esposizione a determinati segmenti di mercato.

Un'altra applicazione è l'analisi di scenari personalizzati, che consente di generare dati sintetici specifici per un cliente o un gruppo di clienti, ad esempio simulando il comportamento creditizio di una nuova categoria di debitori (come i lavoratori autonomi in un'economia digitale) per cui i dati storici sono scarsi, fornendo così una base per valutare il rischio senza dover attendere anni di raccolta dati reali. Inoltre, l'IA generativa può migliorare la spiegabilità dei modelli di rischio: generando scenari e mostrando come diverse combinazioni di variabili influenzano la probabilità di default, si possono offrire spiegazioni più chiare e intuitive agli analisti o ai regolatori,

superando il problema della “scatola nera” tipico di alcuni modelli di machine learning complessi, come le reti neurali profonde.

Un’ulteriore ricaduta positiva è la possibilità di utilizzare i dati sintetici generati per valutare l’accuratezza e la robustezza dei modelli di valutazione del rischio attualmente in uso, confrontando le loro previsioni su scenari realistici ma artificiali con i risultati attesi, un processo che permette di identificare eventuali debolezze o biases nei modelli esistenti e di migliorarne le prestazioni attraverso un ciclo di retroazione continuo.

2 Metodologia e strumenti

Implementare una strategia di valutazione e stima del rischio finanziario basata sulla generazione di dati di sintesi attraverso modelli di intelligenza artificiale generativa, richiede un approccio metodologico strutturato e rigoroso. Di seguito si descrive il percorso metodologico che sottende l'idea alla base dello studio in questione, dettagliandone poi i passi principali.

1. Definizione degli Obiettivi e del Contesto

Identificare gli obiettivi: Determinare quali rischi finanziari si intendono valutare (es. rischio di mercato, di credito, operativo, di liquidità).

Contesto normativo e di mercato: Considerare i requisiti normativi (es. Basel III, Solvency II) e le specificità del mercato di riferimento.

Definizione delle metriche di rischio: Scegliere le metriche appropriate (es. Value at Risk - VaR, Expected Shortfall - ES, stress test).

2. Raccolta e Analisi dei Dati Storici

Raccolta dati: Ottenere dati finanziari storici rilevanti (es. Dati sulla stabilità finanziaria, Costi, Ricavi).

Pulizia e pre-processing: Rimuovere outliers, gestire dati mancanti e normalizzare i dati.

Analisi esplorativa: Studiare le caratteristiche dei dati (distribuzioni, correlazioni, volatilità).

3. Selezione del Modello Generativo

Scelta del modello: Valutare modelli generativi come GAN (Generative Adversarial Networks), VAE (Variational Autoencoders).

Addestramento del modello: Addestrare il modello sui dati storici, assicurandosi che catturi le proprietà statistiche e le dipendenze temporali dei dati reali.

Validazione del modello: Verificare la qualità dei dati sintetici generati confrontandoli con i dati reali, come ad esempio attraverso il test di Kolmogorov-Smirnov (Berger et al., 2014), o effettuando un'analisi delle correlazioni (Corder et al., 2014).

4. Generazione di Dati Sintetici con Filtraggio delle Situazioni di Interesse

Creazione di scenari: Generare dati sintetici per simulare scenari di mercato plausibili, inclusi scenari estremi (stress test).

Diversificazione dei dati: Assicurarsi che i dati sintetici coprano un'ampia gamma di possibili condizioni di mercato.

Definizione dei criteri di filtraggio: Identificare le situazioni di interesse (es. crolli di mercato, picchi di volatilità, correlazioni anomale).

Applicazione di filtri: Utilizzare tecniche di filtraggio (es. soglie statistiche, clustering) per isolare i dati rilevanti.

Analisi delle situazioni filtrate: Studiare le caratteristiche delle situazioni selezionate per comprendere i driver del rischio.

5. Analisi e Interpretazione dei risultati

*Analis*i: Analizzare i risultati della valutazione del rischio e identificare i principali fattori di rischio.

Interpretazione: Formulare una chiara e dettagliata interpretazione dei fenomeni osservati, evidenziando le implicazioni pratiche e suggerendo le azioni necessarie per mitigare i rischi identificati.

2.1 Raccolta dei dati storici

Il processo alla base dell'attività di raccolta dei dati storici di interesse inizia con l'identificazione e la raccolta di un dataset rappresentativo del problema del rischio di credito, necessario per addestrare il modello, un passaggio fondamentale che richiede attenzione alla qualità e alla completezza dei dati, poiché il successo del modello generativo dipende dalla sua capacità di apprendere pattern significativi da un insieme

di informazioni che rispecchi fedelmente la realtà del contesto creditizio di interesse. Questo dataset potrebbe includere variabili come il punteggio di credito dei debitori, il rapporto debito/reddito, la storia dei pagamenti, il tipo di prestito (mutuo, prestito personale, carta di credito), oltre a fattori esterni come il tasso di disoccupazione, l'inflazione e i tassi di interesse, raccolti su un periodo sufficientemente lungo da catturare cicli economici diversi, ad esempio gli ultimi 10-15 anni, includendo sia fasi di crescita che di recessione.

Una volta raccolto il dataset, ed effettuatane una prima attività di pulizia, che si concretizza nell'eliminazione di record associati a dati incongruenti o incompleti, si procede con un'analisi esplorativa dei dati (EDA, Exploratory Data Analysis), un passaggio essenziale per comprendere le caratteristiche statistiche del dataset, come la distribuzione di ciascuna variabile (ad esempio, se il reddito segue una distribuzione log-normale o se l'età dei debitori è multimodale), le correlazioni tra le variabili (ad esempio, tra reddito e probabilità di default), e la presenza di eventuali anomalie o valori mancanti che potrebbero influire sull'addestramento del modello. Questa analisi non è solo un esercizio preliminare, ma serve a determinare la scelta dell'architettura di IA generativa più adatta per la sintesi dei dati: tra le opzioni disponibili, le reti generative antagoniste (GAN) e gli autoencoder (AE) sono due delle candidate principali, ciascuna con punti di forza specifici che dipendono dalla tipologia e dalla complessità dei dati. Le GAN, come il CTGAN, sono particolarmente indicate per dataset tabulari con variabili miste (continue e categoriche) e relazioni non lineari, grazie alla loro capacità di apprendere distribuzioni complesse attraverso la competizione tra generatore e discriminatore, mentre gli autoencoder, che comprimono i dati in una rappresentazione latente e poi li ricostruiscono, possono essere più efficienti per dataset con struttura più semplice o quando l'obiettivo è preservare fedelmente le caratteristiche principali dei dati originali senza necessariamente generare grande diversità.

2.2 Selezione del modello generativo

La scelta tra queste architetture non è banale e richiede una valutazione attenta: se il dataset è altamente eterogeneo e contiene dipendenze complesse, come quelle tra variabili economiche e comportamentali, una GAN potrebbe essere preferibile; se

invece i dati sono più uniformi e l'obiettivo è una ricostruzione accurata, un AE potrebbe essere sufficiente.

Una volta selezionata l'architettura, il modello viene addestrato sui dati storici, un processo che implica l'ottimizzazione dei parametri della rete neurale affinché il generatore apprenda le distribuzioni marginali delle variabili (ad esempio, la distribuzione del reddito) e le loro dipendenze (ad esempio, come il reddito si correla al debito), utilizzando tecniche come la discesa del gradiente stocastico su un insieme di epoche di addestramento, con l'obiettivo di minimizzare una funzione di perdita che bilancia la qualità dei dati generati e la capacità del discriminatore di distinguerli dai dati reali.

Dopo l'addestramento, la qualità dei dati sintetici generati dal modello deve essere controllata rigorosamente, un passaggio critico per garantire che gli scenari prodotti siano effettivamente realistici e utilizzabili per la valutazione del rischio, e che non contengano artefatti o distorsioni che potrebbero compromettere i risultati downstream. Questo controllo si basa su metriche standard nel campo dell'IA generativa, come la somiglianza delle distribuzioni marginali tra dati reali e sintetici (misurata, ad esempio, con il test di Kolmogorov-Smirnov o la divergenza di Kullback-Leibler – o KL), la fedeltà delle correlazioni tra variabili (valutata con la matrice di correlazione di Pearson o Spearman), e l'utilità pratica dei dati sintetici quando sottoposti a un modello di valutazione del rischio, confrontando le prestazioni predittive sui dati generati con quelle sui dati reali. Ad esempio, si potrebbe verificare se la distribuzione dell'età nei dati sintetici segue lo stesso andamento multimodale dei dati storici, o se la correlazione negativa tra reddito e probabilità di default è preservata, assicurando che il modello non abbia appiattito o distorto pattern importanti.

2.3 Generazione di Dati Sintetici con Filtraggio delle situazioni di interesse

Se queste metriche indicano una buona qualità, si procede alla fase successiva, che consiste nel realizzare un filtro a valle della generazione dei dati per selezionare o creare scenari desiderati, un processo che permette di sfruttare la flessibilità del modello generativo per rispondere a esigenze specifiche delle istituzioni finanziarie.

Questo filtro potrebbe essere implementato come un meccanismo condizionale nel caso di un CTGAN, dove si specifica una condizione (ad esempio, "genera solo scenari con disoccupazione superiore al 7%") e il generatore produce campioni che soddisfano tale criterio, oppure come un post-processing dei dati generati, selezionando manualmente gli scenari che corrispondono a determinati intervalli di parametri chiave, come il tasso di interesse o il livello di indebitamento. In ogni caso, i principali meccanismi utilizzabili per implementare le operazioni di filtraggio potrebbero essere:

- *Soglie basate su deviazioni standard*: Filtrare i dati che superano un certo numero di deviazioni standard dalla media (es. dati oltre $\pm 3\sigma$).
- *Percentili*: Selezionare i dati nei percentili estremi (es. 1° o 99° percentile).
- Test di normalità: Identificare dati che si discostano significativamente da una distribuzione normale (es. test di Shapiro-Wilk o Jarque-Bera).
- Analisi della coda della distribuzione: Concentrarsi sulle code delle distribuzioni per identificare eventi rari ma ad alto impatto.
- *Rilevamento di picchi*: Identificare picchi di volatilità o rendimenti anomali utilizzando tecniche come la trasformata di Fourier o filtri passa-basso.
- *Analisi di cambiamento di regime*: Utilizzare modelli come Hidden Markov Models (HMM) per identificare cambiamenti nei regimi di mercato.
- *Filtri di Kalman*: Applicare filtri di Kalman per stimare stati latenti e identificare deviazioni significative.
- *Analisi di correlazione*: Identificare correlazioni insolite o inversioni di tendenza.
- *Copule*: Utilizzare copule per modellare le dipendenze tra variabili e identificare situazioni estreme.

Il modello addestrato diventa così uno strumento dinamico, capace di generare un ampio insieme di scenari creditizi futuri, variando parametri fondamentali in base alla tipologia di soggetti e agli attributi rappresentati nel dataset di riferimento: ad esempio, per un dataset di prestiti personali, si potrebbero modificare il reddito medio dei debitori, il tasso di default atteso o il livello di inflazione, creando scenari che coprono sia situazioni ordinarie che casi estremi, come una crisi economica che colpisce in modo sproporzionato i lavoratori a basso reddito.

2.4 Analisi e Interpretazione dei risultati

Questi scenari generati possono poi essere analizzati dal punto di vista dell'osservazione del comportamento statistico dei principali fattori caratterizzanti oppure sottoposti a uno qualsiasi dei modelli di valutazione del rischio esistenti, un passaggio che permette di valutare l'impatto delle condizioni simulate sulla capacità di discriminazione tra soggetti o situazioni a basso e alto rischio, fornendo una misura concreta della robustezza del portafoglio creditizio e della validità dei modelli di rischio utilizzati. Ad esempio, una banca potrebbe utilizzare un modello logistico per calcolare la probabilità di default su ciascun scenario sintetico, osservando come varia la percentuale di debitori classificati come "ad alto rischio" quando i tassi di interesse aumentano o il PIL diminuisce, e confrontando questi risultati con le soglie di rischio accettabili definite internamente o dai regolatori. Questo processo non solo migliora la precisione della valutazione del rischio, ma consente anche di testare la sensibilità dei modelli a diverse configurazioni di variabili, identificando eventuali punti deboli: se un modello di rischio tende a sovrastimare il rischio in scenari di recessione simulata, ad esempio, ciò potrebbe indicare una dipendenza eccessiva da variabili macroeconomiche o una mancanza di adattabilità a situazioni estreme, suggerendo la necessità di ricalibrazione o di integrazione con ulteriori dati. Inoltre, l'ampia gamma di scenari generati permette di esplorare situazioni che non si sono ancora verificate nei dati storici, ma che potrebbero plausibilmente accadere in futuro, come l'impatto di una nuova regolamentazione sui mutui o l'ingresso di una nuova tecnologia che altera i comportamenti di pagamento dei consumatori, offrendo alle istituzioni finanziarie un vantaggio competitivo nella pianificazione strategica e nella gestione del rischio.

Un altro aspetto significativo è la possibilità di utilizzare i dati sintetici per migliorare la formazione degli analisti o per sviluppare nuovi modelli di rischio, addestrandoli su un insieme più ricco e diversificato di casi rispetto a quello disponibile nei dati reali, un approccio che potrebbe ridurre il rischio di overfitting e aumentare la generalizzabilità delle previsioni. In definitiva, l'intelligenza artificiale generativa, attraverso modelli come il CTGAN o gli AE, trasforma la valutazione del rischio di credito da un processo statico e retrospettivo in uno dinamico e prospettico, capace di anticipare e rispondere a un mondo finanziario sempre più complesso e incerto, con

implicazioni che vanno oltre la semplice ottimizzazione tecnica e toccano la capacità delle istituzioni di navigare con successo le sfide del futuro.

L'intero framework metodologico sviluppato in questo lavoro di tesi è schematizzato nella Figura 1 seguente.

Figura 1: La metodologia di analisi proposta in questo lavoro di tesi.

3 Implementazione “Proof of Concept”

L’approccio adottato per la valutazione del rischio creditizio si basa sulla disponibilità di funzionalità avanzate di generazione di dati sintetici tramite tecniche di intelligenza artificiale generativa, un passaggio che amplia notevolmente le sue potenzialità e lo rende uno strumento indispensabile in contesti dove i dati scarseggiano o sono incompleti. Questo processo non consiste in una semplice duplicazione o manipolazione casuale dei dati esistenti, ma utilizza algoritmi sofisticati, come reti generative antagoniste (GAN), Autoencoders o modelli probabilistici, per creare nuovi campioni che mantengono le proprietà statistiche del dataset originale, garantendo che i dati sintetici siano coerenti con le distribuzioni, le correlazioni e i pattern osservati nei dati reali, senza però replicarli identicamente, il che offre anche un vantaggio in termini di privacy. La disponibilità di una quantità adeguata di dati in grado di rappresentare e descrivere certi fenomeni ci darà la possibilità di rappresentare e costruire modelli che ci permetteranno di studiare in maggiore dettaglio, comprendere e prevedere i fenomeni stessi. Inoltre, la possibilità di filtrare i dati generati in accordo a specifiche situazioni di interesse ci permetterà di dare maggiore enfasi a determinati fenomeni oggetto di attenzione, realizzando una sorta di microscopio statistico che ci può offrire una visione estremamente dettagliata di determinati fenomeni anche quando l’evidenza storica di tali fenomeni ci fornisce informazioni non sempre di estremo dettaglio. Questo nasce dalla capacità intrinseca dei modelli generativi di comprendere le dinamiche fondamentali dei processi per poi essere in grado di replicare gli stessi su scala differente.

In questo lavoro, allo scopo di realizzare una “proof of concept” della metodologia proposta, per la generazione di dati sintetici attraverso modelli di generative AI abbiamo usato una libreria disponibile liberamente, la libreria SDV (Synthetic Data Vault) descritta originariamente in (Patki et al., 2016).

I dati storici utilizzati nel contesto dell’analisi sono stati ricavati dal dataset pubblico American Bankruptcy Dataset, liberamente disponibile su GitHub.

3.1 Il framework SDV

SDV, o Synthetic Data Vault, è una libreria open-source scritta in Python, progettata per generare dati sintetici che replicano le proprietà statistiche e strutturali di dataset

reali, senza comprometterne la privacy o violare normative come il GDPR. Pensata per data scientist, ricercatori e aziende, questa libreria si rivela preziosa in contesti dove i dati reali sono sensibili, limitati o difficili da ottenere, offrendo un'alternativa sicura e versatile. L'obiettivo principale di SDV è consentire la creazione di dataset artificiali di alta qualità, utili per analisi, test di software, machine learning o simulazioni, mantenendo riservate le informazioni originali. Grazie alla sua capacità di imitare correlazioni, distribuzioni e schemi dei dati reali, SDV si posiziona come uno strumento ideale per bilanciare l'accesso ai dati con la protezione della privacy, un'esigenza sempre più sentita in un mondo regolato da leggi stringenti sulla protezione dei dati.

La libreria si distingue per alcune caratteristiche chiave che ne fanno un'opzione potente e flessibile. È ottimizzata per dati tabulari, come quelli in formato CSV o pandas DataFrame, e utilizza una combinazione di modelli statistici e tecniche di deep learning, come i GAN (Generative Adversarial Networks), o gli Autoencoders, per generare dati realistici. SDV gestisce con facilità variabili numeriche, categoriche, date e persino dati mancanti, permettendo agli utenti di personalizzare i modelli per adattarli alle specificità dei propri dataset, controllando vincoli e relazioni tra le variabili. Inoltre, offre strumenti per valutare la qualità dei dati sintetici, confrontandoli con quelli reali in termini di somiglianze statistiche e utilità. Un esempio pratico del suo valore potrebbe essere un'azienda sanitaria che usa SDV per generare dati sintetici di pazienti, utilizzabili per sviluppare modelli predittivi o condividerli con partner esterni senza esporre informazioni sensibili.

Il funzionamento di SDV si basa su un processo chiaro e intuitivo, che parte dall'analisi dei dati reali forniti dall'utente per identificare distribuzioni, correlazioni e schemi tra le variabili, un passaggio cruciale per garantire che i dati sintetici siano rappresentativi. Successivamente, la libreria addestra un modello – come GaussianCopula, CTGAN o TVAE – scelto in base alla complessità del dataset, per "imparare" le sue caratteristiche. Una volta completato l'addestramento, SDV genera un nuovo dataset sintetico, pronto per essere salvato o utilizzato, con un'interfaccia semplice e integrata con pandas. Ad esempio, con poche righe di codice Python, come riportato nel seguito:

```
from sdv.tabular import CTGAN
model = CTGAN()
model.fit(real_data)
synthetic_data = model.sample(1000)
```

è possibile creare mille righe di dati sintetici, rendendo la libreria accessibile anche a chi ha familiarità con gli strumenti di data science più comuni.

Le applicazioni di SDV sono numerose e spaziano dal machine learning, dove si possono creare dataset per addestrare modelli in assenza di dati reali sufficienti, al testing di applicazioni e database, fino alla ricerca, dove i dataset sintetici possono essere condivisi tra team o istituzioni senza violare la privacy. È utile anche per simulazioni, come l'analisi di sistemi finanziari o logistici, grazie alla capacità di generare scenari realistici. Tuttavia, SDV non è priva di limiti: i dati sintetici potrebbero non catturare perfettamente anomalie rare o relazioni estremamente complesse, e la loro qualità dipende dalla scelta del modello e dalla quantità di dati reali disponibili per l'addestramento. Nonostante ciò, la libreria rappresenta una soluzione innovativa e in continua evoluzione, supportata dalla comunità open-source, che la rende sempre più rilevante nel panorama della data science. Per approfondimenti, la documentazione ufficiale su GitHub o il sito di SDV offrono risorse dettagliate per esplorarne tutte le potenzialità.

3.1.1 Gaussian Copula

Il metodo Gaussian Copula è uno degli strumenti principali impiegati dalla libreria Synthetic Data Vault (SDV) per generare dati sintetici, un approccio che si basa su una solida teoria statistica (Durante et al., 2016) per modellare e replicare le relazioni tra variabili in un dataset, permettendo di creare campioni artificiali che mantengono le proprietà statistiche dei dati reali senza copiarli direttamente, un aspetto cruciale per garantire privacy e utilità nei contesti di analisi dei dati. Per comprendere il Gaussian Copula, dobbiamo prima introdurre il concetto generale di copula, una funzione matematica che collega le distribuzioni marginali di singole variabili (cioè le distribuzioni di ciascuna colonna di un dataset tabulare) alla loro distribuzione congiunta, catturando così le dipendenze tra di esse, indipendentemente dalla forma specifica delle marginali stesse. Immaginiamo un dataset con due variabili, come l'età e il reddito di un gruppo di persone: ciascuna variabile ha una propria distribuzione (l'età potrebbe essere approssimativamente uniforme tra 20 e 60 anni, il reddito

potrebbe seguire una distribuzione log-normale), ma ciò che interessa in molti casi non è solo la forma di queste distribuzioni singole, ma come età e reddito variano insieme, ad esempio se età più avanzate tendono a corrispondere a redditi più alti. Le copule risolvono questo problema separando la modellazione delle marginali dalla struttura di dipendenza, e il Gaussian Copula, in particolare, assume che tale struttura possa essere descritta utilizzando una distribuzione normale multivariata, un'ipotesi che semplifica i calcoli e si adatta bene a molti dataset reali, specialmente quelli con correlazioni lineari moderate. In SDV, il Gaussian Copula è implementato come un modello tabulare che analizza un dataset reale per apprendere sia le distribuzioni marginali delle variabili sia la matrice di correlazione che descrive le loro interazioni, per poi generare un nuovo dataset sintetico che riflette queste caratteristiche. Questo approccio è particolarmente utile per dati strutturati in tabelle, come quelli in formato CSV, dove le colonne possono includere variabili numeriche continue (come il reddito), categoriche (come il genere) o persino dati con valori mancanti, e il metodo si distingue per la sua capacità di gestire questa eterogeneità preservando le relazioni tra le variabili, rendendolo una scelta ideale per applicazioni che vanno dalla simulazione di scenari aziendali alla protezione della privacy in contesti sensibili come la sanità.

Il funzionamento del Gaussian Copula in SDV si articola in due fasi principali: l'apprendimento dal dataset reale e la generazione dei dati sintetici, un processo che si basa su una combinazione di trasformazioni statistiche e campionamento da una distribuzione normale multivariata, e che possiamo descrivere passo per passo per chiarirne la logica matematica sottostante. Nella fase di apprendimento, SDV analizza il dataset reale per stimare le distribuzioni marginali di ogni variabile: per le variabili continue, queste possono essere modellate con distribuzioni parametriche (come una normale o una log-normale) o stimate empiricamente usando una funzione di densità cumulativa (CDF); per le variabili categoriche, si calcola la frequenza di ogni categoria e si trattano come distribuzioni discrete. Una volta determinate le marginali, il passo successivo è trasformare i dati originali in una scala uniforme tra 0 e 1, applicando la CDF di ciascuna variabile ai suoi valori: ad esempio, un'età di 30 anni potrebbe essere mappata al valore 0,4 se il 40% della popolazione ha meno di 30 anni, un processo noto come trasformazione in "quantili". Questi valori uniformi vengono poi ulteriormente trasformati in una scala normale standard (media 0, varianza 1)

utilizzando l'inversa della CDF della distribuzione normale standard, nota come funzione quantile o funzione di probit, ottenendo così un insieme di variabili che seguono una distribuzione normale univariata. A questo punto, SDV calcola la matrice di correlazione di Pearson di queste variabili normalizzate, che rappresenta la struttura di dipendenza lineare tra di esse, catturando come le variabili covariano una volta rimosse le specificità delle loro distribuzioni marginali. Nella fase di generazione, SDV campiona nuovi punti da questa distribuzione normale multivariata, li ritrasforma in uniformi e infine applica le inverse delle CDF marginali stimate per ottenere valori sintetici nelle scale originali delle variabili, garantendo che i dati generati rispettino sia le distribuzioni marginali che la struttura di correlazione del dataset reale.

L'implementazione del Gaussian Copula in SDV è progettata per essere pratica e accessibile, integrandosi perfettamente con i dati tabulari e offrendo opzioni per gestire casi complessi come valori mancanti o distribuzioni non standard, un aspetto che lo rende uno dei modelli più robusti della libreria per generare dati sintetici realistici. In pratica, quando un utente carica un dataset in SDV e seleziona il modello GaussianCopula (ad esempio, tramite il comando Python

```
from sdv.tabular import GaussianCopula
model = GaussianCopula()
model.fit(data))
```

la libreria esegue automaticamente i passaggi descritti: stima le marginali, calcola la matrice di correlazione, e genera un nuovo dataset sintetico con lo stesso numero di righe o un numero specificato dall'utente, un processo che può essere completato in poche righe di codice grazie all'integrazione con pandas. Per le variabili categoriche, SDV le trasforma internamente in rappresentazioni numeriche (ad esempio, usando codifiche ordinali) prima di applicare il modello, e ritrasforma i valori sintetici generati in categorie, preservando le proporzioni osservate nei dati reali; per i valori mancanti, il sistema può imputarli o trattarli come una categoria speciale, a seconda delle impostazioni. I vantaggi del Gaussian Copula sono numerosi: è computazionalmente efficiente per dataset di dimensioni moderate, cattura bene le dipendenze lineari tra variabili, e produce dati sintetici che possono essere utilizzati per analisi statistiche o addestramento di modelli senza compromettere la privacy, un aspetto cruciale in settori regolamentati come la finanza o la sanità. Tuttavia, il metodo ha anche dei limiti: l'assunzione di una struttura di dipendenza gaussiana significa che non è ideale per

catturare relazioni non lineari complesse o code pesanti nelle distribuzioni, situazioni in cui modelli più avanzati come CTGAN (basato su reti generative antagoniste) potrebbero essere più appropriati; inoltre, se le distribuzioni marginali sono stimate male o il dataset reale è troppo piccolo, la qualità dei dati sintetici può risentirne. Nonostante queste limitazioni, il Gaussian Copula rimane una scelta popolare in SDV per la sua semplicità e affidabilità, specialmente per dataset con correlazioni moderate e variabili ben comportate, offrendo un equilibrio tra precisione e facilità d'uso che lo rende uno strumento prezioso per generare dati sintetici rappresentativi in una vasta gamma di applicazioni, dalla simulazione di scenari di mercato alla creazione di dataset di test per validare algoritmi.

3.1.2 CTGAN

Il CTGAN, o Conditional Tabular Generative Adversarial Network, è un modello avanzato di intelligenza artificiale generativa implementato nella libreria Synthetic Data Vault (SDV) per generare dati sintetici tabulari, un'evoluzione specifica dei tradizionali GAN progettata per affrontare le sfide uniche poste dai dataset strutturati, come quelli in formato CSV, che contengono una miscela di variabili continue (ad esempio, età o reddito) e categoriche (ad esempio, genere o stato civile), spesso con distribuzioni complesse e relazioni non lineari tra le colonne.

I GAN classici, introdotti in (Goodfellow, 2014), sono stati una rivoluzione nel campo della generazione di dati, noti soprattutto per la loro capacità di creare immagini realistiche, come volti umani o paesaggi, addestrando due reti neurali in competizione: un generatore, che produce dati artificiali a partire da rumore casuale, e un discriminatore, che cerca di distinguere i dati generati da quelli reali, un processo che si affina fino a raggiungere un equilibrio in cui il generatore produce output indistinguibili dai dati autentici. Tuttavia, applicare i GAN standard ai dati tabulari si è rivelato problematico: le immagini sono rappresentazioni continue e spazialmente coerenti, mentre i dati tabulari sono eterogenei, con variabili di natura diversa, distribuzioni multimodali (ad esempio, età con picchi a 20, 40 e 60 anni) e dipendenze complesse che non si adattano bene all'architettura di base dei GAN, progettata per dati continui ad alta dimensionalità. Il CTGAN, sviluppato da (Xu et al., 2019), supera queste limitazioni introducendo modifiche specifiche per i dati tabulari, come la capacità di gestire variabili miste, un meccanismo condizionale per controllare la

generazione di campioni e tecniche di pre-elaborazione per affrontare distribuzioni non gaussiane, rendendolo uno strumento potente in SDV per applicazioni come la creazione di dataset sintetici per la privacy, l'aumento dei dati in machine learning o la simulazione di scenari in ambiti come la sanità o la finanza. In SDV, il CTGAN si distingue dal Gaussian Copula (descritto in precedenza) per la sua capacità di modellare relazioni non lineari e distribuzioni più complesse, offrendo una flessibilità che lo rende ideale per dataset con caratteristiche statistiche difficili da catturare con metodi più semplici, anche se a costo di una maggiore complessità computazionale.

Il funzionamento del CTGAN si basa sull'architettura dei GAN, ma con adattamenti significativi per i dati tabulari, che possiamo esplorare confrontandolo direttamente con il GAN classico per evidenziarne le innovazioni e la parte matematica sottostante, mantenendo un livello di dettaglio accessibile ma rigoroso. Nei GAN tradizionali, il generatore prende come input un vettore di rumore casuale (tipicamente campionato da una distribuzione normale o uniforme) e produce un output continuo, come un'immagine, mentre il discriminatore valuta se il dato che riceve è reale (dal dataset originale) o falso (dal generatore). Il generatore cerca di "ingannare" il discriminatore e il discriminatore cerca di migliorare la sua capacità di distinzione.

Il CTGAN mantiene questa struttura di base, ma introduce tre modifiche chiave. Primo, utilizza un meccanismo condizionale: invece di generare campioni in modo completamente casuale, il generatore è condizionato su specifiche variabili categoriche, permettendo all'utente di controllare la distribuzione dei dati sintetici; ad esempio, si può richiedere che il 50% dei campioni generati sia di genere femminile, un aspetto implementato aggiungendo un vettore condizionale al rumor, così che il generatore produca dati coerenti con la condizione specificata, e il discriminatore valuta i campioni rispetto a questa condizione, migliorando l'utilità dei dati generati per scenari specifici. Secondo, affronta le distribuzioni multimodali delle variabili continue con una tecnica di normalizzazione basata su mode-specific normalization: invece di assumere una distribuzione gaussiana, CTGAN modella ogni variabile continua come una miscela di gaussiane, trasformando i valori in una rappresentazione che cattura i diversi "modi" (picchi) della distribuzione, un processo che implica stimare i pesi, le medie e le varianze dei modi e campionare da questa miscela durante la generazione, una differenza cruciale rispetto ai GAN classici, che trattano i dati

come continui senza considerare multimodalità. Terzo, per le variabili categoriche, CTGAN usa una rappresentazione one-hot encoding durante l'addestramento e applica una funzione softmax nel generatore per produrre probabilità di categoria, garantendo che i dati sintetici rispettino le proporzioni delle categorie reali, mentre i GAN standard non gestiscono direttamente dati discreti. Queste modifiche rendono CTGAN più adatto ai dati tabulari, ma aumentano la complessità dell'addestramento rispetto ai GAN, richiedendo più risorse computazionali e una sintonizzazione attenta degli iperparametri.

Le differenze tra CTGAN e GAN non si limitano alla teoria, ma hanno implicazioni pratiche significative nell'uso all'interno di SDV, dove CTGAN è implementato per generare dati sintetici tabulari con un livello di realismo e flessibilità superiore, insieme a vantaggi e limiti che emergono dal confronto con il GAN classico e da come si posiziona rispetto ad altri modelli come il Gaussian Copula. In termini pratici, l'implementazione di CTGAN in SDV (ad esempio,

```
from sdv.tabular import CTGAN
model = CTGAN()
model.fit(data)
```

consente agli utenti di caricare un dataset, addestrare il modello e generare campioni sintetici che riflettono sia le distribuzioni marginali che le dipendenze non lineari tra le variabili, un processo che può essere controllato specificando condizioni come la distribuzione di una colonna categorica, una caratteristica assente nei GAN tradizionali, che non offrono meccanismi per imporre vincoli sui dati generati e sono più orientati a output non strutturati come immagini. Rispetto al GAN, CTGAN è più lento e computazionalmente intensivo, poiché l'addestramento di reti neurali per dati tabulari richiede di gestire eterogeneità e multidimensionalità ridotta rispetto alle immagini (un dataset tabulare potrebbe avere solo decine di colonne contro migliaia di pixel), ma compensa questa complessità con una maggiore capacità di catturare relazioni complesse: mentre un GAN classico potrebbe fallire nel riprodurre correlazioni tra età e reddito o nel generare valori categorici coerenti, CTGAN eccelle in questi aspetti grazie alle sue tecniche di pre-elaborazione e al design condizionale. Rispetto al Gaussian Copula, CTGAN offre un vantaggio nella modellazione di dipendenze non lineari e distribuzioni non gaussiane, ma perde in efficienza e semplicità, rendendolo più adatto a dataset complessi dove la fedeltà statistica è

prioritaria rispetto alla velocità. I vantaggi di CTGAN in SDV includono la sua versatilità per dati misti, la capacità di generare campioni realistici per applicazioni come l'addestramento di modelli di machine learning o la simulazione di scenari sanitari, e la protezione della privacy, poiché i dati sintetici non replicano direttamente i dati reali. Tuttavia, i limiti sono evidenti: richiede hardware più potente (idealmente una GPU), un tempo di addestramento più lungo e una certa esperienza per ottimizzare parametri come il numero di epoche o la dimensione del batch, a differenza dei GAN classici, che per applicazioni come le immagini possono essere più standardizzati, o del Gaussian Copula, che è più rapido ma meno flessibile. In definitiva, CTGAN rappresenta un'evoluzione mirata dei GAN per i dati tabulari, offrendo in SDV una soluzione sofisticata che bilancia complessità e potenza, adattandosi a esigenze moderne dove la qualità dei dati sintetici è cruciale.

3.1.3 TVAE

I TVAE (Xu et al, 2019) sono un modello di rete generativa basato su autoencoder variazionali (VAE) progettato specificamente per gestire dati tabulari utilizzando la stessa logica di preelaborazione di un VAE tradizionale e modificandone la funzione di loss. Un tipico VAE (Kingma et al., 2013) combina elementi di reti neurali e inferenza probabilistica risultando una soluzione estremamente potente per generare dati sintetici realistici. Sono particolarmente utili in contesti in cui è necessario preservare le relazioni temporali e le dipendenze presenti nei dati originali. Con SDV, puoi facilmente addestrare modelli TVAE e generare dati sintetici di alta qualità per varie applicazioni. A differenza degli autoencoder tradizionali, che apprendono una rappresentazione compatta dei dati, i VAE apprendono una distribuzione di probabilità associata ai dati, permettendo di generare nuovi campioni. In pratica, la funzione di codifica mappa i dati di ingresso in uno spazio latente, rappresentato da una distribuzione di probabilità (di solito una gaussiana multivariata). Il risultato di questa codifica non consisterà in un singolo punto nello spazio latente, ma nei parametri della distribuzione in questione (media e varianza). Viceversa, la funzione di decodifica, partendo da un punto campionato dallo spazio latente, lo mappa di nuovo nello spazio originale, cercando di ricostruire i dati di ingresso. Un elemento cruciale nell'architettura del VAE è la sua funzione di "loss", composta da due componenti fondamentali.

Il primo componente è quello di ricostruzione deputato a misurare quanto bene il decodificatore ricostruisce l'input originale. Spesso è una funzione di valutazione della perdita di precisione (loss) basata su MSE (Mean Squared Error) o cross-entropia. Il secondo componente è deputato alla regolarizzazione (tipicamente basato sulla divergenza KL) il cui scopo è misurare quanto la distribuzione appresa dall'encoder si discosta da una distribuzione gaussiana standard.

Questa funzione incentiva lo spazio latente a essere ben strutturato (continuo e interpretabile, permettendo operazioni come interpolazione tra campioni) e facilita la generazione di nuovi dati. In un TVAE due differenti modelli di rete neurale vengono usati per modellare la distribuzione a priori dello spazio latente cercando di farla "coincidere" per quanto possibile con quella a posteriori degli elementi codificati, che vengono addestrate attraverso n principio di massimizzazione del limite inferiore dell'evidenza, o limite inferiore variazionale (ELBO).

Nel contesto dei VAE, il limite inferiore variazionale si riferisce alla stima del caso peggiore della probabilità (log-likelihood) di una specifica distribuzione a posteriori. In altre parole, le variabili osservabili nei dati di input sono "l'evidenza" delle variabili latenti scoperte dall'autoencoder. Uno specifico output dell'autoencoder, condizionato sia dal termine di perdita della divergenza KL che dal termine di perdita della ricostruzione, si adatta all' "evidenza" dei dati di formazione. In pratica, l'addestramento di un modello per l'inferenza variazionale può essere definito in termini di massimizzazione dell'ELBO.

Strutturati in accordo a questi principi i nostri TVAE possono essere utilizzati per generare dati sintetici che mantengono le relazioni temporali e le dipendenze presenti nei dati originali (tipicamente serie storiche o sequenze di eventi).

Come nel caso dei TGAN sarà necessario prima di tutto addestrare il modello di sintesi dei dati facendogli acquisire conoscenza dai dati campione, per poi generare i dati sintetici in accordo alla cardinalità desiderata.

```
from sdv.tabular import TVAE
model = TVAE()
model.fit(data)
new_data = model.sample(200)
```

3.1.4 Test diagnostici e analisi di qualità dei dati di sintesi

A valle della generazione dei dati, come primo passo, è necessario eseguire una valutazione diagnostica per verificare che i dati di sintesi prodotti siano validi. Il processo diagnostico esegue alcuni controlli di base come ad esempio:

- Tutte le chiavi primarie presenti nei dati devono essere uniche
- I valori continui devono rispettare i valori minimi e massimi dei dati reali.
- Le colonne discrete (se non associate a dati di identificazione personale) devono contenere le stesse categorie dei dati reali.

```
from sdv.evaluation.single_table import run_diagnostic

diagnostic = run_diagnostic(
    real_data=real_data,
    synthetic_data=synthetic_data,
    metadata=metadata
)
Generating report ...

(1/2) Evaluating Data Validity: |██████████| 9/9 [00:00<00:00, 763.82it/s]|
Data Validity Score: 100.0%

(2/2) Evaluating Data Structure: |██████████| 1/1 [00:00<00:00, 221.31it/s]|
Data Structure Score: 100.0%

Overall Score (Average): 100.0%
```

Chiaramente, un punteggio del 100% indica la piena validità dei dati dal punto di vista diagnostico. Successivamente, l'analisi circa la qualità dei dati di sintesi generati verifica la similarità statistica tra i dati reali e quelli sintetici, acquisendo informazioni sulla somiglianza tra la "forma" statistica delle singole colonne (Column Shape), che ne descrive la distribuzione marginale, e i Trend di evoluzione delle varie coppie di colonne (Column Pair Trends) reali e sintetiche, relativi alla correlazione o alla distribuzione bivariata fra le colonne. Questo ultimo elemento ci descrive come due colonne variano una rispetto all'altra. Il valore di qualità stimato per questi parametri può variare fra 0% e 100%. Chiaramente, più alto è il valore stimato, più simili sono le distribuzioni dei dati reali e sintetici.

Un punteggio del 100% è segno che i modelli sono esattamente gli stessi. Ad esempio, se si confrontassero i dati reali con sé stessi (identità), il punteggio ottenuto sarebbe esattamente del 100%. Un punteggio dello 0% indica che i pattern sono il più possibile diversi. Ciò implica che i dati sintetici contengono esplicitamente anti-pattern opposti

ai dati reali. Qualsiasi punteggio intermedio può essere interpretato secondo questa scala. Ad esempio, un punteggio dell'80% significa che i dati sintetici sono simili ai dati reali per circa l'80% - circa l'80% delle tendenze sono simili.

```
from sdv.evaluation.single_table import evaluate_quality

quality_report = evaluate_quality(
    real_data,
    synthetic_data,
    metadata)
Generating report ...

(1/2) Evaluating Column Shapes: |██████████| 9/9 [00:00<00:00, 395.47it/s]|
Column Shapes Score: 89.11%

(2/2) Evaluating Column Pair Trends: |██████████| 36/36 [00:00<00:00,
130.23it/s]|

Column Pair Trends Score: 88.3%
Overall Score (Average): 88.7%
```

Il punteggio di qualità è chiaramente destinato a variare nell'intervallo sopra indicato e nella realtà non si potrà mai raggiungere, a meno di non confrontare i dati reali con se stessi, esattamente il 100% di qualità.

In ogni caso il processo di sintesi dei dati, realizzato attraverso tecniche di AI generativa ha lo scopo di stimare il modello dei dati reali e produrre a partire dallo stesso dati di sintesi che siano quanto più fedeli al modello stesso, il che significa che in alcuni casi il processo può effettuare smoothing, estrapolazione o introdurre un minimo di rumore nei dati di sintesi risultanti. Vediamo ora in dettaglio come vengono stimati i valori di similarità fra dati reali e di sintesi che danno luogo alle conseguenti stime di qualità dei dati generati.

3.1.4.1 Column Shape

Il calcolo di questo parametro utilizza una combinazione della statistica di Kolmogorov-Smirnov (KS) e della distanza di variazione totale (TVD) tra le colonne reali e quelle sintetiche. Per calcolare la statistica KS, si converte una distribuzione numerica nella sua funzione di distribuzione cumulativa (CDF). La statistica KS diventa la differenza massima tra le due CDF. Invece, la distanza di variazione totale (TVD) tra le colonne reali e quelle sintetiche si calcola stimando innanzitutto la frequenza di ciascun valore della categoria ed la esprimendola in forma di probabilità

e successivamente si confrontano le differenze fra le probabilità, in accordo alla formula seguente:

$$\delta(R,S) = \frac{1}{2}\sum_{\omega\in\Omega}|R_\omega - S_\omega|$$

Il valore ω descrive tutte le possibili categorie in una colonna, Ω. *R* e *S* si riferiscono invece alle frequenze reali e sintetiche di queste categorie. Il relativo complemento statistico restituisce 1-TVD, per cui un punteggio più alto significa una qualità superiore.

$$score = 1 - \delta(R,S)$$

3.1.4.2 Column Pair Trends

Questo parametro è stimato come una combinazione fra due componenti di similarità distinte, specificamente *similarità di correlazione* e *similarità di contingenza*. Il primo componente determina, per ogni coppia di colonne, A e B, un coefficiente di correlazione sui dati reali e sintetici, R e S. Si ottengono così due valori di correlazione separati. Il calcolo normalizza e restituisce un punteggio di similarità utilizzando la formula seguente:

$$score = 1 - \frac{|S_{A,B} - R_{A,B}|}{2}$$

Invece, il secondo componente calcola, per ciascuna coppia di colonne, A e B, il test una tabella di contingenza normalizzata per i dati reali e sintetici. Questa tabella descrive la proporzione di righe che hanno ogni possibile combinazione di categorie in A e B.

Quindi, viene calcolata la differenza tra le tabelle di contingenza utilizzando la distanza di variazione totale. Infine, si sottrae la distanza da 1 (complemento statistico) per garantire che un punteggio elevato indichi un'elevata somiglianza. Il processo è riassunto dalla formula seguente.

$$score = 1 - \frac{1}{2}\sum_{\alpha \in A}\sum_{\beta \in B}|S_{\alpha,\beta} - R_{\alpha,\beta}|$$

In ogni caso, la misura complessiva di qualità dei dati di sintesi generati sarà ottenuta mediando i due punteggi associati a Column Shape e Column Pair Trends relativi a tutte le colonne che caratterizzano la struttura record del dataset di interesse.

3.2 American Bankruptcy Dataset

Il dataset American Bankruptcy Dataset disponibile su GitHub è una risorsa preziosa per chi è interessato a studiare i fallimenti delle aziende pubbliche americane quotate nel mercato azionario (Wang, 2024), specificamente sul New York Stock Exchange (NYSE) e sul NASDAQ, coprendo un periodo che va dal 1999 al 2018, un arco temporale significativo che include eventi economici cruciali come la bolla delle dot-com, la crisi finanziaria del 2008 e la successiva ripresa, offrendo così un'opportunità unica per analizzare come le condizioni finanziarie delle aziende si siano evolute in contesti di prosperità e in momenti di crisi. Questo dataset è stato creato con l'obiettivo esplicito di supportare la predizione del fallimento, un tema di crescente rilevanza per investitori, creditori e regolatori che cercano di anticipare i rischi finanziari e proteggere i propri interessi in un mercato volatile come quello statunitense, dove il sistema capitalistico incoraggia l'innovazione ma espone anche le aziende a cicli di espansione e contrazione che possono culminare in insolvenza. La raccolta dei dati è basata su informazioni contabili di 8.262 aziende diverse, estratte da fonti pubbliche come i rapporti annuali depositati presso la Securities Exchange Commission (SEC), l'ente regolatore del mercato azionario americano, che definisce un'azienda come "fallita" in due casi principali: il deposito del Capitolo 11 del Bankruptcy Code, che permette una riorganizzazione del business sotto la supervisione di un tribunale fallimentare mentre l'azienda continua a operare, o il deposito del Capitolo 7, che implica la cessazione totale delle attività e la liquidazione degli asset, due percorsi che riflettono situazioni finanziarie distinte ma convergono nell'indicare un fallimento imminente o in atto.

Gli autori del dataset hanno scelto di etichettare come "fallita" (1) un'azienda nell'anno fiscale precedente il deposito di uno di questi capitoli, mentre le aziende non fallite in quel periodo sono classificate come "attive" (0), una decisione metodologica che consente di analizzare i segnali finanziari che precedono il default, rendendo il dataset particolarmente adatto per modelli predittivi basati su machine learning o analisi statistiche. Con un totale di 78.682 osservazioni annuali azienda-anno, questo dataset si distingue per l'assenza di valori mancanti o sintetici imputati, un vantaggio significativo che elimina la necessità di pre-elaborazioni complesse per gestire dati incompleti, garantendo una base pulita e affidabile per studi accademici o applicazioni

pratiche, come lo sviluppo di algoritmi capaci di identificare precocemente i rischi di insolvenza in un contesto reale e dinamico come il mercato azionario statunitense.

La struttura del dataset è pensata per facilitare l'analisi temporale e predittiva, con un'organizzazione che suddivide i dati in tre subset distinti basati sul periodo temporale: un training set (1999-2011), un validation set (2012-2014) e un test set (2015-2018), una scelta che riflette una pratica comune nel machine learning per garantire che i modelli siano addestrati su dati storici, validati su un periodo intermedio e testati su casi futuri "non visti", simulando così una reale capacità di previsione su dati mai incontrati prima, un aspetto critico per valutare l'efficacia di un sistema predittivo in un contesto operativo. Il dataset contiene 78.682 osservazioni azienda-anno, derivate dalle 8.262 aziende monitorate, il che implica che per ciascuna azienda sono disponibili in media circa 9-10 anni di dati contabili, un livello di dettaglio longitudinale che consente di tracciare l'evoluzione finanziaria nel tempo e di identificare trend o anomalie che potrebbero segnalare un avvicinamento al fallimento, come una riduzione graduale dell'EBITDA o un aumento del rapporto debito/attività.

Sebbene la descrizione ufficiale su GitHub non elenchi esplicitamente tutte le variabili incluse nel file american_bankruptcy_dataset.csv, possiamo dedurre, dalla prima riga del file CSV stesso, che il dataset comprenda metriche finanziarie standard tratte dai bilanci aziendali.

Iniziamo con le attività correnti (**current_assets**), che comprendono tutte le risorse che un'azienda prevede di convertire in liquidità o utilizzare entro un anno o un ciclo operativo, come contanti, crediti verso clienti, scorte di magazzino e investimenti a breve termine, rappresentando una misura della liquidità immediata e della capacità di far fronte agli obblighi a breve termine, come il pagamento di fornitori o stipendi; ad esempio, un'azienda con un alto livello di attività correnti rispetto alle passività correnti è generalmente considerata finanziariamente stabile nel breve periodo, un aspetto che analisti e creditori esaminano attentamente attraverso il calcolo del current ratio (attività correnti divise per passività correnti). Collegato a questo concetto, troviamo le attività totali (**total_assets**), che includono non solo le attività correnti, ma anche quelle a lungo termine, come immobili, macchinari, brevetti e investimenti permanenti, offrendo una visione complessiva del valore patrimoniale dell'azienda in un dato momento, un numero che appare nel totale dell'attivo del bilancio e che serve

come base per valutare la dimensione e la solidità dell'impresa, oltre che per calcolare indici di redditività come il return on assets (ROA), che misura quanto efficacemente l'azienda utilizza le sue risorse per generare profitto. Passando al lato operativo, il costo dei beni venduti (**cost_of_goods_sold**, COGS) rappresenta le spese dirette sostenute per produrre i beni o servizi venduti durante un periodo, includendo materie prime, manodopera diretta e costi di produzione, ma escludendo costi indiretti come marketing o amministrazione; questo valore è cruciale per calcolare il margine lordo e capire quanto l'azienda spende per generare le sue entrate principali, con un COGS elevato che potrebbe indicare inefficienze produttive o pressioni sui costi in un'industria competitiva, come accade spesso nel settore manifatturiero. Sul versante del debito, il debito totale a lungo termine (**total_long_term_debt**) si riferisce agli obblighi finanziari che l'azienda deve rimborsare oltre un anno, come prestiti bancari, obbligazioni o leasing, un dato che riflette la leva finanziaria e il livello di indebitamento strutturale, influenzando la percezione del rischio da parte degli investitori: un debito a lungo termine elevato può essere sostenibile se genera rendimenti superiori al costo degli interessi, ma può anche segnalare vulnerabilità se l'azienda fatica a generare flussi di cassa sufficienti per coprirlo.

Proseguendo con le metriche finanziarie, troviamo ammortamento e deprezzamento (**depreciation_and_amortization**), che rappresentano la riduzione di valore delle attività tangibili (come macchinari) e intangibili (come brevetti o marchi) nel tempo, un costo non monetario che non implica un esborso immediato di cassa, ma che viene registrato per riflettere l'usura o l'obsolescenza delle risorse e per distribuire il loro costo su più esercizi, un processo regolato da principi contabili come i GAAP o gli IFRS; questo valore è fondamentale per calcolare metriche come l'EBITDA, poiché consente di isolare la performance operativa senza l'impatto delle scelte di investimento passate. L'EBIT (earnings before interest and taxes, **ebit**), o utile prima degli interessi e delle imposte, misura la redditività operativa dell'azienda, calcolata sottraendo dai ricavi totali i costi operativi (inclusi COGS e spese generali) ma escludendo gli interessi sul debito e le imposte, offrendo una visione della capacità dell'impresa di generare profitti dalle sue attività principali, indipendentemente dalla struttura finanziaria o dal regime fiscale, un indicatore spesso usato per confrontare aziende dello stesso settore con diverse strategie di indebitamento. Aggiungendo l'ammortamento e il deprezzamento all'EBIT si ottiene l'EBITDA (earnings before

interest, taxes, depreciation, and amortization, **ebitda**), un indicatore ancora più focalizzato sui flussi di cassa operativi, poiché elimina l'effetto delle spese non monetarie, diventando una metrica chiave per valutare la capacità di un'azienda di generare liquidità prima di considerare il costo del capitale o gli investimenti a lungo termine, spesso utilizzata da analisti per stimare il valore di un'impresa o la sua sostenibilità finanziaria, specialmente in settori ad alta intensità di capitale come le telecomunicazioni o l'energia. Tornando ai profitti, il profitto lordo (**gross_profit**) è il risultato della differenza tra i ricavi totali e il COGS, rappresentando la somma disponibile per coprire i costi operativi indiretti e generare utile netto, un valore che indica l'efficienza nella gestione della produzione e dei costi diretti, con un margine lordo elevato che potrebbe suggerire un vantaggio competitivo, come una forte capacità di pricing o una supply chain ottimizzata. Sul lato delle attività correnti, l'inventario (**inventory**) comprende le scorte di materie prime, prodotti in lavorazione e beni finiti pronti per la vendita, un elemento che influisce sulla liquidità e sulla gestione operativa, poiché un inventario eccessivo può legare capitale e aumentare i costi di stoccaggio, mentre un inventario troppo basso può compromettere la capacità di soddisfare la domanda, un equilibrio che aziende come i retailer monitorano attentamente attraverso metriche come il turnover dell'inventario. Le passività correnti totali (**total_current_liabilities**) includono tutti gli obblighi finanziari a breve termine, come debiti verso fornitori, salari da pagare o rate di prestiti in scadenza entro un anno, un dato che, confrontato con le attività correnti, aiuta a valutare la solvibilità immediata dell'azienda, con un rapporto troppo sbilanciato verso le passività che potrebbe segnalare rischi di liquidità, specialmente in periodi di crisi economica.

Completando il quadro finanziario, l'utile netto (**net_income**) rappresenta il profitto finale di un'azienda dopo aver sottratto dai ricavi totali tutti i costi, inclusi COGS, spese operative, interessi, imposte e ammortamenti, un valore che appare come riga finale del conto economico e che sintetizza la redditività complessiva in un determinato periodo, influenzato non solo dalle performance operative ma anche da fattori esterni come le aliquote fiscali o i tassi di interesse sul debito, un indicatore chiave per gli azionisti che desiderano valutare i rendimenti generati dall'impresa e la sua capacità di distribuire dividendi o reinvestire nel business. Parte di questo reddito netto può essere trattenuto dall'azienda sotto forma di utili trattenuti (**retained_earnings**), che rappresentano gli utili accumulati nel tempo non distribuiti

come dividendi, ma reinvestiti per finanziare crescita, ricerca e sviluppo o riduzione del debito, un dato che appare nello stato patrimoniale e che riflette la politica di gestione del capitale dell'azienda, con un livello elevato di utili trattenuti che potrebbe indicare una strategia di espansione a lungo termine o una prudenza finanziaria, mentre un valore basso potrebbe suggerire una preferenza per la distribuzione agli azionisti. I crediti totali (**total_receivables**) si riferiscono alle somme che l'azienda deve ancora incassare dai clienti per beni o servizi venduti a credito, un componente delle attività correnti che misura la dipendenza dalle vendite non ancora liquidate, con un aumento dei crediti che potrebbe indicare una politica di credito più flessibile per stimolare le vendite, ma anche un rischio di insolvenza se i clienti tardano o non pagano, un aspetto che richiede un monitoraggio attento attraverso il **days sales outstanding** (DSO). I ricavi totali (**total_revenue**) rappresentano l'importo complessivo generato dalle vendite di beni o servizi prima di qualsiasi deduzione, un dato che riflette la capacità dell'azienda di attrarre clienti e generare entrate, fondamentale per valutare la crescita e la quota di mercato, con una distinzione importante tra ricavi lordi e netti nel caso di resi o sconti. Il valore di mercato (**market_value**) è una stima del valore dell'azienda basata sul prezzo delle sue azioni moltiplicato per il numero di azioni in circolazione, un indicatore che riflette le aspettative degli investitori sul futuro dell'impresa e che può differire dal valore contabile delle attività totali, influenzato da fattori come la percezione del brand o le prospettive di crescita. Le spese operative totali (**total_operating_expenses**) includono tutti i costi sostenuti per gestire l'azienda al di fuori del COGS, come affitti, stipendi amministrativi, marketing e ricerca, un valore che influisce direttamente sull'EBIT e che riflette l'efficienza nella gestione delle attività quotidiane, con un aumento delle spese che potrebbe indicare investimenti in crescita o, al contrario, una perdita di controllo sui costi. Le passività totali (**total_liabilities**) comprendono tutti i debiti, sia a breve che a lungo termine, offrendo una visione completa degli obblighi finanziari, un dato che, confrontato con le attività totali, permette di calcolare il rapporto debito/patrimonio netto (debt-to-equity ratio), un indicatore di leva finanziaria essenziale per valutare la struttura del capitale. Infine, le vendite nette (**net_sales**) sono i ricavi totali al netto di resi, sconti e abbuoni, una misura più precisa dell'entrata effettiva derivante dalle operazioni principali, utilizzata per analizzare la performance commerciale senza distorsioni.

Come si è visto, le variabili rappresentate nel dataset sono tutte variabili che riflettono la salute finanziaria di un'azienda e che sono comunemente utilizzate per calcolare rapporti finanziari predittivi del fallimento, come il modello di Altman Z-score o indicatori di liquidità e leva finanziaria. L'etichetta binaria di fallimento (1 o 0) è associata a ciascun anno fiscale, con il valore 1 attribuito all'anno immediatamente precedente il deposito del Capitolo 7 o 11 (clausole di fallimento in accordo alla normativa finanziaria USA), una scelta che permette di concentrarsi sui segnali finanziari premonitori piuttosto che sugli effetti post-fallimento, un dettaglio che aumenta l'utilità del dataset per applicazioni predittive.

La suddivisione temporale non è casuale: il training set (1999-2011) copre 13 anni che includono la crisi del 2008, un periodo di forte instabilità che ha visto un picco di fallimenti aziendali (ad esempio, Lehman Brothers nel 2008), offrendo una ricca base di apprendimento per i modelli; il validation set (2012-2014) rappresenta una fase di ripresa economica post-crisi, utile per ottimizzare i parametri; e il test set (2015-2018) include anni di crescita stabile, permettendo di valutare la generalizzabilità dei modelli a contesti più recenti fino al 2018. L'assenza di valori mancanti è un punto di forza, ma potrebbe anche suggerire una selezione rigorosa delle aziende incluse, probabilmente limitata a quelle con dati contabili completi depositati presso la SEC, escludendo potenzialmente piccole imprese o aziende non quotate con registrazioni incomplete, un compromesso che ne limita la rappresentatività ma ne aumenta la qualità per l'analisi quantitativa.

Le applicazioni del dataset sono molteplici e si collocano all'intersezione tra finanza, data science e politica economica, offrendo un terreno fertile per esplorare le dinamiche dei fallimenti aziendali negli Stati Uniti e sviluppare strumenti predittivi che possano avere un impatto reale su decisioni di investimento, gestione del rischio e regolamentazione finanziaria, un potenziale che si realizza grazie alla sua struttura ben definita e alla copertura temporale significativa.

Una delle applicazioni più immediate è la predizione del fallimento aziendale, dove il dataset può essere utilizzato per addestrare modelli di machine learning come Random Forest, reti neurali profonde o il CTGAN di SDV, sfruttando le variabili contabili per identificare pattern che precedono il deposito di un Capitolo 7 o 11, come un calo persistente del margine lordo o un aumento del debito a lungo termine rispetto alle

attività totali, un'attività che potrebbe aiutare banche e investitori a mitigare le perdite evitando esposizioni a aziende a rischio, come dimostrato da studi come quello di Pellegrino et al. (2024), che utilizza LSTM su dati simili per prevedere fallimenti con alta precisione. Un altro uso è l'analisi di resilienza dei portafogli azionari, dove i dati sintetici generati dal dataset (ad esempio, tramite tecniche generative come il Gaussian Copula) possono simulare scenari di stress economico, testando come variazioni nei ricavi totali o nelle passività correnti influenzano la probabilità di default di un gruppo di aziende quotate, un approccio utile per gestori di fondi o regolatori come la SEC nella valutazione della stabilità del mercato.

Inoltre, il dataset può essere impiegato per studi accademici sulle cause dei fallimenti, analizzando come fattori come l'EBITDA o gli utili trattenuti si correlano con l'insolvenza in diversi cicli economici, o per valutare l'impatto di eventi macroeconomici specifici, come la crisi dei mutui subprime, confrontando i dati pre- e post-2008 per identificare segnali di vulnerabilità ricorrenti.

Tuttavia, il dataset presenta alcuni limiti: la copertura si ferma al 2018, il che significa che non include dati recenti come l'impatto della pandemia di COVID-19, che ha causato un'ondata di fallimenti nel 2020-2021 (ad esempio, 7.128 casi di Capitolo 11 nel 2020 secondo l'ABI), un gap che potrebbe essere colmato integrandolo con statistiche aggiornate dai U.S. Courts o dall'American Bankruptcy Institute, disponibili fino al 2024; inoltre, la focalizzazione su aziende pubbliche quotate esclude piccole imprese o entità private, limitando la generalizzabilità a tutto il panorama economico statunitense. Nello specifico del presente studio, essendo i dati analizzati riferiti al solo anno 2018, come riportato nel seguito, questo aspetto assume rilevanza minore.

Una delle applicazioni principali è la predizione del fallimento, dove modelli di machine learning, come reti neurali, Random Forest o il CTGAN di SDV, possono essere addestrati sui dati storici per identificare pattern che precedono il default, come un aumento del rapporto debito/attività o una diminuzione dell'EBITDA, un'attività che banche e investitori usano per valutare la solidità di aziende o clienti, come dimostrato dal lavoro di Pellegrino et al. (2024) che utilizza LSTM su dati contabili per prevedere fallimenti di aziende quotate.

Un'altra applicazione è l'analisi di resilienza dei portafogli creditizi, dove banche simulano scenari estremi (ad esempio, un'impennata dei tassi di interesse) utilizzando dati sintetici generati da modelli come il Gaussian Copula o CTGAN, testando come variabili come il debito a lungo termine o il reddito netto influenzano la probabilità di insolvenza, un approccio che si collega agli stress test regolamentari richiesti dalla Federal Reserve.

Per i fallimenti personali, il dataset può essere usato per studiare disparità sociali, come evidenziato da ProPublica, che ha mostrato che i debitori in aree a maggioranza nera hanno il doppio delle probabilità di vedere i loro casi di Capitolo 13 respinti rispetto a quelli in aree bianche, un'analisi che combina dati finanziari (crediti, passività) con informazioni demografiche per esplorare l'impatto della razza e del reddito sulle opzioni di sollievo dal debito. Inoltre, il dataset può servire a valutare l'efficacia delle leggi fallimentari, confrontando i tassi di successo dei Capitoli 7, 11 e 13 (normative USA) nel tempo e tra stati, o analizzando come le riforme del Bankruptcy Abuse Prevention and Consumer Protection Act (BAPCPA) del 2005 abbiano influenzato i depositi, un tema di interesse per legislatori e accademici.

Dal punto di vista pratico, ottenere un dataset completo richiede l'accesso a fonti come il Public Access to Court Electronic Records (PACER) per dati grezzi caso per caso, o l'aggregazione di statistiche dall'American Bankruptcy Institute, che offre analisi aggiornate al 2025 (ad esempio, 40.271 casi totali a novembre 2024, con un aumento del 6% rispetto al 2023), anche se queste potrebbero non includere dettagli granulari come l'inventario o i crediti totali. I limiti includono la mancanza di dati aggiornati in tempo reale, la difficoltà di integrare variabili non finanziarie (come eventi geopolitici) e la necessità di pulire i dati per gestire valori mancanti o inconsistenze, un processo che richiede tempo ma è essenziale per garantire risultati affidabili. In definitiva, un American Bankruptcy Dataset non è solo una raccolta di numeri, ma uno specchio dell'economia statunitense, che riflette le sfide di aziende e individui in un sistema finanziario complesso, offrendo strumenti per anticipare crisi, proteggere i creditori e informare politiche pubbliche, con implicazioni che si estendono ben oltre i tribunali fallimentari.

L'accesso al dataset è semplice: su GitHub, il repository sowide/bankruptcy_dataset offre il file american_bankruptcy_dataset.csv scaricabile gratuitamente, insieme a un

README che descrive la metodologia e la suddivisione temporale, e il dataset è accompagnato da una licenza consultabile nel file LICENSE.md, un aspetto che lo rende utilizzabile per scopi accademici o commerciali con eventuali restrizioni da verificare. Per utilizzarlo, basta clonare il repository con git clone https://github.com/sowide/bankruptcy_dataset.git e caricare il CSV in un ambiente come Python con pandas

df = pd.read_csv('american_bankruptcy_dataset.csv')

dove le 78.682 righe e le colonne contabili possono essere analizzate o usate per addestrare modelli, un processo che richiede risorse computazionali moderate ma accessibili anche a un laptop standard.

Per la nostra elaborazione "proof of concept" abbiamo scelto di utilizzare soltanto i dati relativi all'anno 2018, essendo questi sufficientemente numerosi, riportanti evidenze relative a 2135 aziende, e ben diversificati. L'obiettivo della proof of concept, infatti è essenzialmente quello di verificare la funzionalità di generazione di dati sintetici e la loro qualità.

3.3 Analisi esplorativa del dataset

In prima istanza è stata effettuata una prima analisi di correlazione fra principali attributi utilizzati calcolando i coefficienti di correlazione dei Pearson per ogni coppia di variabili. I risultati delle analisi evidenziano la presenza di significativi rapporti di correlazione a livello di diverse variabili, come è possibile vedere graficamente dalla heatmap ripostata nel seguito, dove i valori tendenti al rosso (caldo) riportano valori di correlazione elevata mentre quelli tendenti al blu (freddo) sono associati a valori di bassa correlazione.

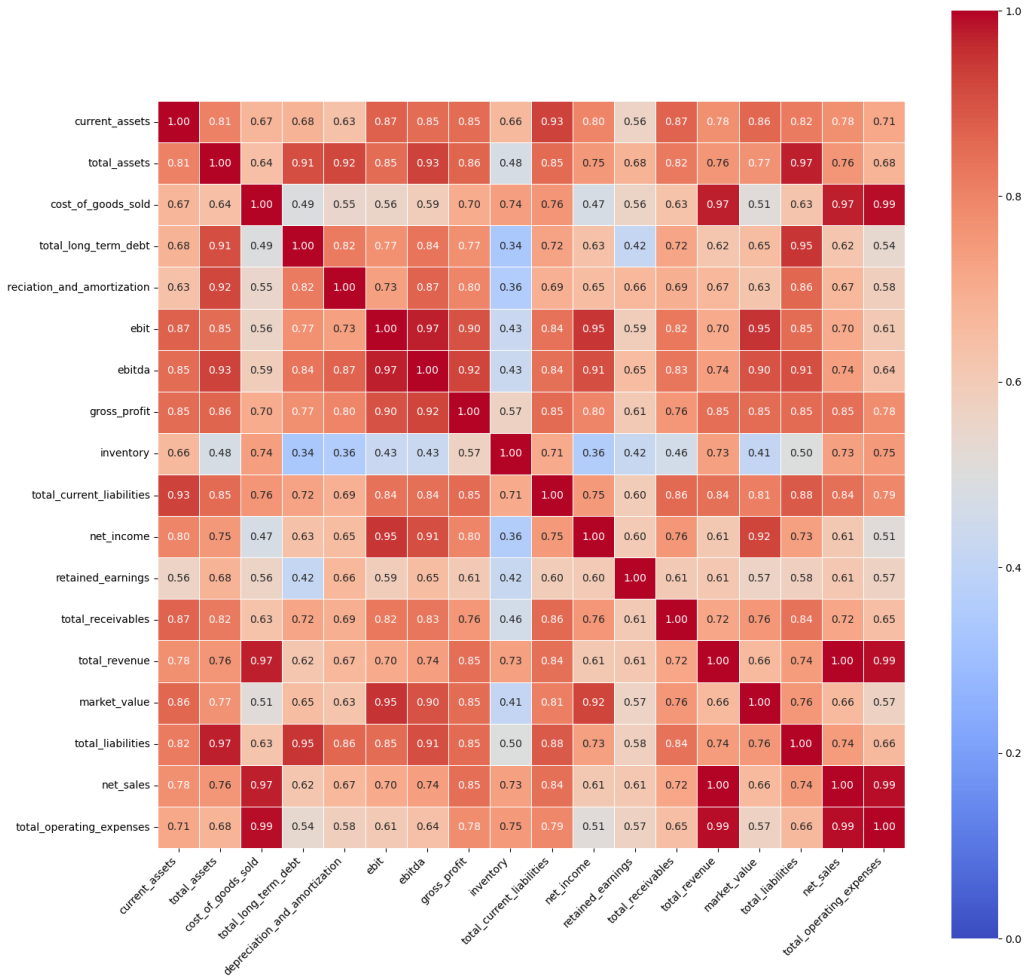

Figura 2: Heatmap di correlazione tra le variabili usate in questo studio.

Sono state inoltre analizzate le distribuzioni di tutte le variabili nel dataset originale. Va notato che i dati evidenziano un notevole intervallo di variazione, che ne complica la visualizzazione. Per questo motivo, si è deciso di escludere le rilevazioni al disopra del 97,5-esimo percentile.

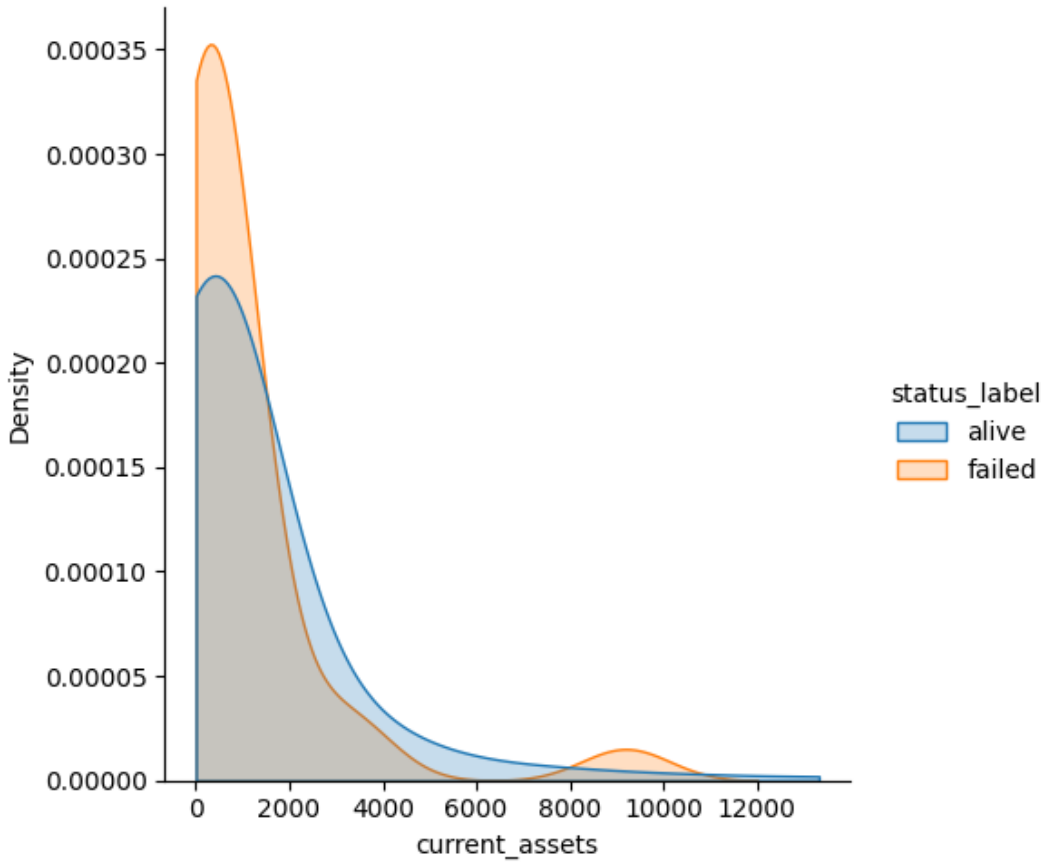

Figura 3: Distribuzione delle attività correnti nel dataset originale.

Il grafico mostra la distribuzione delle attività correnti (current_assets) per le imprese etichettate come "alive" e "failed" nel dataset originale. Le curve di densità evidenziano una marcata asimmetria positiva in entrambe le classi, con una concentrazione prevalente di aziende nei valori più bassi della variabile.

Tuttavia, si osserva una differenza significativa tra i due gruppi: le imprese fallite presentano in media livelli di attività correnti inferiori rispetto a quelle attive, con una curva più ripida e concentrata nei primi intervalli della distribuzione. Al contrario, le imprese attive mostrano una distribuzione leggermente più ampia, suggerendo una maggiore disponibilità di risorse liquide e a breve termine, che può essere interpretata come un indicatore di maggiore solidità finanziaria.

Questa evidenza preliminare conferma la rilevanza delle attività correnti come potenziale variabile discriminante nella predizione del rischio di insolvenza.

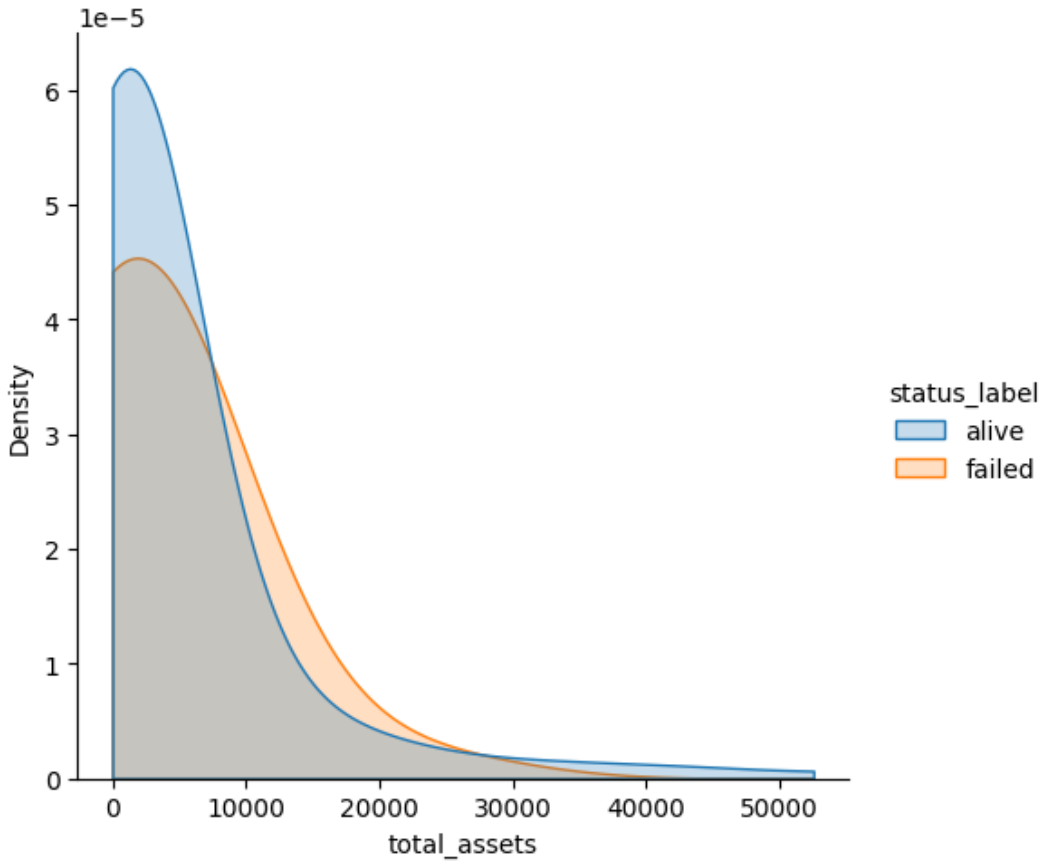

Figura 4: Distribuzione delle attività totali nel dataset originale.

La distribuzione delle attività totali (total_assets) evidenzia una forte concentrazione delle osservazioni nei valori più bassi, con una coda lunga verso destra, tipica delle variabili economico-finanziarie aggregate. Entrambe le classi – aziende attive e fallite – mostrano una distribuzione asimmetrica positiva, ma con importanti differenze di densità nei primi intervalli.

Le imprese attive ("alive") presentano un picco più elevato e concentrato intorno a livelli inferiori di attivo totale, suggerendo che anche realtà aziendali di dimensioni contenute possono sopravvivere, se dotate di equilibrio patrimoniale. Le imprese fallite, invece, si distribuiscono in modo più disperso, con una densità relativamente maggiore in corrispondenza di valori intermedi, a indicare che il fallimento può colpire anche aziende con consistenza patrimoniale significativa.

Questo comportamento suggerisce che la dimensione dell'impresa – intesa come entità patrimoniale – da sola non è sufficiente a spiegare il rischio di default: è necessario considerare anche la qualità degli attivi e l'equilibrio con le passività, rafforzando la necessità di modelli predittivi che integrino indicatori compositi.

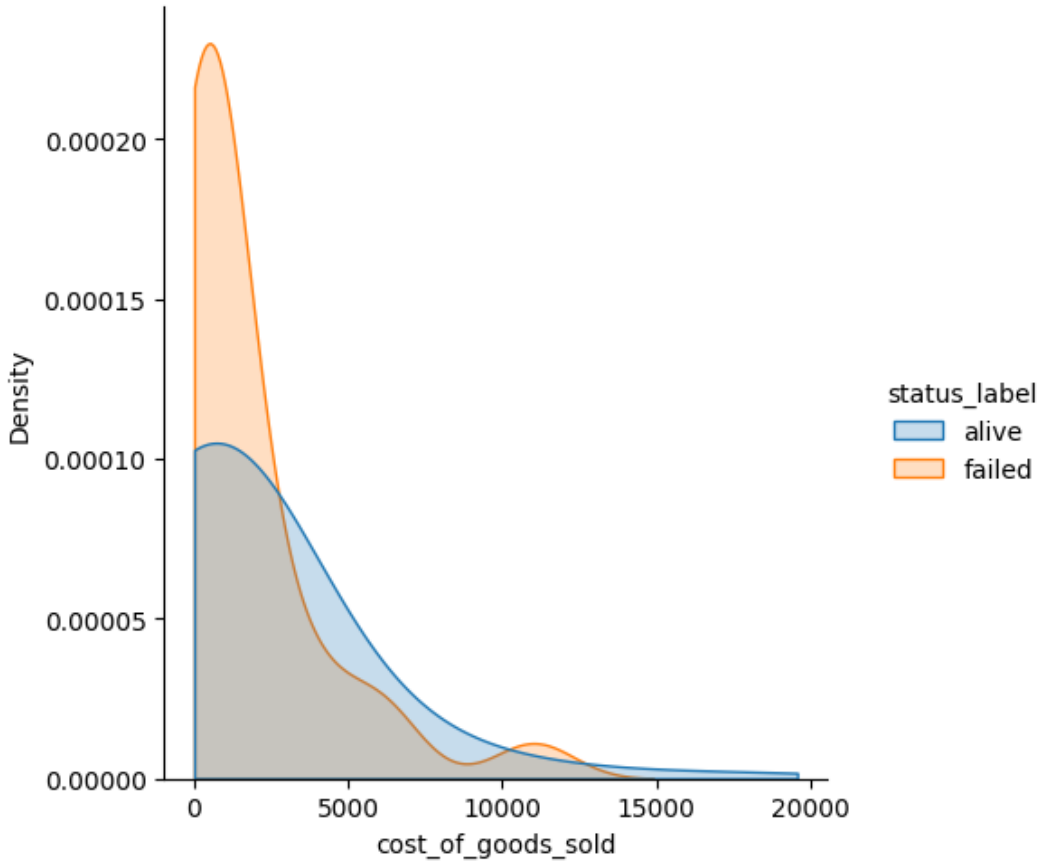

Figura 5: Distribuzione del costo dei beni venduti nel dataset originale.

La distribuzione del costo dei beni venduti (cost_of_goods_sold, COGS) mostra una spiccata asimmetria positiva per entrambe le classi, con la maggior parte delle osservazioni concentrate nei valori più bassi. Tuttavia, la densità per le imprese fallite presenta un picco iniziale più marcato rispetto alle aziende attive, suggerendo che molte imprese in difficoltà operavano con livelli di costo contenuti ma, verosimilmente, anche con ricavi insufficienti a coprirli.

Al contrario, le imprese attive mostrano una distribuzione più ampia, con una coda che si estende maggiormente verso valori elevati. Ciò potrebbe indicare che le aziende sopravvissute sono caratterizzate da volumi operativi più consistenti, coerenti con strutture produttive più ampie e una maggiore capacità di assorbimento dei costi.

Questa evidenza suggerisce che il solo livello assoluto dei costi non è un predittore sufficiente del rischio di default: sarà fondamentale analizzare il rapporto tra COGS e ricavi per derivare indicatori come il margine lordo, più rappresentativi della sostenibilità economica dell'impresa.

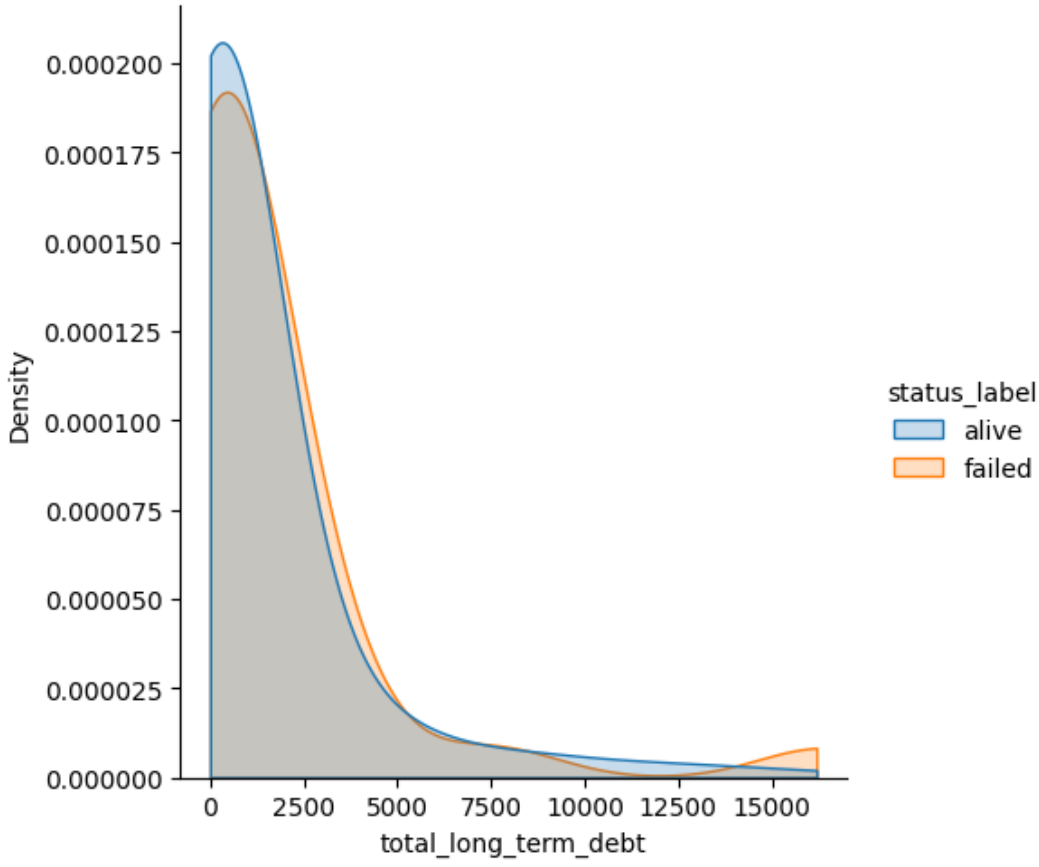

Figura 6: Distribuzione del debito totale a lungo termine nel dataset originale.

Il grafico evidenzia la distribuzione del debito totale a lungo termine (total_long_term_debt) per le imprese classificate come attive e fallite. Anche in

questo caso, la distribuzione è fortemente asimmetrica verso destra, con la maggior parte delle osservazioni concentrate nei valori più bassi.

Le curve relative alle due classi sono quasi sovrapponibili fino a circa 10.000 unità, indicando che il livello di indebitamento a lungo termine, nella fascia bassa e intermedia, non è di per sé sufficiente a discriminare lo stato di salute dell'impresa. Tuttavia, nella coda destra si osserva un picco esclusivo delle aziende fallite, che evidenzia la presenza di soggetti fortemente indebitati che sono successivamente incorsi in insolvenza.

Questa osservazione suggerisce che livelli estremamente elevati di indebitamento a lungo termine possono costituire un indicatore di vulnerabilità finanziaria, soprattutto se non bilanciati da un'adeguata capacità di generazione di cassa. In un'ottica predittiva, sarà dunque utile considerare non solo l'ammontare assoluto del debito, ma anche la sua sostenibilità relativa, attraverso indicatori come il rapporto debito/EBITDA o debito/patrimonio netto.

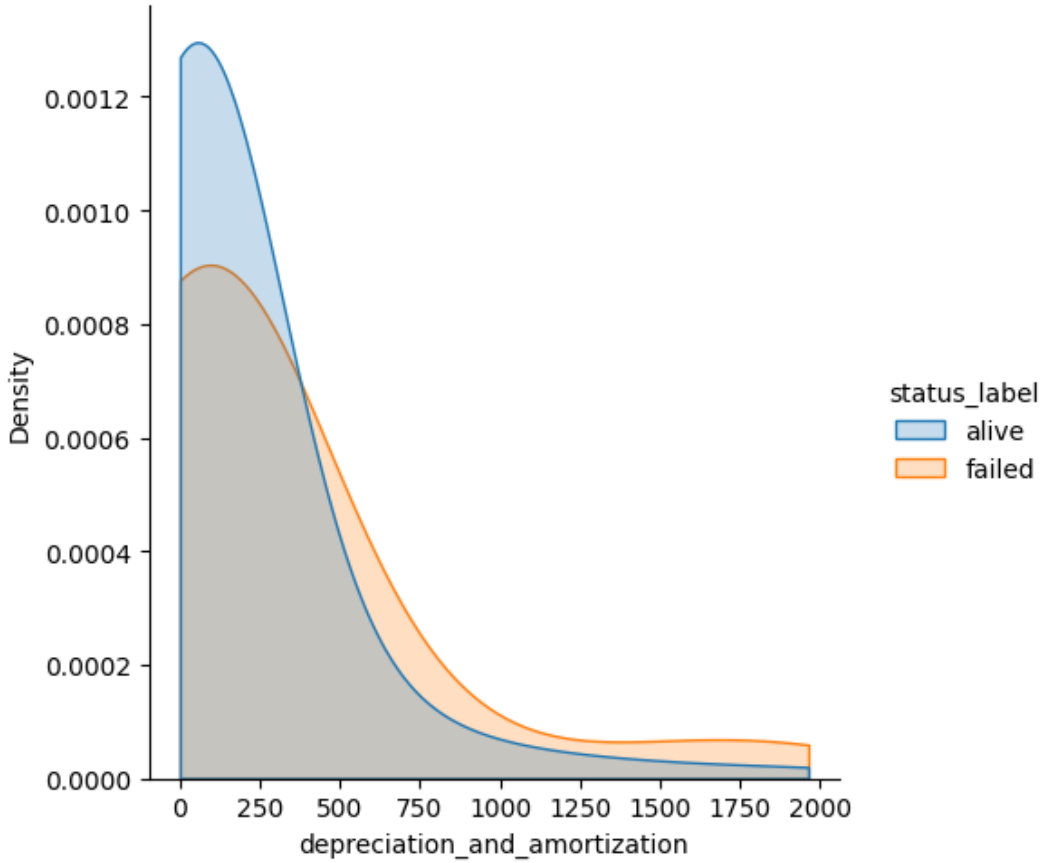

Figura 7: Distribuzione di ammortamento e deprezzamento nel dataset originale.

La distribuzione della variabile ammortamento e deprezzamento (depreciation_and_amortization) evidenzia una concentrazione prevalente nei valori inferiori, con una classica asimmetria positiva. In particolare, le imprese attive evidenziano un picco più accentuato nella fascia compresa tra 0 e 250, suggerendo un utilizzo contenuto ma costante di asset soggetti ad ammortamento.

Le imprese fallite presentano invece una distribuzione più dispersa, con una maggiore densità relativa nei livelli intermedi e alti. Questo potrebbe riflettere due scenari: da un lato, aziende con asset più obsoleti e onerosi da gestire; dall'altro, realtà che hanno effettuato importanti investimenti in immobilizzazioni, poi rivelatisi insostenibili in termini di ritorno operativo.

L'ammortamento, pur essendo una voce non monetaria, è un indicatore rilevante della struttura patrimoniale e della storia degli investimenti aziendali. Una sua valutazione

congiunta con altre variabili – come l'EBIT o il cash flow operativo – può offrire indicazioni importanti sulla sostenibilità della gestione e sulla capacità di assorbire shock economici.

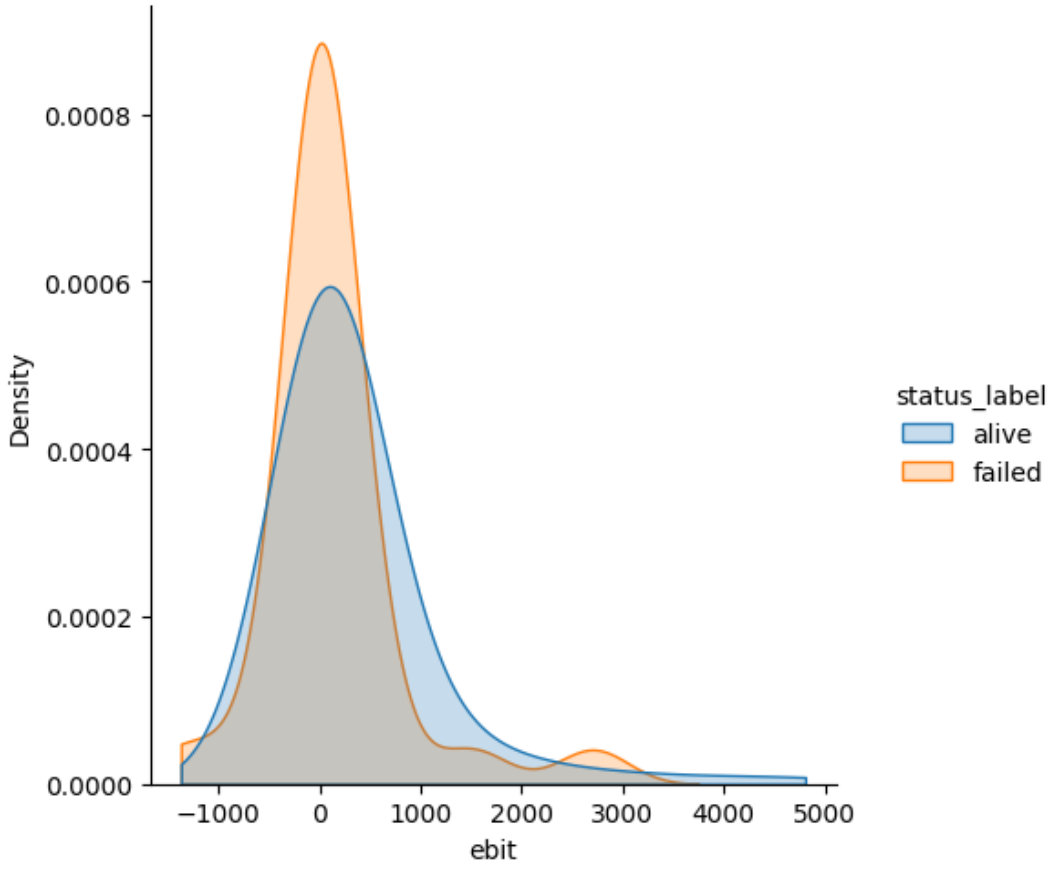

Figura 8: Distribuzione dell'EBIT nel dataset originale.

La distribuzione dell'EBIT (Earnings Before Interest and Taxes) evidenzia una differenza significativa tra le imprese attive e quelle fallite. Entrambe le distribuzioni sono asimmetriche, ma mentre le aziende attive presentano una maggiore densità a destra dello zero – con un picco in corrispondenza di valori positivi – le imprese fallite evidenziano una netta concentrazione di osservazioni intorno e al di sotto dello zero.

Questa evidenza suggerisce un'elevata capacità discriminante dell'EBIT nella previsione del rischio di default: le aziende fallite tendono infatti a presentare redditività operativa negativa o marginale, mentre quelle attive registrano performance

migliori, anche se spesso contenute. La presenza di valori negativi anche tra le imprese attive evidenzia come occasionali flessioni dell'EBIT non siano da sole predittive del fallimento, ma diventino significative se persistenti o combinate ad altri segnali di stress finanziario.

L'EBIT rappresenta una delle variabili chiave per la valutazione della sostenibilità operativa dell'impresa, essendo indipendente da scelte finanziarie e fiscali. La sua analisi, integrata ad altri indicatori come il leverage o la liquidità, consente di costruire modelli predittivi più robusti e affidabili.

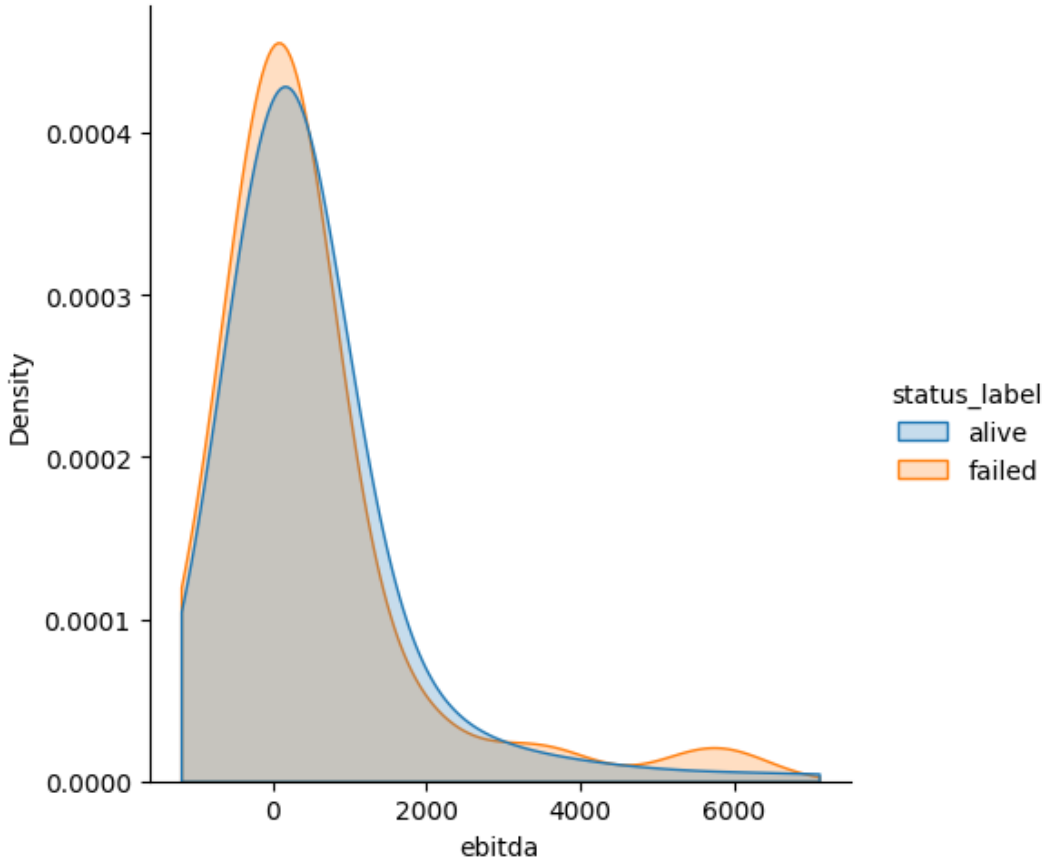

Figura 9: Distribuzione dell'EBITDA nel dataset originale.

La distribuzione dell'EBITDA (Earnings Before Interest, Taxes, Depreciation and Amortization) evidenzia una forma asimmetrica positiva per entrambe le classi di imprese, con una concentrazione prevalente nei valori compresi tra 0 e 2000. Le curve

relative ad aziende attive e fallite risultano molto simili, soprattutto nella parte centrale della distribuzione, indicando una ridotta capacità discriminante dell'indicatore in termini assoluti.

Tuttavia, si rileva una leggera prevalenza di valori negativi nelle imprese fallite e una maggiore dispersione nei valori alti, anch'essi associati alla classe "failed". Questo comportamento apparentemente controintuitivo può essere attribuito a situazioni in cui elevati livelli di EBITDA non si traducono in effettiva redditività netta o sostenibilità operativa – ad esempio, per via di una struttura di costi fissi elevata, debiti insostenibili o investimenti errati.

L'EBITDA rappresenta un indicatore utile per valutare la performance operativa "pura", priva delle distorsioni legate ad ammortamenti e scelte finanziarie. Tuttavia, nel contesto dell'analisi del rischio di default, la sua efficacia aumenta se utilizzato in rapporto ad altre voci di bilancio (come il debt/EBITDA) o all'interno di modelli predittivi multivariati.

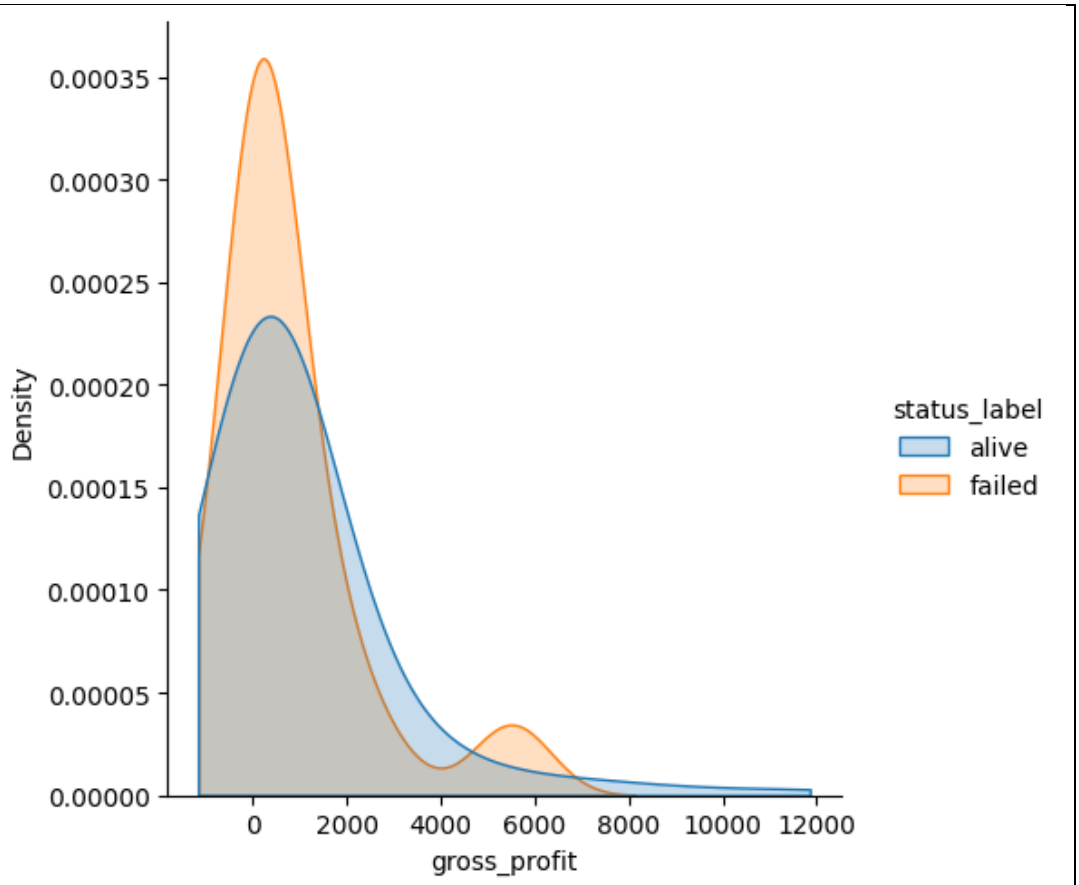

Figura 10: Distribuzione del margine lordo nel dataset originale.

La distribuzione del margine lordo (gross_profit) evidenzia un pattern simile a quello osservato per altre variabili reddituali, con una forte asimmetria positiva e una netta concentrazione di valori nelle fasce inferiori. Tuttavia, emerge una differenza significativa tra le imprese attive e quelle fallite: queste ultime mostrano una densità maggiore in corrispondenza di valori prossimi allo zero o negativi, mentre le imprese attive si distribuiscono più ampiamente lungo la coda positiva.

Questa evidenza suggerisce che un margine lordo positivo – seppur contenuto – rappresenti una condizione minima necessaria per la sopravvivenza aziendale. Le imprese fallite, infatti, appaiono spesso caratterizzate da margini insufficienti a coprire i costi operativi, il che le rende vulnerabili a shock esogeni o a tensioni di liquidità.

Dal punto di vista analitico, il margine lordo costituisce un indicatore cruciale della redditività operativa lorda e della capacità dell'impresa di generare valore aggiunto

nella fase produttiva. La sua inclusione in modelli predittivi è particolarmente rilevante, specie se considerato in rapporto ai ricavi (margine lordo percentuale), poiché consente di isolare l'efficienza industriale dalle dinamiche contabili e finanziarie.

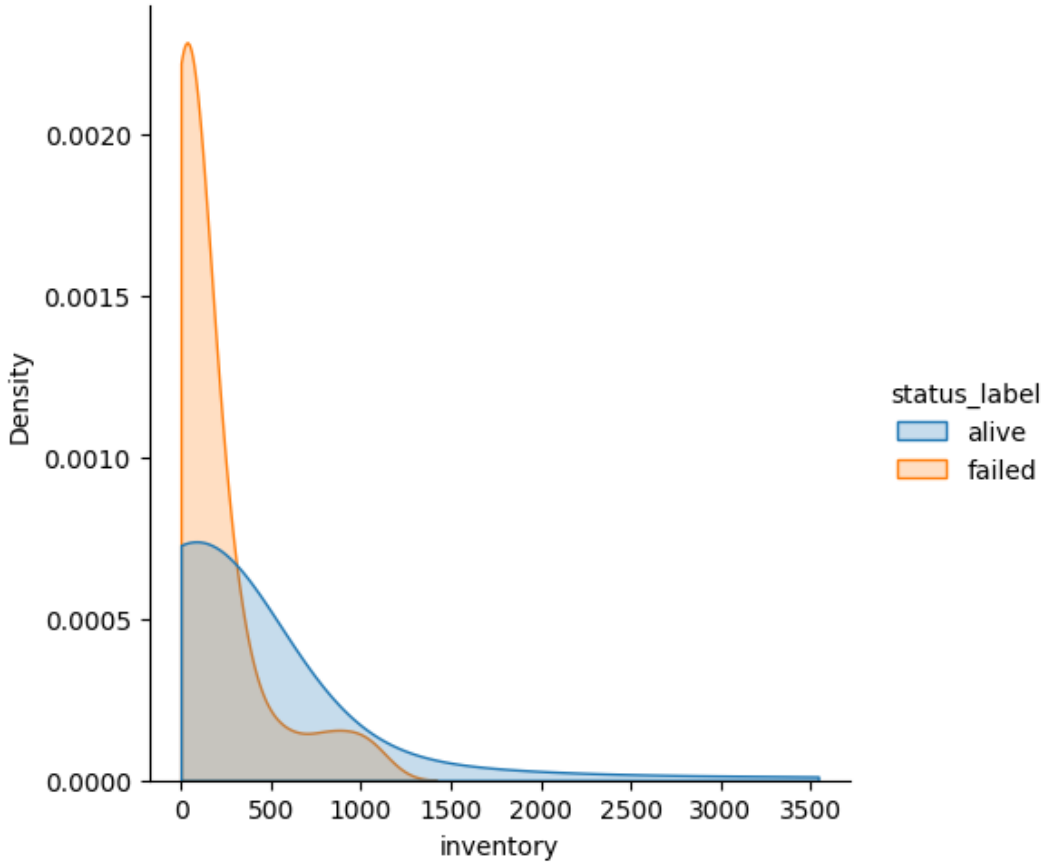

Figura 11: Distribuzione dell'inventario nel dataset originale.

La distribuzione della variabile inventario (inventory) evidenzia una chiara concentrazione di valori nelle fasce più basse, con una marcata asimmetria positiva per entrambe le classi. Tuttavia, la differenza tra le due distribuzioni è particolarmente significativa: le imprese fallite presentano un picco iniziale molto più elevato e stretto, mentre le imprese attive mostrano una curva più ampia e distribuita anche nei valori intermedi e alti.

Questo comportamento suggerisce che livelli estremamente contenuti di inventario – tipici di realtà a bassa operatività o con problemi di approvvigionamento/produzione – siano più frequenti tra le imprese che successivamente falliscono. Al contrario, un livello più consistente di scorte può essere indicativo di una struttura operativa attiva e funzionante, tipica delle imprese sopravvissute.

L'inventario, pur rappresentando una voce dell'attivo corrente, riflette indirettamente la vitalità della catena produttiva e della gestione operativa. Tuttavia, valori eccessivamente elevati potrebbero anche indicare inefficienze o problemi di rotazione. Pertanto, nel contesto dell'analisi predittiva, questa variabile dovrebbe essere considerata in combinazione con indicatori dinamici come il turnover delle scorte.

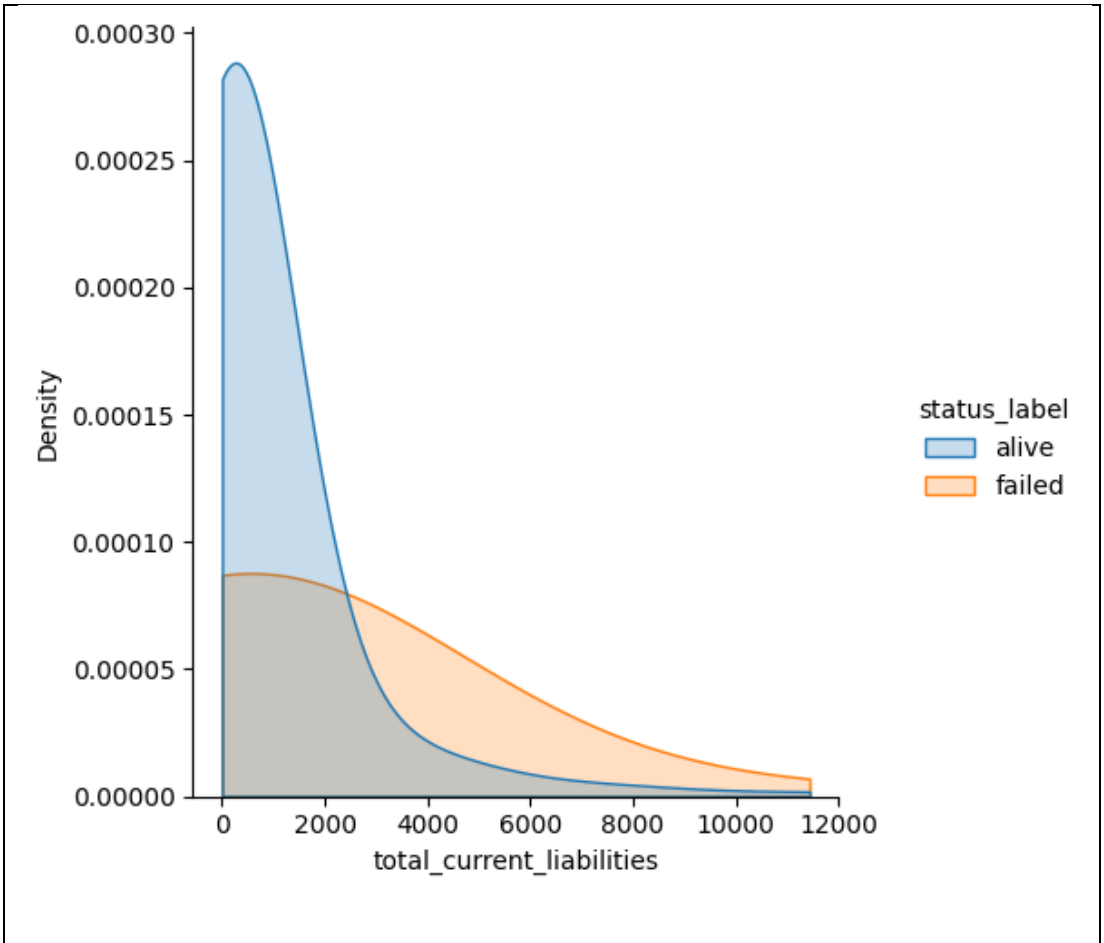

Figura 12: Distribuzione delle passività correnti totali nel dataset originale.

La distribuzione delle passività correnti totali (total_current_liabilities) evidenzia un pattern particolarmente interessante: mentre le imprese attive mostrano una forte concentrazione nei valori più bassi, con una curva ripida e una rapida decrescita, le

imprese fallite si distribuiscono in maniera molto più ampia, con una densità relativamente più elevata anche nei livelli medi e alti della variabile.

Questa evidenza suggerisce che le imprese fallite tendono ad accumulare maggiori debiti a breve termine, potenzialmente in risposta a situazioni di tensione finanziaria o per far fronte a squilibri di liquidità. L'eccessiva esposizione verso passività correnti può rappresentare un importante fattore di vulnerabilità, in quanto riflette obblighi imminenti che l'impresa potrebbe non essere in grado di onorare in assenza di adeguata copertura da parte dell'attivo corrente.

Nel contesto dell'analisi del rischio di default, questa variabile assume quindi un ruolo centrale, in particolare se considerata all'interno di indicatori di equilibrio finanziario di breve periodo, come l'indice di liquidità corrente o il quick ratio. La sua inclusione in modelli predittivi appare giustificata dalla chiara capacità discriminante mostrata dal grafico.

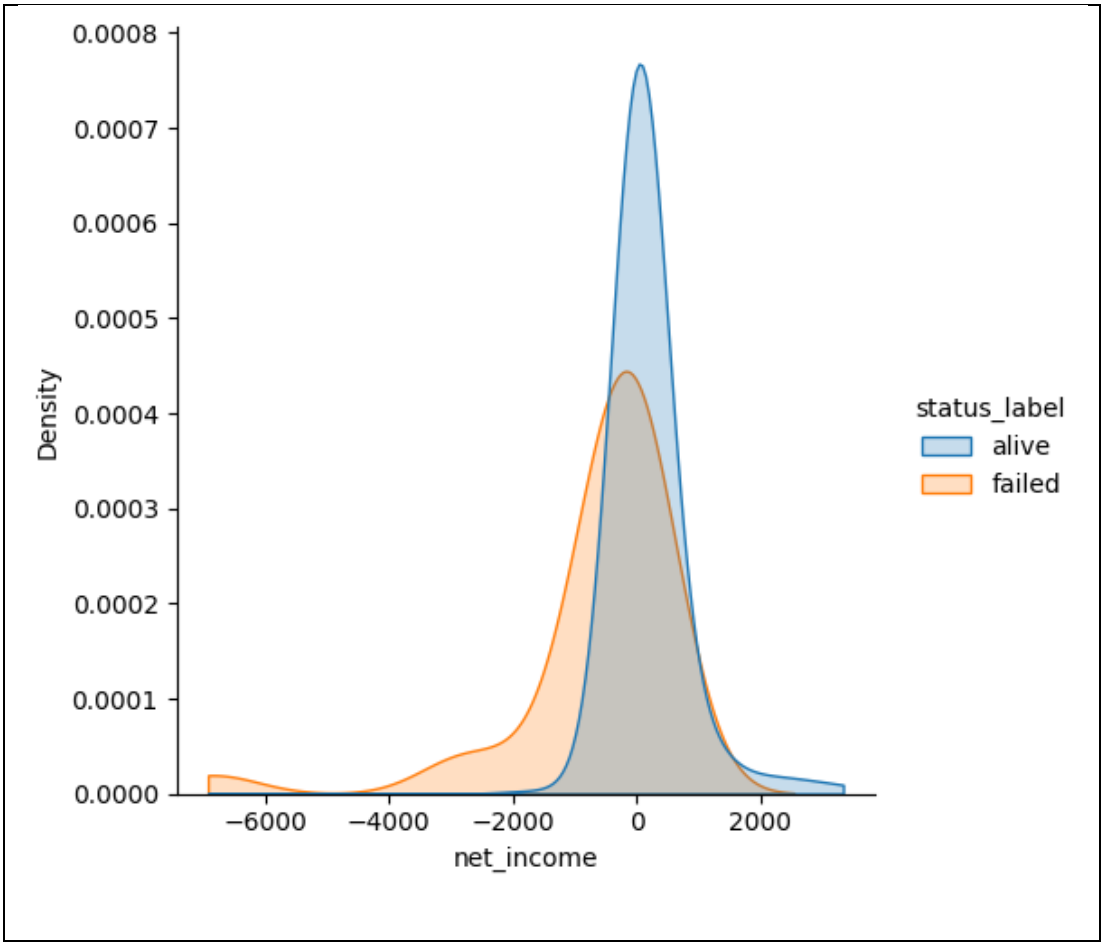

Figura 13: Distribuzione dell'utile netto nel dataset originale.

La distribuzione dell'utile netto (net_income) rappresenta un indicatore diretto e sintetico della performance aziendale complessiva. Il grafico evidenzia una netta differenza tra le due classi: le imprese attive mostrano una distribuzione centrata su valori positivi, con un picco molto accentuato attorno allo zero, mentre le imprese fallite tendono a concentrarsi su valori negativi, con una maggiore dispersione verso sinistra.

Questa evidenza conferma l'intuizione secondo cui il fallimento aziendale è spesso preceduto da una fase prolungata di perdite nette. L'asimmetria nella coda sinistra della distribuzione delle imprese fallite suggerisce inoltre la presenza di casi di dissesto grave, con perdite rilevanti che compromettono la sostenibilità economico-finanziaria dell'impresa.

L'utile netto, pur essendo influenzato da componenti straordinarie e da politiche contabili, rimane uno degli indicatori più significativi in termini di capacità predittiva del rischio di default. La sua combinazione con misure di cassa e altri indici di redditività permette una valutazione più completa della performance e della resilienza aziendale.

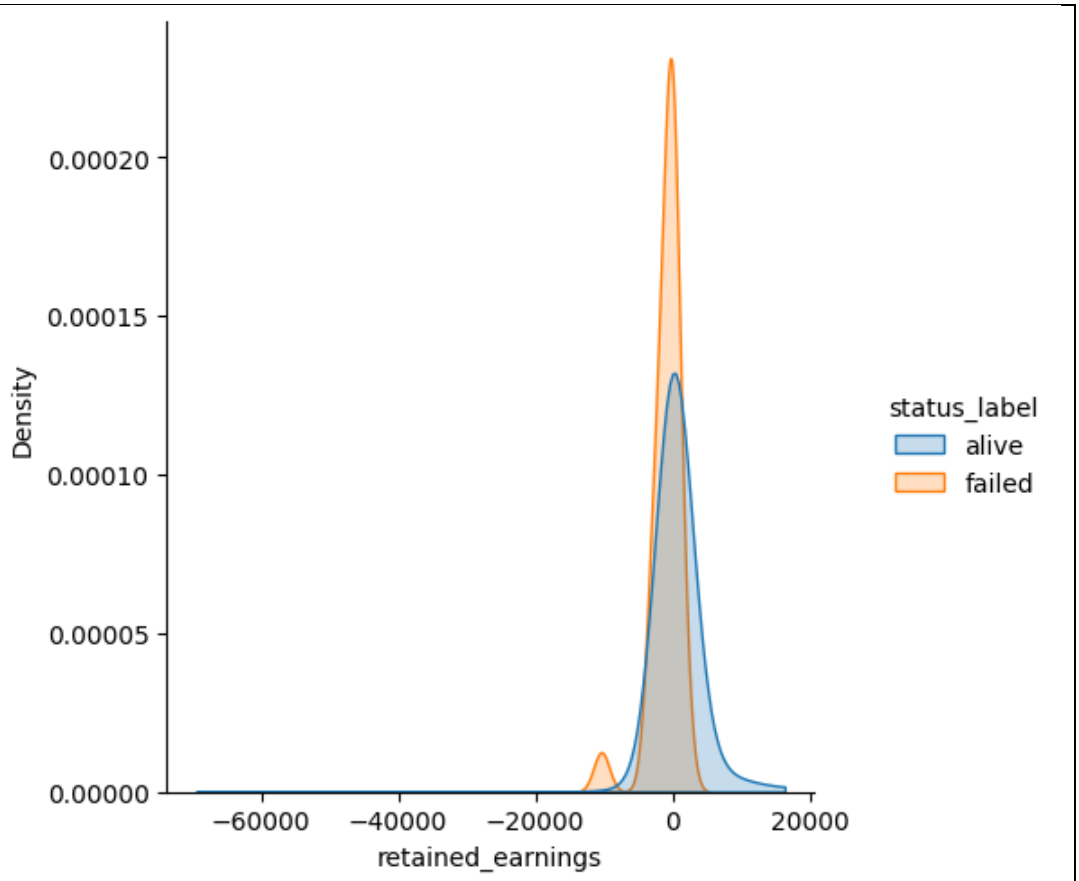

Figura 14: Distribuzione degli utili trattenuti nel dataset originale.

La distribuzione degli utili trattenuti (retained_earnings) nel dataset originale rivela una chiara capacità discriminante tra le imprese attive e quelle fallite. La curva delle imprese fallite si concentra prevalentemente intorno a valori molto prossimi allo zero o negativi, con un picco stretto e pronunciato. Questo suggerisce che le aziende che non sopravvivono tendano ad avere una storia contabile di perdite cumulate o di mancata accumulazione di capitale interno nel tempo.

Al contrario, la curva delle imprese attive mostra una distribuzione lievemente più ampia e spostata verso destra, con valori di utili trattenuti positivi, indicando una maggiore capacità storica di generare utili e reinvestirli nel ciclo economico. Questo comportamento riflette un'implicita relazione tra continuità aziendale e capacità di generare valore nel lungo termine.

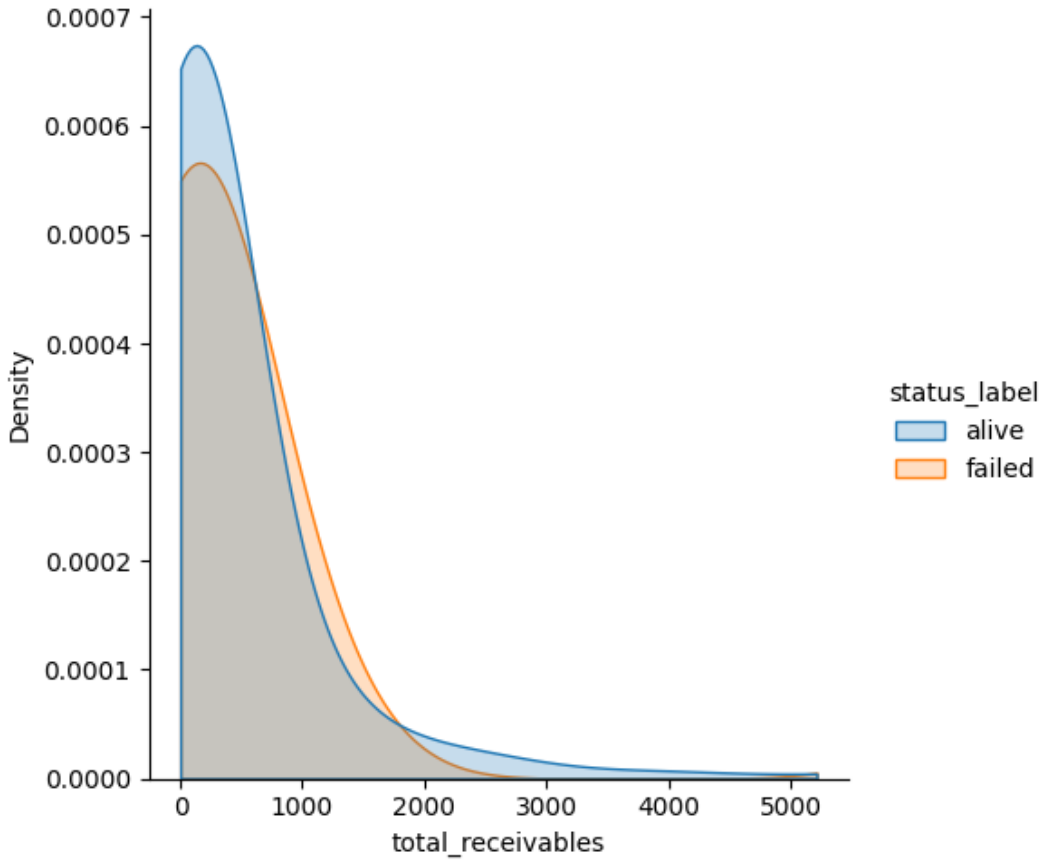

Figura 15: Distribuzione dei crediti totali nel dataset originale.

La distribuzione dei crediti totali (total_receivables) evidenzia una marcata asimmetria positiva, con una forte concentrazione nei valori più bassi e una coda lunga verso destra per entrambe le classi di imprese. Tuttavia, le aziende attive evidenziano una maggiore presenza nella fascia alta della distribuzione, rispetto a quelle fallite.

Questo comportamento suggerisce che le imprese sopravvissute siano in grado di sostenere rapporti commerciali più estesi, esprimendo un livello di attività economica più intenso e continuativo. Le imprese fallite, invece, presentano una distribuzione più ristretta e centrata su valori contenuti, indicando una ridotta capacità di generare credito commerciale o una scarsa attività operativa prima del fallimento.

Pur essendo una voce dell'attivo, il livello assoluto dei crediti deve essere interpretato con cautela, poiché può anche riflettere inefficienze nella gestione dell'incasso. La sua

reale utilità predittiva emerge quando è considerato in rapporto ai ricavi (giorni di incasso) o integrato in indicatori compositi di performance commerciale.

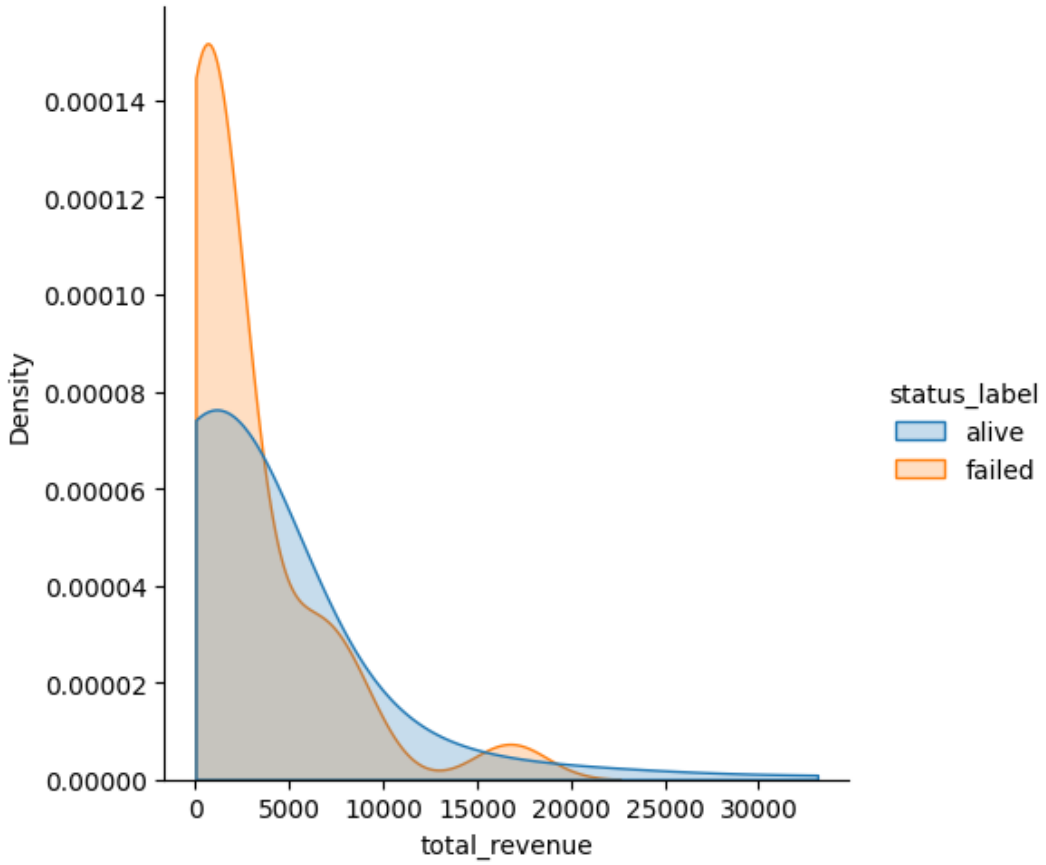

Figura 16: Distribuzione dei ricavi totali nel dataset originale.

La distribuzione dei ricavi totali (total_revenue) evidenzia una spiccata asimmetria positiva, con un'elevata concentrazione nei valori più bassi e una coda lunga verso destra. Le aziende attive presentano una maggiore densità nei valori medi e alti, mentre le imprese fallite si concentrano prevalentemente nei livelli di ricavo più contenuti.

Questa differenza suggerisce una chiara associazione tra il livello di fatturato e la capacità di sopravvivenza aziendale: le imprese fallite tendono infatti a essere caratterizzate da una scala operativa ridotta, potenzialmente sintomo di una limitata capacità competitiva o di una posizione di mercato fragile. Al contrario, le imprese attive mostrano ricavi più consistenti, indice di un'attività più stabile e sostenuta.

Sebbene il valore assoluto del fatturato non sia di per sé un predittore esaustivo del rischio di default, il grafico conferma la sua importanza come proxy della vitalità economica dell'impresa. La sua combinazione con altre variabili – in particolare con i costi operativi e la redditività – risulta essenziale per costruire modelli predittivi affidabili e robusti.

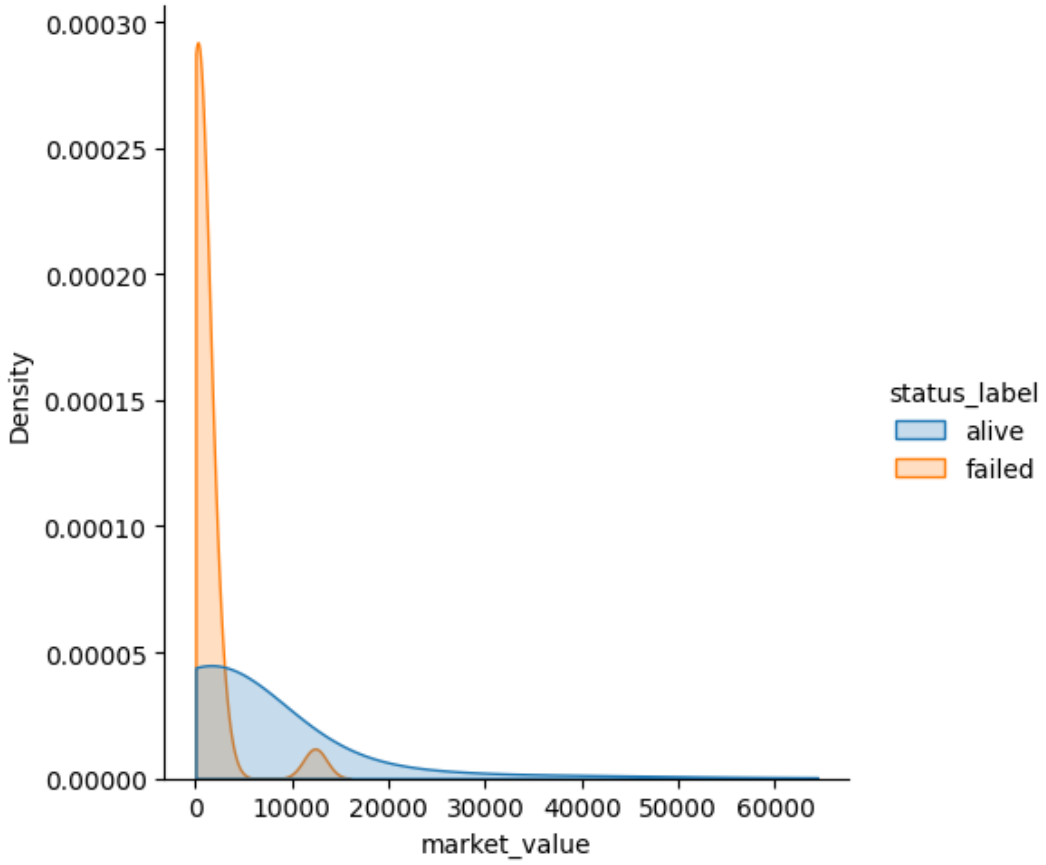

Figura 17: Distribuzione del valore di mercato nel dataset originale.

La distribuzione del valore di mercato (market_value) conferma le tendenze osservate per altre variabili di scala aziendale, con una netta asimmetria positiva e una concentrazione nei valori più bassi. Le imprese fallite risultano fortemente raggruppate nella fascia iniziale, mentre le aziende attive mostrano una distribuzione più estesa, con maggiore presenza anche nelle fasce intermedie e alte.

Questa differenza riflette il fatto che le imprese sopravvissute tendono ad avere una capitalizzazione più robusta, spesso correlata alla fiducia del mercato nella capacità futura di generare valore. Al contrario, le imprese fallite si caratterizzano per una bassa valutazione di mercato, sintomo di aspettative negative da parte degli investitori e di vulnerabilità economica.

Il valore di mercato rappresenta un indicatore sintetico della percezione esterna della solidità aziendale ed è spesso il risultato di molteplici fattori interni, quali redditività, indebitamento, prospettive di crescita e contesto settoriale. Per questo motivo, la sua integrazione nei modelli predittivi di default può aggiungere una dimensione di valore segnaletico particolarmente utile.

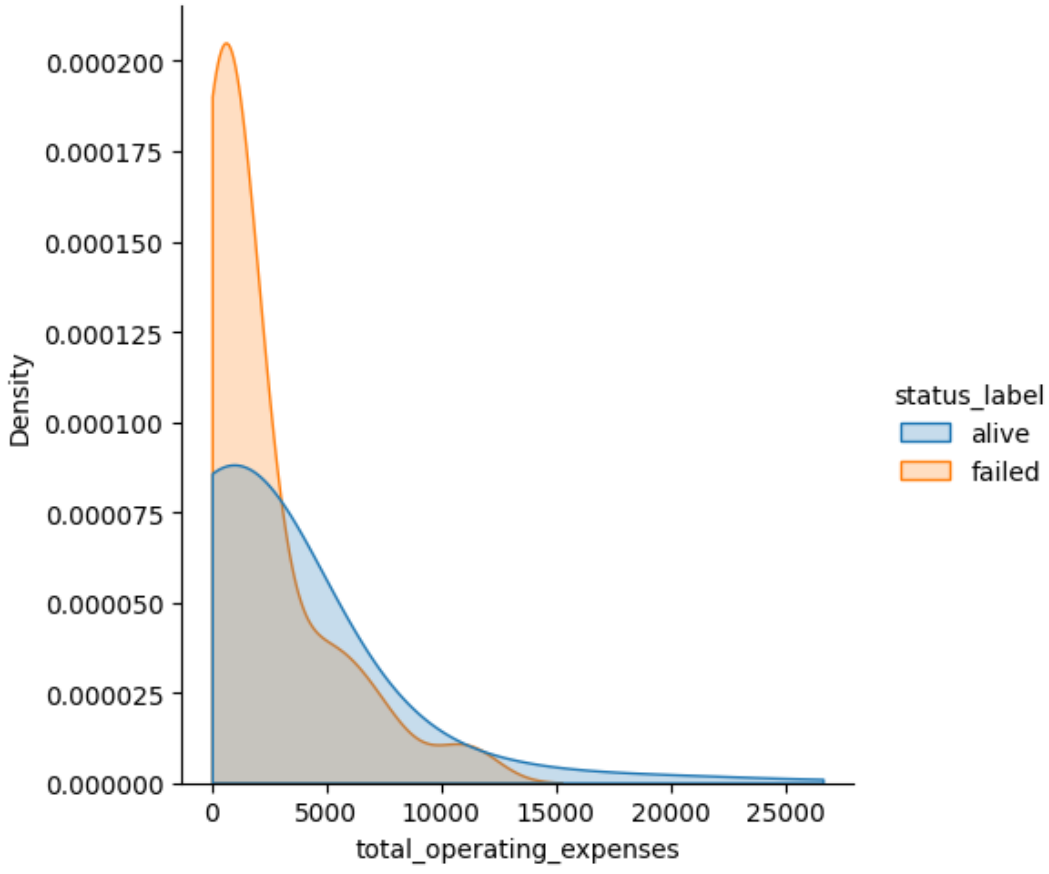

Figura 18: Distribuzione delle spese operative totali nel dataset originale.

La distribuzione delle spese operative totali (total_operating_expenses) conferma la tendenza osservata in altre variabili di struttura e scala aziendale. La curva di densità delle imprese fallite presenta un picco molto pronunciato nei valori più bassi, mentre le imprese attive mostrano una distribuzione più ampia e progressivamente decrescente lungo tutta la coda destra.

Questa differenza suggerisce che le imprese fallite operino tendenzialmente su scala ridotta, sostenendo livelli di spesa contenuti, coerenti con una minore attività produttiva. Le imprese attive, al contrario, sostengono costi operativi più elevati, indicativi di una maggiore dimensione organizzativa e capacità di generare valore attraverso la gestione operativa.

Le spese operative, se analizzate in isolamento, non permettono di determinare l'efficienza gestionale, ma assumono particolare significato quando messe in relazione con i ricavi, l'EBIT o il margine operativo. La loro integrazione in modelli predittivi consente di valutare la sostenibilità del modello di business aziendale e il bilanciamento tra costi e performance economiche.

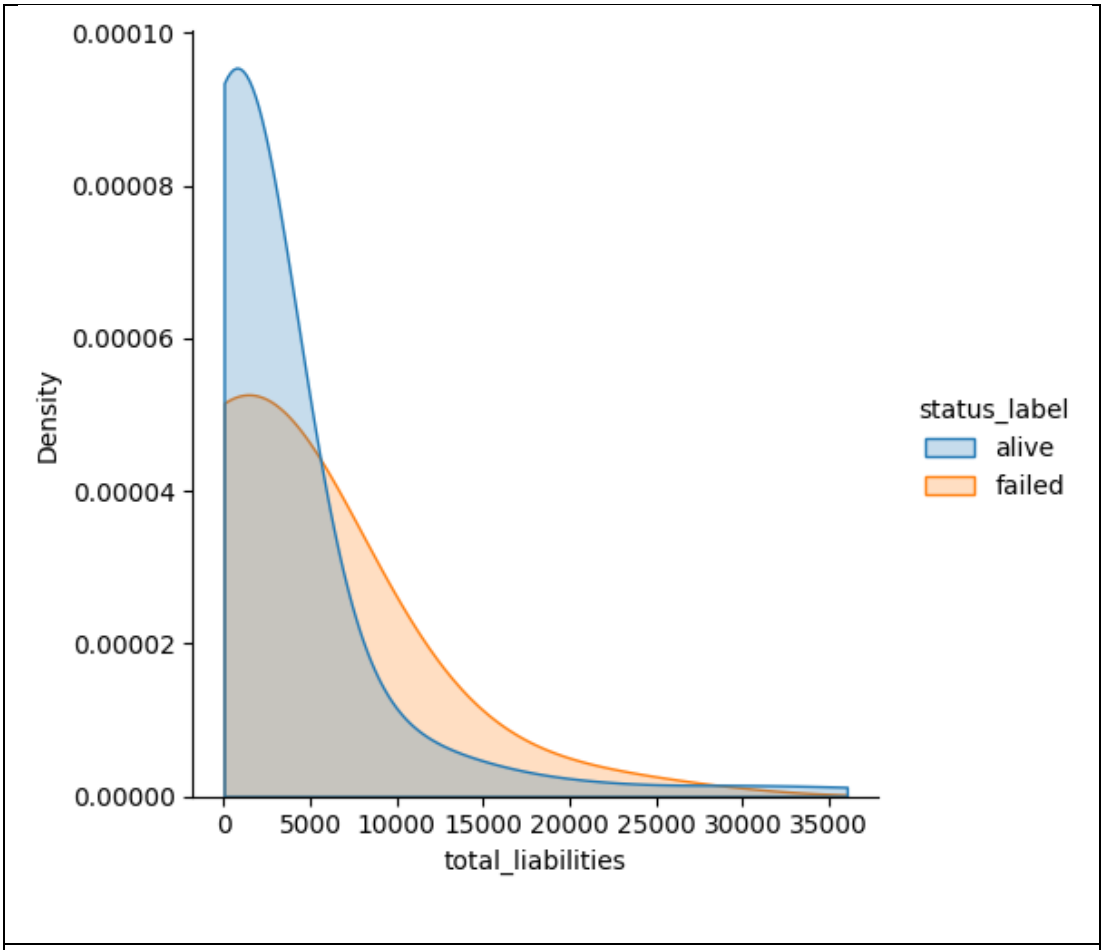

Figura 19: Distribuzione delle passività totali nel dataset originale.

La distribuzione delle passività totali (total_liabilities) rivela una differenza evidente tra le due classi di imprese. Le aziende attive mostrano una concentrazione nei valori più bassi, con una curva fortemente decrescente. Le imprese fallite, invece, presentano una distribuzione più ampia, con una densità significativamente maggiore anche in corrispondenza di livelli di indebitamento elevati.

Questa evidenza conferma che un elevato ammontare di debiti totali costituisce una condizione ricorrente tra le imprese che entrano in stato di default. L'indebitamento complessivo rappresenta, infatti, un fattore di rischio strutturale, che può compromettere la sostenibilità finanziaria dell'azienda, soprattutto in presenza di marginalità ridotte o di una gestione operativa inefficiente.

Nel contesto dell'analisi predittiva, la variabile total_liabilities risulta particolarmente informativa se combinata con altre voci di bilancio, come l'attivo totale o l'EBITDA,

per costruire indicatori sintetici come il leverage o l'indice di copertura del debito. La sua capacità discriminante, già evidente in questa analisi univariata, sarà ulteriormente valorizzata all'interno di modelli multivariati.

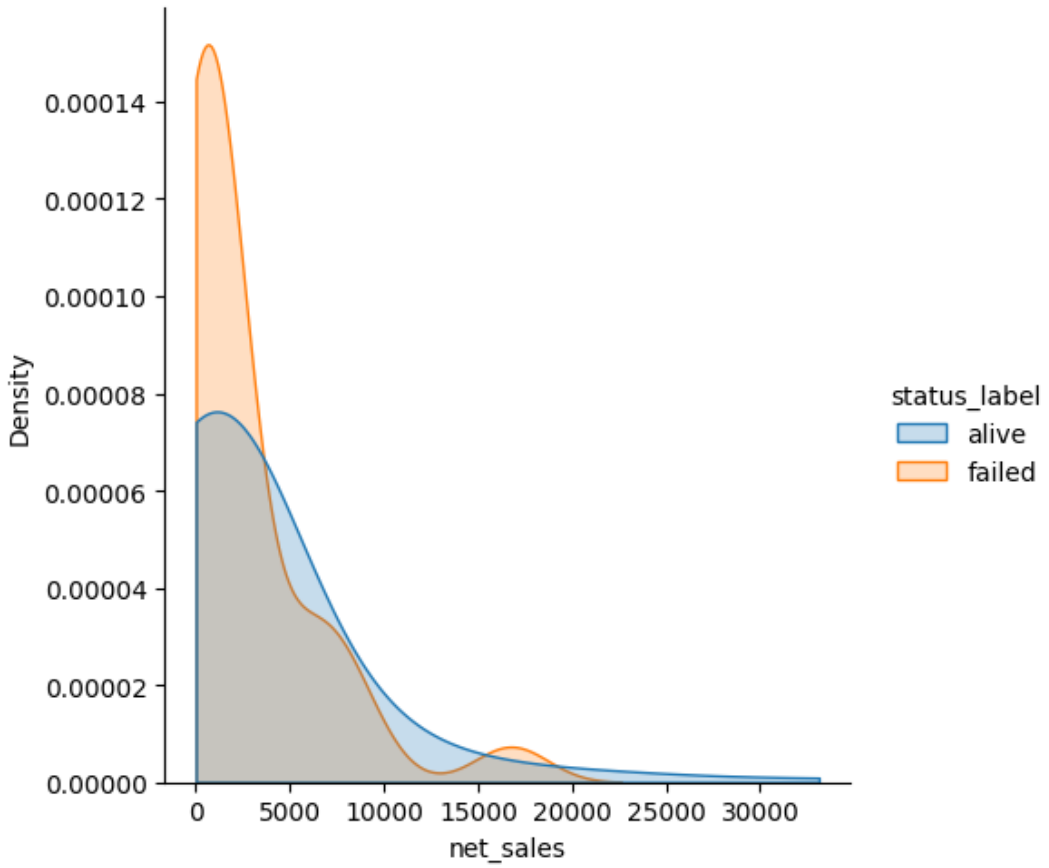

Figura 20: Distribuzione delle vendite nette nel dataset originale.

La distribuzione delle vendite nette (net_sales) evidenzia una dinamica simile a quella dei ricavi totali, con un andamento fortemente asimmetrico e una concentrazione nei valori inferiori. Le imprese fallite mostrano un picco iniziale più pronunciato, mentre le imprese attive evidenziano una coda più estesa verso destra, con maggiore densità nei livelli medi e alti.

Questo comportamento riflette una correlazione tra volume di vendite e capacità di sopravvivenza: le imprese attive tendono ad avere una base clienti più ampia, una maggiore continuità commerciale e, di conseguenza, una maggiore stabilità dei flussi operativi. Le imprese fallite, al contrario, mostrano una struttura di vendite ridotta, potenzialmente sintomatica di una perdita di competitività o di difficoltà nel posizionamento sul mercato.

Le vendite nette costituiscono un indicatore cruciale dell'attività operativa reale, spesso più affidabile dei ricavi totali in quanto depurato da componenti straordinarie o discontinuità contabili. In modelli predittivi, questa variabile assume particolare valore se utilizzata in combinazione con la marginalità, la rotazione degli attivi o gli indici di performance commerciale.

3.4 Generazione di dati sintetici

È stata effettuata la generazione sia con una Gaussian Copula che con CTGAN e TVAE. I risultati migliori in termini di qualità dei dati di sintesi sono stati conseguiti attraverso l'uso di un modello basato su TVAE anche se con un minimo scarto in termini di qualità rispetto alla Copula Gaussiana. In ogni caso, la maggiore velocità della Gaussian Copula, la rende preferibile soprattutto in contesti caratterizzati dalla presenza di grandi quantità di dati. In generale, tuttavia, la capacità di TVAE e CTGAN di riprodurre associazioni non lineari consiglia tuttavia di considerarne sempre l'uso, in dipendenza dalle particolari condizioni e dagli obiettivi. I risultati delle verifiche di congruenza diagnostica e di qualità effettuate in accordo ai controlli resi disponibili nel workflow standard di SDV (le due funzionalità, run_diagnostic) ed evaluate_quality() precedentemente citate) sono riportati nelle due tabelle seguenti.

MODELLO	VALIDITÀ DEI DATI	STRUTTURAZIONE	COMPLESSIVO
GAUSSIAN COPULA	100%	100%	100%
CTGAN	100%	100%	100%
TVAE	100%	100%	100%

Tabella1: Congruenza Diagnostica dei dati sintetici generati

MODELLO	COLUMN SHAPE	PAIR TRENDS	COMPLESSIVO
GAUSSIAN COPULA	85.42%	86.04%	85.73%
CTGAN	67.6%	70.28%	68.94%
TVAE	85.45%	92.83%	89.14%

Tabella2: Qualità dei dati sintetici generati

La scelta finale per le successive fasi dell'analisi è ricaduta sulla Gaussian Copula in quanto la generazione di dati sintetici ad essa associati ha dato luogo a un maggior numero di record validi ottenuti a seguito delle operazioni di filtraggio.

Per verificare la qualità della generazione dei dati sintetici, una volta selezionalo il sottoinsieme del dataset sul quale operare, si è osservata, la distribuzione delle variabili nei dati sintetici, confrontandola con quella rilevabile nel dataset originale. In questo modo, è stato possibile verificare che le associazioni tra le variabili siano state correttamente riprodotte nei dati sintetici. Andando oltre, la stessa verifica è stata effettuata applicando anche un filtro sui dati sintetici, così da adattarli ad un contesto diverso, per poter studiare come e in quale misura i cambiamenti di una variabile si riflettano sulle altre. Quest'ultima analisi è descritta in dettaglio nelle sezioni successive.

3.5 Selezione e Filtraggio

Per verificare il funzionamento dell'approccio ci si è basato su una logica di filtraggio *ex post*, realizzato attraverso combinazioni di semplici meccanismi di soglia.

Da 2000 istanze di dati sintetici generati, sono state successivamente selezionate nel contesto del processo di filtraggio, soltanto quelle che avevano il costo dei beni venduti compreso tra 500$ e 15000$.

Sono state ottenuta a valle del processo di filtraggio 917 righe. Di esse sono state calcolate e graficate le distribuzioni usando lo stesso criterio, lo scarto del 2,5% superiore dei dati, adottato nell'analisi esplorativa.

Per quanto riguarda la variabile sulla quali siamo intervenuti, come si vede, la Figura 23 evidenzia il picco poco sotto ai 10000$ per le aziende fallite, picco che era già visibile nella Figura 5, anche se lì si collocava intorno agli 11000$.

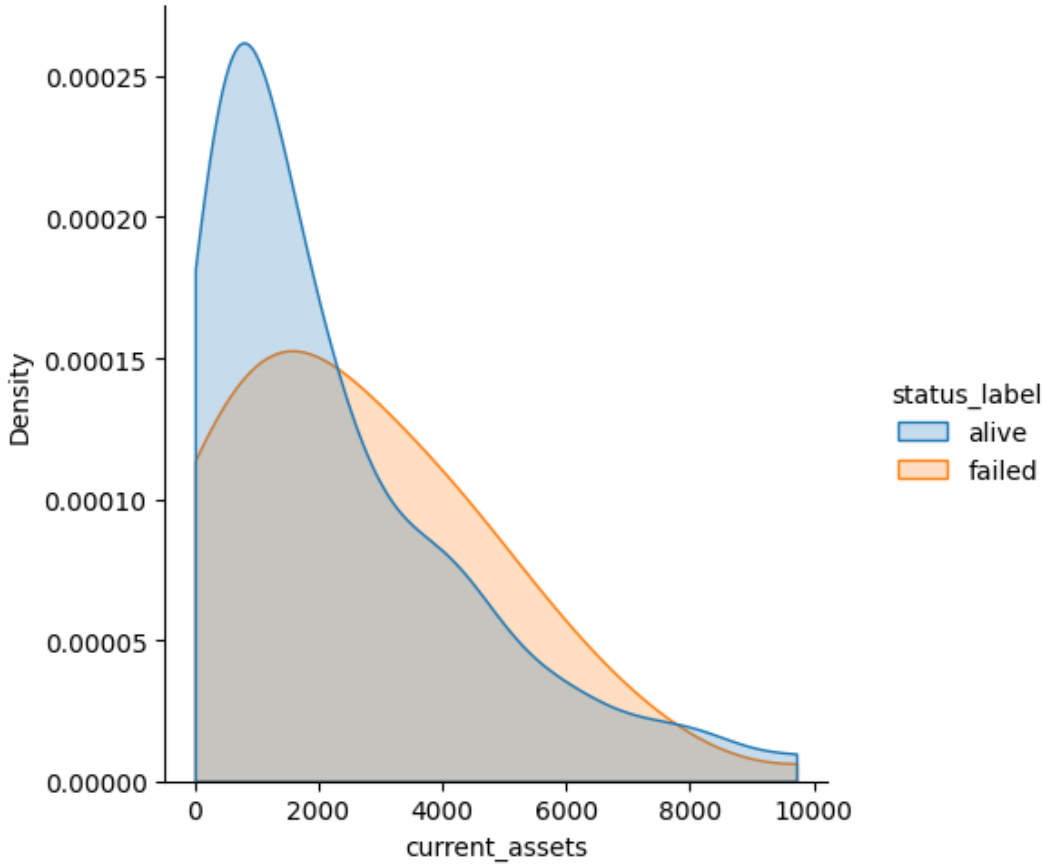

Figura 21: Distribuzione delle attività correnti nel dataset sintetico.

Confrontando le figure da 21 a 38 con le corrispondenti (da 3 a 20) si osserva come alcune distribuzioni per i dati sintetici (depreciation_and_amortization, ebit, ebitda, gross_profit, retained_earnings, total_assets, total_current_liabilities, total_liabilities, total_long_term_debt, total_receivables, total_revenue) restano simili a quelle dei dati originali e al più possono essere considerate un ingrandimento di una regione della distribuzione originale, mentre altre mostrano cambiamenti notevoli (ovviamente cost_of_goods_sold, ma anche current_assets, inventory, market_value, net_income, net_sales, total_operating_expenses). Ciò suggerisce un legame tra la variable modificata (cost_of_goods_sold) e quelle del secondo gruppo.

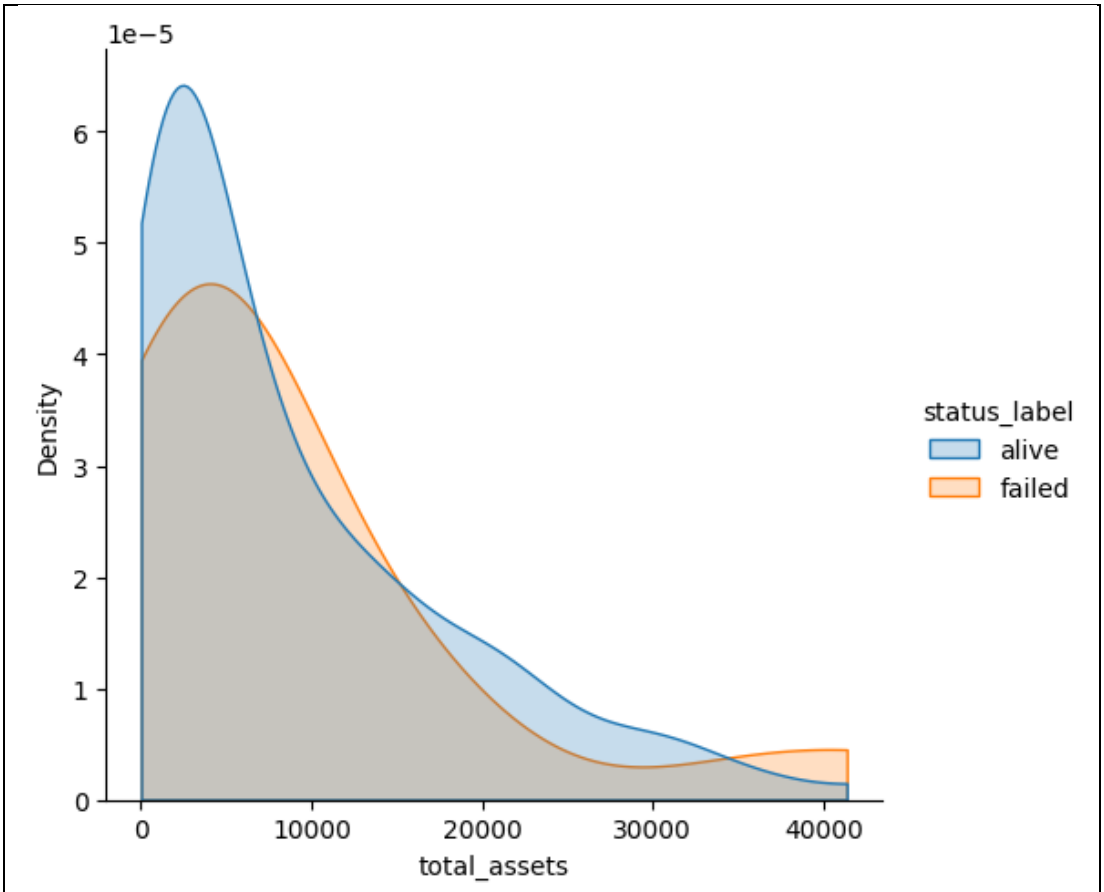

Figura 22: Distribuzione delle attività totali nel dataset sintetico.

La distribuzione delle attività totali (total_assets) nel dataset sintetico evidenzia una forma complessivamente simile a quella osservata nel dataset originale, ma con alcune differenze rilevanti. Si osserva una maggiore sovrapposizione tra le due classi ("alive" e "failed"), con curve di densità che divergono solo parzialmente, specialmente nella coda destra.

Le imprese attive mantengono una maggiore densità nei valori più elevati di attivo totale, ma le imprese fallite nel dataset sintetico si estendono in modo più significativo anche verso valori alti, suggerendo che il processo di generazione sintetica ha introdotto una maggiore varietà di dimensione aziendale tra le imprese fallite. Questo comportamento potrebbe avere lo scopo di arricchire la variabilità della classe "failed", migliorando l'addestrabilità dei modelli predittivi attraverso una rappresentazione più ampia degli scenari possibili.

Dal punto di vista metodologico, il confronto tra la distribuzione nel dataset sintetico e quella originale è essenziale per valutare la qualità della sintesi e l'aderenza ai pattern economico-finanziari reali. In particolare, il fatto che la differenza tra le classi rimanga visibile – pur con maggiore sovrapposizione – indica che l'informazione discriminante non è stata persa nel processo di generazione artificiale dei dati.

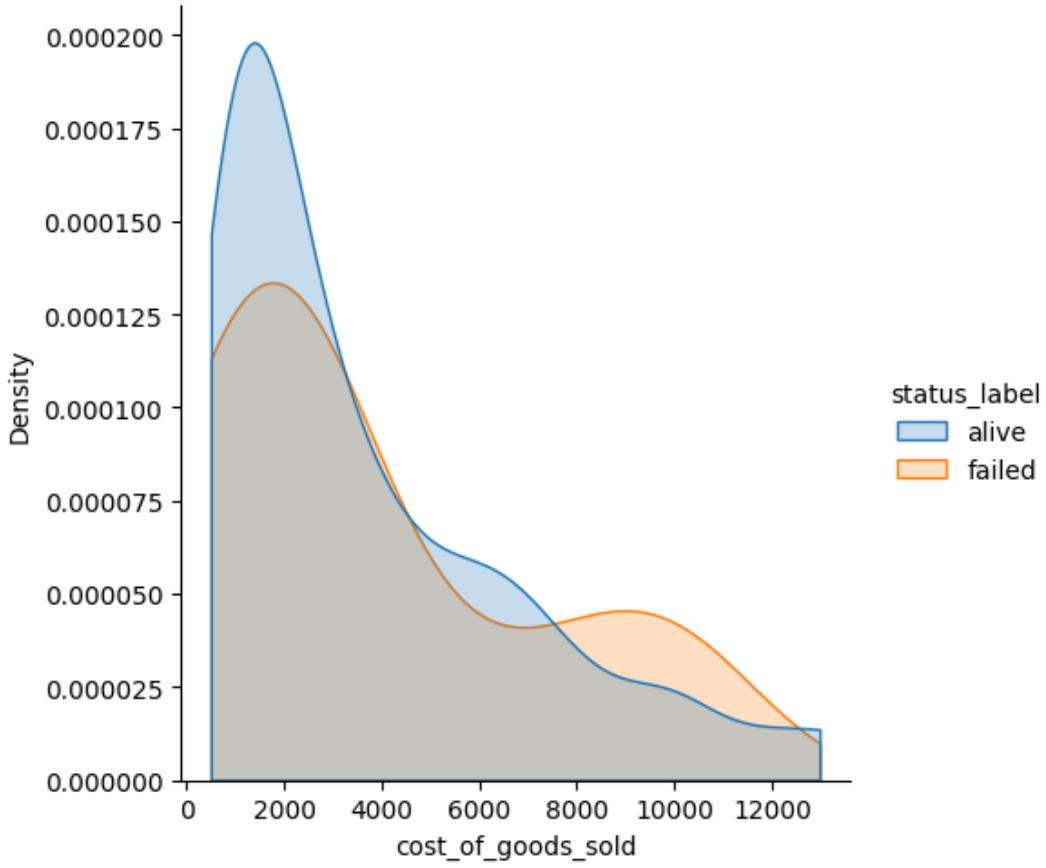

Figura 23: Distribuzione del costo dei beni venduti nel dataset sintetico.

Nel dataset sintetico, la distribuzione del costo dei beni venduti (cost_of_goods_sold) evidenzia un comportamento interessante rispetto al dataset originale. Le curve di densità mostrano un andamento simile nella fascia bassa della distribuzione, ma divergono progressivamente nei valori medi e alti: le imprese fallite nel dataset sintetico presentano una densità più elevata nei costi superiori a 6.000, mentre le imprese attive sono maggiormente concentrate sotto tale soglia.

Questa configurazione suggerisce che, nel processo di generazione sintetica, siano stati inseriti scenari in cui imprese fallite sostengono livelli elevati di costi operativi – potenzialmente sproporzionati rispetto alla capacità di generare ricavi – accentuando così una delle possibili dinamiche di dissesto.

Rispetto al dataset originale, in cui la discriminazione tra le due classi era meno marcata in questa variabile, il dataset sintetico sembra amplificare il valore predittivo del costo dei beni venduti. Tale scelta può contribuire a rendere i modelli di apprendimento automatico più sensibili a squilibri strutturali nella gestione industriale, migliorando la capacità di identificare pattern di rischio.

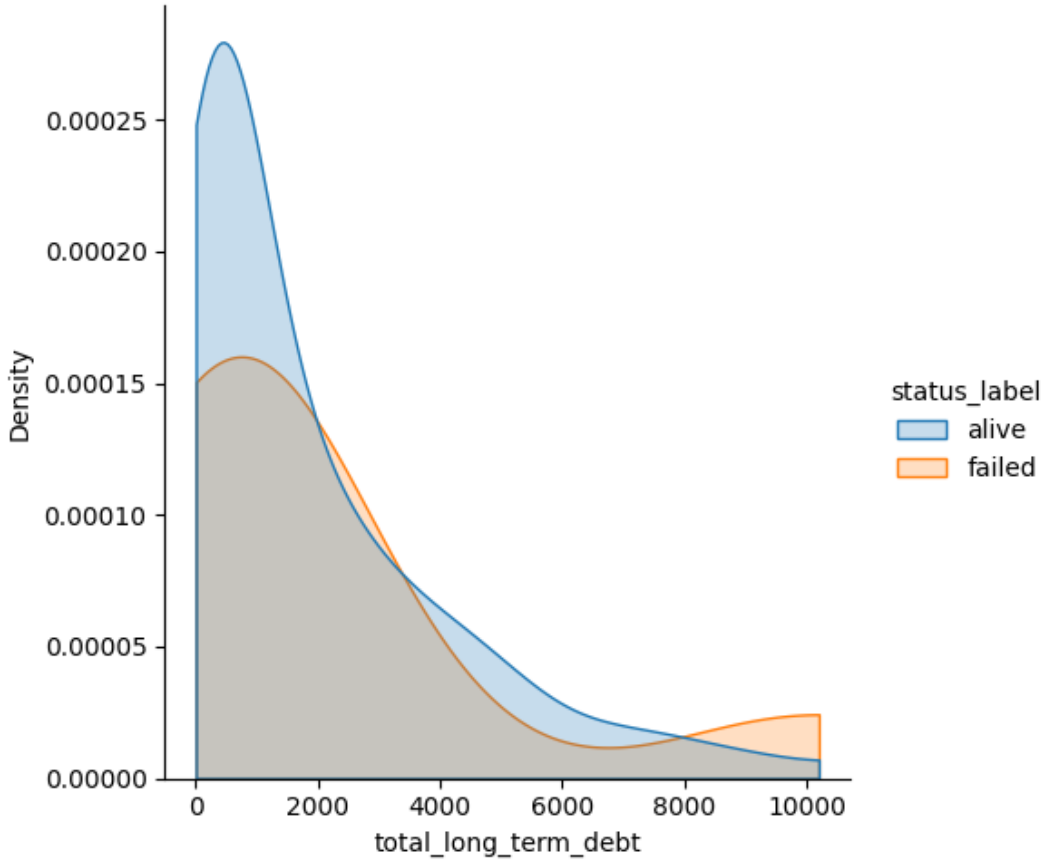

Figura 24: Distribuzione del debito totale a lungo termine nel dataset sintetico.

La distribuzione del debito totale a lungo termine (total_long_term_debt) nel dataset sintetico evidenzia un profilo più differenziato rispetto a quello osservato nel dataset originale. Le imprese attive risultano concentrate prevalentemente nei valori più bassi del debito, mentre le imprese fallite si distribuiscono in maniera più ampia, con un incremento della densità relativa nelle fasce comprese tra i 6.000 e i 10.000.

Questa configurazione lascia intendere che il dataset sintetico abbia accentuato il legame tra indebitamento strutturale elevato e probabilità di fallimento, aumentando l'incidenza di scenari in cui il debito di lungo periodo rappresenta un elemento critico. Si tratta di una strategia plausibile nei processi di oversampling, volta a fornire al modello di apprendimento automatico esempi più variegati di dissesto.

La presenza di un picco nella coda destra della distribuzione delle imprese fallite – che nel dataset originale era molto meno marcato – segnala un potenziale rafforzamento della capacità predittiva di questa variabile all'interno del modello. L'ipotesi economico-finanziaria implicita è che un eccessivo ricorso al debito a lungo termine, se non accompagnato da un'adeguata redditività o liquidità, aumenti significativamente il rischio di default.

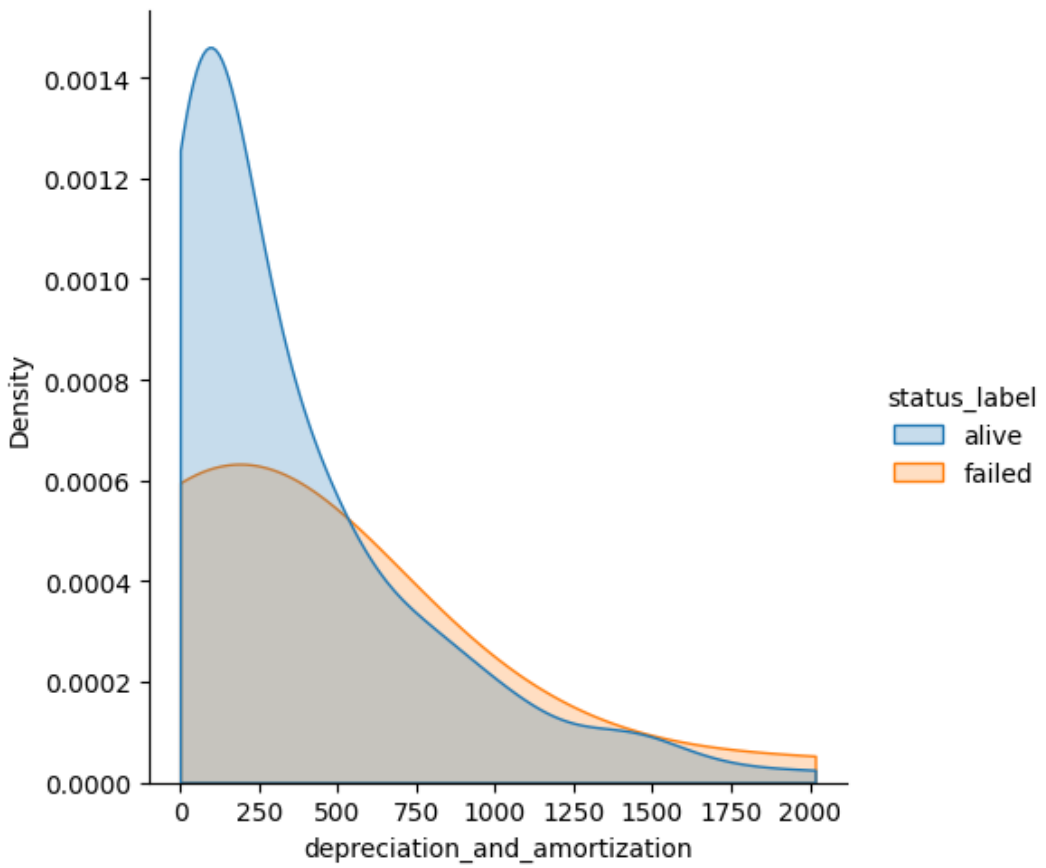

Figura 25: Distribuzione di ammortamento e deprezzamento nel dataset sintetico.

La distribuzione della voce di ammortamento e deprezzamento (depreciation_and_amortization) nel dataset sintetico evidenzia una separazione più marcata tra le classi rispetto a quanto osservato nel dataset originale. Le imprese attive risultano concentrate nei livelli più bassi, con un picco molto pronunciato entro i primi 250 unità, mentre le imprese fallite presentano una distribuzione più uniforme, con maggiore densità nei valori medi e alti.

Questa modifica nella distribuzione suggerisce che, nel processo di generazione sintetica, si sia voluto enfatizzare il peso dell'ammortamento come potenziale fattore di stress economico. Valori elevati di questa voce, infatti, possono riflettere una struttura di asset pesante, investimenti eccessivi o non efficienti, o ancora la necessità di rettifiche per riduzione di valore in contesti di bassa redditività.

Rispetto al dataset originale – dove le due curve erano più sovrapposte – l'informazione contenuta nella variabile risulta qui più discriminante. L'obiettivo di una tale configurazione potrebbe essere quello di rafforzare l'efficacia predittiva della variabile nel training dei modelli, simulando situazioni in cui l'ammortamento diventa un indicatore di deterioramento strutturale dei fondamentali aziendali.

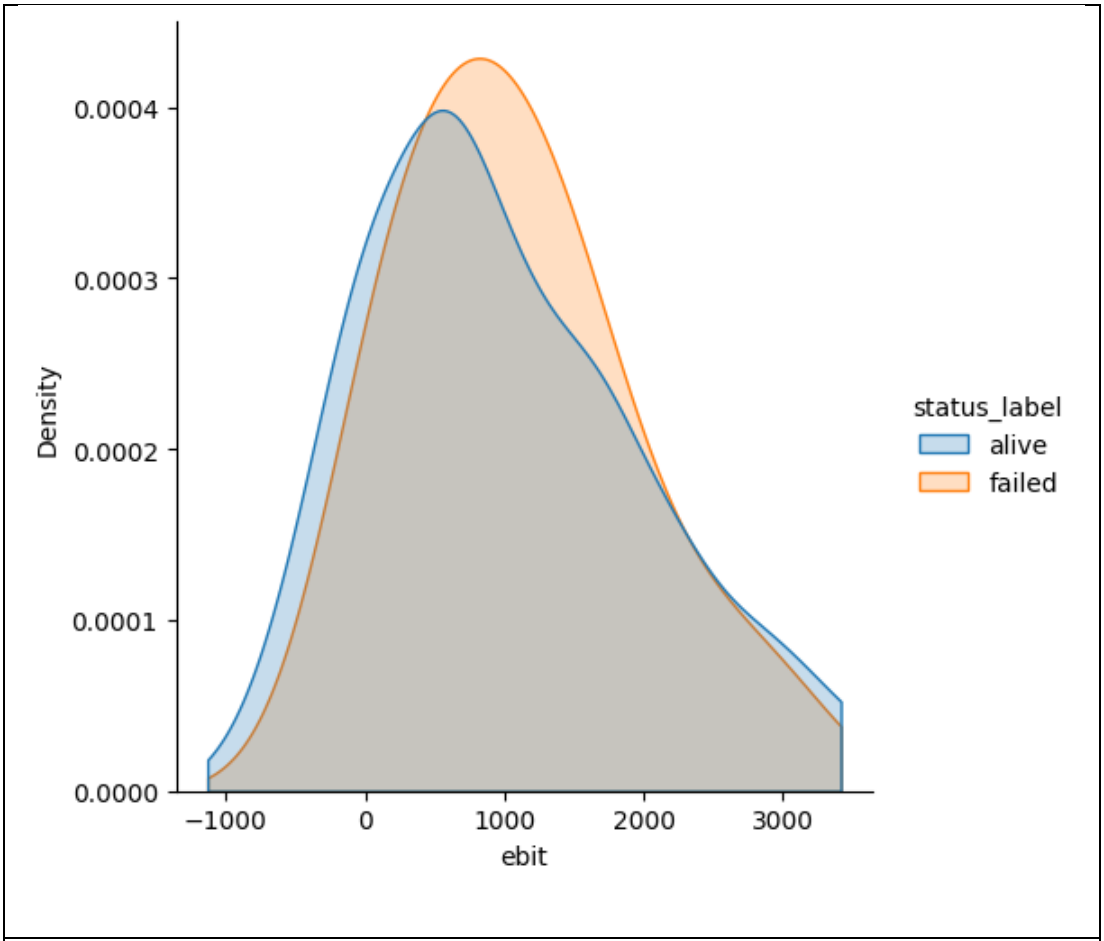

Figura 26: Distribuzione dell'EBIT nel dataset sintetico.

Nel dataset sintetico, la distribuzione dell'EBIT (Earnings Before Interest and Taxes) evidenzia un'evidente convergenza tra le due classi, con una sovrapposizione quasi completa per la maggior parte della distribuzione. Tuttavia, si nota un lieve spostamento della curva delle imprese fallite verso destra rispetto a quella delle imprese attive, suggerendo un comportamento in parte controintuitivo rispetto ai dati reali.

Nel dataset originale, l'EBIT si era mostrato una variabile ad elevata capacità discriminante, con le imprese fallite concentrate su valori negativi o molto bassi. Al contrario, nel dataset sintetico, il picco per le imprese fallite si colloca attorno a valori positivi (circa 1000), con una densità perfino maggiore di quella delle imprese attive nella stessa fascia.

Tale configurazione potrebbe essere frutto di una generazione artificiale volta ad aumentare la varietà dei casi di default, includendo scenari di fallimento che si verificano nonostante una redditività operativa momentaneamente positiva. Ciò può essere coerente con logiche modellistiche più avanzate, che intendono simulare crisi indotte da altri fattori (es. eccesso di debito, shock di liquidità, eventi straordinari).

In ogni caso, la differente struttura della distribuzione suggerisce cautela nell'interpretazione dell'EBIT come predittore diretto nel dataset sintetico, e sottolinea l'importanza di combinare le variabili all'interno di modelli multivariati.

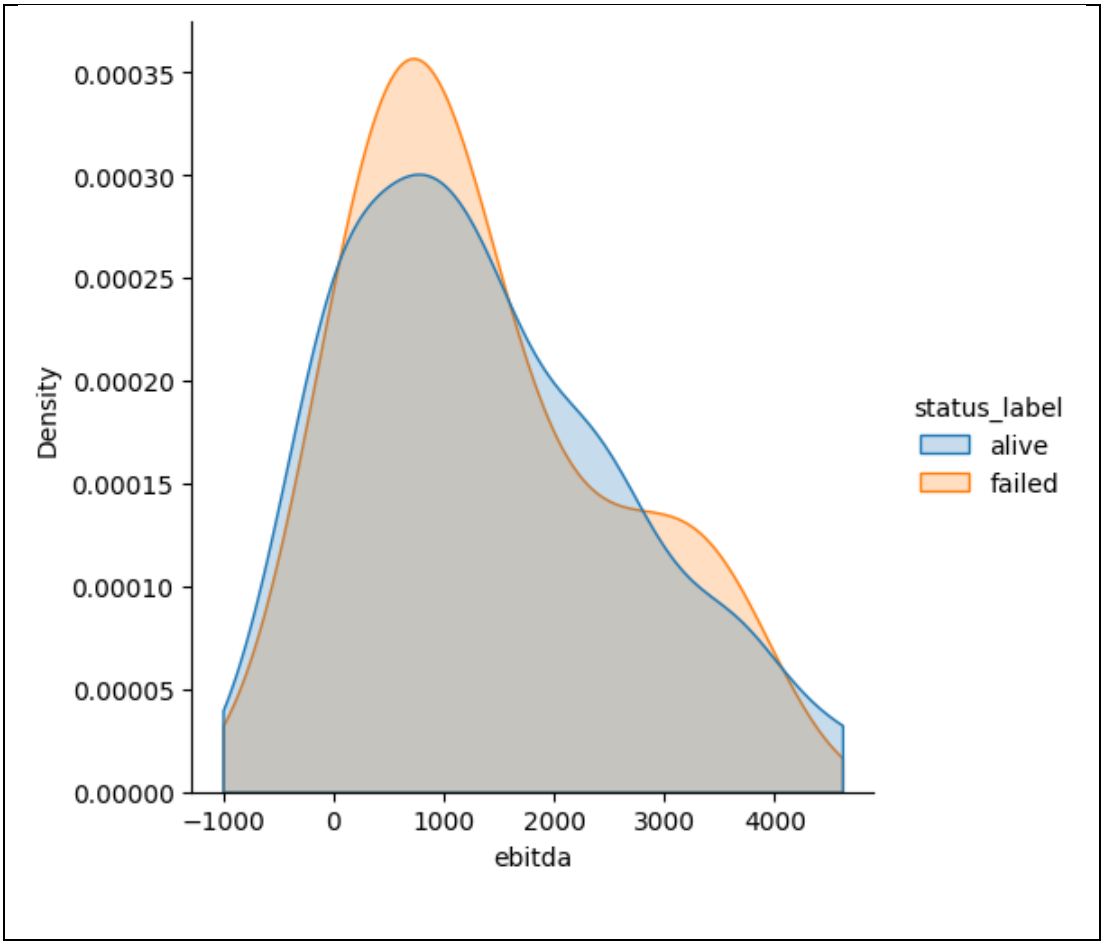

Figura 27: Distribuzione dell'EBITDA nel dataset sintetico.

La distribuzione dell'EBITDA (Earnings Before Interest, Taxes, Depreciation and Amortization) nel dataset sintetico si presenta con una forte sovrapposizione tra le due classi, ma con alcune differenze degne di nota. Le imprese fallite mostrano un picco leggermente più marcato e spostato verso sinistra rispetto alle imprese attive, che invece si distribuiscono più uniformemente nella fascia medio-alta della variabile.

Questa configurazione, sebbene non radicalmente diversa da quella dell'EBIT nel dataset sintetico (Figura 26), riflette una maggiore concentrazione di valori EBITDA contenuti tra le imprese fallite, suggerendo una generazione di scenari in cui l'azienda presenta un flusso operativo lordo positivo ma non sufficientemente elevato da garantirne la sopravvivenza nel lungo periodo.

Nel dataset reale, la variabile EBITDA non si era mostrata altamente discriminante. Qui, invece, il dataset sintetico sembra voler introdurre una maggiore eterogeneità nei fallimenti, includendo anche casi "meno estremi" dal punto di vista reddituale. Questo potrebbe migliorare la capacità del modello di riconoscere imprese vulnerabili anche quando non presentano segnali operativi gravemente negativi.

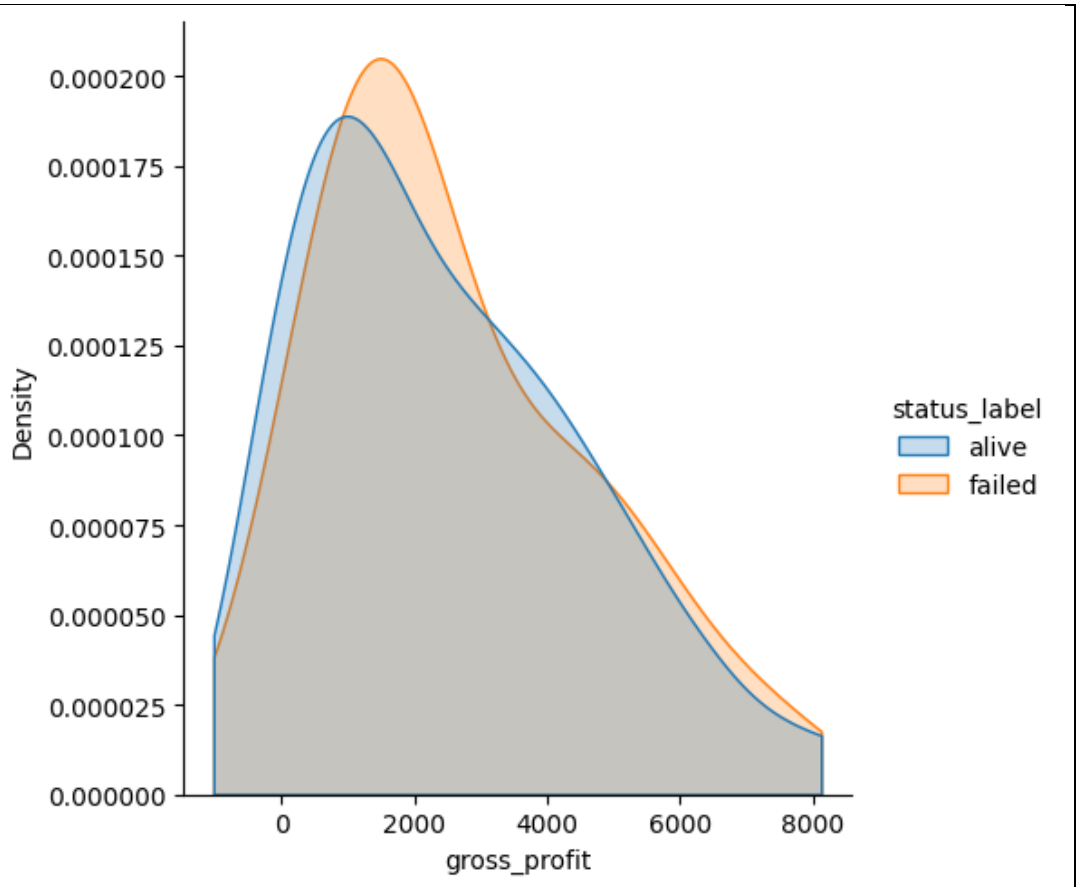

Figura 28: Distribuzione del margine lordo nel dataset sintetico.

La distribuzione del margine lordo (gross_profit) nel dataset sintetico evidenzia una sovrapposizione piuttosto estesa tra le due classi, ma con una caratteristica peculiare: la curva delle imprese fallite è leggermente spostata verso destra rispetto a quella delle imprese attive, evidenziando un picco più marcato nella fascia attorno ai 2.000.

Questo pattern è atipico rispetto a quanto osservato nel dataset originale, dove il margine lordo era tendenzialmente più basso nelle imprese fallite. Nel dataset sintetico, invece, sembrerebbe che siano stati generati numerosi casi di fallimento anche in presenza di margini lordi positivi o mediamente elevati. Tale scelta potrebbe essere giustificata dalla volontà di rappresentare scenari di fallimento dovuti a inefficienze nella gestione operativa o finanziaria, piuttosto che a problemi di profitto lordo in sé.

Questa configurazione arricchisce la varietà dei dati fallimentari, consentendo al modello predittivo di apprendere che un buon margine lordo non è condizione

sufficiente per evitare il default, soprattutto se l'azienda presenta alti costi fissi, debiti eccessivi, o flussi di cassa negativi.

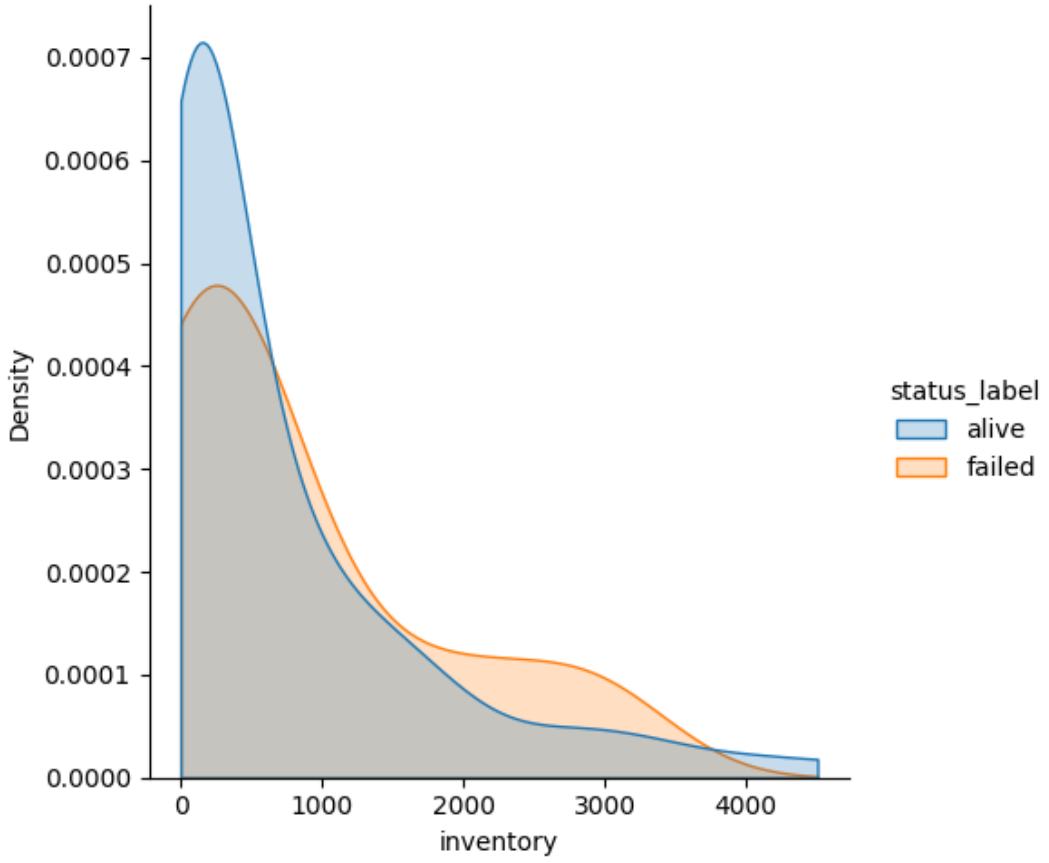

Figura 29: Distribuzione dell'inventario nel dataset sintetico.

La distribuzione dell'inventario (inventory) nel dataset sintetico evidenzia un'interessante inversione rispetto a quanto osservato nei dati reali. Le imprese attive risultano fortemente concentrate nei livelli più bassi di inventario, mentre le imprese fallite presentano una distribuzione più ampia, con una maggiore densità anche in corrispondenza di valori elevati.

Questo pattern suggerisce che, nel processo di generazione sintetica, sia stato enfatizzato il ruolo potenzialmente problematico di un elevato livello di rimanenze.

Nella letteratura economico-finanziaria, un inventario eccessivo è spesso interpretato come segnale di inefficienze operative, difficoltà di smaltimento del prodotto o scarsa domanda, tutti elementi potenzialmente riconducibili a situazioni di crisi aziendale.

La forma della curva delle imprese fallite evidenzia, inoltre, un secondo picco oltre i 2000, assente nella distribuzione delle imprese attive: questa caratteristica può essere interpretata come un tentativo deliberato di arricchire il dataset con profili di fallimento legati a un sovra-accumulo di magazzino.

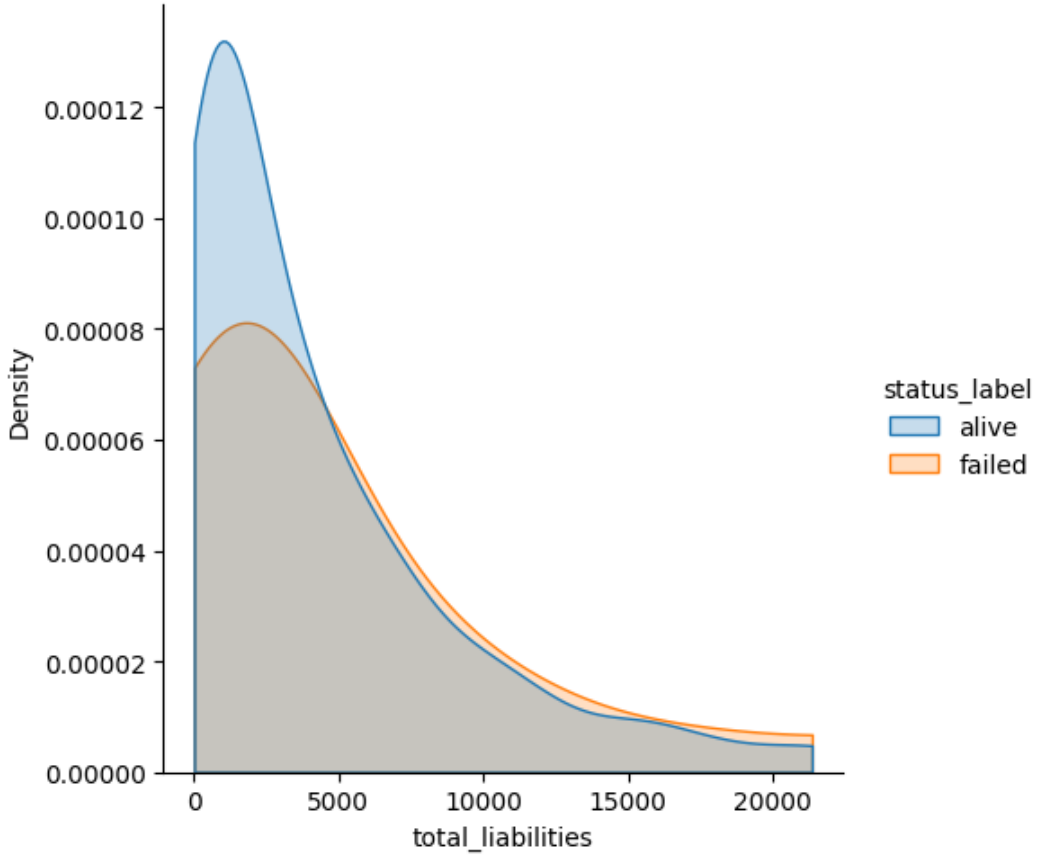

Figura 30: Distribuzione delle passività correnti totali nel dataset sintetico.

Nel dataset sintetico, la distribuzione delle passività totali (total_liabilities) mostra una tendenza analoga a quella osservata nei dati reali, ma con differenze meno marcate tra le due classi. Le imprese attive presentano un picco netto nei livelli più bassi di indebitamento, mentre le imprese fallite risultano leggermente più distribuite lungo la

coda destra della distribuzione, indicando una maggiore presenza di soggetti con passività elevate.

Tuttavia, rispetto al dataset originale, la separazione tra le due classi è qui più attenuata. La sovrapposizione tra le curve è infatti significativa per la maggior parte del dominio, e solo nella fascia compresa tra 5.000 e 10.000 si osserva un lieve scostamento a favore delle imprese fallite.

Questa struttura può indicare che il dataset sintetico è stato progettato per includere casi di default anche con livelli di passività relativamente contenuti, allo scopo di non rendere troppo dipendente la classificazione da una singola variabile. Ciò contribuisce a rafforzare la complessità dei dati di addestramento, rendendo il modello più sensibile alla combinazione di variabili piuttosto che a soglie rigide.

Nel complesso, pur mantenendo un comportamento coerente con la teoria economico-finanziaria, la variabile total_liabilities appare nel dataset sintetico meno discriminante rispetto ad altri indicatori, ma utile in chiave multivariata per rafforzare pattern di rischio associati all'indebitamento.

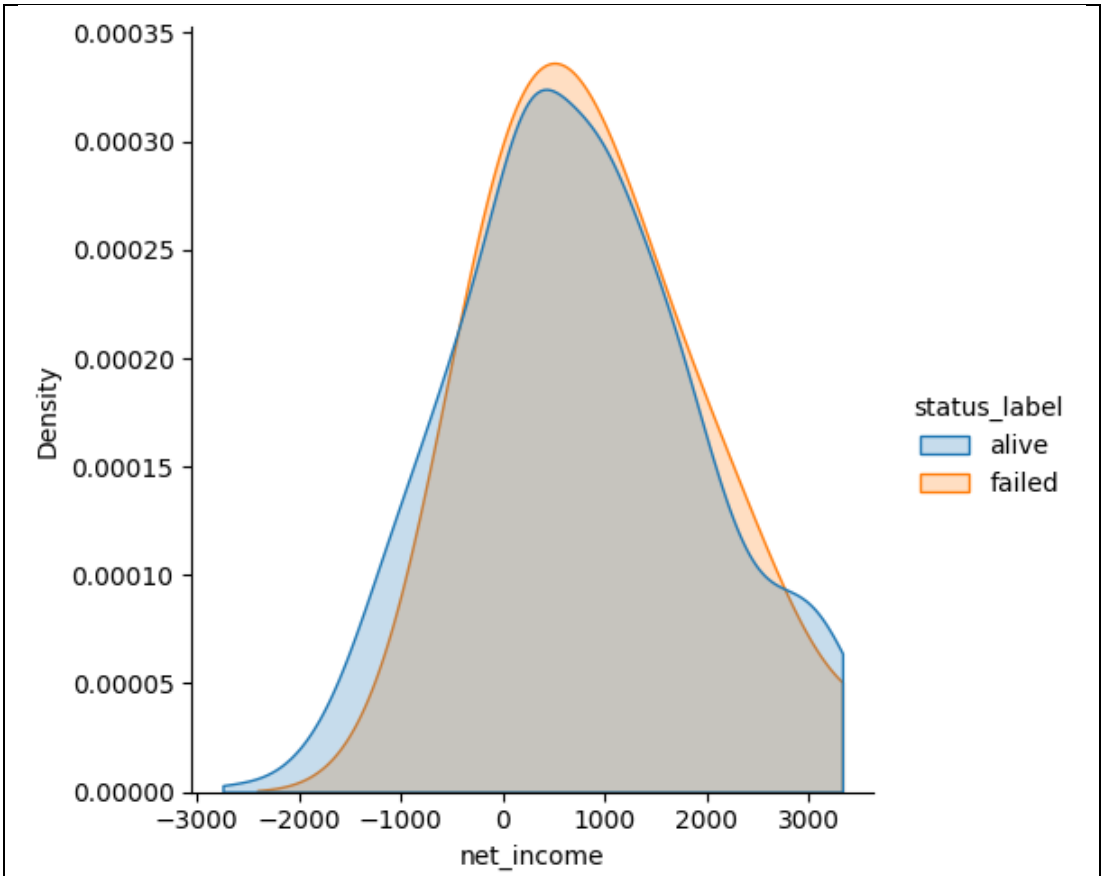

Figura 31: Distribuzione dell'utile netto nel dataset sintetico.

Entrambe le classi evidenziano una concentrazione centrale attorno allo zero, con un leggero picco maggiore per le imprese fallite nei pressi del valore mediano. Questa configurazione differisce da quella del dataset reale, dove le imprese fallite tendevano ad accumularsi in corrispondenza di valori negativi più estremi.

Nel dataset sintetico, la distribuzione è stata chiaramente costruita per includere casi di fallimento non necessariamente associati a perdite nette significative, simulando scenari in cui il deterioramento aziendale avviene per cause diverse dal risultato economico finale (es. tensioni di liquidità, indebitamento eccessivo, instabilità patrimoniale).

Tale approccio consente di arricchire la complessità del modello predittivo, offrendo al sistema di apprendimento esempi più eterogenei di imprese in crisi, anche in presenza di profitti contabili positivi o in linea con la media del campione.

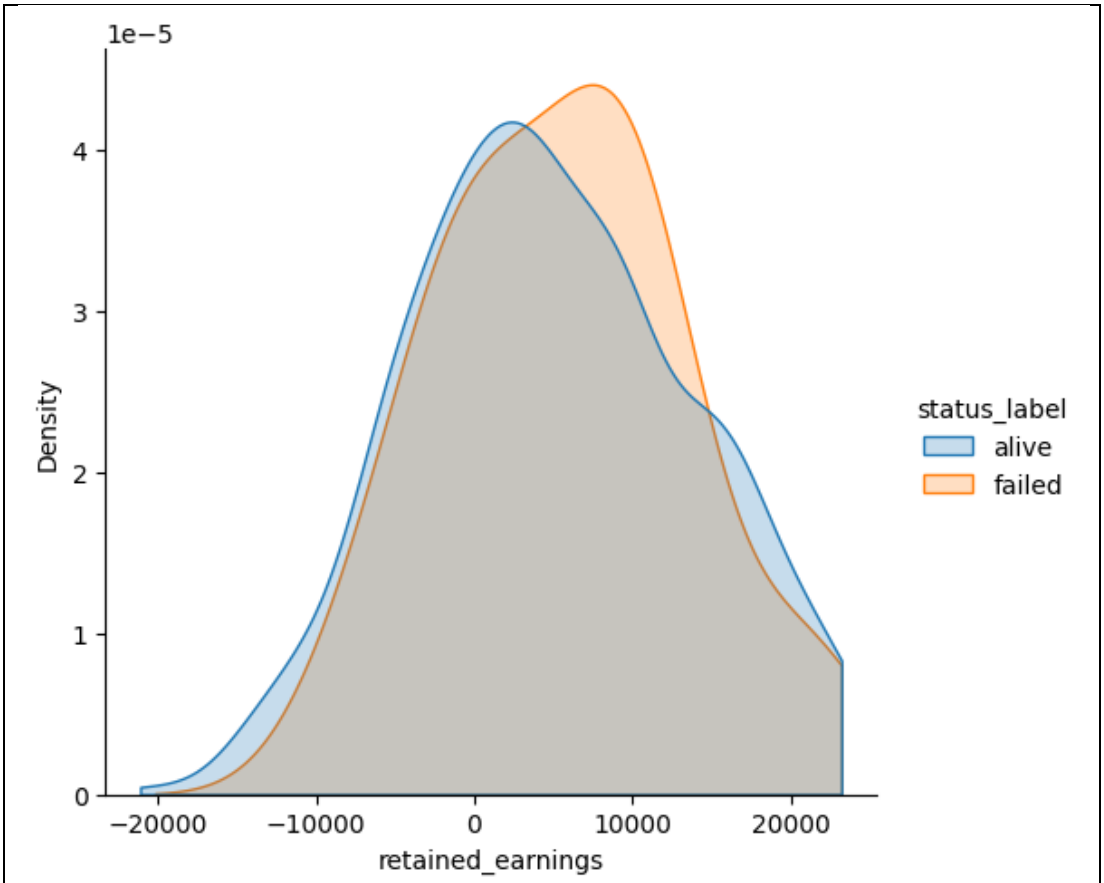

Figura 32: Distribuzione degli utili trattenuti nel dataset sintetico.

La distribuzione degli utili trattenuti (retained_earnings) nel dataset sintetico evidenzia una sovrapposizione significativa tra imprese attive e fallite, ma con una leggera predominanza di valori positivi tra le imprese fallite, il che risulta parzialmente controintuitivo.

Nel dataset reale, le imprese fallite tendevano a presentare utili trattenuti negativi o molto contenuti, riflettendo un percorso aziendale caratterizzato da perdite cumulate e assenza di reinvestimento interno. Nel dataset sintetico, al contrario, il picco della curva per le imprese fallite si sposta verso destra, indicando una maggiore frequenza di fallimenti anche tra le aziende con storici contabili formalmente solidi. Questa impostazione potrebbe derivare dall'intenzione di simulare scenari di crisi improvvise o non strettamente legate al passato contabile, ma piuttosto a shock recenti, crisi di

liquidità o cambiamenti strutturali che non si riflettono immediatamente sulla voce "retained earnings".

Inoltre, la distribuzione troncata e simmetrica – con limiti artificiali evidenti a circa ±20.000 – è coerente con la natura sintetica del dataset e con la necessità di contenere la variabilità estrema.

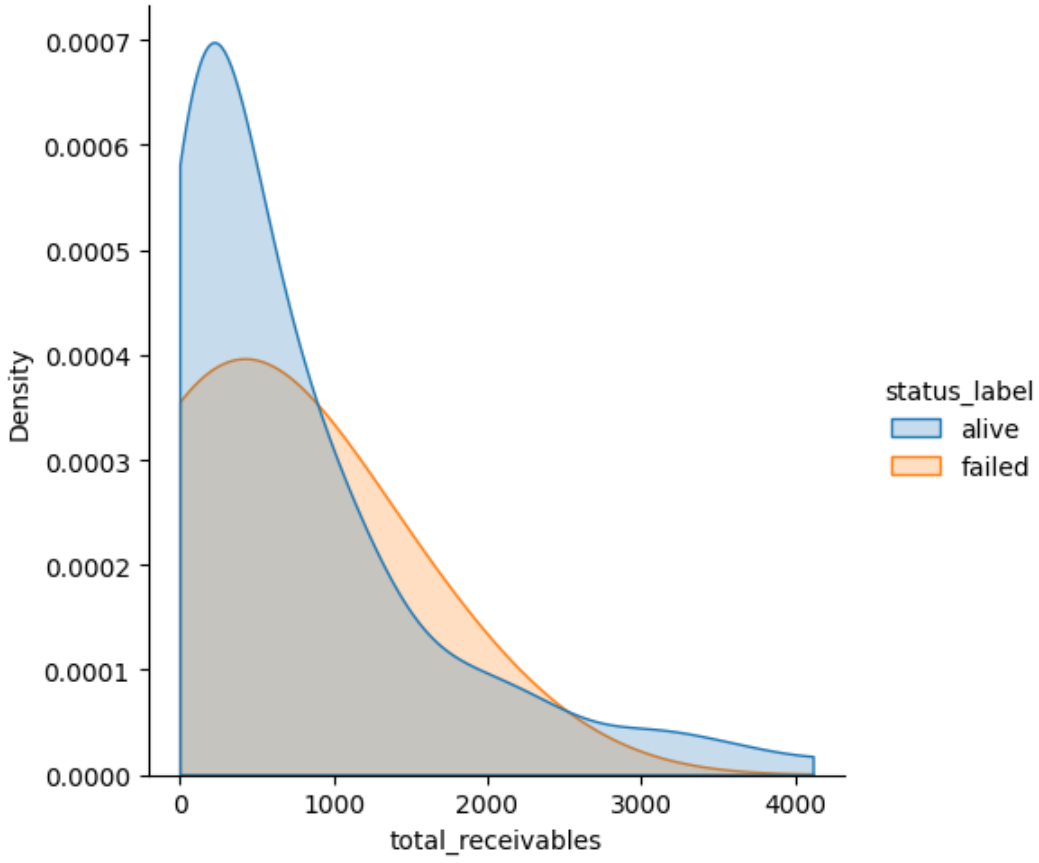

Figura 33: Distribuzione dei crediti totali nel dataset sintetico.

La distribuzione dei crediti totali (total_receivables) nel dataset sintetico evidenzia un andamento differenziato tra imprese attive e fallite. Le imprese attive mostrano una chiara concentrazione di valori più contenuti, con un picco nella fascia tra 0 e 500, mentre le imprese fallite si distribuiscono in maniera più ampia e piatta, con una maggiore densità relativa nei valori medi e medio-alti.

Questo comportamento suggerisce che nel dataset sintetico è stato associato un rischio maggiore di fallimento ad aziende con crediti elevati, un possibile segnale di inefficienza nella gestione del capitale circolante o di esposizione verso clienti poco affidabili. In tal senso, un eccessivo accumulo di crediti potrebbe riflettere una scarsa capacità di incasso o una dipendenza da clienti insolventi, condizioni che spesso anticipano situazioni di crisi di liquidità.

La distanza tra le due curve è più pronunciata rispetto a quella del dataset reale (Figura 15), dove le differenze erano meno evidenti. Questo implica che nel dataset sintetico si è voluta rafforzare la dimensione predittiva di questa variabile, trasformandola in un potenziale indicatore di vulnerabilità finanziaria.

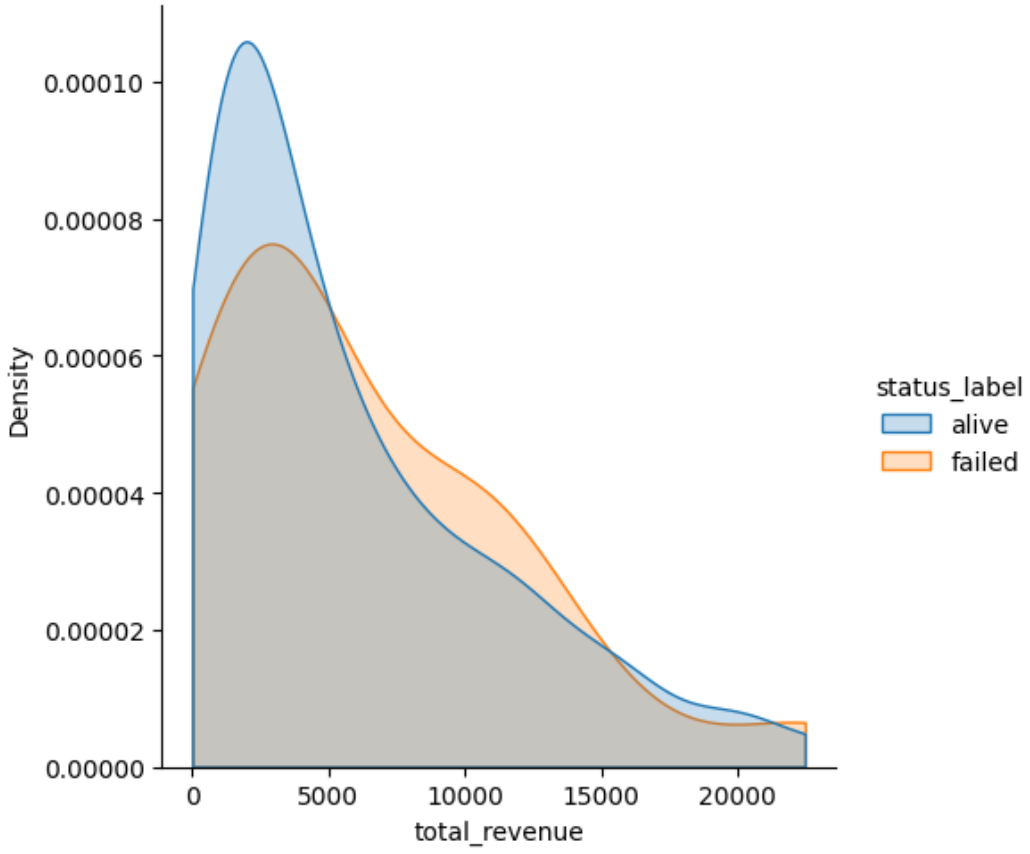

Figura 34: Distribuzione dei ricavi totali nel dataset sintetico.

Nel dataset sintetico, la distribuzione dei ricavi totali (total_revenue) evidenzia una netta distinzione tra le imprese attive e quelle fallite, con la curva delle imprese attive (in azzurro) caratterizzata da un picco molto marcato nei valori bassi (tra 1.000 e 3.000) e una rapida decrescita, mentre la curva delle imprese fallite (in arancione) appare più spostata verso destra, suggerendo una maggiore frequenza relativa tra imprese fallite con ricavi più alti.

Questa osservazione, apparentemente controintuitiva, sembra riflettere l'intento di simulare casi di "falsa sicurezza dimensionale": imprese che, pur registrando alti livelli di fatturato, presentano fragilità strutturali o inefficienze che le conducono comunque al fallimento. Il messaggio implicito è che i ricavi, presi isolatamente, non sono una misura sufficiente della solidità aziendale, e anzi, un elevato fatturato non accompagnato da marginalità o controllo dei costi può rappresentare un rischio.

Rispetto al dataset originale (Figura 16), dove le imprese fallite tendevano ad avere ricavi mediamente più bassi, il dataset sintetico rovescia parzialmente l'interpretazione, rendendo il total revenue una variabile ambigua, la cui utilità emerge solo se analizzata in relazione ad altri indicatori economico-finanziari (es. margine operativo, EBITDA, incidenza dei costi fissi).

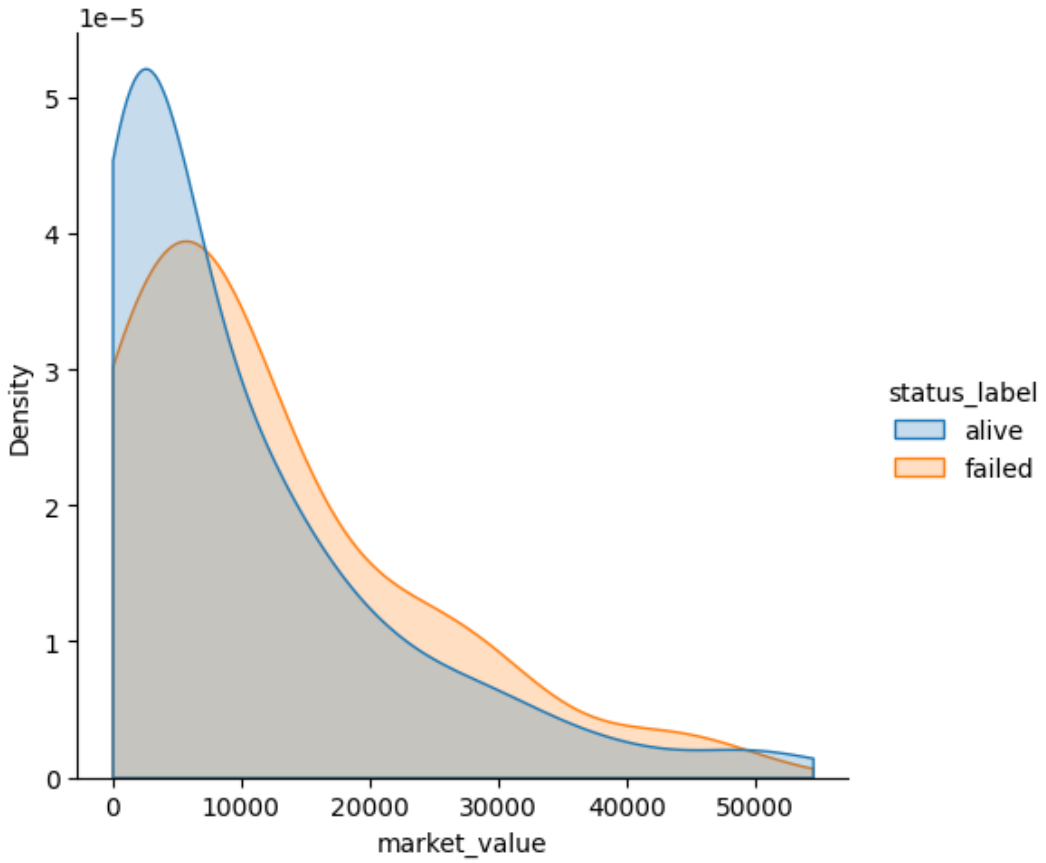

Figura 35: Distribuzione del valore di mercato nel dataset sintetico.

La distribuzione del valore di mercato (market_value) nel dataset sintetico evidenzia una significativa sovrapposizione tra le imprese attive e fallite, ma con una distinzione interessante nelle code della distribuzione. La curva relativa alle imprese fallite (in arancione) mostra una coda più lunga verso i valori elevati rispetto a quella delle imprese attive (in azzurro), suggerendo che nel dataset sintetico un valore di mercato elevato non esclude la possibilità di fallimento.

Questa impostazione rompe con la visione tradizionale secondo cui un'alta capitalizzazione rifletterebbe la fiducia del mercato e, quindi, una minore probabilità di default. Al contrario, qui si ipotizza che anche aziende apparentemente solide sul piano del valore di mercato possano incorrere in crisi, forse per squilibri interni non immediatamente percepiti dagli investitori (es. leva finanziaria, incapacità di generare cash flow, contingenze esterne).

La curva delle imprese attive, invece, tende a concentrarsi su valori più contenuti, con un picco visibile attorno ai 5.000–7.000, segno che il modello sintetico ha voluto riflettere una maggiore prudenza nei valori di mercato delle aziende sane.

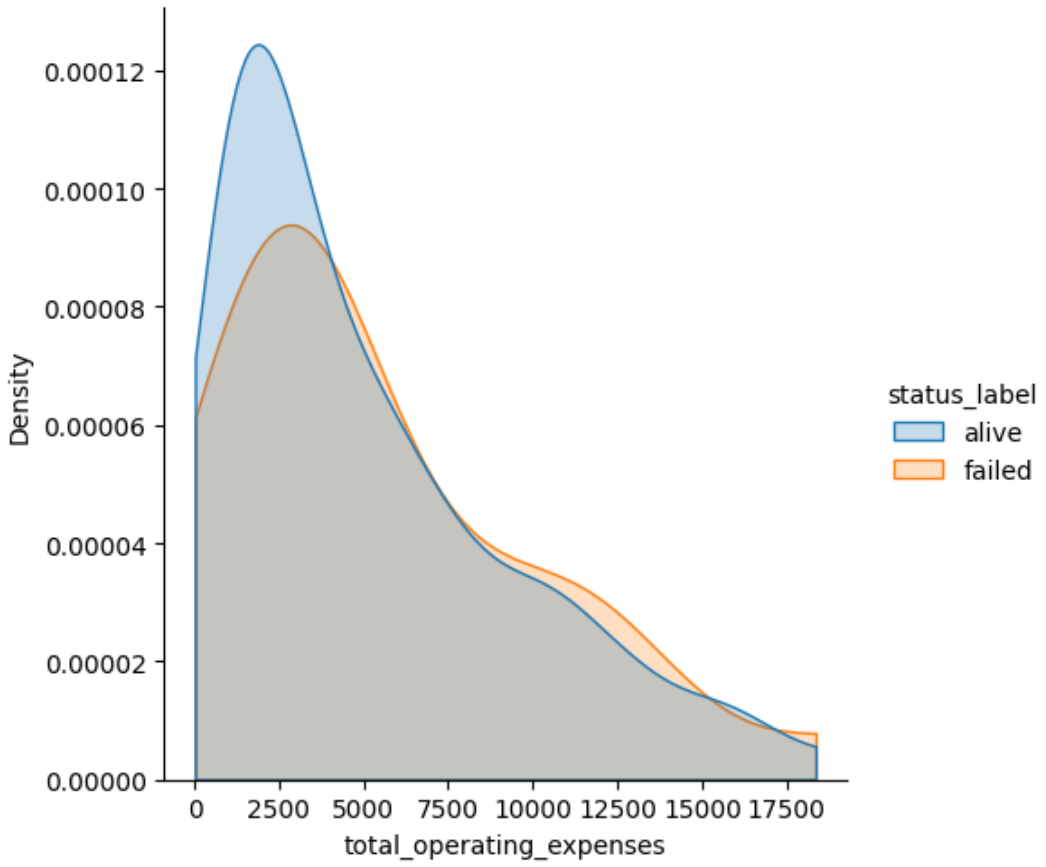

Figura 36: Distribuzione delle spese operative totali nel dataset sintetico.

La distribuzione delle spese operative totali (total_operating_expenses) nel dataset sintetico evidenzia un comportamento coerente con quanto osservato in molte altre variabili: le curve delle imprese attive (azzurre) e fallite (arancioni) si sovrappongono per la maggior parte del dominio, ma con divergenze evidenti nella coda destra, ovvero per valori più elevati.

In particolare, il ***picco della densità*** è leggermente più pronunciato e spostato verso sinistra per le imprese attive, suggerendo che esse tendono a sostenere spese operative mediamente inferiori, più contenute e forse più efficienti. Al contrario, le imprese

fallite mostrano una distribuzione più ampia, con maggiore densità anche per livelli elevati di spesa. Questo comportamento può essere interpretato come indicativo di inefficienze operative o diseconomie di scala, che contribuiscono all'insorgere di criticità finanziarie.

Un ulteriore elemento interessante è la presenza di una doppia modalità (bimodalità) nelle curve, in particolare per le imprese fallite. Questo potrebbe indicare l'esistenza di due cluster distinti: uno caratterizzato da spese più contenute, l'altro da spese particolarmente elevate, suggerendo che le cause di fallimento possono derivare da dinamiche molto diverse (eccesso di spesa o squilibri strutturali su scala maggiore).

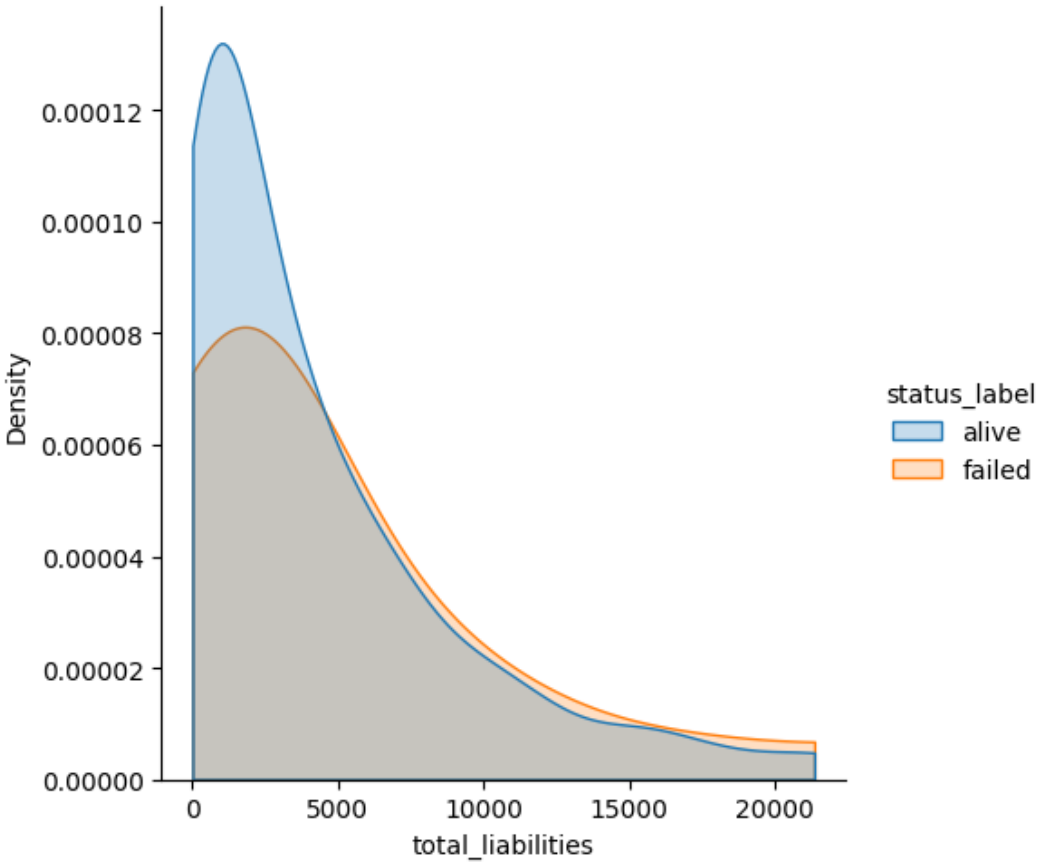

Figura 37: Distribuzione delle passività totali nel dataset sintetico.

La distribuzione della variabile total_liabilities nel dataset sintetico evidenzia un comportamento simile a quello osservato per altre voci di bilancio: le curve delle

imprese attive (in blu) e fallite (in arancione) presentano un'ampia sovrapposizione ma rivelano alcune differenze significative in termini di dispersione e picco di densità.

Nel dettaglio, la curva delle imprese attive presenta un picco più elevato e più stretto, situato in prossimità di valori relativamente bassi di passività totali. Ciò suggerisce che nel campione sintetico le imprese attive tendano a mantenere un profilo di indebitamento più contenuto e stabile. Al contrario, la distribuzione delle imprese fallite si estende in modo più marcato verso destra, indicando una maggiore variabilità e una presenza più consistente di osservazioni con elevati livelli di passività.

Questo comportamento è coerente con l'ipotesi secondo cui un eccessivo livello di indebitamento può rappresentare un fattore di rischio per la sopravvivenza aziendale, anche se la distribuzione delle due classi non è nettamente separata. La leggera prevalenza della curva arancione nei livelli di passività più elevati rafforza tale lettura.

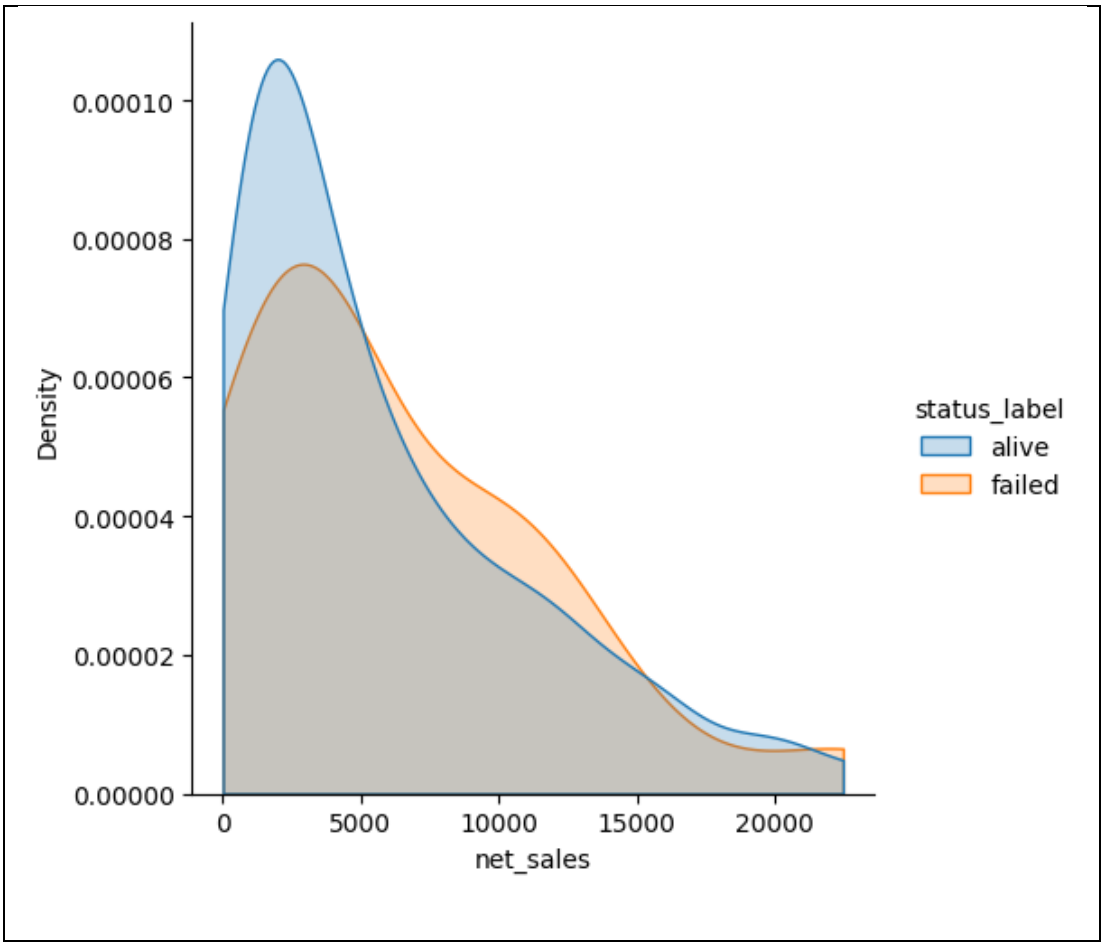

Figura 38: Distribuzione delle vendite nette nel dataset sintetico.

La distribuzione della variabile net_sales nel dataset sintetico evidenzia un andamento coerente con quanto osservato nelle altre grandezze economiche simulate. Anche in questo caso, le curve di densità delle imprese sopravvissute (alive) e fallite (failed) mostrano una forte sovrapposizione, ma con alcune differenze strutturali interessanti.

La curva relativa alle imprese attive (in blu) si distingue per un picco più accentuato e concentrato su valori di vendite nette contenuti, ma significativamente superiori a zero. Questa conformazione suggerisce che, nel dataset sintetico, le imprese che risultano in attività tendono ad avere una base di vendite più stabile e un livello minimo di fatturato, condizione che può rappresentare una soglia di sopravvivenza.

Al contrario, la curva delle imprese fallite (in arancione) mostra una maggiore dispersione e una coda lunga verso destra, indicando la presenza di casi con livelli elevati di vendite, ma che non hanno comunque garantito la continuità aziendale. Questo potrebbe suggerire che un volume di vendite elevato, da solo, non sia sufficiente a prevenire il fallimento, se non accompagnato da una gestione efficiente dei costi e delle passività.

Rispetto al dataset originale, si nota un maggiore equilibrio tra le due distribuzioni, segno che il modello sintetico ha teso a normalizzare il comportamento di questa variabile, riducendo lievemente le estremità estreme della distribuzione reale.

4 Conclusioni

L'uso di dati sintetici generati da modelli di intelligenza artificiale rappresenta una frontiera promettente per la valutazione del rischio finanziario. Questo approccio non solo migliora la capacità di prevedere e gestire scenari complessi, ma apre anche nuove opportunità per l'innovazione metodologica e pratica. Tuttavia, è essenziale affrontare una serie di sfide tecnologiche, etiche e normative per garantire che questi strumenti siano utilizzati in modo responsabile ed efficace. Va sottolineato inoltre che in un mondo sempre più complesso e interconnesso, l'adozione di approcci innovativi come quello descritto potrebbe rappresentare un punto di svolta, contribuendo a creare sistemi finanziari più resilienti e sostenibili.

L'implementazione della strategia sopra citata attraverso una semplice prova di concetto ha dimostrato la fattibilità pratica di un approccio innovativo e potenzialmente rivoluzionario che, rispetto a schemi e soluzioni classiche offre numerosi vantaggi, tra cui:

- Ampliamento del set di dati: La disponibilità di dati sintetici realistici, generabili in grandi quantità, permette di superare i limiti dei dati storici, generando scenari eventualmente critici e potenzialmente difficili da studiare che potrebbero non essere presenti nei dataset reali.
- Flessibilità: La generazione di dati consente di simulare una vasta gamma di condizioni di mercato, inclusi scenari estremi e stress test, che sono fondamentali per una valutazione robusta del rischio.
- Riduzione del rischio di overfitting: Utilizzando dati sintetici, è possibile testare modelli di rischio in condizioni diverse da quelle storiche, riducendo il rischio di overfitting e migliorando la generalizzazione.
- Innovazione metodologica: L'uso di modelli generativi avanzati (es. TGAN, TVAE) apre nuove frontiere nella modellazione finanziaria, consentendo di catturare relazioni complesse e non lineari tra variabili.

Tuttavia, è importante riconoscere alcune limitazioni e sfide:

- Qualità dei dati sintetici: La validità dei risultati dipende dalla capacità del modello generativo di produrre dati realistici e rappresentativi.
- Complessità computazionale: L'addestramento di modelli generativi e l'applicazione di tecniche di filtraggio possono richiedere risorse computazionali significative.
- Rischi etici e normativi: L'uso di dati sintetici deve essere gestito con attenzione per evitare bias o violazioni normative.

Nonostante questi problemi, l'approccio proposto dimostra un potenziale significativo per migliorare, specialmente in contesti molto dinamici e incerti.

I lavori futuri dovrebbero concentrarsi sul miglioramento dei modelli generativi, sull'affinamento delle tecniche di filtraggio e sulla validazione rigorosa dei risultati. Inoltre, l'integrazione di dati sintetici in nuove applicazioni pratiche e l'estensione a nuovi settori saranno fondamentali per massimizzare il successo di questa tecnologia. Di seguito sono elencate alcune direzioni future di ricerca e sviluppo futuri:

1. Miglioramento dei Modelli Generativi
- Uso di Modelli ibridi: Combinare diverse architetture generative (es. GAN con VAE) per migliorare la qualità e la diversità dei dati sintetici.
- Sviluppo di Modelli specifici per il dominio finanziario: Sviluppare modelli generativi progettati specificamente per catturare le peculiarità dei dati finanziari, come la non-stazionarietà e le code pesanti.
- Incorporazione di conoscenza esterna: Integrare informazioni macroeconomiche o di mercato nei modelli generativi per migliorare la rilevanza dei dati sintetici.

2. Affinamento delle Tecniche di Filtraggio
- Filtraggio adattivo: Sviluppare algoritmi di filtraggio che si adattano dinamicamente alle condizioni di mercato.
- Filtraggio basato su apprendimento profondo: Utilizzare reti neurali avanzate (Deep Neural Networks) per identificare pattern più complessi e particolari situazioni di interesse.

- Integrazione di più criteri: Combinare filtraggio statistico con quello basato su machine learning e regole di dominio per una selezione più robusta dei dati.

3. Nuove Applicazioni Pratiche

- Gestione del portafoglio: Utilizzare dati sintetici per ottimizzare la gestione del portafoglio in condizioni di mercato estreme.
- Pianificazione strategica: Integrare i dati sintetici nei processi di pianificazione aziendale per valutare l'impatto di scenari futuri.
- Regolamentazione e compliance: Sviluppare strumenti basati su dati sintetici per supportare le istituzioni finanziarie nel rispetto delle normative.

4. Aspetti Etici e Normativi

- Linee guida per l'uso etico: Definire best practice per l'uso responsabile dei dati sintetici nella valutazione del rischio.
- Conformità normativa: Studiare come i dati sintetici possano essere integrati nei framework normativi esistenti (es. Basel III, Solvency II).

5. Estensione ad Altri Settori

- Assicurativo: Applicare tecniche simili per la valutazione del rischio assicurativo, come la stima delle perdite catastrofali.
- Energia e Commodities: Utilizzare dati sintetici per modellare il rischio nei mercati energetici e delle materie prime.
- Fintech: Integrare dati sintetici nelle piattaforme fintech per migliorare la gestione del rischio e l'offerta di servizi personalizzati.

5 Bibliografia

1. Berger, Vance W., and YanYan Zhou. (2014). Kolmogorov–smirnov test: Overview. Wiley statsref: Statistics reference online.
2. Corder, G. W. & Foreman, D. I. (2014). Nonparametric Statistics: A Step-by-Step Approach, Wiley. ISBN 978-1118840313.
3. Durante, F., & Sempi, C. (2016). Principles of copula theory (Vol. 474). Boca Raton, FL: CRC press.
4. Goodfellow, Ian J., et al. (2014). Generative adversarial nets. Advances in neural information processing systems, 27.
5. Kasztelnik, K., & Campbell, S. (2024). Revolutionizing Financial Health Predictions: The Integration of GenAI and Advanced Machine Learning Techniques. Journal of Applied Business & Economics, 26(6).
6. Kingma, D. P. and Welling, M. (2013) Auto-encoding variational bayes. In International Conference on Learning Representations.
7. Lombardo, G., Pellegrino, M., Adosoglou, G., Cagnoni, S., Pardalos, P. M., & Poggi, A. (2022). Machine learning for bankruptcy prediction in the American stock market: dataset and benchmarks. Future Internet, 14(8), 244.
8. Patki, N., Wedge, R., & Veeramachaneni, K. (2016). The synthetic data vault. In 2016 IEEE international conference on data science and advanced analytics (DSAA) (pp. 399-410). IEEE.
9. Pellegrino, M., Lombardo, G., Adosoglou, G., Cagnoni, S., Pardalos, P. M., & Poggi, A. (2024). A Multi-Head LSTM Architecture for Bankruptcy Prediction with Time Series Accounting Data. Future Internet, 16(3), 79.
10. Wang, X., Kräussl, Z., & Brorsson, M. (2024). Datasets for Advanced Bankruptcy Prediction: A survey and Taxonomy. arXiv preprint arXiv:2411.01928.
11. Xu, L., Skoularidou, M., Cuesta-Infante, A., & Veeramachaneni, K. (2019). Modeling tabular data using conditional GAN. Advances in neural information processing systems, 32.