Architetture Dati
Relazione Progetto

A.A. 2019-2020, Appello 23/09/2020

Componenti gruppo:

- Christian Bernasconi 816423
- Marco Ripamonti 806785

Indice

1 Introduzione **1**

2 Record Linkage **2**

3 Datasets e Tools **4**
3.1 Datasets 4
3.2 Tools 5

4 Workflow **6**
4.1 Preprocessing 6
4.2 Blocking 7
4.3 Compare 8
4.4 Classification 9
4.5 Evaluation 9

5 Conclusioni **13**

1 Introduzione

Questo elaborato tratta il tema del *Record Linkage*. In particolare, esso è stato affrontato sotto diversi punti di vista e coinvolgendo alcune tecniche di *Machine Learning*. Il contesto in cui si svolge questo lavoro è quello della *FEIII 2016 Challenge*[1] e riguarda il linking di istitutuzioni finanziarie provenienti da tre sorgenti di dati diverse.

Al capitolo 2 vengono presentate alcune nozioni base riguardo alle diverse tecniche di *Record Linkage*. Nel capitolo 3 sono mostrate le caratterstiche dei datasets e gli strumenti utilizzati per svolgere il progetto. Proseguendo al capitolo 4, viene illustrato il workflow con le varie fasi affrontate nel processo di linking, accompagnato dall'analisi dei risultati ottenuti. Infine, al capitolo 5 si possono trovare le considerazioni finali.

[1] `https://ir.nist.gov/feiii/2016-challenge.html`

2 Record Linkage

Il *Record Linkage* consiste nel processo di confronto di dati provenienti da diverse sorgenti, presentando quindi diversi aspetti di eterogeneità, con lo scopo di determinare quali record corrispondono alla stessa entità del mondo reale.
Nel contesto del *Record Linkage* si parla invece di *Deduplicazione* quando si dispone di un unico dataset proveniente da una singola sorgente e si vogliono identificare i record duplicati.
Il processo di *Record Linkage* è strettamente necessario quando si incorre nella necessità di unire i dati provenienti da diverse fonti attraverso varie strategie e tecniche di *Data Fusion*.
Ciò che rende complicato questo processo è la presenza, nel mondo reale, di dati di bassa qualità con errori e diverse forme con cui uno stesso dato è rappresentato.
Esistono diversi approcci per affrontare il problema del *Record Linkage* [1]:

- **Record Linkage Deterministico**: è il più semplice degli approcci e consiste nel generare i link basandosi sul numero di attributi che coincidono l'uno con l'altro, valutandone la distanza con una qualche misura di similarità. Ad esempio si possono considerare coppie di records come *Matching records* quelle dove quattro o più attributi coincidono [1].

- **Record Linkage Probabilistico**: considera più attributi e per ognuno di questi computa un peso basandosi sulla sua capacità di identificare un match o un non match. Usando poi questi pesi viene calcolata la probabilità che dati due record essi riconducano alla stessa entità. I record con probabilità superiore ad una certa soglia saranno considerati come *"Matches"*, coppie la cui probabilità cade sotto un'altra soglia saranno considerate come *"Not Matching"*, infine, le coppie con probabilità compresa tra le due soglie saranno classificate come *"Possible Matches"*. Questo approccio, a differenza di quello deterministico che potrebbe richiedere un intervento umano molto più esigente riguardante la programmazione di regole anche complesse, risulta essere molto più automatico [1].

- **Record Linkage con Machine Lerning**: in anni più recenti si sono sviluppate tecniche di Machine Learning utili anche nel contesto del linking. Il record Linkage Probabilistico assume gli attributi come tutti indipendenti cosa che in realtà non è sempre vera. Con diverse tecniche e modelli di Machine Learning si possono superare alcuni delle limitazioni imposte dai due approcci precedenti [1].
  Esistono tre diversi approcci per quanto riguarda il Record Linkage con Machine Learning:

  - Apprendimento supervisionato: se si dispone di un dataset etichettato, ossia, avendo la conoscenza di quali record sono i *"Matching"* è possibi utilizzare un modello per classificazione supervisionata. Questa informazione è assolutamente necessaria per addestrare il modello e conseguentemente poterlo utilizzare per la classificazione di nuovi dati. Questo approccio, pur essendo efficiente ed efficace, richiede una conoscenza che non sempre si dispone in un contesto reale che è quella di un training set.
  - Apprendimento non supervisionato: a differenza del precedente approccio, questo non richiede alcuna conoscenza di un training set. I modelli di apprendimento non supervisionato cercano di individuare i pattern nascosti nei dati in input e

identificare i diversi gruppi di punti che soddisfano una determinata caratteristica. Questi fanno uso di algoritmi di *clustering* e nel caso specifico del *Record Linkage* l'obiettivo è individuare i gruppi corrispondenti ai *Matching Records* e ai *Not Matching Records*.

- Apprendimento ibrido: questo approccio cerca di combinare i vantaggi sia delle tecniche di apprendimento supervisionato che di quelle di apprendimento non supervisionato.
  In particolare l'apprendimento supervisionato risulta essere più preciso e accurato dell'apprendimento non supervisionato, ma non sempre si dispone del training set assolutamente necessario per addestrare il modello. L'apprendimento non supervisionato cerca di risolvere questa limitazione applicando la tecnica a un gruppo limitato di esempi cercando di prevedere la classe degli esempi non etichettati generando quindi un training set.

In questo progetto è stato utilizzato sia un semplice approcio deterministico per il Record Linkage che un approccio con il Machine Learning con tecniche di apprendimento supervisionato e non supervisionato.

3 Datasets e Tools

In questa sezione vengono presentati i datasets e i tools utilizzati per lo sviluppo di un processo di Rercord Linkage.

3.1 Datasets

I datasets utilizzati sono stati ottenuti dalla *FEI Challenge 2016* e sono reperibili al seguente link `https://ir.nist.gov/feiii/2016-challenge.html`. I dataset forniti sono in formato *csv* e sono:

- **FFIEC.csv:** il file proviene dal *Federal Financial Institution Examination Council*[2] e fornisce informazioni riguardo banche e altre istituzioni finanziarie che sono regolate da agenzie affiliate con il consiglio.
  Il dataset è composto da 6,652 records ognuno dei quali descritto da 14 attributi.
- **LEI.csv:** il file contiene *Legal Entities Identifiers*[3] per una vasta gamma di istituzioni.
  Il dataset è composto da 53,958 records ognuno dei quali descritto da 39 attributi.
- **SEC.csv:** il file proviene dal *Security and Exchange Commision*[4] e contiene informazioni riguardo entità registrate con SEC.
  Il dataset è composto da 129,312 records ognuno dei quali descritto da 24 attributi.

Di seguito vengono riportate le coppie di attributi confrontati per FFIEC e LEI:

- Financial Institution Name Cleaned - LegalNameCleaned
- Financial Institution Address - LegalAddress_Line_Cleaned
- Financial Institution City - LegalAddress_City
- Financial Institution State - LegalAddress_Region_2
- Financial Institution Zip Code 5 - LegalAddress_PostalCode_5

Nell'elenco seguente, invece, vengono riportate le coppie di attributi confrontati per FFIEC e SEC:

- Financial Institution Name Cleaned - CONFORMED_NAME
- Financial Institution Address - B_STREET
- Financial Institution City - B_CITY
- Financial Institution State - B_STPR
- Financial Institution Zip Code 5 - B_POSTAL

[2] `https://en.wikipedia.org/wiki/Federal_Financial_Institutions_Examination_Council`
[3] `https://en.wikipedia.org/wiki/Legal_Entity_Identifier`
[4] `https://en.wikipedia.org/wiki/U.S._Securities_and_Exchange_Commission`

Inoltre sono state fornite le *Ground Truth* parziali contenenti le coppie di *Matching Records*, *Not Matching Records* e *Ambiguous Records*.
In particolare le ground truth sono state costruite con un algoritmo il cui output è uno score per ciascuna coppia di record. È stato scelto un threshold per stabilire i *possibili match* e i *possibili non match*. Tutti i possibili match e un campione di non match sono stati riesaminati da un team di esperti per aggiudicare la riposta corretta [2]. Le ground truth rese disponibili in file *csv* sono:

- **FFIEC-LEI-GroundTruth.csv:** ground truth per valutare i risultati ottenuti dal *Record Linkage* tra le sorgenti *FFIEC* e *LEI*. Contiene 496 TP (*Matching Records*), 38 Ambiguous (*Record Ambigui*) e i restanti TN più gli esclusi dalla ground truth reputati come record *Not Matching*.
- **FFIEC-SEC-GroundTruth.csv:** ground truth per valutare i risultati ottenuti dal *Record Linkage* tra le sorgenti *FFIEC* e *SEC*. Contiene 230 TP (*Matching Records*), 95 Ambiguous (*Record Ambigui*) e i restanti TN èiù gli esclusi dalla ground truth reputati come record *Not Matching*.

3.2 Tools

Di seguito vengono descritti i tool e le librerie utilizzate per lo sviluppo di un processo di *Record Linkage*.

Python È un linguaggio di programmazione interpretato, di alto livello e utilizzato in diversi contesti scientifici. In particolare è stato utilizzato per la sua facilità nel maneggiare dataset ed eseguire esperimenti con diversi modelli di Machine Learning.

Python Record Linkage Toolkit [5] È una libreria di Python che permette diverse operazioni legate strettamente al processo di Record Linkage. La libreria fornisce metodi *cleaning e preprocessing*, metodi di *indexing e blocking*, metodi per la *comparazione di record* con diverse misure e metriche e infine offre metodi per la *classificazione e valutazione* dei *Matching Records* e *Not Matching Record*.

Sklearn [6] È una libreria di Python che offre funzioni per l'analisi dei dati e modelli di Machine Learning. In particolare abbiamo utilizzato modelli per apprendimento non supervisionato come *K-Means* ed *ECM* e un modello di apprendimento supervisionato *Random Forest*. Inoltre per la valutazione di quest'ultimo è stata usata una *Stratified k-fold Cross Validation*.

[5] `https://recordlinkage.readthedocs.io/en/latest/ref-compare.html`
[6] `https://scikit-learn.org/stable/`

4 Workflow

Il processo di *Record Linkage* adottato per questo progetto è suddiviso in cinque fasi principali:

1. **Preprocessing**: in questa fase si è cercato di pulire i dati disponibili ad esempio normalizzando codici postali o nomi presenti in diverse forme.
2. **Blocking**: sono state sperimentate diverse strategie di blocking per limitare il numero di confronti nella fase successiva.
3. **Compare**: le coppie di record ottenute dalla fase di blocking sono state confrontate applicando diverse misure di similarità.
4. **Classification**: diversi approcci sono stati adottati per classificare, sulla base dei risultati ottenuti dai confronti, quali record appartenessero alla categoria *Matching* piuttosto che *Not Matching*.
5. **Evaluation**: infine sono state calcolate alcune misure di performance per valutare i risultati ottenuti.

4.1 Preprocessing

La fase di preprocessing è stata molto importante per effettuare una pulizia generale dei dati e per uniformare i valori provenienti dalle diverse sorgenti.
Prima di procedere alle operazioni sui valori dei campi scelti, sono state innanzitutto rimosse le entità non valide in quanto aventi lo stesso identificativo univoco[7].
Dopo una prima analisi dei datasets si sono potuti individuare gli attributi d'interesse comune rispetto alle varie tabelle. Essi sono quelli relativi al nome, indirizzo, città, stato e zip code degli istituti finanziari. Su questi attributi sono state poi osservate le caratteristiche e le differenze tra alcuni esempi di match presenti nelle ground truth parziali.
La prima operazione effettuata sui valori di tutti i campi è stata quella di rendere tutte le stringhe lower case per poter rendere più agevoli i passaggi successivi di pulizia e più precise le fasi di blocking e compare.
Partendo da qui, sono state definite delle funzioni per la pulizia e la normalizzazione di questi campi che tengono conto dei seguenti aspetti:

- **nome:** per i nomi degli istituti finanziari sono state innanzitutto rimosse stopwords e punteggiatura e rimpiazzati i caratteri '&' con 'and'. I nomi, inoltre, contengono anche della conoscenza semantica implicita che indica il tipo di istituto finanaziario, il quale viene riportato con sigle o per esteso a seconda della fonte del dato. Per questo è stata prevista la possibilità di normalizzare rimuovendo i modifiers dai nomi e spostarli in un attributo a parte per eventuali controlli successivi. I modifiers sono stati individuati sia consultando fonti esterne[8] che andando ad individuare unigrammi, bigrammi e trigrammi di parole con frequenza maggiore presenti nei nomi. Essi sono consultabili in tabella 1.

[7] sono state rimosse 9 istanze dalla tabella LEI con ID pari a 9.598E+19 per un errore che potrebbe essere dovuto all'esportazione del dataset

[8] `https://en.wikipedia.org/wiki/List_of_legal_entity_types_by_country`

- **indirizzo:** per quanto riguarda gli indirizzi sono state notate delle grandi difformità tra i formati delle diverse tabelle. Non disponendo di una chiave per le API di maps che potesse soddisfare il numero totale di normalizzazioni necessarie a coprire tutti i records, sono state effettuate delle operazioni che hanno permesso di ridurre in parte queste differenze. Innanzitutto sono stati uniformati i modi per indicare l'orientamento di una strada. Dopodichè sono stati uniformati i modi per indicare una tipologia di strada. I valori con cui sono stati effettuati questi due passaggi sono disponibili rispettivamente alla tabelle 2 e 3. Inoltre sono state rimosse le informazioni relative ai P.O. BOX, in quanto non presenti in tutte le tabelle.

- **zip code:** per gli zip code si è osservata la presenza di molti valori non validi o registrati come *nan*. Inoltre, per la tabella SEC lo zip code è riportato con il formato 5+4 digits ed è stato necessario estrarre solamente la prima parte. Per tutte e tre le tabelle sono stati rilevati anche dei codici con numero di cifre inferiore a 5. La causa di ciò era dovuta all'omissione degli zeri iniziali che è stata poi corretta.

La funzione che si occupa del preprocessing è stata resa configurabile per poter decidere quali delle singole operazioni effettuare al fine di poter verificare i miglioramenti ottenuti tramite gli steps di pulizia e normalizzazione.

Full Name	**Abbr.**
corporation	corp
company	co
federal savings bank	fsb
national association	na
incorporated	inc
trust association	ta

Tabella 1: Tabella normalizzazione modifiers

Full Name	**Abbr.**
north	n
south	s
east	e
west	w
northeast	ne
northwest	nw
southeast	se
southwest	sw

Tabella 2: Tabella normalizzazione orientamenti

Full Name	**Abbr.**
street	st
road	rd
avenue	ave
square	sq
lane	la
suite	su
plaza	pl

Tabella 3: Tabella normalizzazione tipi strada

4.2 Blocking

Per evitare di effettuare un confronto su tutte le possibili coppie di valori, ossia effettuare il prodotto cartesiano tra le due tabelle tramite *Full Indexing*, esistono diverse strategie di *Blocking* che permettono di limitare significativamente il numero di coppie da considerare per la verifica.
Tra i vari approcci possiamo trovare quelli più semplici come *Exact Blocking*, passando poi per *Sorted Neighbourhood*, *Bigram Indexing* e *Canopy Clustering*, arrivando infine a strategie più complesse come quelle basate su metodi *LSH* (locality-sensitive hashing), non applicabili a dataset di grandi dimensioni [3, 4].

Tra le strategie offerte dalla libreria utilizzata per l'analisi troviamo quelle di *Exact Blocking* e di *Sorted Neighbourhood Blocking*. La prima, dopo aver scelto quali campi tra le due tabelle utilizzare per accoppiare le istanze, crea dei link candidati in base all'uguaglianza

sintattica dei valori considerati. La seconda, invece, effettua un ordinamento dei valori sulla base dei campi chiave scelti ed utilizzando una finestra di dimensione n configurabile tiene in considerazione anche i records del vicinato. Ciò permette di considerare un numero di link candidati superiore rispetto alla strategia precedente. Il fatto di tenere in considerazione il vicinato permette una migliore efficacia laddove vi sia la presenza di molti errori di spelling.

Nel codice sono state implementate delle funzioni che permettono di configurare i vari tipi di blocking con diversi parametri, con lo scopo di effettuare varie prove e verificare quale strategia sia più efficacie in questo contesto.

4.3 Compare

Nella fase di *Compare* è stato possibile confrontare le coppie di record ottenute dalla fase di *Blocking* mediante diversi metodi e metriche. Il risultato di questa fase sarà una tabella dove per ogni coppia di record individuata verrà riportato un vettore contentente, per ogni coppia di attributi, il corrispettivo valore di similarità secondo la metrica stabilita. Il valore di similarità assume valore pari a 1 nel caso in cui due attributi sono uguali o superino una soglia di similarità, 0 altrimenti. La libreria *Python Record Linkage Toolkit* mette a disposizione diverse modalità di confronto tra cui:

- **Exact Matching:** è possibile confrontare il valore di due attributi in modo esatto, in particolare la similarità sarà pari a 1 nel caso di match completamente esatto, 0 altrimenti.
- **String Matching:** è possibile confrontare due valori di tipo stringa usando diverse metriche. Vengono fornite svariate metodologie basate sulla distanza di edit tra due stringhe, ad esempio *Levenshtein Distance*, metrica che conteggia il numero di modifiche per trasformare una stringa nell'altra (conteggiando inserimenti, cancellazioni, o sostituzioni), e una sua variante *Damerau-Levenshtein* che concede la trasposizione di simboli adiacenti. È inoltre possibile utilizzare la metrica di *Jaro* con una sua variante *Jaro-Wrinkler*, o ancora metriche basate su *n-grammi* e sul *coseno*. Con questa modalità di confronto è necessario specificare una soglia sopra la quale due stringhe vengono considerate come approssimativamente equivalenti. Se il risultato ottenuto da una di queste metriche supera la soglia stabilita il valore di similarità sarà pari a 1, 0 altrimenti.

Sono inoltre disponibili modalità che non abbiamo considerato perchè non utili per il confronto dei dati disponibili come:

- **Numeric Matching:** per confrontare valori numerici, anche in questo caso sono disponibili diversi metodi di confronto.
- **Geographic Matching:** per computare la similarità parziale tra valori WGS84 di coordinate geografiche.
- **Date Matching:** per computare la similarità tra due date.

È stata definita una funzione che permette di stabilire come i record devono essere confrontati tra di loro, specificando quali metriche e soglie utilizzare per verificare come diversi confronti possono influenzare lo step di classificazione.

4.4 Classification

L'obiettivo della fase di *Classification* consiste nel determinare quali coppie di record identificano una stessa entità del mondo reale (*Matching Records*) e quali, invece, rappresentano entità distinte (*Not Matching Records*). Sono stati utilizzati diversi approcci per la classificazione:

- **Deterministico:** l'approccio deterministo adottato consiste nel ritenere *Matching Records* tutti quei record il cui numero di attributi equivalenti, secondo una metrica adottata nella fase di *Compare*, è maggiore di una determinata soglia.
- **Unsupervised Machine Learning:** un secondo approccio riguarda l'utilizzo di modelli di *Machine Learning* per l'apprendimento non supervisionato. Questi modelli non richiedono alcuna presenza di un training set e in particolare sono stati scelti algoritmi di clustering come *K-Means* e *ECM* (*Expectation/Conditional Maximization Algorithm*). Nel caso specifico del record linkage l'obiettivo consiste nel determinare i gruppi dei *Matching Records* e *Not Matching Records* analizzando la matrice ottenuta dalla fase di *Compare*.
- **Supervised Machine Learning:** l'ultimo degli approcci considerati consiste nell'utilizzare sempre tecniche di *Machine Learning*, ma che coinvolgono questa volta modelli per l'apprendimento supervisionato. L'apprendimento supervisionato necessita di un training set con cui allenare un modello utilizzato poi per classificare nuovi dati. Grazie alla disponibilità di una *Ground Truth* è stato selezionato come training set un campione bilanciato tra le classi di *Matching Records*, *Not Matching Records* e *Ambiguous Records*. Un modello *Random Forest* è stato scelto per svolgere questo task di classificazione.

4.5 Evaluation

In questa sezione vengono valutati i risultati ottenuti effettuando Record Linkage attraverso diversi approcci e configurazioni. In particolare sono state osservate le differenze di performance al variare del grado di preprocessing e di alcune strategie di blocking. É bene sottolineare che diverse metriche di comparazione sulle stringhe sono state testate senza ottenere differenze sostanziali.

FFIEC-LEI Per quanto riguarda il Record Linkage tra le tabelle FFIEC e LEI sono stati ottenuti dei risultati molto buoni rispetto alla Ground Truth fornita. In tabella 4 vengono mostrate le misure di *Precision*, *Recall* e *F-Score* sia per l'approccio deterministico (DET) che per l'approccio machine-learning (ML). Le colonne *Modifier*, *Stopwords* e *Address* indicano se sia stato effettuato o meno il preprocessing relativo. Infine, vi è indicato anche il numero di *Candidate Links* ottenuto tramite Exact Blocking.

Dalla tabella si può osservare come l'effettuare tutti gli steps del preprocessing permetta di individuare un numero più alto di links candidati al matching e di avere le performance migliori. In questo caso, rispetto alla Ground Truth sono stati individuati 474 su 496 matches. Tra i 22 matches non individuati, 15 di essi sono stati esclusi per via del blocking. Inoltre, sono stati individuati 12 possibili matches errati, i quali sono stati classificati come ambigui nella Ground Truth.

Modifier	Stopwords	Address	#Candidate Links	Precision DET	Recall DET	F-Score DET	Precision ML	Recall ML	F-Score ML
Yes	Yes	Yes	1178	0,975	0,955	0,965	0,969	0,963	0,966
Yes	Yes	No	1178	0,982	0,913	0,946	0,969	0,963	0,966
No	Yes	Yes	977	0,976	0,941	0,958	0,971	0,949	0,960
No	No	Yes	898	0,981	0,877	0,926	0,977	0,883	0,927
No	No	No	898	0,988	0,840	0,908	0,975	0,883	0,926

Tabella 4: Tabella risultati linking FFIEC-LEI con approccio deterministico e machine-learning utilizzando Exact Blocking al variare del preprocessing

Modifier	Stopwords	Address	#Candidate Links	Precision DET	Recall DET	F-Score DET	Precision ML	Recall ML	F-Score ML
Yes	Yes	Yes	624	0,774	0,700	0,735	0,620	0,752	0,679
Yes	Yes	No	624	0,772	0,695	0,732	0,620	0,752	0,679
No	Yes	Yes	257	0,925	0,430	0,587	0,886	0,443	0,591
No	No	Yes	251	0,933	0,430	0,589	0,894	0,443	0,593
No	No	No	251	0,925	0,430	0,587	0,730	0,447	0,555

Tabella 5: Tabella risultati linking FFIEC-SEC con approccio deterministico e machine-learning utilizzando Exact Blocking al variare del preprocessing

FFIEC-SEC Le medesime analisi sono state condotte sulla coppia di tabelle FFIEC e SEC. Come in precedenza, in tabella 5 sono disponibili i risultati ottenuti.
Anche in questo caso la configurazione che porta a migliori risultati è quella del preprocessing completo. Rispetto alla tabella precedente, i valori delle performance sono generalmente più bassi. Ciò è dovuto al fatto che la tabella SEC presenta molte più difformità nella rappresentazione dei dati e rende il processo più complicato. Riguardo alla configurazione migliore si riscontra la presenza di 161 matches rispetto ai 230 della Ground Truth. Dei 69 mancati matches, 51 sono stati esclusi durante il blocking. Passando ai matches classificati erroneamente come positivi, essi sono 47 di cui 22 considerati ambigui, 24 negativi e 1 non presente nella Ground Truth.

Su questa coppia di dataset è stato anche verificato l'utilizzo di un modello di machine learning supervisionato. Come training set è stata utilizzata la Ground Truth dopo aver effettuato un bilanciamento tra le classi Matches, Not Matches e Ambiguous, di cui è stata presa una porzione pari al 75%. Il modello utilizzato è quello del Random Forest, il quale ha prodotto la matrice di confusione riportata in tabella 6. Come si può notare, Matches e Not Matches riescono ad essere classificati con più facilità rispetto alle coppie di records Ambiguous.
Dato il numero limitato di esempi etichettati disponibili come training set, è stata effettuata una Stratified 10-Fold Cross-Validation per poter fornire delle misure più accurate. I risultati ottenuti mostrano Precision, Recall e F-Score medie rispettivamente pari a 0.591, 0.591 e 0.580 .

Considerazioni Ground Truths É bene sottolineare che le Ground Truths rese disponibili dalla *FEII 2016 Challenge* sono in realtà delle tabelle parziali. Le tabelle complete, integrate sulla base dei Matches non presenti nella Ground Truth iniziale individuati dai diversi teams, sarebbero state rese disponibili esclusivamente ai partecipanti della sfida durante una seconda fase.

	Not Match (Predicted)	Match (Predicted)	Ambiguous (Predicted)
Not Match	17	5	4
Match	3	20	5
Ambiguous	9	12	8

Tabella 6: Matrice confusione 10-fold stratified cross-validation con Random Forest

Name	Address	City	State	Zip Code
OLD NATIONAL BANK	1 MAIN STREET	EVANSVILLE	IN	47708
Old National Bancorp	1 Main Street	Evansville	IN	47708
NORTHERN TRUST COMPANY	50 SOUTH LA SALLE STREET	CHICAGO	IL	60603
Northern Trust Collective Canada Index Fund - Non-Lending	50 South Lasalle Street	Chicago	IL	60603
INTRUST BANK, NATIONAL ASSOCIATION	105 NORTH MAIN STREET	WICHITA	KS	67202
Intrust Financial Corporation	105 North Main	Wichita	KS	67202
BLACKROCK INSTITUTIONAL TRUST COMPANY, NATIONAL ASSOCIATION	400 HOWARD STREET	SAN FRANCISCO	CA	94105
BlackRock Institutional Trust Company, National Association Investment Funds for Employee Benefit Trusts - Global Allocation Collective Fund	400 Howard Street	San Francisco	CA	94105
FIDELITY PERSONAL TRUST COMPANY, F.S.B.	245 SUMMER STREET, 2ND FLOOR	BOSTON	MA	2210
Fidelity Puritan Trust - Fidelity Low-Priced Stock Fund	245 Summer Street	Boston	MA	2210

Tabella 7: Esempi di possibili matches che non fanno parte della ground truth FFIEC-LEI. (Bianco = FFIEC, Azzurro = LEI)

Per individuare ulteriori possibili matches sono stati selezionati i candidati attraverso la strategia di *Sorted Neighbourhood Blocking* utilizzando una finestra di dimensione 3.
Per la coppia FFIEC-LEI sono stati individuati 62 links reputati Matches non facenti parte della Ground Truth, mentre per FFIEC-SEC ne sono stati rilevati 279. Nelle tabelle 7 e 8 sono riportati alcuni esempi estratti a campione in cui il matching è molto plausibile.

Name	Address	City	State	Zip Code
SKAGIT BANK	301 EAST FAIRHAVEN AVENUE	BURLINGTON	WA	98233
SKAGIT STATE BANCORP, INC.	301 E. FAIRHAVEN AVENUE	BURLINGTON	WA	98233
SUNNYSIDE FEDERAL SAVINGS AND LOAN ASSOCIATION OF IRVINGTON	56 MAIN ST	IRVINGTON	NY	10533
SUNNYSIDE BANCORP, INC.	56 MAIN STREET	IRVINGTON	NY	10533
NORTH VALLEY BANK	2775 MAYSVILLE PIKE	ZANESVILLE	OH	43701
NORTH VALLEY BANCSHARES, INC.	2775 MAYSVILLE PIKE	ZANESVILLE	OH	43701
FIRST LIBERTY BANK	9601 NORTH MAY AVENUE	OKLAHOMA CITY	OK	73120
FIRST LIBERTY HOLDINGS LLC	9601 N MAY AVENUE	OKLAHOMA CITY	OK	73120
FRATERNITY FEDERAL SAVINGS AND LOAN ASSOCIATION	764 WASHINGTON BOULEVARD	BALTIMORE	MD	21230
FRATERNITY COMMUNITY BANCORP INC	764 WASHINGTON BOULEVARD	BALTIMORE	MD	21230

Tabella 8: Esempi di possibili matches che non fanno parte della ground truth FFIEC-SEC. (Bianco = FFIEC, Azzurro = SEC)

5 Conclusioni

Una prima considerazione che si può fare su quanto ottenuto è quella relativa al proprocessing. Si è potuto osservare che eseguendo tutti gli steps di pulizia e normalizzazione previsti sono stati raggiunti i risultati migliori, in quanto ciò ha permesso di individuare un numero più elevato di possibili matches. Questo è particolarmente utile nel caso si adotti una strategia di *Exact Blocking*.

Come seconda considerazione, si può osservare come i risultati ottenuti da entrambi gli approcci deterministico e di Machine Learning non supervisionato non presentino rilevanti differenze. Per quanto riguarda l'approccio supervisionato, invece, sono stati ottenuti scarsi risultati nella classificazione delle coppie considerate ambigue. Questo potrebbe essere dovuto al fatto che in molti casi è necessaria una valutazione umana da parte di un esperto del settore per stabilire l'effettivo matching.

Infine, sono stati individuati nuovi possibili matches non presenti nelle Ground Truths parziali messe a disposizione dalla *FEIII 2016 Challenge*. Il fatto di avere a disposizione le Ground Truths non finali ha portato ad avere un numero elevato di coppie classificate come matches la cui correttezza non è stata verificabile, sebbene osservando i valori sia molto probabile.

Riferimenti bibliografici

[1] Wikipedia contributors, "Record linkage — Wikipedia, the free encyclopedia." `https://en.wikipedia.org/w/index.php?title=Record_linkage&oldid=963403193`, 2020.

[2] M. Flood, J. Grant, H. Luo, L. Raschid, I. Soboroff, and K. Yoo, "Financial entity identification and information integration (feiii) challenge: the report of the organizing committee," in *Proceedings of the Second International Workshop on Data Science for Macro-Modeling*, pp. 1–4, 2016.

[3] R. C. Steorts, S. L. Ventura, M. Sadinle, and S. E. Fienberg, "A comparison of blocking methods for record linkage," in *International Conference on Privacy in Statistical Databases*, pp. 253–268, Springer, 2014.

[4] R. Baxter, P. Christen, and T. Churches, "A comparison of fast blocking methods for record linkage; erschienen in: Proceedings of the workshop on data cleaning, record linkage and object consolidation at the ninth acm sigkdd international conference on knowledge discovery and data mining; washington dc; 2003; o."