PROGETTO DATA MINING

Team 7
Gianfranco Sapia, 223954
Giovanni Iannuzzi, 214900
Andrea De Seta, 227755
Paolo Falvo, 223974
Domenico Sestito, 223962

This page intentionally left blank

Sommario

Business Understanding

Background

Un consorzio formato da 130 ospedali diversi situati sul suolo americano ha reso disponibile i dati raccolti nel corso di 10 anni (1999-2008) riguardanti le cure cliniche offerte ai propri pazienti. Da questi dati sono stati estratte quelle informazioni che possono essere utili ai fini di un'analisi delle condizioni dei pazienti dopo le cure, in particolare la lunghezza della degenza, test eseguiti sul paziente, farmaci somministrati e diagnosi riconducibili a una forma di diabete.

Il dataset risultante è composto da 50 attributi riguardanti le informazioni dei pazienti, le loro condizioni pregresse, le diagnosi e l'esito della cura.

Business Reason

È importante sapere, con l'ausilio di esperienze pregresse, se il paziente riceverà o meno le cure adeguate.

Business Goal

Stabilire se un paziente affetto da diabete sarà riammesso in ospedale entro 30 giorni (o meno) dalla sua dimissione.

Business Success Criteria

Riuscire a predire con un'accuratezza del 90% se una cura prescritta a un paziente avrà successo o meno. Il successo o meno è determinato dal ritorno del paziente in ospedale non prima dei 30 giorni dalla sua dimissione.

Inventory of Resources

I seguenti strumenti tecnologici, tra cui librerie, software e linguaggi di programmazione, sono stati utilizzati ai fini della risoluzione del problema:

- Jupyter-lab;
- Colab;
- Pycharm;
- Python;
- Numpy;
- Pandas;
- Matplotlib;
- Scikitlearn;
- Seaborn.

Data-Mining Goal

Si tratta di un problema di classificazione binaria il cui obiettivo è quello di trovare un modello in grado di desumere il ritorno entro 30 giorni di un paziente sulla base dei suoi attributi. Il risultato finale vedrà la valorizzazione dell'attributo "**readmitted**" con una combinazione binaria di valori, nello specifico 1 nel caso in cui il paziente ritornerà entro i 30 giorni, 0 altrimenti.

Project Plan

La seguente tabella riassume come è stato preparato e affrontato il lavoro in termini di tempo.

Fase	Time
Business understanding	1 settimana
Data understanding	1 settimana
Data preparation	1 settimana
Modeling	1 settimana
Evaluation	1 settimana

Data Understanding

Data Structure

I dati da analizzare sono stati forniti attraverso il dataset "*diabetic_data.csv*", che presenta le seguenti caratteristiche:

- multivariato;
- 101.766 righe, che rappresentano il numero di record;
- 50 colonne, che rappresentano il numero di attributi nel dataset;
- 5.088.300 dati totali.

Insieme al dataset in formato csv è stato fornito un paper riassuntivo di una precedente analisi sul medesimo set di dati. Grazie a questo paper si è riusciti ad avere delle informazioni precise e dettagliate su ogni attributo.

In particolare, per ognuno di essi, è stato fornito il tipo di dato, una breve descrizione e un elenco di dati univoci dove presenti.

Feature Name	Type	Description and values
Encounter ID	Nominal	Unique identifier of an encounter
Patient number	Numeric	Unique identifier of a patient
Race	Nominal	Values: Caucasian, Asian, African American, Hispanic, and other
Gender	Nominal	Values: male, female, and unknown/invalid
Age	Nominal	Grouped in 10-year intervals: [0, 10), [10, 20), ... , (90, 100)
Weight	Nominal	Grouped in 25-pound: [0-25),[25-50),… , >200
Admission type	Nominal	Integer identifier corresponding to 9 distinct values, for example, emergency, urgent, elective, newborn, and not available
Discharge disposition	Nominal	Integer identifier corresponding to 29 distinct values, for example, discharged to home, expired, and not available
Admission source	Nominal	Integer identifier corresponding to 21 distinct values, for example, physician referral, emergency room, and transfer from a hospital
Time in hospital	Numeric	Integer number of days between admission and discharge

Payer code	Nominal	Integer identifier corresponding to 23 distinct values, for example, Blue Cross\Blue Shield, Medicare, and self-pay
Medicai specialty	Nominal	Integer identifier of a specialty of the admitting physician, corresponding to 84 distinct values, for example, cardiology, internal medicine, family\general practice, and surgeon
Number of lab procedures	Numeric	Number of lab tests performed during the encounter
Number of procedures	Numeric	Number of procedures (other than lab tests) performed during the encounter
Number of medications	Numeric	Number of distinct generic names administered during the encounter
Number of outpatient visits	Numeric	Number of outpatient visits of the patient in the year preceding the encounter
Number of emergency visits	Numeric	Number of emergency visits of the patient in the year preceding the encounter
Number of inpatient visits	Numeric	Number of inpatient visits of the patient in the year preceding the encounter
Dìagnosis I	Nominal	The primary diagnosis (coded as first three digits of ICD9); 848 distinct values
Dìagnosis 2	Nominal	Secondary diagnosis (coded as first three digits of ICD9); 923 distinct values
Dìagnosis 3	Nominal	Additional secondary diagnosis (coded as first three digits of ICD9); 954 distinct values
Number of diagnoses	Numeric	Number of diagnoses entered to the system
Glucose serum test result	Nominal	Indicates the range of the result or if the test was not taken. Values: ">200;· ">300", "normal" and "none" if not measured
Aie test result	Nominal	Indicates the range of the result or if the test was not taken. Values: ">8" if the result was greater than 8%, "> 7" if the result was greater than 7% but less than 8%, "normal" if the result was less than 7%, and "none" if not measured.
Change of medications		Indicates if there was a change in diabeti e medications (either dosage or generic name). Values: "change" and "no change"

Dìabetes medications	Nominal	Indicates if there was any diabetic medication prescribed. Values: "yes" and "no"
24 features for medications	Nominal	For the generic names: metformin, repaglinide, nateglinide, chlorpropamide, glimepiride, acetohexamide, glipizide, glyburide, tolbutamide, pioglitazone, rosiglitazone, acarbose, miglitol, troglitazone, tolazamide, examide, sitagliptin, insulin, glyburide-metformin, glipizide-metformin, glimepiride-pioglitazone, metformìn-rosiglitazone, and metformin-pioglitazone, the feature indicates whether the drug was prescribed or there was a change in the dosage. Values: "up" if the dosage was increased during the encounter, "down" if the dosage was decreased, "steady" if the dosage did not change, and "no" if the drug was not prescribed
Readmitted	Nominal	Days to inpatient readmission. Values: "<30" if the patient was readmitted in less than 30 days, ">30" ifthe patient was readmìtted in more than 30 days, and "No" for no record of readmìssion.

Di questi 50 attributi, si ha la seguente divisione:

- 41 attributi nominali;
- 9 attributi numerici.

Il tipo di alcuni attributi, letti tramite l'utilizzo della libreria Pandas, non risultava coerente con quello descritto nella precedente tabella. Per tanto, si è proseguito ad una conversione dei tipi di dati errati a quelli corretti. Gli attributi interessati a questa conversione sono i seguenti:

- *encounter_id*;
- *admission_type_id*;
- *discharge_disposition_id*;
- *admission_source_id*;

Questi attributi, che risultavano come "numerici", sono stati convertiti a "nominali". Un ulteriore attributo, ovvero "*readmitted*", è stato convertito in "categorico" poiché è un tipo di dato adatto per la class label.

A seguito di questa operazione il tipo degli attributi risulta corretto come si evince nella seguente tabella, la quale riporta anche la quantità di dati presenti (**Non-Null**) per ogni attributo.

Attributo	Valori Non-null	Tipo
Encounter ID	101766 non-null	object
Patient number	101766 non-null	int64
Race	99493 non-null	object
Gender	101766 non-null	object
Age	101766 non-null	object
Weight	3197 non-null	object
Admission type	101766 non-null	object
Discharge disposition	101766 non-null	object
Admission source	101766 non-null	object
Time in hospital	101766 non-null	int64
Payer code	61510 non-null	object
Medicai specialty	51817 non-null	object
Number of lab procedures	101766 non-null	int64
Number of procedures	101766 non-null	int64
Number of medications	101766 non-null	int64
Number of outpatient visits	101766 non-null	int64
Number of emergency visits	101766 non-null	int64
Number of inpatient visits	101766 non-null	int64
Dìagnosis I	101745 non-null	object
Dìagnosis 2	101408 non-null	object
Dìagnosis 3	100343 non-null	object
Number of diagnoses	101766 non-null	int64
Glucose serum test result	101766 non-null	object
Aie test result	101766 non-null	object
Change of medications	101766 non-null	object
Dìabetes medications	101766 non-null	object
24 features for medications	101766 non-null	object
Readmitted	101766 non-null	category

Data Quality

Nella fase di Data Quality si è proseguiti alla verifica della presenza di valori **null** nel dataset, la quale ha riportato i seguenti risultati per alcuni attributi:

Attributo	Num. valori mancanti	% valori mancanti
race	2273	2,23
weight	98569	96,86
payer_code	40256	39,56
medical_specialty	49949	49,08
diag_1	21	0,02
diag_2	358	0,35
diag_3	1423	1,40

Data Exploration

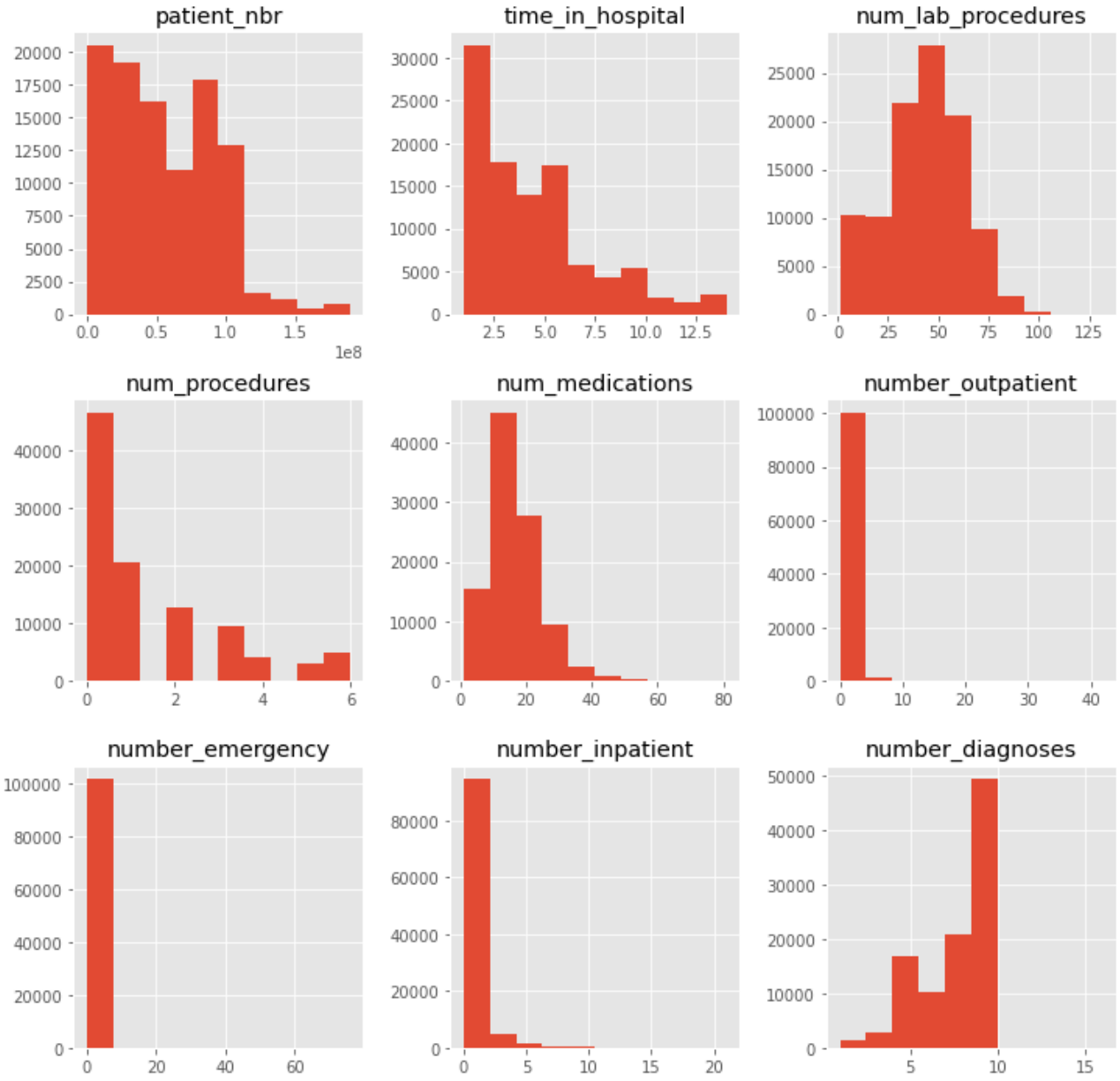
Per quanto riguarda la class label, l'esplorazione ha riguardato la distribuzione dei valori in quanto era importante riuscire a capire se il dataset fosse bilanciato o meno.

Dalle analisi si è potuto osservare la caratterizzazione in tre differenti valori:

- **"NO"**, pazienti che non sono tornati in ospedale al momento della raccolta dei dati. Sono presenti 54864 record con questo valore;
- **">30"**, pazienti che sono tornati in ospedale dopo più di 30 giorni. Sono presenti 35545 record con questo valore;
- **"<30"**, pazienti che sono tornati in ospedale entro 30 giorni. Sono presenti 11357 record con questo valore.

Grafici su attributi numerici

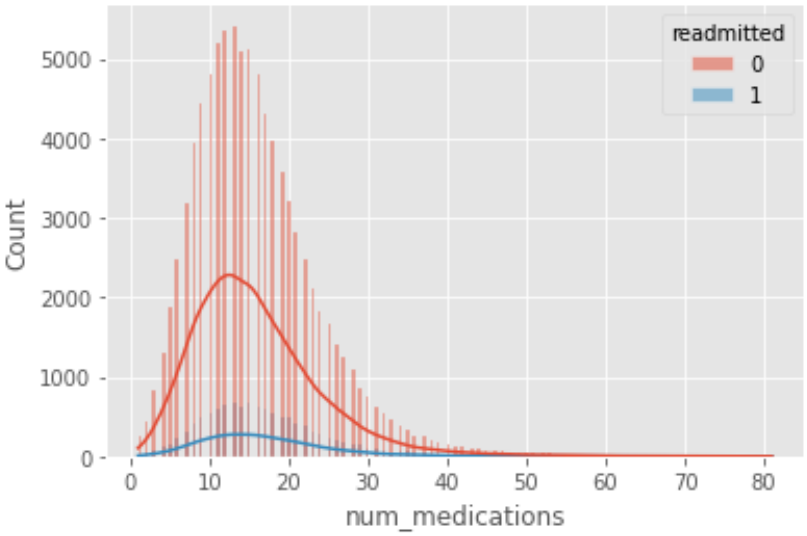
Una prima visualizzazione dei dati è stata fatta sugli attributi numerici utilizzando gli istogrammi.



Da questi grafici si può evincere sui tre attributi *number_outpatient*, *number_emergency* e *number_inpatient* che il valore più frequente è 0. Una possibile soluzione, durante la fase di Data Preparation, potrebbe essere quella di discretizzare i valori. Sui restanti valori non sono presenti anomalie evidenti da tenere poi in considerazione.

Continuando l'analisi sugli attributi numerici, in particolar modo sfruttando la libreria *seaborn*, si è messo a confronto la distribuzione delle class label per i vari istogrammi. Da qui non emerge nessun'altra osservazione particolare.

Di seguito alcuni grafici di esempio:

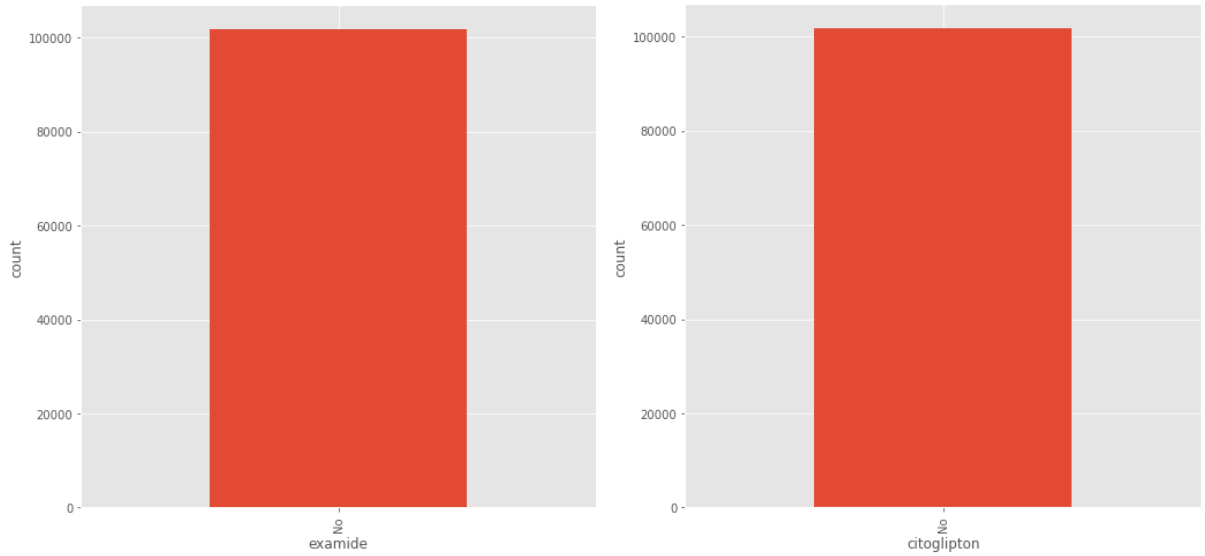


Grafici su attributi nominali

Dall'analisi fatta sugli attributi nominali sono stati riportati di seguito solo quelli ritenuti interessanti ai fini di una successiva lavorazione.

Come descritto nella fase di Data Understanding sono presenti 24 attributi rappresentanti diversi tipi di medicinale. Questi attributi possono assumere un determinato valore a seconda dal dosaggio somministrato a ogni paziente.

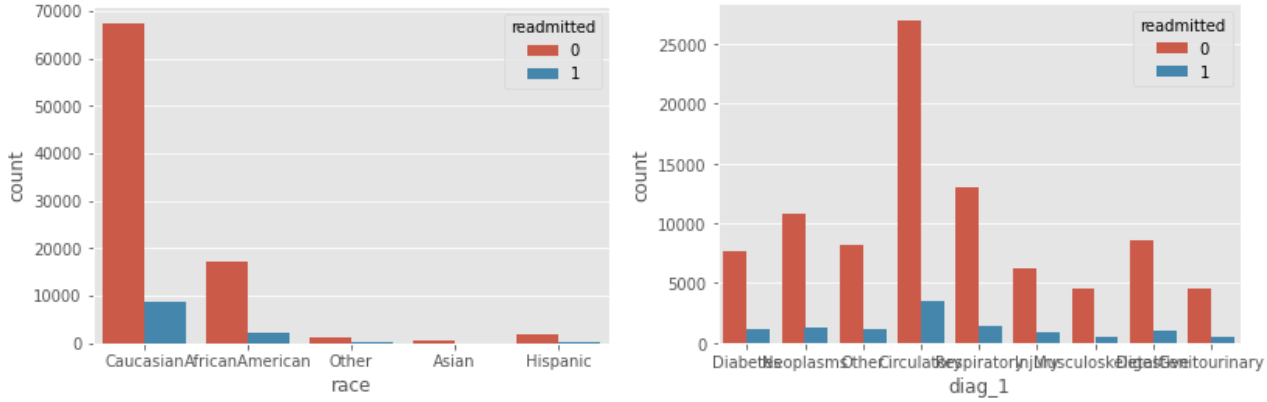
In generale è presente una buona varietà di valori assunti su tutti questi attributi tranne che per i seguenti:



Da questi grafici emerge che per tutto il dataset il valore di "*examide*" e "*citoglipton*" è sempre costante. La soluzione applicabile successivamente potrà essere quella di rimuovere l'attributo dal dataset.

Anche per gli attributi nominali si è verificato, sempre attraverso l'utilizzo della libreria *seaborn*, se il valore di alcuni attributi fosse discriminante. Da questi grafici non è stato possibile verificarlo, poiché il dataset è fortemente sbilanciato.

Di seguito alcuni grafici di esempio:



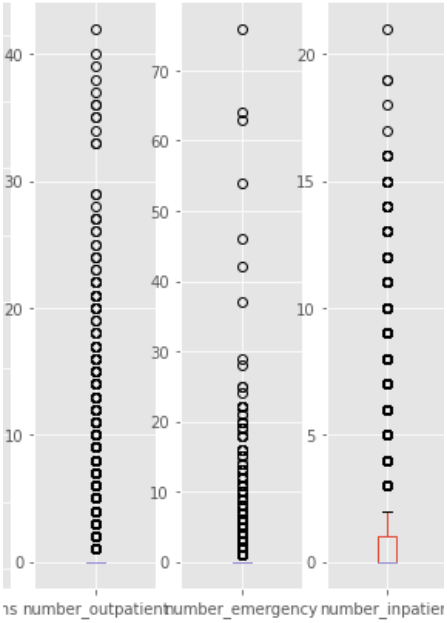
Boxplot

Tramite i boxplot si può osservare la distribuzione dei valori degli attributi numerici e i relativi outliers.

Il seguente grafico conferma che sugli attributi "*number_outpatient*", "*number_inpatient*" e "*number_emergency*" la distribuzione non è omogenea, infatti la quantità di istanze valorizzate a 0 è molto maggiore rispetto agli altri valori.

Di conseguenza, una possibile soluzione potrebbe essere la discretizzazione delle colonne citate precedentemente:

- per "*number_outpatient*" in "0" e ">0";
- per "*number_emergency*" in "0" e ">0";
- per "*number_inpatient*" in "0", "1", ">1".



Scatter plot

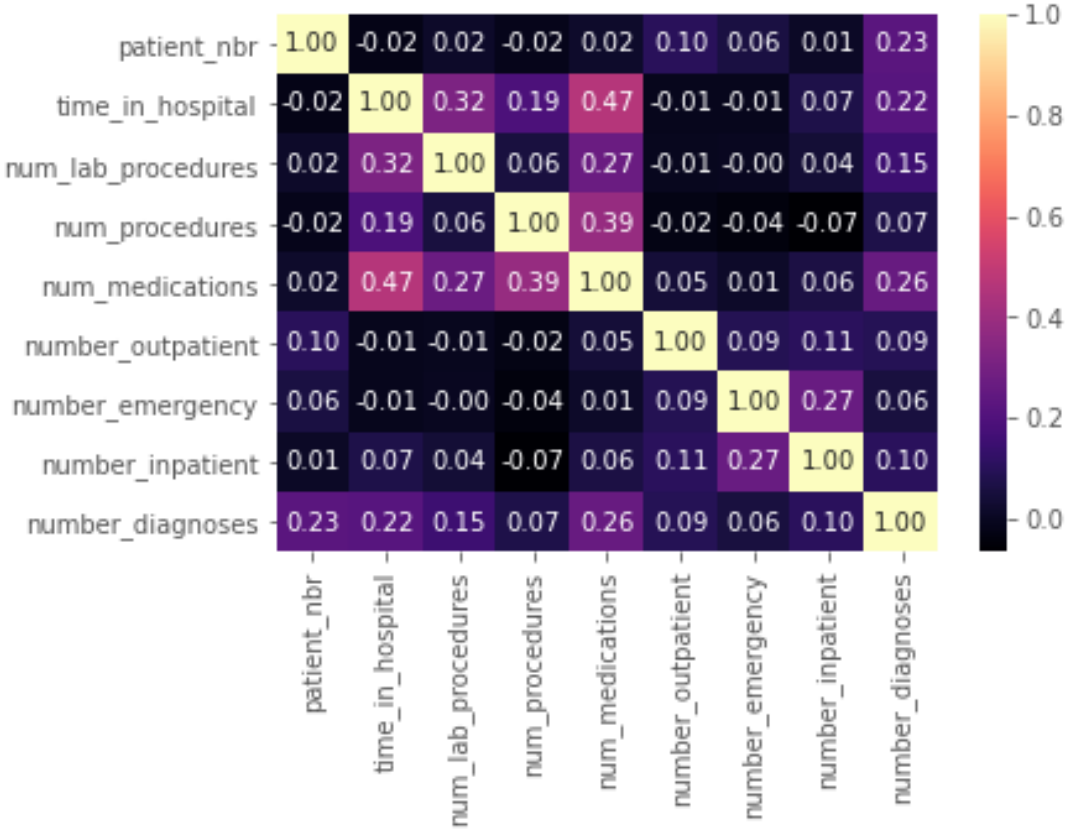
L'utilizzo dello scatterplot ha permesso di capire se tra gli attributi ci fosse una certa correlazione. Il risultato di questo grafico ha mostrato una leggera correlazione tra le seguenti coppie di attributi:

- "*num_medications*" e "*time_in_hospital*";
- "*num_medications*" e "*num_procedures*";
- "*num_lab_procedures*" e "*time_in_hospital*";
- "*num_medications*" e "*num_lab_procedures*".


time_in_hospital
num_lab_procedures
num_procedures
num_medications
number_diagnoses
readmitted
0
1

Heatmap

Attraverso l'heatmap le correlazioni precedentemente indicate sono state confermate e dal risultato del grafico si può dire che non si ha una forte correlazione tra i vari attributi.



Data Preparation

Data Selection

Alla luce dell'analisi fatte sui dati si è scelto di rimuovere alcuni attributi poiché ininfluenti per le fasi successive. Di seguito si riportano gli attributi che sono rimossi con le rispettive motivazioni:

- "*weight*", in quanto non era valorizzato nel 96.86% dell'intero dataset;
- "*encounter_id*", in quanto identificativo di ogni record e non utile nelle fasi successive;
- "*patient_nbr*", in quanto identificativo di ogni paziente e non utile nelle fasi successive;
- "*examine*" e "*citoglipton*", in quanto per l'intero dataset hanno sempre lo stesso valore;
- "*payer_code*", in quanto identificativo del metodo di pagamento usato e non utile nelle fasi successive;
- "*medical_specialty*", in quanto il 49.08% di dati era assente nel dataset e inoltre quelli presenti avevano numerosi valori diversi e quindi difficile da raggruppare.

Prima di rimuovere l'attributo "*medical_specialty*" si è deciso di approfondire la conoscenza del dominio, chiedendo anche il parere di un esperto nel campo medico, che ha confermato che ai fini della nostra valutazione non è utile sapere qual è la specializzazione medica del dottore che ha ricoverato il paziente.

Data Cleaning

Durante la fase di Data Cleaning sono stati risolti i problemi individuati durante il task di Data Quality per quanto riguarda gli attributi di "*race*", "*diag_1*", "*diag_2*" e "*diag_3*".

Le opzioni che sono state considerate per l'attributo "*race*" sono le seguenti:

- fare il drop dei record con valore "*null*";
- sostituire il valore "*null*" con il valore di moda, che in questo caso si tratta del valore "***Caucasian***";
- sostituire il valore "*null*" con il valore "***Others***";
- sostituire il valore "*null*" con un nuovo valore "***missing***", che indica l'assenza del dato nel record.

A fronte di queste opzioni si è deciso di sostituire ai valori mancati il nuovo valore "***missing***". È stata effettuata questa scelta in quanto è stato ritenuto un numero di record sufficientemente importante per non poterli eliminare (circa 2000 record),

mentre le altre due opzioni (sostituire i valori "*null*" con "*Caucasian*" o "*Others*") avrebbero potuto "*sporcare*" la distribuzione dei dati.

Per quanto riguarda le diagnosi è stata necessaria una analisi ulteriore: inizialmente si era ipotizzato che l'assenza della diagnosi primaria implicasse l'inutilità del record anche in presenza delle diagnosi secondarie.

Nuovamente il parare dell'esperto è stato determinante. Infatti, ci ha permesso di comprendere a pieno il significato di questi attributi: è stato appreso che ogni diagnosi era indipendente dall'altra, quindi ad esempio, la mancanza di valore sulla diagnosi primaria non aveva nessuna ripercussione sulle altre diagnosi presenti.

Di conseguenza si è scelto di eliminare solo i record dove erano contemporaneamente assenti i valori di tutte e tre le diagnosi, e di sostituirle con il valore "***missing***" nei casi in cui era assente una diagnosi ma presente almeno una tra le altre.

Data Constructing

In questa fase si è trasformato i valori di alcuni attributi sulla base di osservazioni fatte durante la fase di Data Exploration. In particolare, sono stati modificati i valori di:

- "*diag_1*", "*diag_2*", "*diag_3*";
- "*number_outpatient*";
- "*number_emergency*";
- "*number_inpatient*";
- "*readmitted*".

Per quanto riguarda i valori delle diagnosi, che rappresentavano i codici ICD9, sono stati sostituiti dalla macrocategoria alla quale il codice appartiene. Questa discretizzazione è stata fatta per ridurre il dominio di questi attributi e semplificare il lavoro agli algoritmi di machine learning.

Come già anticipato nella fase di Data Exploration sono state eseguite le seguenti modifiche:

- per l'attributo "*number_outpatient*" è stata fatta una discretizzazione, dopo l'analisi fatta sul boxplot, raggruppando i valori in due categorie: i valori uguali a zero con "0" e i valori maggiori di zero con ">0".
- per l'attributo "*number_emergency*" è stata fatta una discretizzazione, dopo l'analisi fatta sul boxplot, raggruppando i valori in due categorie: i valori uguali a zero con "0" e i valori maggiori di zero con ">0".

- per l'attributo "*number_inpatient*" si è adottata la stessa procedura, suddividendo nelle seguenti categorie: i valori uguali a zero con "0", i valori uguale a uno con "1" e i valori maggiori di uno con ">1".

Essendo un problema di classificazione binaria è stato deciso di discretizzare la class label "*readmitted*" in due possibili valori: 1 e 0. Questa colonna presentava come valori ">30", "<30" e "NO", quindi si è deciso di raggruppare in due categorie: "<30" con valore 1 e i valori ">30" e "NO" con il valore 0.

Un ulteriore operazione da compiere, prima di passare alla fase di modellazione, è quella di bilanciare il dataset. Come visto precedentemente, il dataset risulta sbilanciato verso i valori con "0" nella class label con 90409 record su 101766 record totali nel dataset.

Per effettuare questo bilanciamento è stato scelto di ricampionare il dataset estraendo con rimpiazzo 90409 record con class label uguale a "1".

In poche parole, è stato applicato un resampling del dataset, nel particolare un oversampling: in modo random sono state selezionate delle righe della minority class (attribute target pari a 1) e sono state copiate all'interno del dataset fino a quando il numero di righe con attribute target pari a 1 uguagliasse le righe con attribute target pari a 0.

In questo modo anche il training set su cui si "addestrerà" l'algoritmo sarà bilanciato e permetterà una migliore previsione.

L'ultima operazione prepara il dataset alla modellazione per la fase successiva. L'operazione eseguita è stata quella di fare la binarizzazione di tutti gli attributi nominali. La binarizzazione è stata fatta sul dataset utilizzando la funzione *resample* della libreria Pandas sugli attributi categorici.

Modeling

Selecting Modeling Techniques

Il primo passo da fare per proseguire nella fase di modellazione è quello di valutare quali tecniche possono essere utilizzate per la classificazione del nostro problema. Come già detto in precedenza si tratta di un problema di classificazione binaria, quindi le tecniche prese in considerazione per andare ad operare su di esso sono le seguenti:

- *DecisionTree*;
- *Naive Bayes Classifier*;
- *RandomForestClassifier*;
- *AdaBoostClassifier*;

Entrambe le tecniche "*DecisionTree*" e "*RandomForestClassifier*" sono state utilizzate con due diversi criteri: "*entropy*" e "*gini index*".

Generate Test Design

È stato deciso di procedere con due approcci differenti. Questo perché vi era la possibilità di dividere la fase di training e validation su due dataset diversi: bilanciato e sbilanciato.

Per la fase di training l'unica scelta sensata possibile in realtà era quella dell'utilizzo del bilanciato, infatti utilizzando quello sbilanciato poteva portare a dei fenomeni di overfitting o di apprendimento di una distribuzione sbilanciata.

È stata comunque svolta fase di training anche su quello sbilanciato con il fine di evidenziarne le differenze sui risultati ottenuti.

In tutti e due gli approcci partendo dal relativo dataset creato nella fase di Data Preparation è stato utilizzato l'80% di esso come training set e il 20% come test set. Il modello è stato testato utilizzando la tecnica di K-fold con il numero di split pari a dieci.

La metrica utilizzata per valutare qual è stata la miglior tecnica è l'accuratezza.

Di seguito riportiamo i risultati ottenuti:

Tecnica	Accuratezza (dataset sbilanciato)	Accuratezza (dataset bilanciato)
DecisionTreeClassifier('entropy')	0.801160	0.922539
DecisionTreeClassifier('gini')	0.799255	0.922311
NaiveBayes	0.153147	0.51615
RandomForestClassifier('gini')	0.888358	0.980650
RandomForestClassifier('entropy')	0.888358	0.979095
AdaBoostClassifier	0.887695	0.623578

Evaluation

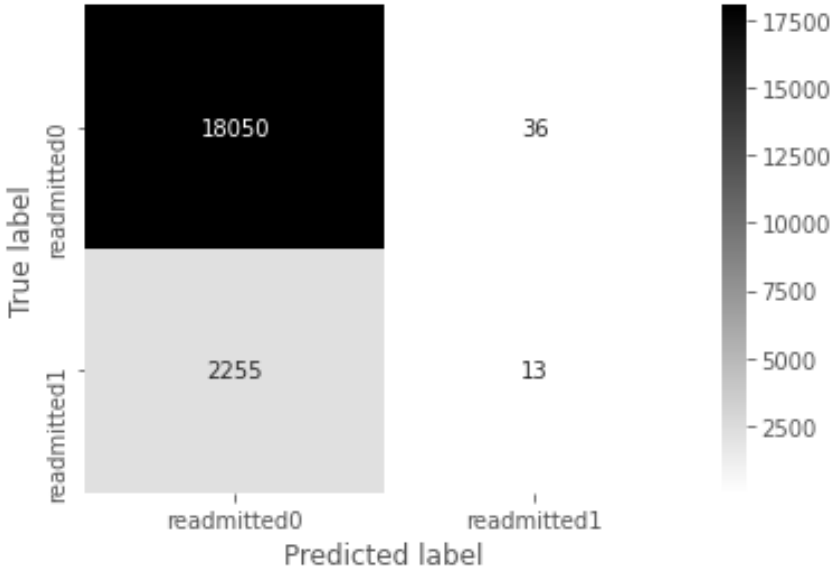
Evaluate results

Utilizzando due approcci differenti nella fase di modeling, nella fase di evaluation sono stati valutati entrambi col fine di evidenziarne le differenze.

Come si evince dalla tabella della precedente fase, sia nel primo approccio e sia nel secondo, la tecnica che ha ottenuto un risultato migliore è stata quella del "*RandomForest*" con il criterio "*gini index*". Dunque, questa tecnica è stata scelta per la prediction in entrambi gli approcci.

Primo approccio

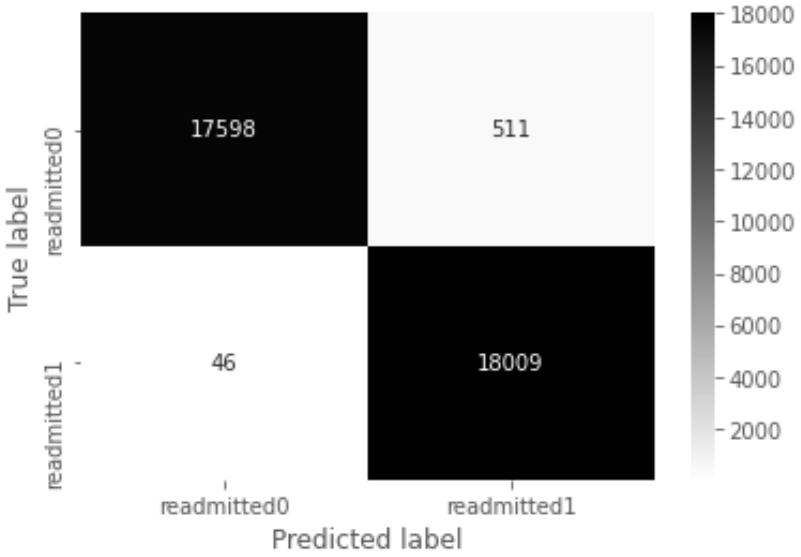
Nel caso del modello allenato sul dataset sbilanciato (cioè il dataset originale) la prediction sul test set ha ottenuto un'accuratezza dell'89%. A prima vista questo sembrerebbe un buon risultato, ma in realtà è una numerica ingannevole, come si evince dalla confusion matrix:



Infatti, siamo di fronte all'Accuracy Paradox: modellazione con accuratezza alta ma con un forte overfitting poiché tendenzialmente il modello ha imparato a predire il valore 0 dell'attribute class in quanto è il valore più presente (90% dei casi).

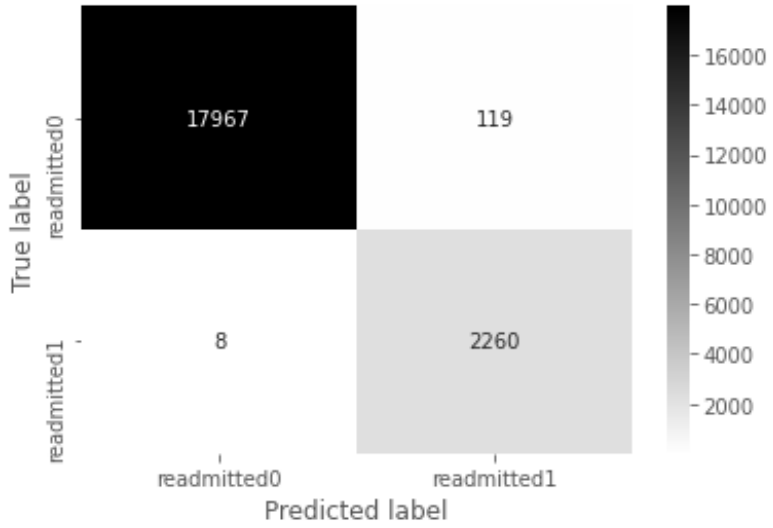
Secondo approccio

Nel caso del modello allenato sul dataset bilanciato i risultati ottenuti sono stati decisamente migliori con un'accuratezza del 98%. Questi risultati vengono confermati dalla confusion matrix:



Nell'operazione precedente è stato utilizzato come test set il 20% del dataset bilanciato, ma per completezza e per confronto di risultati è stato deciso di utilizzare questo modello anche sul 20% del dataset sbilanciato. Anche in questo caso i risultati ottenuti sono stati ottimi, con un'accuratezza del 99%.

Di seguito la confusion matrix:

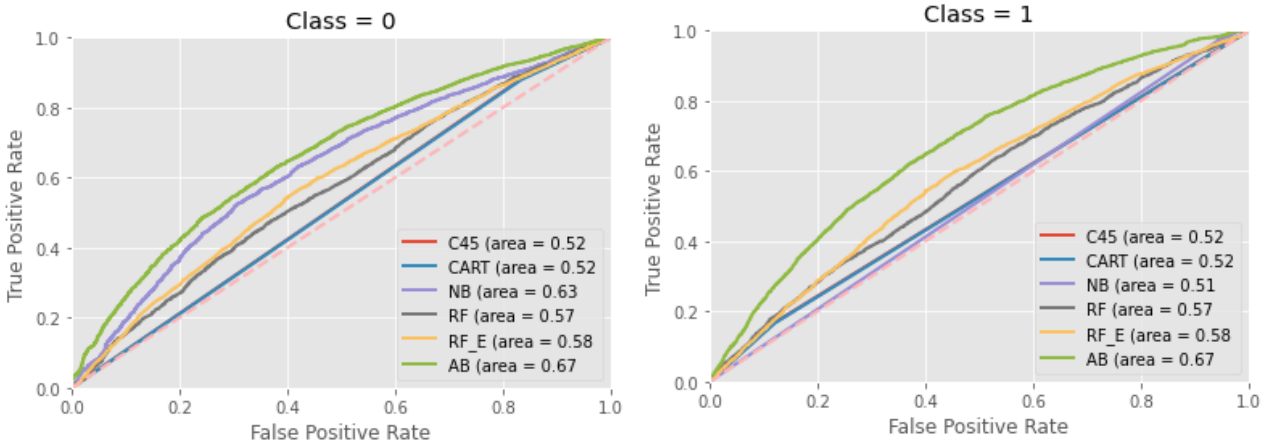


Roc curve

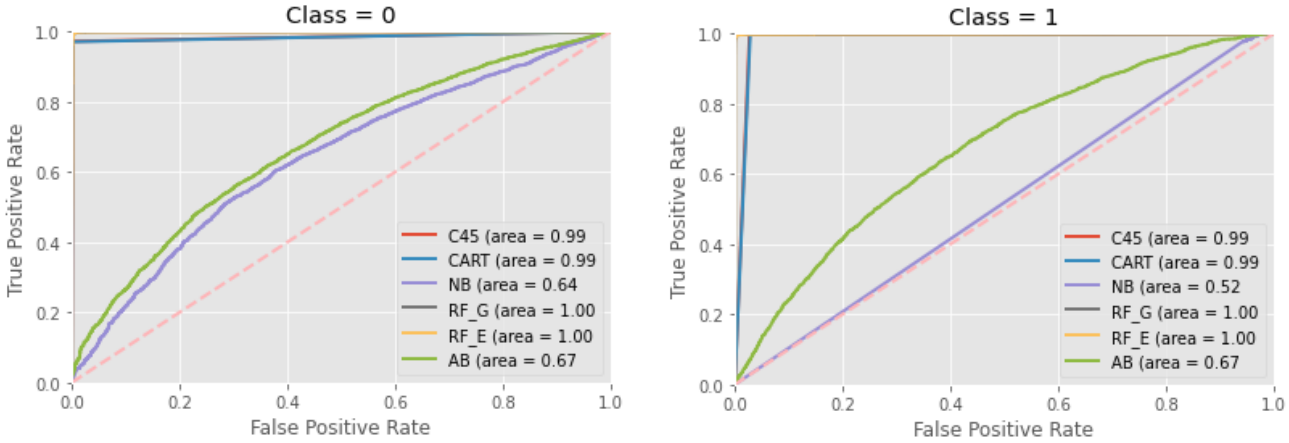
Un ulteriore metodo di analisi per verificarne la correttezza dei risultati è quello delle Roc Curve (*Receiver Operating Characteristic*). Le Roc Curve vengono utilizzate per evidenziare i rapporti fra allarmi veri e falsi allarmi e vengono create tracciando il valore del True Positive Rate (TPR, frazione di veri positivi) rispetto al False Positive Rate (FPR, frazione di falsi positivi).

Utilizzando il primo approccio di modellazione, si può notare come i risultati ottenuti per ogni algoritmo utilizzato siano vicini alla retta soglia, cioè i valori di AUC (Area Under The Curve) sono prossimi al valore di 0,5.

Questo significa che il modello potrebbe non avere capacità discriminatoria per distinguere tra classe positiva e classe negativa.



Seguendo il secondo approccio (cioè testando i modelli sul dataset bilanciato) i risultati ottenuti sono stati decisamente migliori. Come si può notare dal grafico, i valori di AUC del RandomForest sono praticamente perfetti (pari a 1), dunque ci troviamo nella situazione ideale: il modello ha una misura ideale di separabilità, è perfettamente in grado di distinguere tra classe positiva e classe negativa.



Conclusione

In conclusione, possiamo affermare che il dataset si è rilevato essere molto pulito dal punto di vista della qualità del dato ma pessimo per quanto riguarda la distribuzione dei valori della class label.

Questo ha reso la fase di Data Understanding e Data Preparation abbastanza veloce e intuitiva nonostante abbia comunque richiesto uno studio approfondito delle caratteristiche degli attributi e del loro significato.

La pessima distribuzione ha invece fortemente influito sulle ultime due fasi, rendendole poco lineari e obbligandoci a conti fatti a modificare il dataset originale.

Al termine di tutte le analisi i risultati ottenuti sono più che soddisfacenti essendo riusciti anche a superare con un buon margine l'obbiettivo prefissato, in fase di progettazione, nel Business Understanding.