Progetto Icon 2021-2022

Sistema diagnostico dei problemi cardiaci

Angelo Cassano 648427

Luca Depalo 646958

Il progetto consiste nell'utilizzare un dataset contenente i vari indicatori chiave associati ai problemi al cuore di un campione di persone. Il dataset specifica anche coloro che hanno effettivamente riscontrato questo tipo di problemi. Il progetto prevede l'uso di una parte di questi dati per addestrare un classificatore utilizzando diversi modelli e utilizzando una parte dei dati come test per calcolare e confrontare le performance dei modelli utilizzati. Di seguito verranno illustrate caratteristiche e scopi del dataset.

Il dataset di riferimento è disponibile qui: link

Cosa contiene questo dataset?

Questo dataset contiene i risultati del sondaggio annuale (2020) del Center for Disease Control and prevention (CDC) su un campione di quattrocentomila americani adulti relativo al loro stato di salute

Di cosa tratta questo dataset?

Secondo il CDC, le patologie cardiache sono una delle principali cause di morte per le persone di varie etnie presenti negli Stati Uniti (afroamericani, nativi americani, nativi d'Alaska e caucasici). Quasi la metà degli americani (47%) presenta almeno uno dei tre fattori chiave di rischio di patologie cardiache: ipertensione, ipercolesterolemia e tabagismo. Altri fattori chiave includono diabete, obesità, scarsa attività fisica e consumo elevato di bevande alcoliche. Determinare e prevenire quei fattori che hanno il maggior impatto nelle patologie cardiache è molto importante in ambito sanitario. Per questo motivo, negli ultimi anni parte della ricerca in ambito informatico si è proposta di sviluppare soluzioni innovative che permettano l'applicazione di metodi di machine learning al fine di determinare "pattern" dai dati che possano predire le condizioni di un paziente.

Da dove proviene il dataset e quali operazioni sono state applicate?

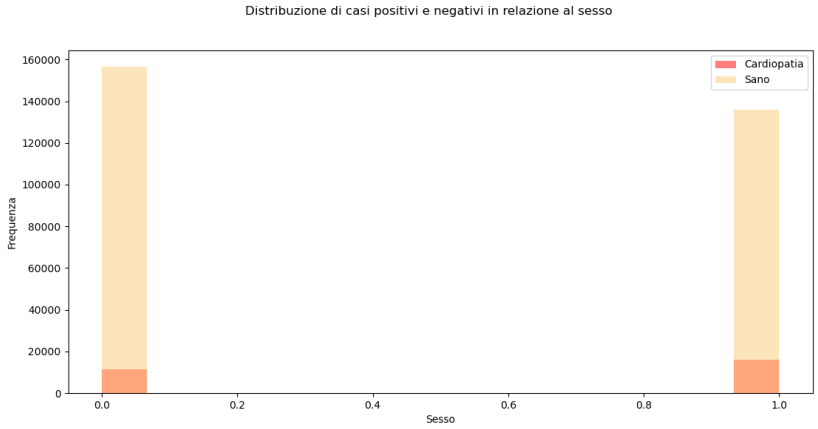
Il dataset originale proviene dal CDC ed è frutto del lavoro del Behavioral Risk Factor Surveillance System (BRFSS), che conduce sondaggi telefonici annuali per raccogliere dati sullo stato di salute dei cittadini degli U.S.A. Come affermato dal CDC: "Stabilito nel 1984 in 15 stati, il BRFSS adesso raccoglie dati in tutti i 50 stati. Il BRFSS compila più di 400,000 interviste ogni anno, rendendolo il più grande e il più costante sistema di questionari al mondo.". Il dataset più recente include dati del 2020. Consiste di 401.958 righe e 279 colonne. La maggioranza delle colonne sono domande poste agli intervistati circa la loro situazione di salute, ad esempio "Hai seri problemi a camminare o a salire le scale?" o "Hai fumato almeno 100 sigarette nella tua vita? ". In questo dataset, si notano diversi fattori che direttamente o indirettamente influenzano le patologie cardiache, e per questo si è deciso di selezionare le variabili più rilevanti al fine di rendere il dataset più usabile per un progetto di machine learning.

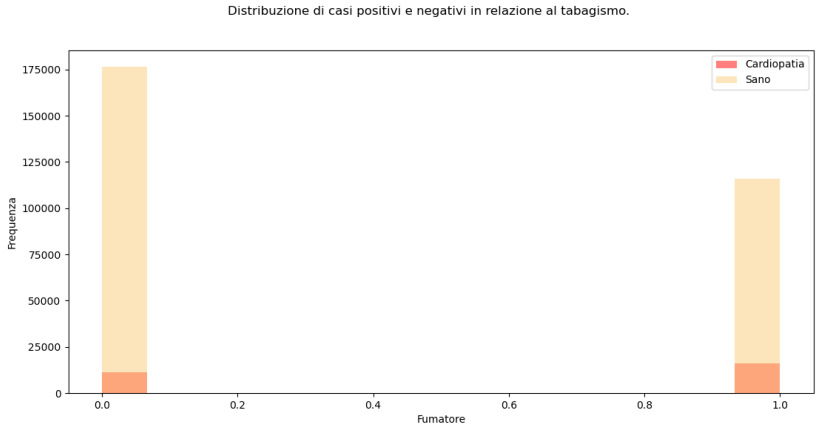
Cosa si può fare con questo dataset?

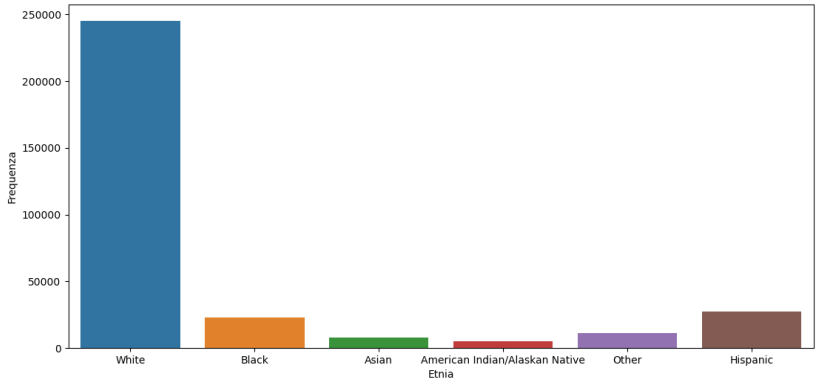
Come descritto in precedenza, il dataset originale di quasi trecento variabili è stato ridotto alle venti variabili più significative. In aggiunta all'analisi esploratoria dei dati, questo dataset può essere utilizzato su diversi metodi di machine learning, in particolar modo classificatori. È necessario trattare la variabile "HeartDisease"

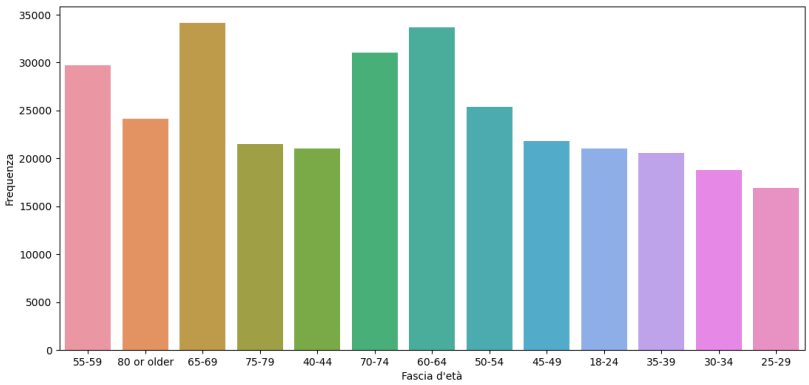
come booleana ("Yes" - la persona che ha risposto in tal modo soffre di patologie cardiache; "No" - la persona che ha risposto così non soffre di patologie cardiache).

Di seguito una panoramica dei dati estratti analizzando il dataset e come essi sono correlati fra loro e con la presenza di patologie cardiache:

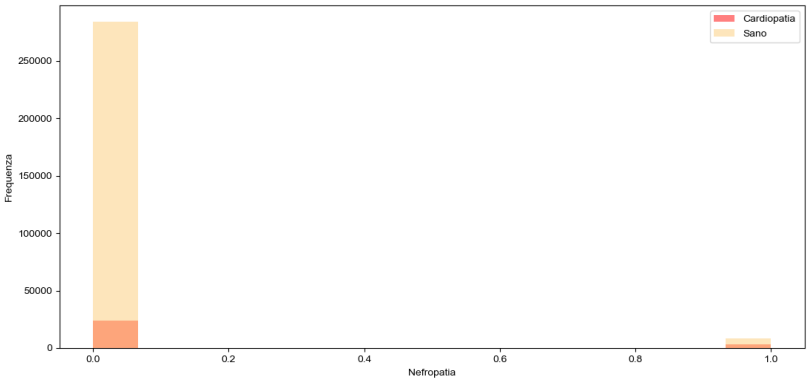




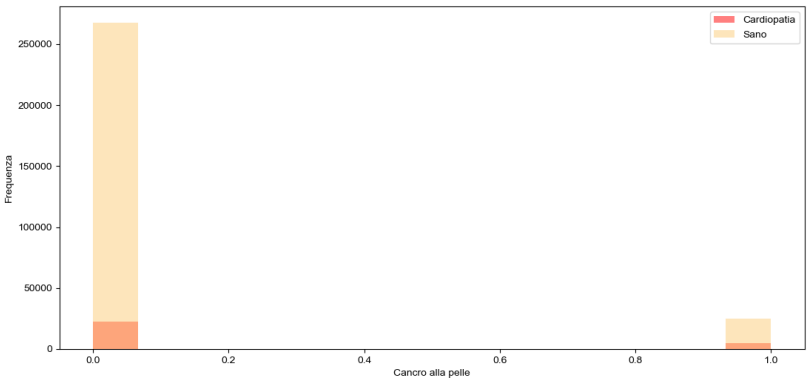




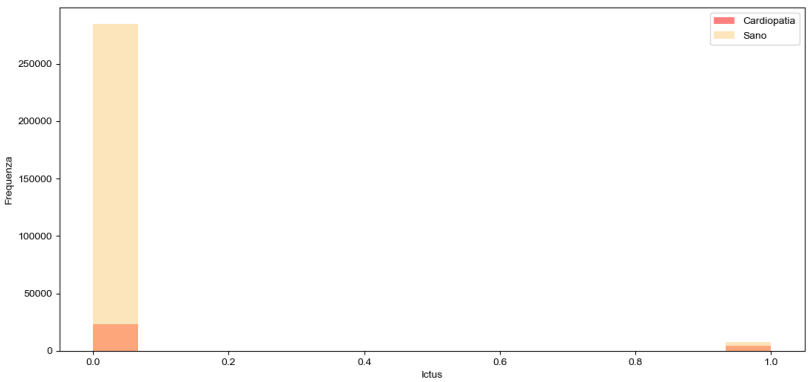
Distribuzione di casi positivi e negativi in relazione alle nefropatie



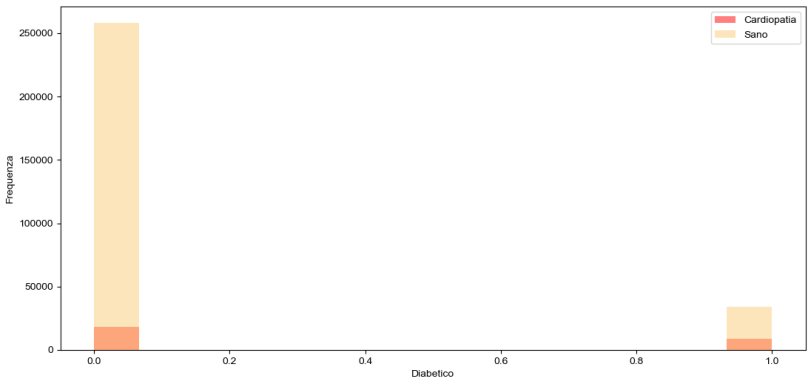
Distribuzione di casi positivi e negativi in relazione al cancro alla pelle

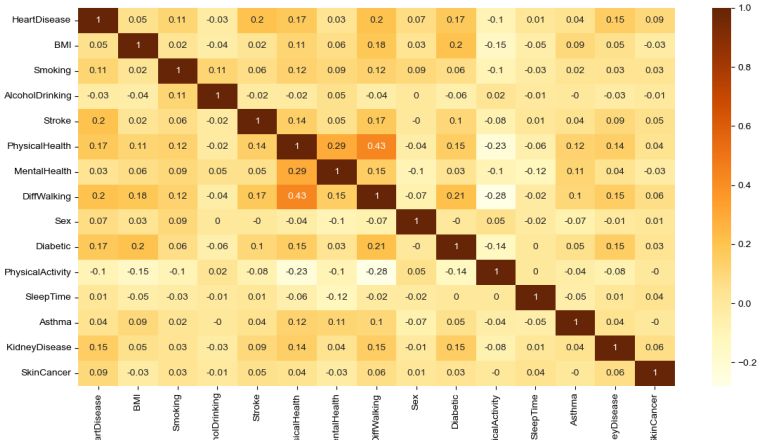


Distribuzione di casi positivi e negativi in relazione a ictus

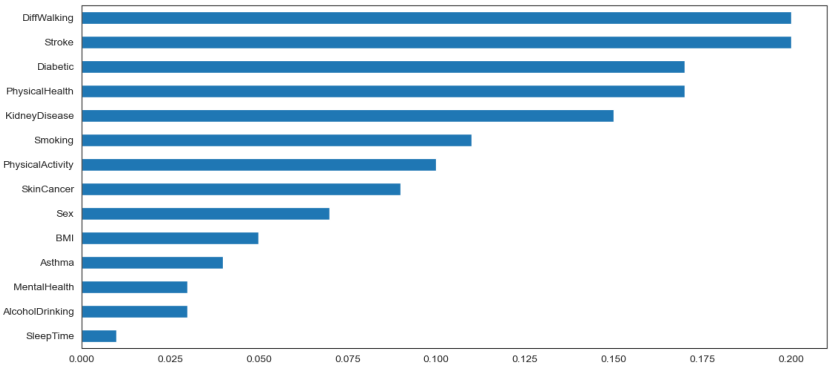


Distribuzione di casi positivi e negativi in relazione al diabete

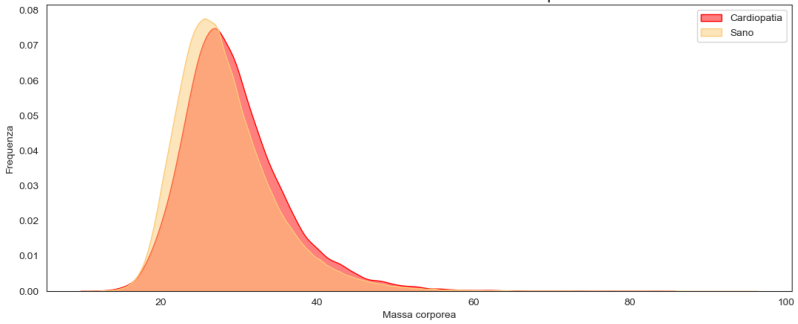


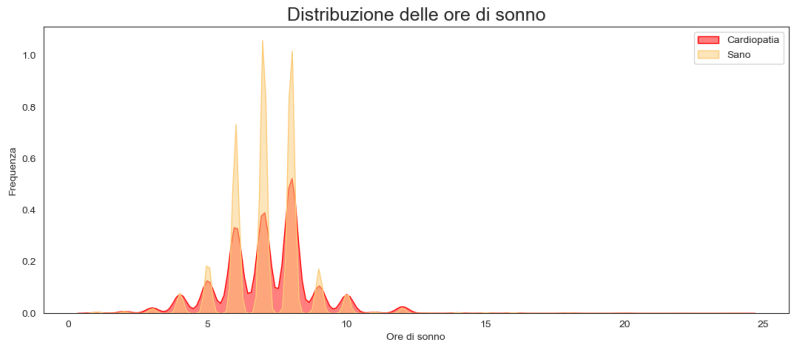


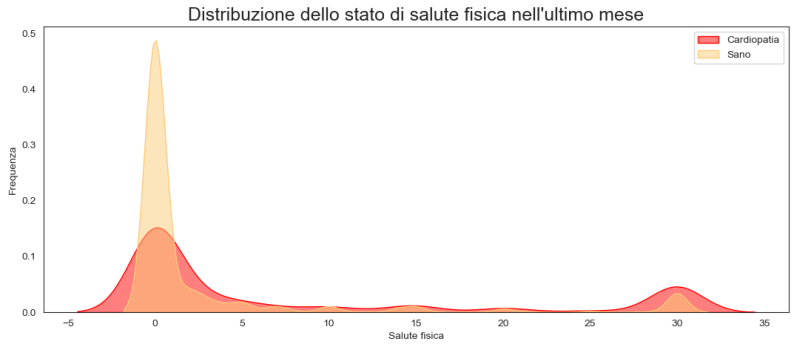
Distribuzione della correlazione di features

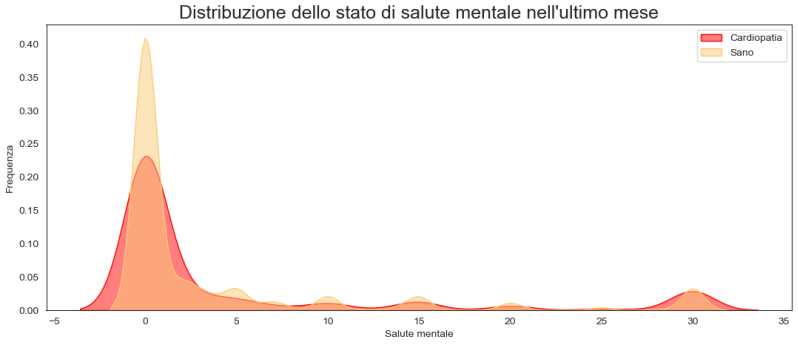


Distribuzione dell'indice di massa corporea









Quali modelli di machine learning sono stati utilizzati su questo dataset?

Per classificare i casi fra positivi e negativi in merito alla presenza di patologie cardiache sono state utilizzate le seguenti metodiche:

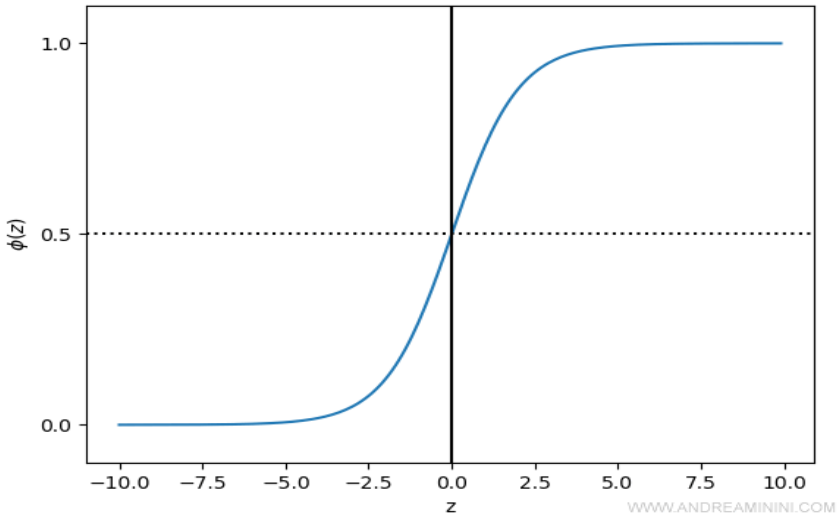
- Regressione logistica
- Classificatore bayesiano naive gaussiano
- Classificatore K nearest neighbors
- Classificatore random forest
- Classificatore AdaBoost

Di seguito una spiegazione di ciascun algoritmo:

REGRESSIONE LOGISTICA:

La regressione logistica è un algoritmo di apprendimento supervisionato che costruisce un modello probabilistico lineare di classificazione dei dati. E' usata nel machine learning per l'addestramento di un algoritmo nella classificazione supervisionata dei dati. La classificazione può essere binaria (due classi) o multiclasse (più classi). Nella fase di addestramento l'algoritmo di regressione logistica prende in input n esempi da un insieme di training (X). Ogni singolo esempio è composto da m attributi e dalla classe corretta

di appartenenza. Durante l'addestramento l'algoritmo elabora una distribuzione di pesi (W) che permetta di classificare correttamente gli esempi con le classi corrette. Poi calcola la combinazione lineare z del vettore dei pesi W e degli attributi xm z = → x · → w + b z=x→·w→+b z = x 0 · w 0 + . . . + x m · w m + b z=x0·w0+...+xm·wm+b La combinazione lineare z viene passata alla funzione logistica (sigmoid) che calcola la probabilità di appartenenza del campione alle classi del modello. ϕ (z) = 1 1 + e − z ϕ(z)=11+e−z La funzione sigmoid calcola la probabilità di appartenenza del campione alle classi. E' la tipica curva a S della distribuzione delle probabilità.



GAUSSIAN NAIVE BAYES

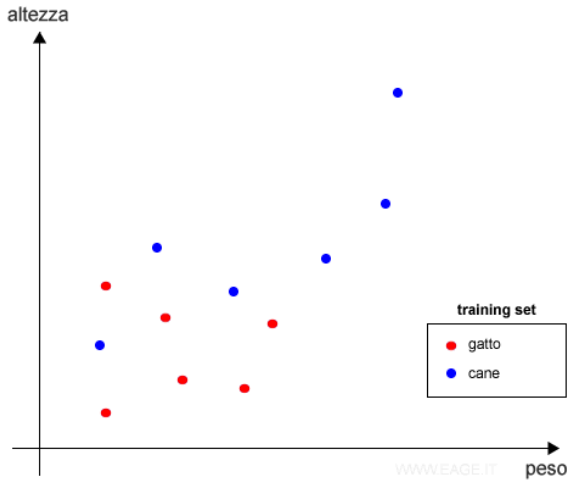
Naive Bayes è un algoritmo per risolvere problemi di classificazione e apprendimento automatico (machine learning) che utilizza il teorema di Bayes. Il teorema di Bayes permette di calcolare per ogni istanza la probabilità di appartenenza a una classe. Il teorema di Bayes calcola la probabilità di un evento A condizionata a un altro evento B.

P(A|B) = P(B|A) * P(A) / P(B)

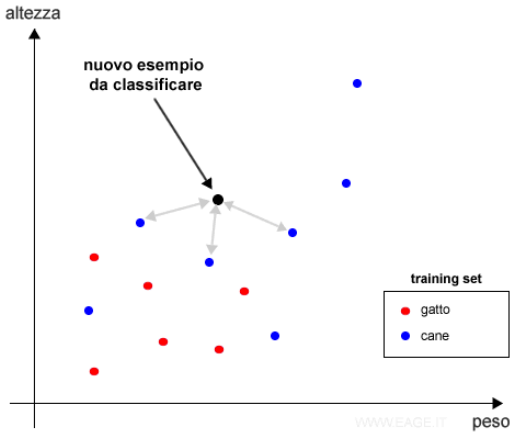
Si usa per calcolare la probabilità condizionata di un evento, in base alle informazioni disponibili su altri eventi correlati. Per praticità si assume l'indipendenza degli eventi (da cui Naive) e la distribuzione gaussiana di essi.

K-NEAREST NEIGHBORS

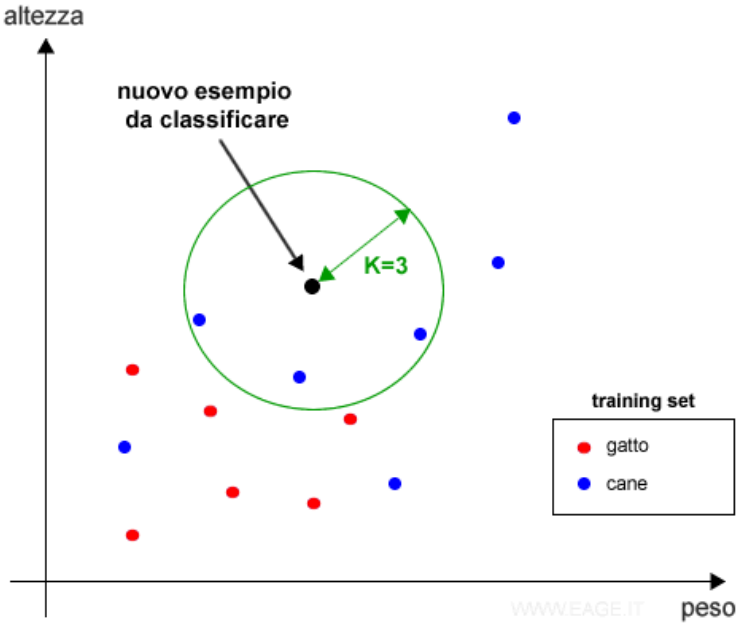
K-NN (K-Nearest Neighbors) è un algoritmo di riconoscimento dei pattern basato sulla vicinanza dei dati. Se un oggetto A ha caratteristiche simili a un oggetto B, probabilmente A appartiene alla stessa classe (categoria) di B. E' usato nei problemi di classificazione. Si tratta di un algoritmo di apprendimento supervised ML perché necessita di un gruppo di esempi di training già classificati (training set).



Quando l'algoritmo analizza i nuovi dati (non classificati) li classifica in base alla distanza rispetto agli esempi del training set.



Per il calcolo della distanza si possono adottare diversi criteri: la distanza euclidea o la distanza Manhattan per i dati numerici, distanza di Hamming per le stringhe. Se le caratteristiche sono soltanto due, l'algoritmo prende in considerazione i K esempi di training più vicini all'esempio da classificare.



La classe più frequente nei K esempi più vicini determina la classificazione del nuovo esempio.

Nell'algoritmo K-NN la configurazione del parametro K è fondamentale:

- Se K è troppo piccolo, l'esempio è assegnato in base ai pochi esempi più vicini. La classificazione è influenzata dal rumore nei dati.

- Se K è troppo grande, l'esempio viene classificato nella categoria più numerosa nella popolazione. Inoltre, la complessità computazionale dell'algoritmo aumenta perché il calcolo della distanza è un'operazione molto onerosa in termini di risorse e tempo.

RANDOM FOREST

Il modello random forest è composto da un certo numero di alberi di decisione (da cui la denominazione "foresta"), al fine di aumentare le performance di classificazione del modello ad albero di decisione introducendo della casualità. La sua caratteristica principale è di addestrare un numero predefinito di alberi di decisione sul dataset in modo che ciascun albero faccia la sua predizione per ciascun esempio; a questo punto si combinano le diverse predizioni effettuate per ciascun esempio a formare la predizione finale per tale esempio. Per ogni albero viene usato un sottoinsieme casuale delle feature e ad ogni nodo viene scelta la migliore feature in un insieme ristretto; perciò, ciascun albero opera su sottoinsiemi del training set contenenti ciascuno esempi estratti casualmente con rimpiazzo (bagging) dal dataset originale. Il random forest è tanto più efficace quanto più sono diversi gli alberi generati in quanto essi forniranno diverse predizioni e al momento del confronto si potrà stabilire la bontà della predizione finale effettuando la media fra le singole.

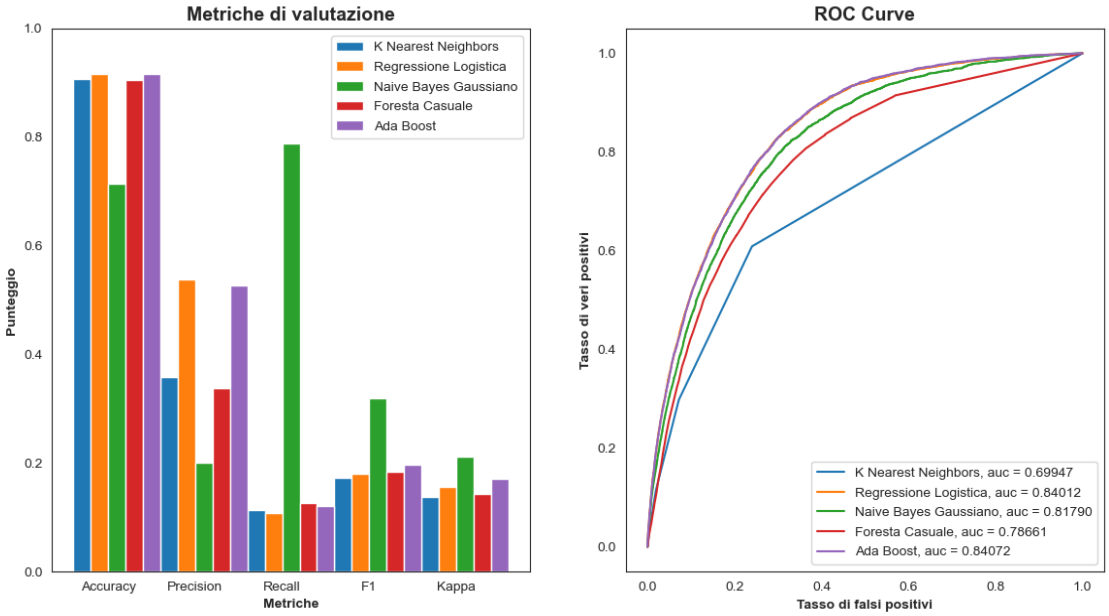
ADABOOST

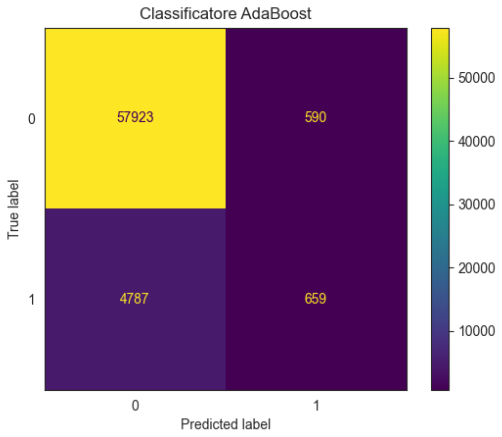
Il modello AdaBoost (adaptive boosting) consiste nell'utilizzare la tecnica di boosting, ossia impiegare un classificatore base sul dataset iterando il procedimento un prefissato numero di volte e avendo cura di ripetere il passaggio successivo sugli esempi su cui sono stati riscontrati degli errori di classificazione nel passaggio precedente, ordinando gli esempi in base all'errore; la predizione finale avverrà aggregando (con la media, la somma o con un'altra operazione ritenuta adatta allo scopo e al tipo di predizione) le predizioni dei singoli modelli prodotti nelle iterazioni susseguitesi. In particolare, nell'AdaBoost si usa la funzione d'errore $E(x)=e^{-y\,\hat{Y}(x)}$ e un classificatore di base ad albero di decisione.

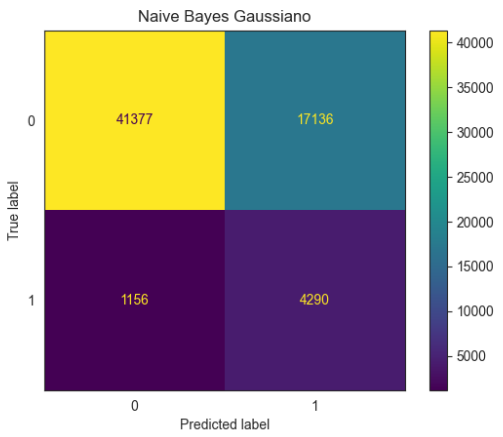
CONFRONTO DELLE TECNICHE UTILIZZATE

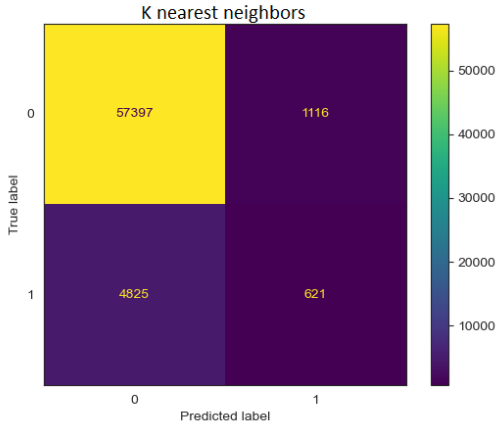
Di seguito un confronto delle performance e le rispettive matrici di confusione:

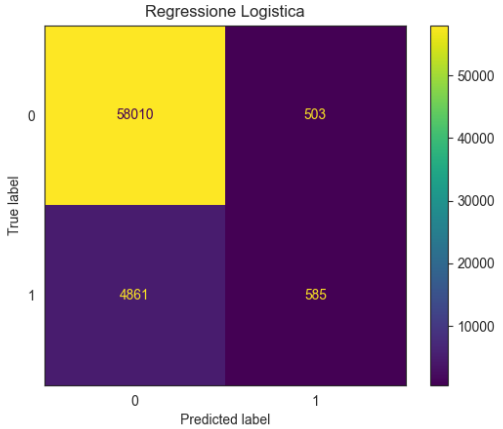
Confronto dei modelli

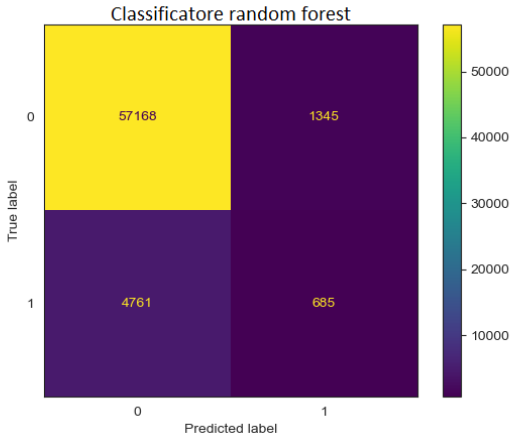



Classificatore AdaBoost
57923
590
4787
659
0
1
True label
Predicted label
50000
40000
30000
20000
10000


Naive Bayes Gaussiano
41377
17136
1156
4290
0
1
True label
Predicted label
40000
35000
30000
25000
20000
15000
10000
5000


K nearest neighbors
57397
1116
4825
621
0
1
True label
Predicted label
50000
40000
30000
20000
10000


Regressione Logistica
58010
503
4861
585
0
1
True label
Predicted label
50000
40000
30000
20000
10000


Classificatore random forest
57168
1345
4761
685
0
1
True label
Predicted label
50000
40000
30000
20000
10000