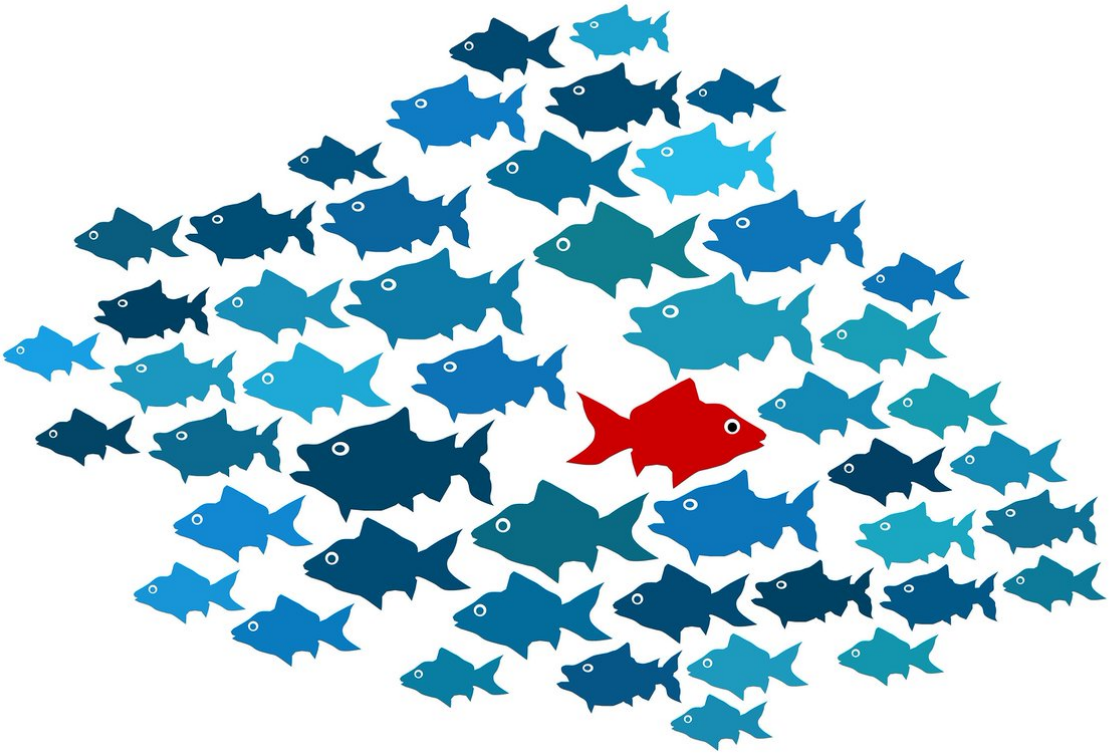
Università degli Studi di Padova
Corso di Ingegneria del Software 2021/2022



Login Warrior

Riconoscere e visualizzare i tentativi di accesso non autorizzati

ZUCCHETTI
IL SOFTWARE CHE CREA SUCCESSO

Oggetto dell'appalto

Le applicazioni cloud in tecnologia web stanno gradualmente sostituendo le applicazioni tradizionali "on premise".
La velocità della rete Internet, sia quella legata alla connessione via cavo telefonico che quella con tecnologie wireless, è ormai così elevata e così diffusa che gli utenti non percepiscono differenze significative nell'esperienza d'uso dei programmi.
La pandemia, con l'introduzione generalizzata dello "smart working", ha ulteriormente accelerato questa tendenza. Il lavoratore che si trova a volte in ufficio a volte a casa è più produttivo con procedure nel cloud.

Il vantaggio in termini di TCO (Total Cost of Ownership) è evidente: non occorre più gestire l'installazione e l'aggiornamento di server dedicati, di connessioni di rete specifiche, di sistemi "client": le applicazioni cloud funzionano con quello che oggi viene considerata la dotazione standard di ogni computer, telefono o tablet, cioè un browser e una connessione ad Internet veloce e con bassa latenza.

Ma il cloud non ha solo aspetti positivi, un lato negativo è la possibilità di subire attacchi da parte di script kiddies, hacker e criminali informatici. I sistemi esposti in rete infatti possono ricevere per loro natura connessioni da tutto il mondo, anche da robot o utenti umani che non sono esattamente quelli per cui si è rilasciata la propria procedura.
Gli attacchi sono così frequenti ed importanti da aver guadagnato la prima pagina dei media, basta ricordare l'attacco di quest'estate dove la Regione Lazio ha perso gran parte dei dati contenuti nei propri sistemi compresi quelli relativi alla pandemia Covid-19 e alle relative vaccinazioni.

E' quindi molto importante riuscire a capire se chi si presente all'ingresso delle nostre applicazioni è un malintenzionato o un utente legittimo. La porta di ingresso è la "login", dove possono essere messe in atto varie metodologie per identificare la persona: richiesta di utente e password, secondo fattore con chiavette o SMS che forniscono OTP (One Time Password), dati biometrici come impronta digitale o riconoscimento del volto.
Ma chiedere all'utente questi fattori può essere fastidioso, quindi riuscendo a distinguere con un certo grado di sicurezza se l'utente è legittimo ci permette di avere porte di ingresso semplificate che diventano complesse solo quando necessario.In poche parole se capisco che l'utente non è un criminale chiedo solo utente e password, se l'accesso è sospetto proseguo con OTP o altre protezioni.

La Zucchetti da tempo propone ai propri clienti procedure nel cloud sviluppate con tecnologia web. Ogni giorno accedono centinaia di migliaia di utenti alle procedure erogate dalla nostra server farm, e ogni giorno abbiamo centinaia di migliaia di tentativi di attacco.

I nostri sistemi sono già dotati di tutte le precauzioni descritte, ma al fine di migliorare l'esperienza d'uso dei nostri utenti vogliamo costruire di un sistema di analisi esplorativa dei dati ottenuti dalle login per poter costantemente studiare i pattern d'uso regolare e i pattern di attacco.

L’ analisi dei dati avviene sfruttando tecniche di statistica, machine learning e intelligenza artificiale. Possiamo distinguere quattro fasi fondamentali: l'acquisizione dei dati, la loro pulizia, l'analisi dei dati e infine l’interpretazione.
Nell'interpretazione dei dati distinguiamo due momenti: la prima esplorazione, chiamata EDA (Exploratory Data Analysis), e quindi la modellazione dei fenomeni trovati.

Il tipico strumento dell’EDA è la visualizzazione grafica del dato. Molto spesso un’immagine ed un grafico risultano molto più esplicativi che non una serie di numeri, infatti l’occhio umano percepisce senza difficoltà forme, gruppi e punti isolati.

Il presente capitolato ha per oggetto l’affidamento della fornitura per la realizzazione di un’applicazione di visualizzazione di dati di login a supporto della fase esplorativa del riconoscimento e distinzione di attività lecite e attività illecite.

Caratteristiche e Requisiti Obbligatori

L’applicazione “Login Warrior” di visualizzazione dei dati di login sarà sviluppata prevalentemente in tecnologia HTML/CSS/JavaScript utilizzando la libreria D3.js.

I dati da presentare verranno forniti dall’azienda in formato CSV. Oltre a questi dati potranno essere utilizzati anche altri dataset simili recuperati da archivi aperti.

L’applicazione dovrà presentare i dati con degli appositi grafici che aiutino a distinguere i casi sospetti dai casi leciti.

“Login Warrior” dovrà presentare almeno le seguenti visualizzazioni:

1. Scatter plot
2. Parallel coordinates
3. Force-directed graph
4. Sankey diagram

Queste 4 visualizzazioni sono disponibili nella libreria D3.js.

Lo “Scatter plot” permette di visualizzare dati su due dimensioni, ma utilizzando che la forma e la dimensione e colore del punto si arriva a 5. Se il grafico viene animato è possibile visualizzare anche una sesta dimensione. Riteniamo che l’occhio umano sia in grado di valutare casi particolari con questo grafico.
Come esempio riportiamo https://observablehq.com/@mbostock/the-wealth-health-of-nations dove si notano facilmente gli effetti delle due gruerre mondiali.

Il grafico “Parallel coordinates” permette di visualizzare più assi e con gli opportuni ordinamenti facilita l’identificazione di casi che non rientrano nella normalità.

Il grafico “Force-directed” si basa sul concetto di distanza tra i punti. Ogni elemento viene posizionato vicino ad altri con le stesse caratteristiche e allontanato da quelli diversi. Ci aspettiamo che le login lecite costituiscano degli aggregati uniformi e le illecite si allontanino da questi gruppi.

Il diagramma di “Sankey” ordina gli effetti di ogni dimensione esplorata, evidenziando gli effetti in cascata. Nel caso delle login informazioni come l’ora in cui avviene l’accesso o la locazione geografica possono essere elementi che distinguono in modo forte tra i gruppi leciti e non, quindi posizionandoli nelle parti più a sinistra ci aspettiamo che vengano distinti i due casi.

Requisiti Opzionali

Il tema proposto è un capitolo importante del Machine Learning, in particolare si tratta di un problema non supervisionato di “Anomaly Detection”. Le richieste obbligatorie sono relative alla prima fase esplorativa dei dati (EDA), dove puntiamo più alla visualizzazione per aiutare l’esperto umano a proseguire con altri strumenti di analisi.
Ma anche in questa fase gli algoritmi di Machine Learning possono aiutare a derivare dati più adatti alla visualizzazione.

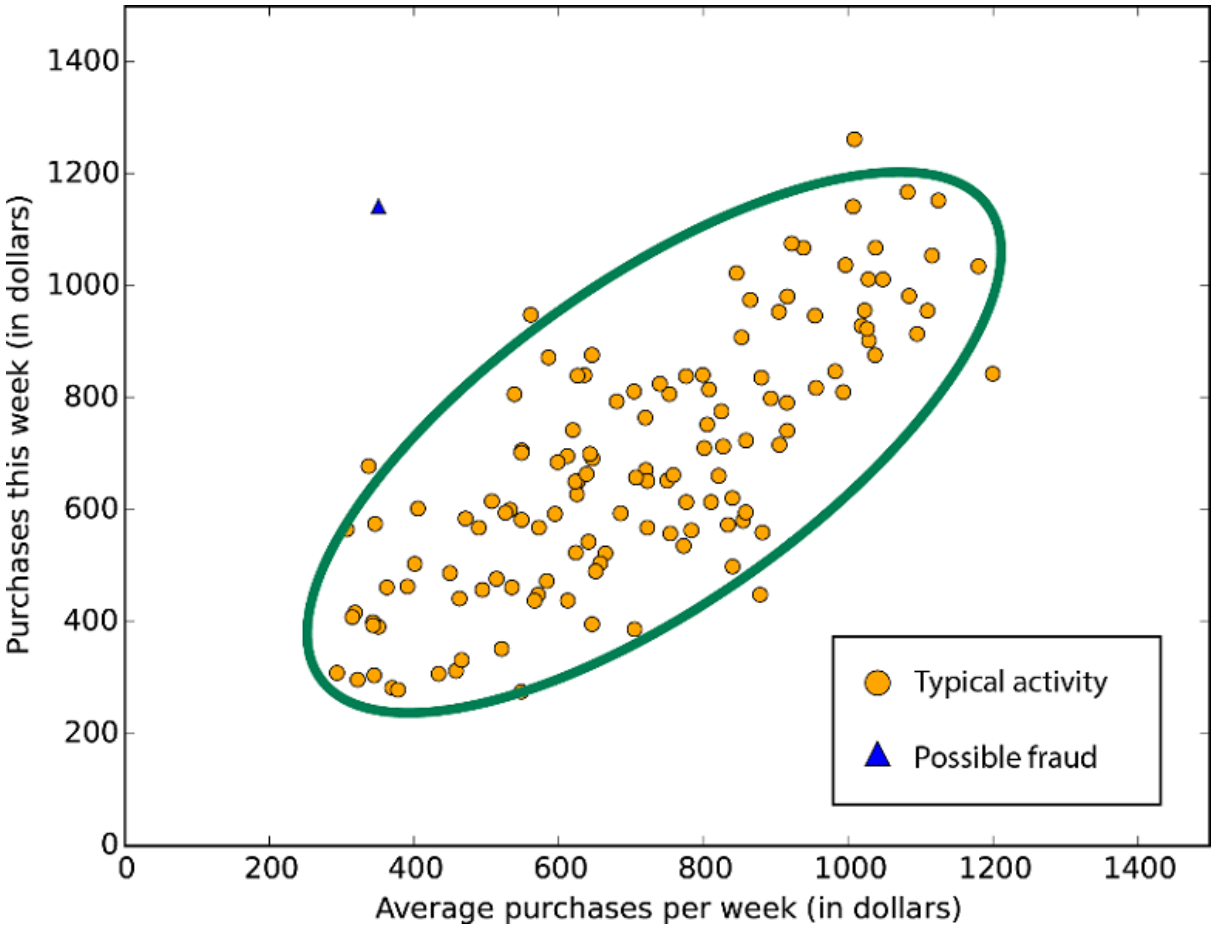
Sono quindi requisiti opzionali i seguenti:

1. Algoritmi di preparazione del dato per la visualizzazione. Per questo requisito opzionale l’azienda può mettere a disposizione librerie per gli algoritmi:
    a. t-SNE
    b. UMAP
    c. Self Organizing Map
2. Algoritmi di analisi del dato. Per questo requisito opzionale l’azienda può mettere a disposizione librerie per gli algoritmi:
    a. Isolation Forest
    b. One Class SVM
    c. Gaussian MCD
    d. Robust Gaussian
    e. HDBSCAN
3. Altri algoritmi di analisi del dati da librerie non fornite dall’azienda.
4. Altre visualizzazioni dalla libreria D3.js o da altre librerie.

Qualunque altra proposta del fornitore verrà valutata dall’azienda e eventualmente accettata come requisito opzionale se ritenuta valida per lo scopo del progetto.

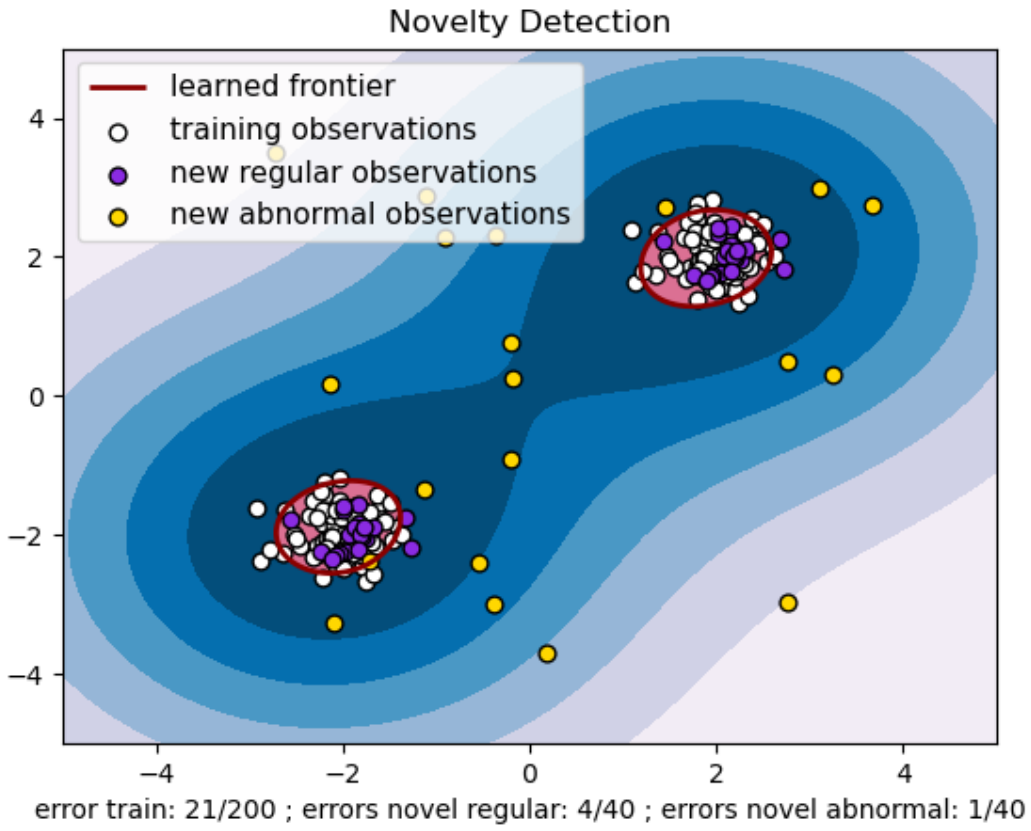
Suggerimenti

Il problema proposto è uno dei temi più importanti del Machine Learning non supervisionato. Come già spiegato la Zucchetti è interessata alla visualizzazione del dato prima delle analisi, ma parecchi spunti possono essere presi dalle tecniche utilizzate in questo campo.
Si possono trovare molte ispirazioni dalle tecniche di "anomaly detection" che sono di fatto la fase che segue quella di esplorazione dei dati, ma anche altre tecniche toccano argomenti simili.
Nell'immagine seguente vediamo una tipica visualizzazione per l'identificazione delle frodi. In questo caso i dati sono chiaramente raggruppati con pochi casi fuori dalla normalità. Il punto completamente isolato è quello incriminato.

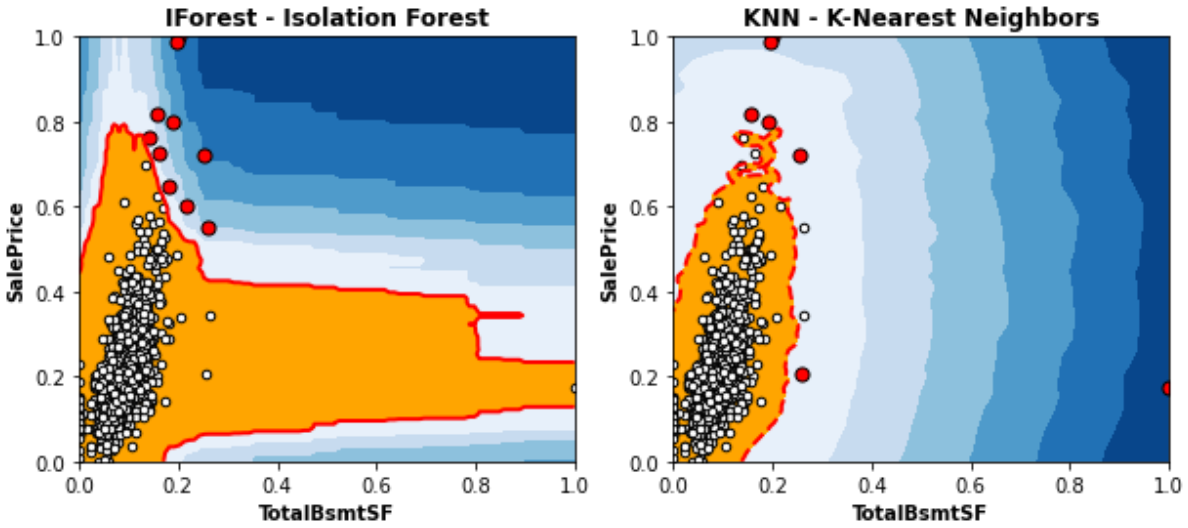


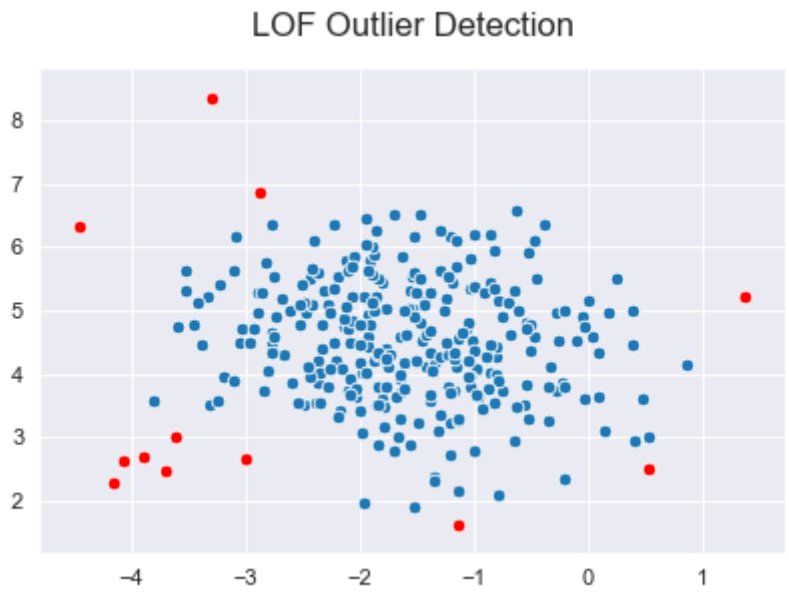
Con la sola visualizzazione l'occhio umano avrebbe percepito il caso particolare, con algoritmi di Machine Learning si possono attribuire i colori ed il giudizio di caso di frode.

Nell'immagine precedente l'analisi è stata fatta con approssimazioni gaussiane, nella seguente con Supper Vector Machines ad una sola classe.

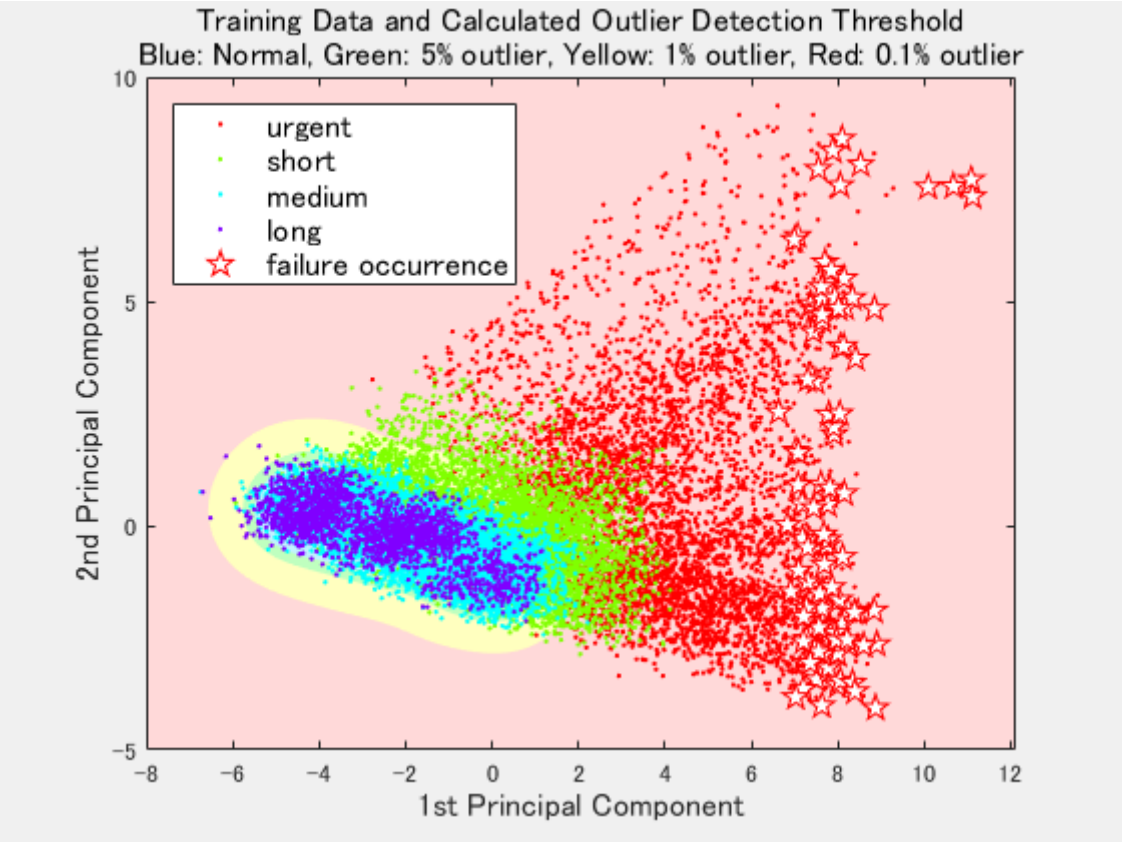


Altri temi simili del ML che possono fornire spunti sono la"Domain Approximation", la "Novelty Detection" e l'"Outlier detection".





Non sempre i dati si presentano come negli esempi, l'immagine che segue mostra un caso in cui ad un nucleo corposo è associata una "coda" molto larga e qui diventa più difficile convincersi di quali possono essere i casi sospetti.



ZUCCHETTI
IL SOFTWARE CHE CREA SUCCESSO

Variazioni ai requisiti

In corso d'opera non sarà possibile variare/modificare i requisiti minimi (obbligatori per accettare il prodotto). Sarà invece possibile variare i requisiti opzionali, in quanto saranno i gruppi vincitori dell'appalto a modificarli / eliminarli / aggiungerli.

Documentazione

Il progetto dovrà essere supportato dalla documentazione minima richiesta per il corso di Ingegneria del software e dovrà essere fornito un manuale per l'utilizzo ed un manuale per chiunque voglia estendere l'applicazione.

Garanzia e Manutenzione

L'azienda Zucchetti SPA è interessata a questo progetto come dimostrazione della fattibilità dell'obiettivo utilizzando le tecnologie web. Costituirà titolo preferenziale nella valutazione delle proposte la pubblicazione del progetto sul sito “github.com” o altri repository pubblici, in conformità con i relativi requisiti di natura open-source, per favorire la continuità del prodotto risultante.

Rinvio

Per tutto quanto non previsto nel presente capitolato, sono applicabili le disposizioni contenute nelle leggi e nei collegati per la gestione degli appalti pubblici.