# Riassunto Tesi Magistrale: *Predizione di difettosità nello sviluppo software attraverso machine learning*

Jacopo Nasi [s255320]

21 febbraio 2020

## 1 Introduzione

Lo sviluppo software non si presenta molto differente dallo sviluppo di qualsiasi altro prodotto, dopo una fase iniziale di progettazione lo sviluppo del codice può avere inizio, durante esso emergeranno sistematicamente dei problemi che dovranno essere risolti prima della consegna della versione finale.
Ogni progetto software è costituito da diversi commit per giorno, ognugno di essi contiene innumerevoli informazioni le quali possono essere utilizzate per analisi statistiche. La predizione della difettosità può migliorare enormemente il processo di sviluppo, allocando un corretto numero di sviluppatori per risolvere le problematiche e riducendo quindi le tempistiche per la correzione. Anche il machine learning può essere utilizzato per la predizione dei difetti.

## 2 Data

L'applicazione di modelli di apprendimento automatico necessità di dati relativi ad esperienze pregresse al fine di utilizzarli per la fase di allenamento dei modelli che verranno successivamente utilizzati per la predizione.
L'accesso e la disponibilità dei dati non è sufficiente a garantire dei buoni mo-

delli, è necessario prima pulire, ordinare ed eventualmente aggregarli. Questa fase viene chiamata pre-processamento e viene utilizzata per migliorare i dati prima dell'allenamento dei modelli.
Il lavoro di tesi è basato su un base dati chiamata SEOSS33, una collezione di problematiche, commenti e modifiche di 33 progetti software open source. I dati sono organizzati attorno alle issue, concetto base dello processo di sviluppo. Direttamente correlati alle issue vi sono innumerevoli altre informazioni come commenti, dati temporali, numeri di versione e altro.
La pre-elaborazione di questi dati prevede la rimozione di tutto ciò che non verrà utilizzato, l'aggiunta della seniority degli sviluppatori, un lista di parole dei componenti modificati, la versione ed altre informazioni di carattere più matematico. Infine verrà sostiuità l'etichetta testuale della priorità con un valore numerico corrispondende al valor medio della distribuzione della durata di quella etichetta, questo valore prenderà il nome di severity. I dati verranno poi aggregati per settimana calcolando la differenza tra la severity delle issue aperte e quelle chiuse, questo valore rappresenterà l'obbiettivo della predizione. La generazione di questi dati verrà fatta per differenti finestre temporali, da 1 a 52 settimane.

# 3 Forecasting

Una volta che i dati per l'allenamento sono stati generati gli algoritmi statistici possono essere applicati. L'obbiettivo è quello di prevedere la futura difettosità in uno spazio temporale da 1 a 52 settimane.
Sono stati scelti tre differenti modelli per la previsione: Random Forest, Gradient Boosting e Reti Neurali, verranno inoltre applicati modelli predefiniti di alcuni framework per la valutazione dei risultati. La fase di allenamento è stata applicata utilizzando tre differenti metodologie: la prima allena e predice utilizzando lo stesso filone di dati, la seconda, cross-version, prevede che il modello venga allenato su dati relativi ad alcune versione del progetto per poi effettuare la predizione sulle successive, la terza, cross-project, allena il modello con dati

relativi ad un progetto per poi prevedere l'andamento di uno differente.
Tutti le tipologie ottengono dei buoni risultati, il migliore è quello cross-project che riesce ad ottenere una precisione maggiore del 90% fino a quattro settimane e comunque maggiore del 70% fino a 20 settimane.