

Analisi dei dati per problemi di medicina legale

Alessandro Beranti - 855489

Febbraio 2020

Il tirocinio è stato svolto internamente all'Università. Lo scopo del tirocinio era quello di applicare il paradigma del Machine Learning per riuscire a classificare che tipo di mezzo avesse investito una persona deceduta a causa dell'impatto con un veicolo. Il lavoro si è inizialmente concentrato sull'apprendere le basi del Machine Learning, per poi passare all'applicazione pratica usando un dataset fornito da medici legali.

Le ragioni che mi hanno spinto a svolgere un tirocinio su questo argomento, e di conseguenza una tesi, sono principalmente due: la necessità da parte di medici legali di conoscere il prima possibile che tipo di veicolo, vettura o mezzo pesante, avesse investito il malcapitato, ma soprattutto sono sempre stato affascinato dallo studio dei dati e dal Machine Learning in generale.

Il codice per eseguire gli esperimenti descritti nella tesi è stato scritto usando il linguaggio *Python* 3.7.4, unitamente ai seguenti pacchetti:

- *skikit-learn*, versione 0.22.1
- *numpy*, versione 1.18.1
- *pandas*, versione 0.25.3

Ho inoltre utilizzato *Jupyter* come ambiente di lavoro.

La tesi si compone di tre capitoli.

- Capitolo 1: viene introdotto il concetto di Machine Learning, in particolare viene spiegato come funziona e quali sono gli approcci esistenti. Successivamente vengono trattati i modelli che ho usato durante il tirocinio, descrivendone gli specifici dettagli.
- Capitolo 2: viene inizialmente spiegato il dataset che ho usato per prendere conoscenza e familiarità con le tecniche e gli algoritmi del Machine Learning. Successivamente viene illustrato il dataset che ho utilizzato durante il tirocinio, raccolto dai medici legali, ovvero una raccolta di dati legati a vittime di investimento, unitamente a una descrizione delle lesioni che hanno riportato. Vengono poi spiegate

due tecniche usate per ridurre la dimensionalità, pratica utile nel processo di Machine Learning per migliorare le performance degli algoritmi, quando questi sono utilizzati su dataset di grandi dimensioni.

- Capitolo 3: si entra nello specifico del processo che ho usato per passare dal dataset a dei classificatori che predicono il tipo di veicolo che ha causato l'incidente, valutandoli in base all'accuratezza ottenuta e al variare dei diversi algoritmi di apprendimento discussi al Capitolo 1. Il lavoro si chiude con un'analisi dei risultati e un breve chiarimento delle tecnologie impiegate.

Il risultato raggiunto è da considerarsi molto buono in quanto personalmente ho imparato il processo da applicare a un dataset per produrre modelli che ben si adattano ai dati. Ho inoltre imparato come si imposta un lavoro di questo genere che ha richiesto mesi di tempo. In generale l'accuratezza ottenuta dai modelli non è altissima ma ho tenuto conto solo dei livelli di dettaglio meno specifici, come descritto nel Capitolo 2. Un possibile sviluppo futuro potrebbe essere quello di considerare tutte le 350 dimensioni del dataset e non solo le 30 usate durante il tirocinio.