Scuola di Scienze Matematiche, Fisiche e Naturali
Corso di Laurea in Informatica

Tesi di Laurea

TITOLO ITALIANO

TITOLO INGLESE

ALESSANDRO PISCOPO

Relatore: *Tommaso Zoppi*
Correlatore: *Correlatore*

Anno Accademico 2022-2023

INDICE

ELENCO DELLE FIGURE

"Inserire citazione"
— Inserire autore citazione

1

INTRODUZIONE

Nell'era in cui viviamo l'analisi dei dati riveste un ruolo cruciale in molti settori della nostra società. In questo contesto il Machine Learning si è dimostrato un potente strumento per estrarre informazioni utili e conoscenza da dati complessi e voluminosi. Il Machine Learning (ML) è un sottoinsieme dell'intelligenza artificiale (AI) che si occupa di creare sistemi che apprendono e migliorano le proprie performance in base ai dati che utilizzano.

Esistono principalmente due categorie di modelli di apprendimento automatico: machine learning supervisionato e machine learning non supervisionato. L'apprendimento supervisionato utilizza set di dati etichettati per addestrare gli algoritmi per classificare o prevedere i risultati in modo accurato. L'apprendimento non supervisionato, utilizza gli algoritmi di machine learning per analizzare e organizzare in cluster i set di dati senza etichette.

Nel presente lavoro di Tesi sono stati utilizzati dei modelli di machine learning supervisionato, avendo a disposizione dei set di dati etichettati per addestrare gli algoritmi. In particolare lo scopo del lavoro di Tesi è stato analizzare due approcci diversi all'apprendimento automatico: un approccio classico e un approccio time series. Una time series può essere definita come un insieme di osservazioni ordinate rispetto al tempo. La differenza sostanziale tra i due approcci è che con un approccio time series si hanno informazioni non solo sull'istante di tempo corrente, ma anche su un numero di istanti di tempo precedenti scelto arbitrariamente. Il fine ultimo del presente lavoro di Tesi è stato quello di comparare le performance di 4 modelli in condizioni diverse: Logistic Regression, Linear Discriminant Analysis, Random Forest, XGBoost. Le condizioni diverse sopracitate sono state date dall'addestramento dei modelli su set di dati differenti in base all'approccio utilizzato, classico o time series. L'ipotesi che abbiamo voluto verificare è quella che un approccio time series migliori le performance dei modelli rispetto ad un approccio

classico, avendo a dispozione informazioni su una finestra di istanti di tempo precedenti e quindi più dati disponibili durante l'apprendimento.

Il lavoro è organizzato nel seguente modo:

- Capitolo 2: fornisce una panoramica su Machine Learning, Anomaly Detection e Time Series
- Capitolo 3: descrive le metodologie e le strategie utilizzate durante gli esperimenti
- Capitolo 4: sono riportati gli esperimenti realizzati
- Capitolo 5: analisi dei risultati ottenuti
- Capitolo 6: conclusioni della Tesi

2

STATO DELL'ARTE

2.1 MACHINE LEARNING

Il Machine Learning (ML) è una branca dell'intelligenza artificiale che si è sviluppata negli'ultimi decenni del XX secolo. Nel campo dell'informatica, l'apprendimento automatico è una variante alla programmazione tradizionale nella quale in una macchina si predispone l'abilità di apprendere dai dati in maniera autonoma, senza istruzioni esplicite. I metodi principali per l'apprendimento automatico sono due:

- **Apprendimento Supervisionato**: vengono utilizzati dati etichettati per effettuare il training del modello di ML. Viene quindi fornita la corrispondenza tra input e output durante la fase di training.
- **Apprendimento Non Supervisionato**: vengono utilizzati dati non etichettati durante l'addestramento. Non viene resa esplicita quindi nessuna relazione tra input e output, sarà l'algoritmo ad estrarre le informazioni necessarie a classificare o predire i risultati attesi.

Nel nostro caso, come già menzionato nell'Introduzione, abbiamo utilizzato l'Apprendimento Supervisionato avendo a disposizione dei set di dati etichettati. Esistono principalmente due tipi di Apprendimento Supervisionato:

- **Classificazione**: un algoritmo (classificatore) è addestrato a classificare i dati di input su variabili discrete. Durante l'addestramento, gli algoritmi ricevono dati di input con un'etichetta "classe" e dovranno essere in grado, una volta addestrati, di restituire la classe di appartenenza di nuovi input forniti al modello.
- **Regressione**: un algoritmo (regressore) deve individuare una relazione funzionale tra i parametri di input e l'output. Il valore di

output non è discreto come nella classificazione, ma è una funzione dei parametri di input.

2.1.1 *Algoritmi utilizzati*

Gli algortitmi utilizzati nel presente lavoro di tesi sono tutti classificatori, avendo preso come caso di studio un problema di classificazione. Sono stati utilizzati in totale 4 algoritmi di classificazione, 2 per ciascuna classe:

- **Modelli Lineari**: Logistic Regression, Linear Discriminant Analysis
- **Modelli basati su Alberi Decisionali**: Random Forest, XGBoost

Modelli Lineari

Modelli basati su Alberi Decisionali

2.1.2 *Metriche di valutazione (+ speed detection score)*

2.2 ANOMALY DETECTION

2.3 TIME SERIES

3

METODOLOGIE

3.1 DESCRIZIONE DATASET

3.1.1 *Dataset Università*

3.1.2 *Dataset All3*

3.2 PREPROCESSING

3.3 STRATEGIE

3.3.1 *Approccio Classico*

3.3.2 *Approccio Time Series con Differenze*

3.3.3 *Approccio Time Series con Media Mobile*

3.4 WINDOW E SHUFFLE

3.4.1 *Window*

3.4.2 *Shuffle*

4

ESPERIMENTI

4.1 WINDOW 5 CON SHUFFLE

4.2 WINDOW 5 SENZA SHUFFLE

4.3 WINDOW 4 SENZA SHUFFLE

4.4 WINDOW 3 SENZA SHUFFLE

4.5 WINDOW 2 SENZA SHUFFLE

5

ANALISI RISULTATI

NULL

6

CONCLUSIONI

NULL

BIBLIOGRAFIA

[1] Autore - *titolo*

[2] Autore - *Titolo* - altre informazioni