Immagine che contiene testo, moneta

Descrizione generata automaticamente

|  |
| --- |
| Immagine che contiene testo  Descrizione generata automaticamente  Scuola Politecnica e delle Scienze di Base  Corso di Laurea Magistrale in Ingegneria Informatica |

|  |
| --- |
| **Big Data Engineering**  Anno Accademico 2021/2022 |

Professore:

**Sperlì Giancarlo**

Studenti:

**Lamboglia Anna M63001219**

**Perrot Raffaella M63001135**

**Ricchiuti Fausto M63001144**

Sommario

[Big Data Engineering I](#_Toc105066503)

[Capitolo 1 3](#_Toc105066504)

[1.1 Descrizione della challenge 3](#_Toc105066505)

[1.2 Dataset utilizzato 3](#_Toc105066506)

[1.3 Strumenti Utilizzati 5](#_Toc105066507)

[1.3.1 MongoDB 5](#_Toc105066508)

[1.3.2 Python 5](#_Toc105066509)

[1.3.3 Spark 6](#_Toc105066510)

[1.3.4 Excel 7](#_Toc105066511)

[Capitolo 2 8](#_Toc105066512)

[2.1 Trend delle carriere universitarie 8](#_Toc105066513)

[Capitolo 3 9](#_Toc105066514)

[3.1 ANALISI DEI fuori sede 9](#_Toc105066515)

[Capitolo 4 10](#_Toc105066516)

[4.1 predizione dei trend di iscrizione 10](#_Toc105066517)

[Conclusioni 11](#_Toc105066518)

# Capitolo 1

### Descrizione della challenge

L’obiettivo della challenge è quello di effettuare uno studio sui trend nelle università in Italia.

In particolare, ci si è focalizzati su tre punti:

* Analizzare e descrivere i trend delle carriere universitarie. Sono state estratte informazioni significative e interessanti utilizzando dati demografici, ambiti di studio, background personale e opportunità di lavoro;
* Analizzare e descrivere i trend del decentramento attraverso le scelte delle sedi universitarie. Sono state analizzate le scelte degli studenti fuorisede;
* Prevedere il trend di iscrizione in base ai dati storici e fattori come l’istruzione, la demografia o altri fattori che si possono scoprire.

### Dataset utilizzato

Al fine di risolvere i tre punti sopraelencati, è stata svolta una ricerca per poter trovare dei dati che siano quanto più consoni possibile alle analisi da effettuare.

In particolare, sono stati scaricati quattro file *.csv* messi a disposizione dal MIUR mediante questo link: <http://dati.ustat.miur.it/organization/ace58834-5a0b-40f6-9b0e-ed6c34ea8de0?tags=Universit%C3%A0&tags=Studenti>

e dall’ISTAT mediante questo link:

<http://dati.istat.it/Index.aspx?DataSetCode=INDUNIV> .

I quattro file in questione sono i seguenti ed è possibile trovarli all’interno della seguente repository github (INSERIRE IL LINK) :

* *Immatricolati\_Anno\_Gruppo.csv*. Il file è composto da cinque colonne*:*

1. *AnnoA,* l’anno accademico;
2. *GruppoNOME*, nella qualesono definite le tipologie di università;
3. *Sesso;*
4. *Imm*, nella quale sono indicati il numero di immatricolati;
5. *Note* per definire delle note particolari.

Tale file permette di comprendere il numero di immatricolati, dato l’anno, il tipo di università e il sesso.

Il file ha una dimensione pari a 96 Kb e presenta 360 righe;

* *Iscritti\_Anno\_Gruppo.csv.* Il file è composto da cinque colonne*:*

1. *\_id;*
2. *AnnoA,* l’anno accademico;
3. *GruppoNOME*, nella quale sono definite le tipologie di università;
4. *Sesso;*
5. *Isc*, nella quale sono indicati il numero di iscritti.

Tale file permette di comprendere il numero di iscritti di un determinato gruppo universitario, dato l’anno e il sesso.

Il file ha una dimensione pari a 89 Kb e presenta 330 righe;

* *Immatricolati\_Residenza\_Ateneo.sv*. Il file è composto da otto colonne*:*

1. *\_id;*
2. *AnnoA,* l’anno accademico;
3. *AteneoCOD,* il codice associato ad un determinato ateneo;
4. *SedeP*, ilcodice relativo ad una determinata provincia;
5. *ResidenzaR,* la regione di residenza;
6. *ResidenzaP,* la provincia di residenza;
7. *Imm*, nella quale sono indicatiil numero di immatricolati.

Tale file permette di comprendere la residenza degli immatricolati associati ad una determinata università.

Il file ha una dimensione pari a 1.805 Kb e presenta 32000 righe;

* *Popolazione Province.csv*. Il file è composto da tredici colonne:

1. *ITTER107*, id del territorio;
2. *Territorio*, la provincia o la regione;
3. *TIPO\_DATO15*, il codice del periodo di riferimento;
4. *Tipo dato*, popolazione del primo gennaio;
5. *ETA1*, classe di età (totale);
6. *Classe di età*, totale;
7. *SEXISTAT1*, il codice per identificare il sesso;
8. *Sesso*;
9. *CITTADINANZA*, la sigla della nazionalità;
10. *Cittadinanza*, la sigla della nazionalità;
11. *TIME*, l’anno di riferimento;
12. *Seleziona periodo*, l’anno di riferimento;
13. *Value*, la popolazione.

Il file ha una dimensione pari a 502 Kb e presenta 4051 righe.

### Strumenti Utilizzati

Per effettuare le analisi sono stati utilizzati i seguenti strumenti:

* *Mongo DB;*
* *Python;*
* *Pyspark;*
* *Excel.*

#### 1.3.1 MongoDB

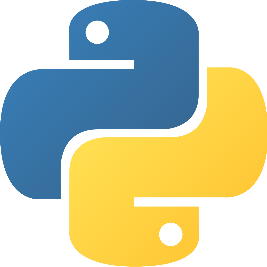
Per effetuare l’analisi del dataset è stato deciso di utilizzare *MongoDB*. Esso è un DBMS non relazionale sviluppato in C++, open-source, document-oriented e scalabile. Esso è stato realizzato in maniera tale da avere alte prestazioni, sia in lettura che in scrittura. Tra i vantaggi, si osserva che le letture più consistenti possono essere distribuite in più server replicati e le interrogazioni sono pi`u semplici e veloci grazie all’approccio ai documenti che rende possibile la rappresentazione di relazioni gerarchiche complesse attraverso documenti nidificati e array. Le caratteristiche principali di MongoDB sono le seguenti:

* **Database document-oriented:** i dati vengono archiviati sotto forma di documenti in formato JSON;
* **Supporto completo agli indici**: indicizzazione di qualsiasi attributo;
* **Replicazione**: facilità nella replicazione dei dati attraverso la rete e alta scalabilità;
* **Sharding**: scalabilità orizzontale senza compromettere nessuna funzionalità.



#### 1.3.2 Python

Python è un linguaggio di programmazione di "alto livello", orientato a oggetti, adatto, tra gli altri usi, a sviluppare applicazioni distribuite, scripting, computazione numerica e system testing.



#### 1.3.3 Spark

Per poter lavorare sui dati ed applicare tecniche di Machine Learning è stata utilizzato *Pyspark*, ossia una libreria che consente l’utilizzo in Python di Apache Spark.

*Apache Spark* è un framework di elaborazione parallela open source, che supporta l’elaborazione in memoria al fine di migliorare le prestazioni delle applicazioni che analizzano Big Data.

Le caratteristiche principali di Spark sono le seguenti:

* **Fast processing:** Spark contiene Resilient Distributed Dataset (RDD), che consentono di risparmiare tempo necessario per le operazioni di lettura e scrittura, permettendo di ottenere un tempo di elaborazione più veloce rispetto ad Hadoop, di uno o due ordini di grandezza;
* **In-Memory computing:** in Spark i dati sono salvati nella RAM (Random Access Memory), in questo modo è possibile accedere a quest’ultimi più velocemente ed accelerare la velocità della fase di analytics;
* **Flexible**: Spark supporta diversi linguaggi e consente agli sviluppatori di realizzare applicazioni in Java, Scala, R o Python;
* **Fault tolerance.** Spark contiene Resilient Distributed Dataset (RDD), progettati per gestire il guasto di un qualsiasi nodo worker del cluster. In questo modo, si garantisce che la perdita di dati sia pressocché pari a zero.
* **Better Analytics**: Spark ha a disposizione una ricca serie di funzionalità, come la possibilità di effettuare query SQL oppure applicare algoritmi di machine learning, che consentono di migliorare le performance in fase di analisi.

#### 1.3.4 Excel

*Microsoft Excel* è un programma prodotto da Microsoft, dedicato alla produzione ed alla gestione di fogli elettronici. È parte della suite di software di produttività personale Microsoft Office, ed è disponibile per i sistemi operativi Windows e macOS. È il programma per la produzione e gestione di fogli elettronici più utilizzato.



# Capitolo 2

### Trend delle carriere universitarie

Per risolvere il primo punto definito dalla challenge è stato deciso di utilizzare il database NOSQL MongoDB. In particolare, per poter analizzare i dati delle carriere sono stati utilizzati i 2 file Immatricolati\_Anno\_Gruppo.csv, Iscritti\_Anno\_Gruppo.csv.

Innanzitutto è stato creato un DB contente una collection per ogni file .csv in modo da poter effettuare delle projection dei dati di interesse.

IscrizioniCollection = Database['Iscrizioni']

ImmatricolazioniCollection = Database['Immatricolazioni']

Innanzitutto è stato creato un DB contente una collection per ogni file .csv in modo da poter effettuare delle projection dei dati di interesse

# Capitolo 3

### ANALISI DEI fuori sede

# Capitolo 4

### 4.1 predizione dei trend di iscrizione

Per poter portare a termine questo task è stata utilizzata la libreria python Pyspark, in particolare è stato sfruttato il componente Mllib che permette di effettuare operazioni di Machine Learning.

L’obiettivo è creare un modello in grado di predirre l’andamento del numero di iscrizioni nelle varie università italiane.

Prima di tutto però è necessario definire il dataset da quale partire, infatti sono stati effettuati vari ragionamenti per capire quali potessero essere delle features di interesse da poter utilizzare per addestrare il modello predittivo.

**AnnoAccademico, Ateneo, PopolazioneP, AteneiP, PopolazioneR, AteneiR, Fuorisede, Immatricolati**

# Conclusioni