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

[Capitolo 1 3](#_Toc105060056)

[1.1 Descrizione della challenge 3](#_Toc105060057)

[1.2 Dataset utilizzato 3](#_Toc105060058)

[1.2.1 Pulizia del dataset 5](#_Toc105060059)

[Capitolo 2 6](#_Toc105060060)

[2.1 Modellazione dei dati 6](#_Toc105060061)

[Capitolo 3 7](#_Toc105060062)

[3.1 Modello Predittivo 7](#_Toc105060063)

[Capitolo 4 8](#_Toc105060064)

[4.1 Modellazione dei dati 8](#_Toc105060065)

[Conclusioni 9](#_Toc105060066)

# 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. Csv*. 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 determinatata 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.

#### 1.2.1 Pulizia del dataset

Per poter lavorare con questo dataset è stato deciso di utilizzare il database NOSQL MongoDB (scrivere cos’è Mongo). In particolare è stato creato un DB contente una collection per ogni file .csv in modo da poter effettuare delle projection dei dati di interesse.

File utilizzati:

1. Immatricolazioni: File Immatricolati Anno Gruppo 96 Kb
2. Iscrizioni: File Iscritti Anno Gruppo 89 Kb
3. Collection Fuorisede: Immatricolati Residenza Ateneo 1.805 Kb
4. Istat: Popolazione province 502 kb

# Capitolo 2

### Modellazione dei dati

# Capitolo 3

### Modello Predittivo

# Capitolo 4

### 4.1 Modellazione dei dati

# Conclusioni