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: Homework 1 I](#_Toc104910622)

[Capitolo 1 3](#_Toc104910623)

[1.1 Problema 3](#_Toc104910624)

[1.2 DataSet 3](#_Toc104910625)

[Capitolo 2 4](#_Toc104910626)

[2.1 Modellazione dei dati 4](#_Toc104910627)

[Capitolo 3 5](#_Toc104910628)

[3.1 Modello Predittivo 5](#_Toc104910629)

[Queries and Related Output 6](#_Toc104910630)

## 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.

### Problema

### Dataset utilizzato

Per poter risolvere i tre punti definiti nel paragrafo precedente è stato fatto un lavoro di ricerca per poter trovare dei dati che siano quanto più consoni alle analisi che si vuole effettuare.

In particolare sono stati scaricati 4 file .csv messi a disposizione dal MIUR tramite questo link (inserire il link) e dall’ISTAT tramite questo link (INSERIRE IL LINK).

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 *: AnnoA* ossia l’anno accademico, *GruppoNOME* dove sono definite le tipologie di università, *Sesso, Imm* dove sono indicati il numero di immatricolati e la colonna *Note* per definire delle note particolari.

Dunque, questo file permette di comprendere il numero di immatricolati dato l’anno, tipo di università e 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 *: \_id, AnnoA* ossia l’anno accademico, *GruppoNOME* dove sono definite le tipologie di università, *Sesso, Isc* dove sono indicati il numero di iscritti.

Tale file consente 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 8 colonne *: \_id, AnnoA* ossia l’anno accademico, *AteneoCOD* ossia il codice associato ad un determinato ateneo, *SedeP* codice relativo ad una determinatat provincia*, ResidenzaR* è la regione di residenza, *ResidenzaP* è la provincia di residenza ed *Imm* indicata il numero di immatricolati.

Con questo file è possibile comprendere la residenza degli immatricolati 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 ITTER107, Territorio, TIPO\_DATO15, Tipo dato, ETA1, Classe di età, SEXISTAT1, Sesso, CITTADINANZA, Cittadinanza, TIME, Seleziona periodo, Value

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

### 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