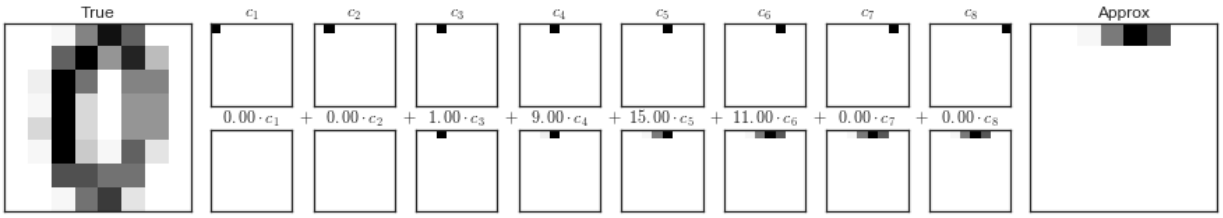
Università di Palermo
Corso di Laurea in Informatica
Esame di "Fondamenti di Data Science"
Prova pratica - Appello del 26 Giugno 2023

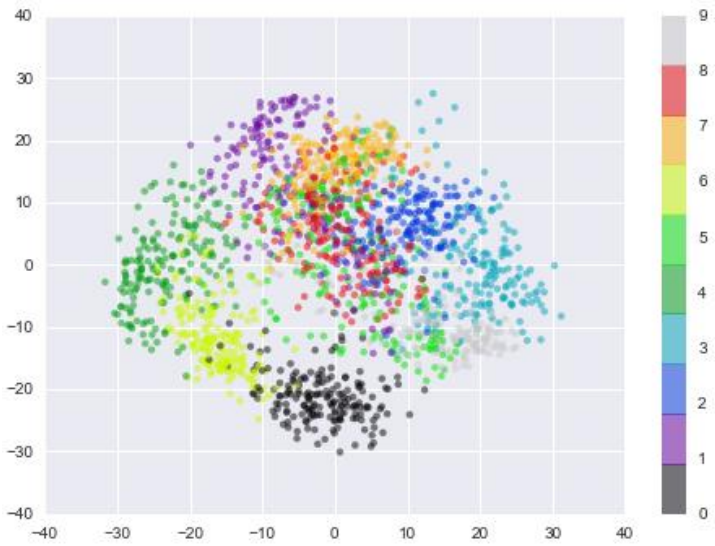
Sia dato il dataset `digits.csv` che contiene le immagini di numeri scritti a mano ($8 \times 8 = 64$ pixel), inseriti come righe del file, in cui l'ultima colonna di ogni riga rappresenta il target.



https://www.educative.io/answers/what-is-datasetsloaddigits-in-sklearn

Realizzare un **Progetto di Data Analysis** che:

- Prevede l'utilizzo della PCA per proiettare le 64 dimensioni sulle prime due componenti principali, questi punti sono la proiezione di ciascun punto dati lungo le direzioni con la varianza maggiore (riprodurre la seguente immagine).
- Prevede la clusterizzazione dei punti ($k = 10$) tramite tecnica K- Means (basato sul centroide).
- Ripete la clusterizzazione sulle proiezioni sulle prime N componenti principali, con N che varia tra 2 e 6.
- Per ogni clusterizzazione produce una matrice di confusione che metta a confronto l'accuracy nella classificazione al variare del numero di componenti principali scelte.



Output:

- **Relazione di progetto** con la descrizione dell'analisi progettuale.
  Questa deve includere la descrizione delle features, l'intera pipeline di analysis, e risultati ottenuti.
  Il codice deve essere consegnato su file separati, la relazione può comunque riportare eventuali funzioni di importanza e le relative descrizioni.
- **Codice sorgente del progetto**.

Note: il progetto va inviato all'indirizzo ***domenico.garlisi@unipa.it***, si suggerisce di specificare nell'oggetto ***NOME_GRUPPO-ID-PROGETTO***, allegando:

- PDF della relazione;
- ZIP file contenente i codici sorgente in Python.