

An abstract geometric pattern in the top right corner of the slide. It consists of various blue lines, some solid and some dashed, intersecting and forming a network. Several arrows are integrated into these lines, pointing in different directions, suggesting a flow or a complex system. The pattern is dense and occupies the upper right portion of the slide.

STATISTICA E ANALISI DEI DATI

Progetti di Corso

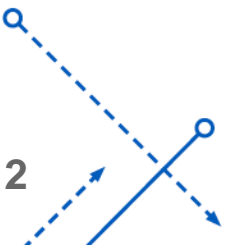
—
Dott. Stefano Cirillo
Dott. Luigi Di Biasi

a.a. 2024-2025

OBIETTIVI GENERALI DEL PROGETTO

- L'obiettivo principale del progetto di corso è quello di verificare le conoscenze apprese durante tutto il corso

- 1.Apprendimento dei Concetti Chiave:** Dimostrare una solida comprensione dei principali concetti del corso
- 2.Applicazione Pratica:** Applicare le competenze statistiche acquisite a situazioni reali
- 3.Risoluzione di Problemi:** Sviluppare la capacità di affrontare l'analisi statistica applicata per diversi obiettivi
- 4.Pensiero Critico:** Sviluppare il pensiero critico e la capacità di analizzare e valutare dati
- 5.Conoscenza di Abilità Specifiche:** Acquisire competenze e conoscenze di analisi statistica trasversali
- 6.Valutazione e Autovalutazione:** Essere in grado di valutare il proprio apprendimento e i propri risultati
- 7.Comunicazione Efficace:** Fornire competenze di comunicazione scritte ed orali



TASK DEL PROGETTO

- Task 1: **Individuazione del database/dataset**

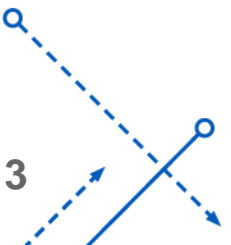
- Ogni gruppo/singolo dovrà:

- Individuare un **argomento** di interesse tra quelli disponibili

- Gli argomenti selezionati per il progetto sono disponibili al seguente link: <https://shorturl.at/pX7Dj>

https://unisalerno-my.sharepoint.com/:x:/g/personal/scirillo_unisa_it/ETkJksBeLI5BvcUZeM3TxrEBmtqvaO1crfWwg8KR6SwpDQ?e=bBe8zL

- Individuare un database/dataset dalla piattaforma su cui voler condurre le analisi statistiche



TASK DEL PROGETTO

- Task 1: **Individuazione del database/dataset**

- Ogni gruppo/singolo dovrà:

- Individuare un **argomento** di interesse tra quelli disponibili

- Gli argomenti selezionati per il progetto sono disponibili al seguente link: <https://shorturl.at/pX7Dj>

https://unisalerno-my.sharepoint.com/:x:/g/personal/scirillo_unisa_it/ETkJksBeLI5BvcUZeM3TxrEBmtqvaO1crfWwq8KR6SwpDQ?e=bBe8zL

- Individuare un database/dataset dalla piattaforma su cui voler condurre le analisi statistiche

- Task 2: **Selezione e conferma del Progetto**

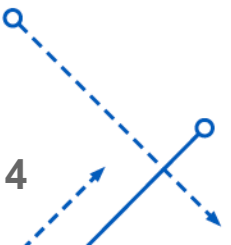
- Ogni gruppo/singolo dovrà:

- Compilare il file excel disponibile al seguente link per confermare la selezione dell'argomento di interesse (se ancora disponibile)

- Nel file sono riportati i gruppi/singoli che hanno risposto al form condiviso in precedenza

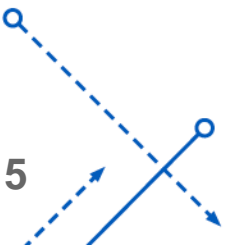
- Nota 1: Non possono esserci 2 gruppi che selezionano lo stesso dataset

- Nota 2: In caso l'argomento di interesse non risultasse più disponibile, scegliere un altro argomento



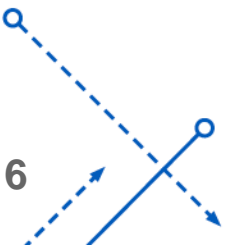
OBIETTIVI DETTAGLIATI DEL PROGETTO

- **OB1**. Selezionare un dataset reale:
 - **Studiare**:
 - L'utilità delle feature presenti.
 - La necessità di aggiungere nuove feature o di raccogliere più dati.
 - L'individuazione di anomalie nei dati.
 - Distribuzione e dipendenza delle variabili: valutare la struttura di dipendenze tra le variabili (ad es. correlazioni) e identificare possibili anomalie o bias
 - ...
 - **Pre-processing il dataset**:
 - **Visualizzazione dei dati**: Creazione di grafici (istogrammi, scatter plot, ecc.) per identificare pattern visivi
 - **Pulizia dei dati**: Trattamento di **valori mancanti** e **outliers**
 - **Analisi delle feature**: Valutare l'importanza e la rilevanza delle feature esistenti
 - Identificare eventuali necessità di **nuove feature** da inferire da quelle esistenti

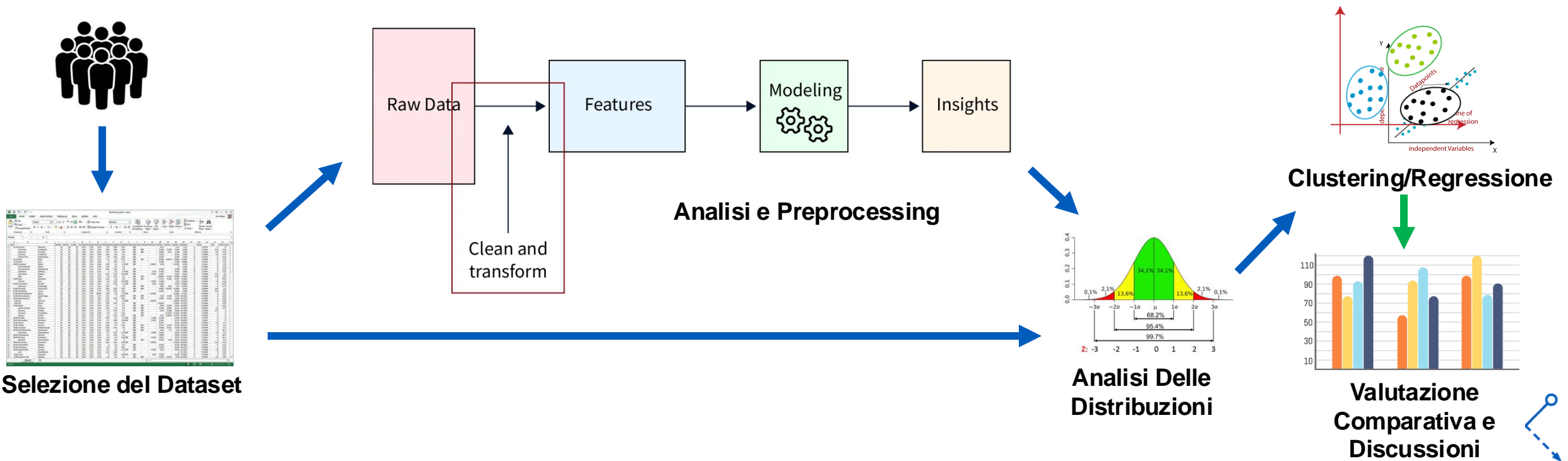


OBIETTIVI DETTAGLIATI DEL PROGETTO

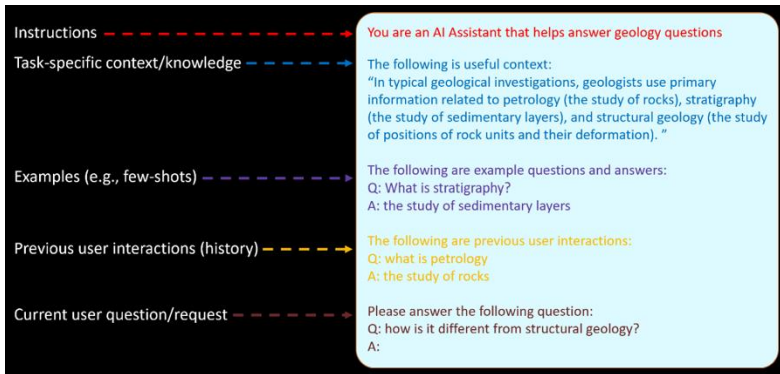
- **OB2.** Utilizzare **Large Language Model** (GPT-4, Gemini, Claude) per generare dati sintetici dello stesso tipo del dataset reale
 - **Selezione del LLM:**
 - Scelta del/i modelli preaddestrati (GPT-4, Gemini, Claude)
 - E' possibile anche trainare modelli generativi ad-hoc (Non è obbligatorio o necessario per il progetto)
 - **Generazione dei Dati:**
 - Definire una strategia di prompt engineering (Almeno un template di interazione)
 - Capire se fornire al modello un contesto o meno è significativo (es. se fornire uno
 - Creare dati con caratteristiche simili al dataset originale
 - **Valutazione Statistica dei Dati Sintetici:**
 - Confrontare le statistiche descrittive dei dati reali e sintetici
 - Valutare le somiglianze e differenze nelle distribuzioni delle feature
 - Valutare se i dati sintetici sono correlati a **distribuzioni statistiche** note (es. Poisson, Normale, ecc)



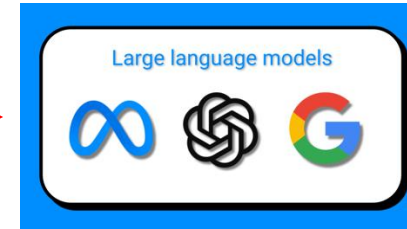
PIPELINE



PIPELINE



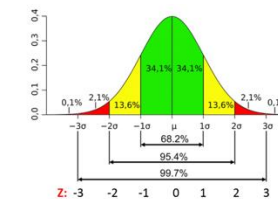
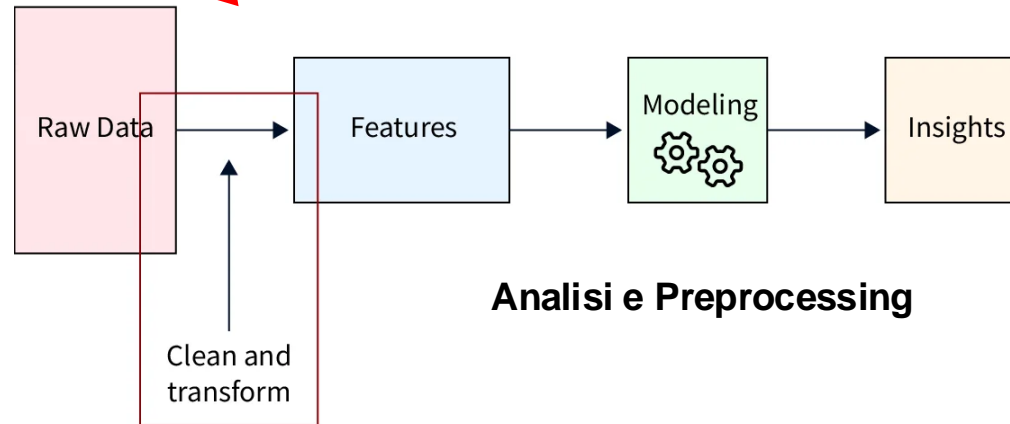
Prompt Engineering



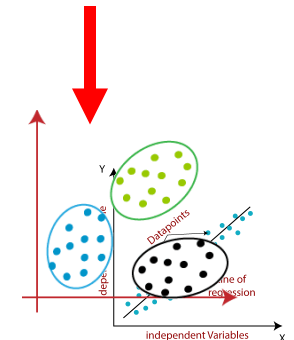
Dataset Generato



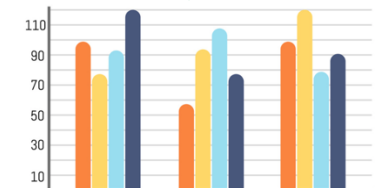
Selezione del Dataset



Analisi Delle Distribuzioni



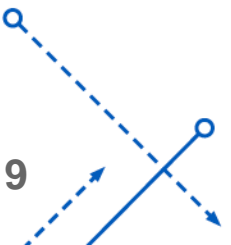
Clustering/Regressione



Valutazione Comparativa e Discussioni

RESEARCH QUESTION

- Lo studio condotto deve avere alla base delle **Research Question** precisi che delineano le analisi da condurre
- Esempi di RQ sono:
 - RQ: I dati sintetici generati dai Large Language Model mantengono le stesse proprietà statistiche dei dati reali?
 - RQ: Le feature generate dai LLM possono essere ricondotte a distribuzioni statistiche note?
 - RQ: Quali sono le differenze principali tra dati sintetici generati e dati reali in contesti di regressione o clustering?
 - RQ: I dati sintetici generati sono adeguati per sostituire o integrare i dati reali in scenari applicativi pratici (es. regressione, clustering)?
 - RQ: Qual è l'impatto delle anomalie presenti nei dati reali sui dati sintetici generati dagli LLM?
 - RQ: In che misura l'aggiunta di nuove feature/tuple nei dati reali influenza la qualità dei dati sintetici generati dagli LLM?
 - RQ: È possibile utilizzare i dati generati dai LLM per migliorare la qualità di dataset reali?
 - ...



TASK DEL PROGETTO

- Task 4: **Consegna progetto e Documentazione**

- Ogni gruppo/singolo dovrà:

- Consegnare un documento PDF che dovrà:

- Contenere una discussione introduttiva sul dominio del problema affrontare

- Esplicitare gli obiettivi dell'analisi statistica

- Dove necessario, le opportune formule e calcoli necessari per le analisi effettuate

- Evitare snippet di codice nel contenuto del testo (il codice può essere messo in Appendice con opportuni riferimenti)

- Per questioni di strutturazione del documento, vi consigliamo di utilizzare LaTeX (**non obbligatorio**)

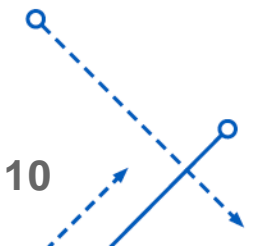
- Di seguito il riferimento ad un template pubblico disponibile su Overleaf:

- <https://www.overleaf.com/latex/templates/math-notes-template/kfqdrzrzpww>

- Consegnare il codice R o i notebook del progetto opportunamente commentati e suddivisi in base delle analisi effettuate

- Nota 1: Verrà stabilita una data ultima di consegna prima di ogni appello di esame (a partire da Gennaio 2025)

- Nota 2: La piattaforma di sottomissione del progetto sarà resa disponibile tempestivamente rispetto all'appello





STATISTICA E ANALISI DEI DATI

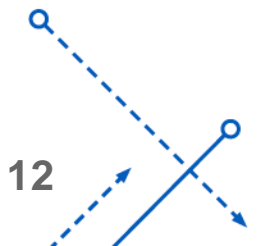
Google Cloud Vertex AI: Guida Completa

Dott. Stefano Cirillo
Dott. Luigi Di Biasi

a.a. 2023-2024

Creazione Account Google Cloud

1. Visitare **cloud.google.com**
2. Cliccare su «**Inizia Gratuitamente**»
3. Accedere con il proprio account **Google**
4. Inserire i dati della carta di credito
 - (necessaria per verifica, **non verrà addebitato nulla**)
5. Riceverai \$300 di credito gratuito valido per 90 giorni



Attivazione Free Trial

1. Confermare i dettagli dell'account
2. Specificare Organizzazione: «Università»
3. Specificare Ruolo: «Studente»
4. Accettare i termini di servizio
5. Verificare l'attivazione nel menu principale

Stai usando la **prova gratuita**



0 di 270 € crediti utilizzati

Scadenza: 16 dicembre 2024

[Cosa succede quando termina il periodo di prova?](#)

ATTIVA ACCOUNT COMPLETO

Creazione Primo Progetto

1. Una volta loggato, accedi alla **Google Cloud Console**.
2. Clicca su **"Select a project"** in alto a sinistra
 - Poi su **"New Project"**.
3. Dai un **nome** al progetto
4. Seleziona la **Località** «Nessuna Organizzazione»
5. Clicca **"Crea"**.

Nuovo progetto



Nella quota rimangono 22 projects. Richiedi un aumento o elimina dei progetti. [Ulteriori informazioni](#)

[MANAGE QUOTAS](#)


Nome progetto *

My Project 89423



ID progetto: warm-particle-438810-a3. Non può essere modificato in un secondo momento. [MODIFICA](#)

Località *

 Nessuna organizzazione

[SFOGLIA](#)

Organizzazione o cartella padre

CREA

ANNULLA

Abilitare l'API Vertex AI

1. Premi il menu di navigazione della Cloud Console
2. Vai su "**APIs & Services**" e clicca su "**Library**".
3. Cerca "**Vertex AI**" e **abilita** l'API.

Dettagli del prodotto



Vertex AI API

[Google Enterprise API](#)

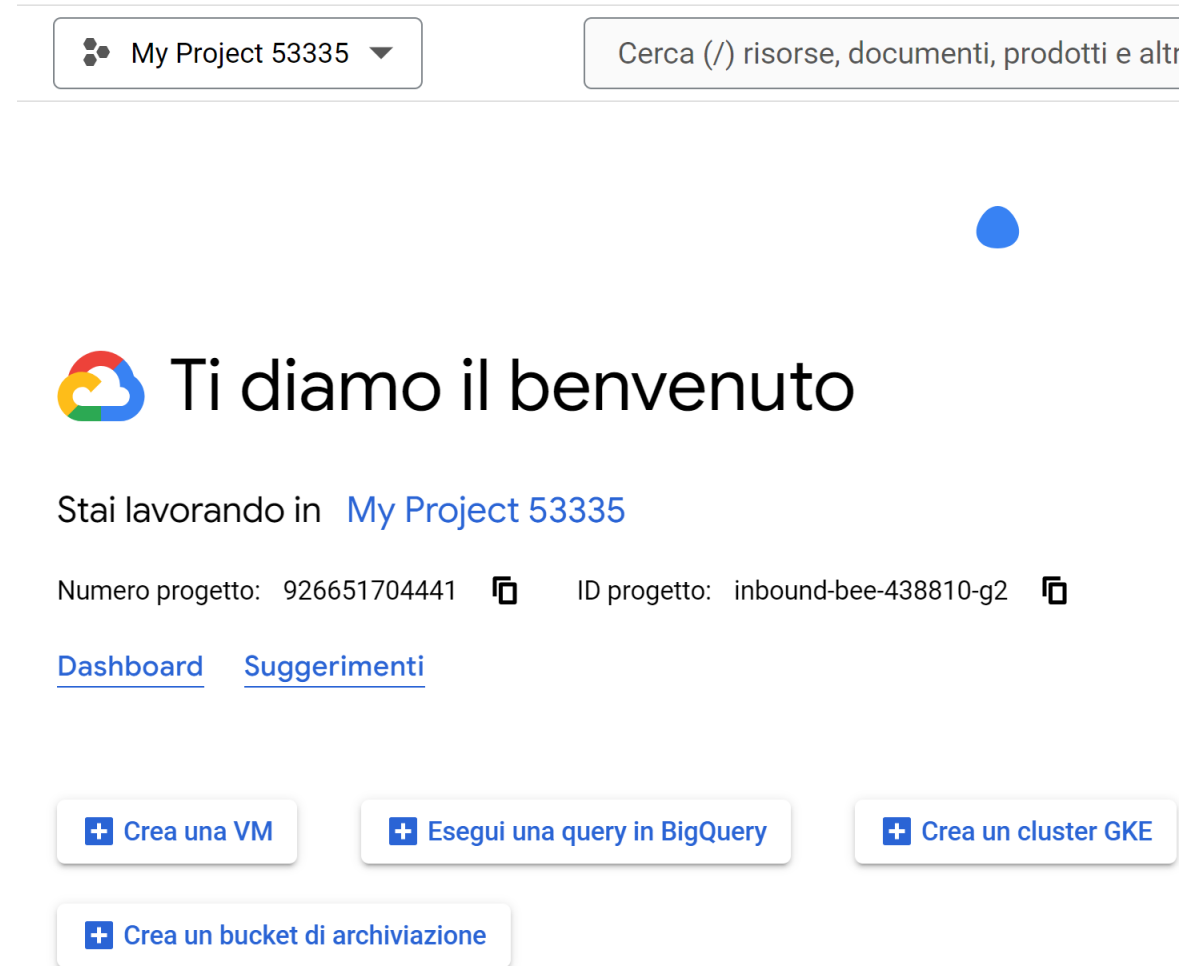
Train high-quality custom machine learning models with minimal machine learning expertise and...

ABILITA

PROVA QUESTA API [↗](#)

Creare un Bucket in Google Cloud Storage

1. Clicca su **"Select a project"** in alto a sinistra
2. Clicca su **"Create Bucket"**.
3. Dai un **nome univoco** al bucket
4. Scegli una **regione** (europe-west8 (Milano))
5. Configura il controllo degli accessi e altre opzioni
(puoi mantenere i valori di default).
6. Clicca **"Create"**.





The screenshot shows the Google Cloud console interface. At the top, there is a dropdown menu for "My Project 53335" and a search bar. Below this, a blue circle indicates the current view. The main heading is "Ti diamo il benvenuto" (We welcome you) with the Google Cloud logo. Below the heading, it says "Stai lavorando in My Project 53335". Further down, it displays the "Numero progetto: 926651704441" and "ID progetto: inbound-bee-438810-g2". There are links for "Dashboard" and "Suggerimenti". At the bottom, there are four buttons: "Crea una VM", "Esegui una query in BigQuery", "Crea un cluster GKE", and "Crea un bucket di archiviazione".

My Project 53335 ▼

Cerca (/) risorse, documenti, prodotti e altri

Ti diamo il benvenuto

Stai lavorando in [My Project 53335](#)

Numero progetto: 926651704441  ID progetto: inbound-bee-438810-g2 

[Dashboard](#) [Suggerimenti](#)

+ Crea una VM

+ Esegui una query in BigQuery

+ Crea un cluster GKE

+ Crea un bucket di archiviazione

Caricare dati nel Bucket

1. Vai al tuo bucket creato.
2. Clicca su **"Upload Files"** per caricare il materiale di lavoro (dataset, modelli preaddestrati, ecc.).

