POLITECNICO DI MILANO





Linee guida per il progetto

Statistica

per Ingegneria Matematica, classe A-L

A.A. 2018-2019

Linee guida per il progetto

1) **Presentazione del problema**

2) **Presentazione dei dati**

3) **Analisi:**

   1) Esplorazione grafica dei dati e statistica descrittiva

   2) Inferenza statistica

4) **Conclusioni**

1) Presentazione del problema

Descrivere il problema che si sta considerando dal punto di vista ingegneristico / economico / fisico / sperimentale / ecc… , specificando quale ruolo il consulente statistico (cioè voi) è chiamato ad assumere e quindi gli obiettivi del progetto.

Esempio:

L'azienda vinicola Bacco è interessata ad analizzare la sua produzione di vino. Il vino proviene da due coltivazioni diverse. Il contenuto nominale di alcool (da etichetta) dovrebbe essere del 14% per la prima e 12% per la seconda. Una quantità importante per la genuinità del vino è anche la prolina (l'amminoacido maggiormente presente). L'azienda Bacco pone le seguenti domande:

- *La concentrazione di alcool delle due coltivazioni rispetta quella nominale?*
- *La concentrazione di prolina è la stessa per le due coltivazioni?*
- *Si può ipotizzare un qualche legame tra la concentrazione di alcool e quella di prolina?*

2) Presentazione dei dati

Spiegare nel dettaglio da dove provengono i dati che verranno utilizzati per affrontare il problema illustrato in precedenza. Descrivere la struttura del dataset e le quantità che vi compaiono (definizione, unità di misura…)

I dati possono:

- essere raccolti direttamente (e.g. laboratori di altri corsi, ma NON tramite interviste!)
- essere richiesti ad aziende, associazioni, enti pubblici, istituti di ricerca, ecc…
- essere trovati su siti dedicati (Uci machine learning, Kaggle, …)
  - UCI Machine Learning Repository
  - Kaggle
  - NASA Surface Metheorology and Solar Energy
  - Open data comune di Milano

2) Presentazione dei dati

Spiegare nel dettaglio da dove provengono i dati che verranno utilizzati per affrontare il problema illustrato in precedenza. Descrivere la struttura del dataset e le quantità che vi compaiono (definizione, unità di misura…)

Esempio: *I dati per l'esempio dell'azienda Bacco provengono da un set di dati vinicoli messo a disposizione online dall'Istituto di Chimica e tecnologie farmaceutiche ed alimentari di Genova.*

Misurazioni della concentrazione di alcool e prolina da 59 bottiglie della prima coltivazione.

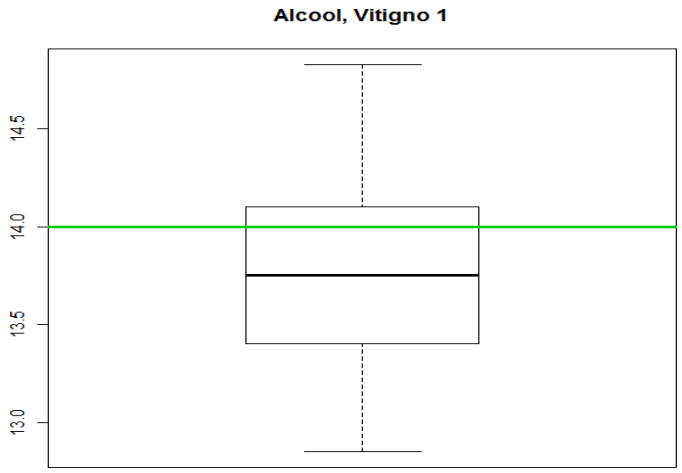
Misurazioni della concentrazione di alcool e prolina da 71 bottiglie della seconda coltivazione.

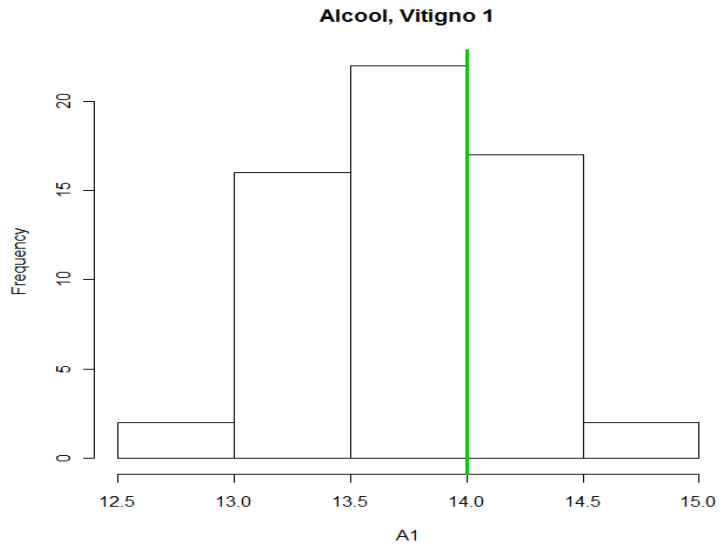
3.1) Esplorazione grafica dei dati e statistica descrittiva

Applicare ai dati gli strumenti di statistica descrittiva più adatti visti nel corso, senza perdere di vista qual è il problema che si vuole affrontare.

Esempio:
La concentrazione di alcool delle due coltivazioni rispetta quella nominale?

Alcool, prima coltivazione:



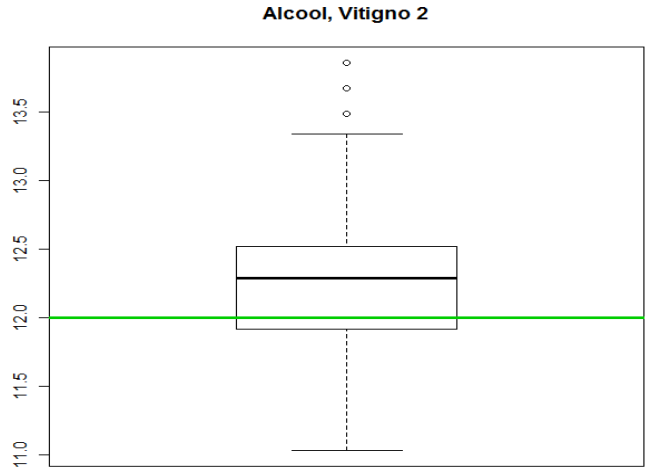


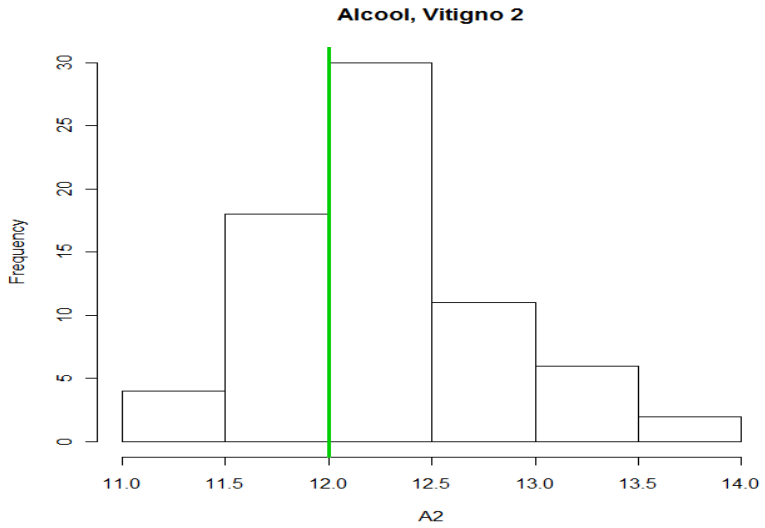
3.1) Esplorazione grafica dei dati e statistica descrittiva

Applicare ai dati gli strumenti di statistica descrittiva più adatti visti nel corso, senza perdere di vista qual è il problema che si vuole affrontare.

Esempio:
La concentrazione di alcool delle due coltivazioni rispetta quella nominale?

Alcool, seconda coltivazione:

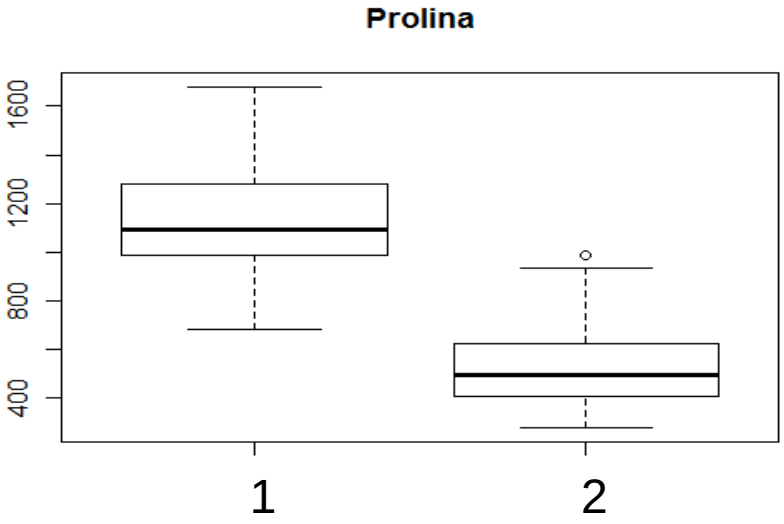


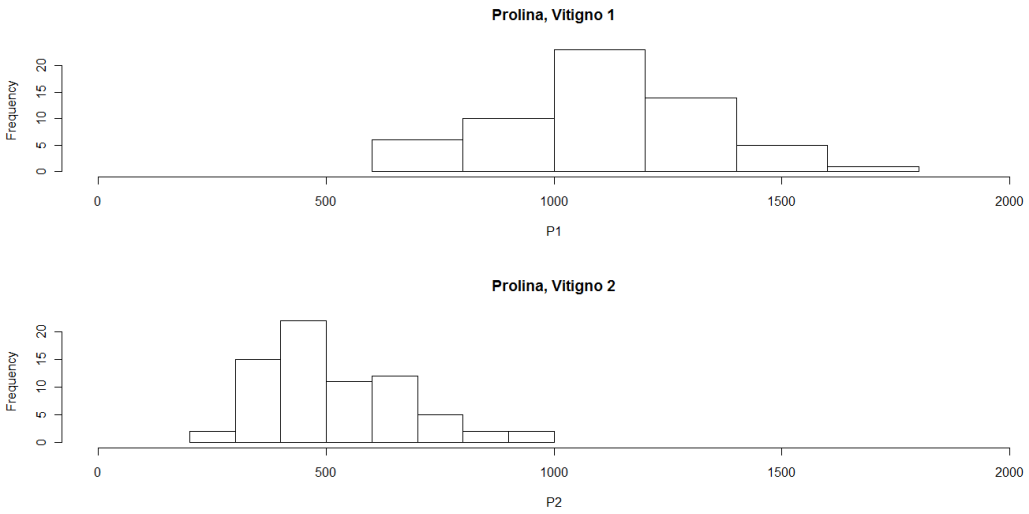


3.1) Esplorazione grafica dei dati e statistica descrittiva

Applicare ai dati gli strumenti di statistica descrittiva più adatti visti nel corso, senza perdere di vista qual è il problema che si vuole affrontare.

Esempio:
La concentrazione di prolina è la stessa per le due coltivazioni?



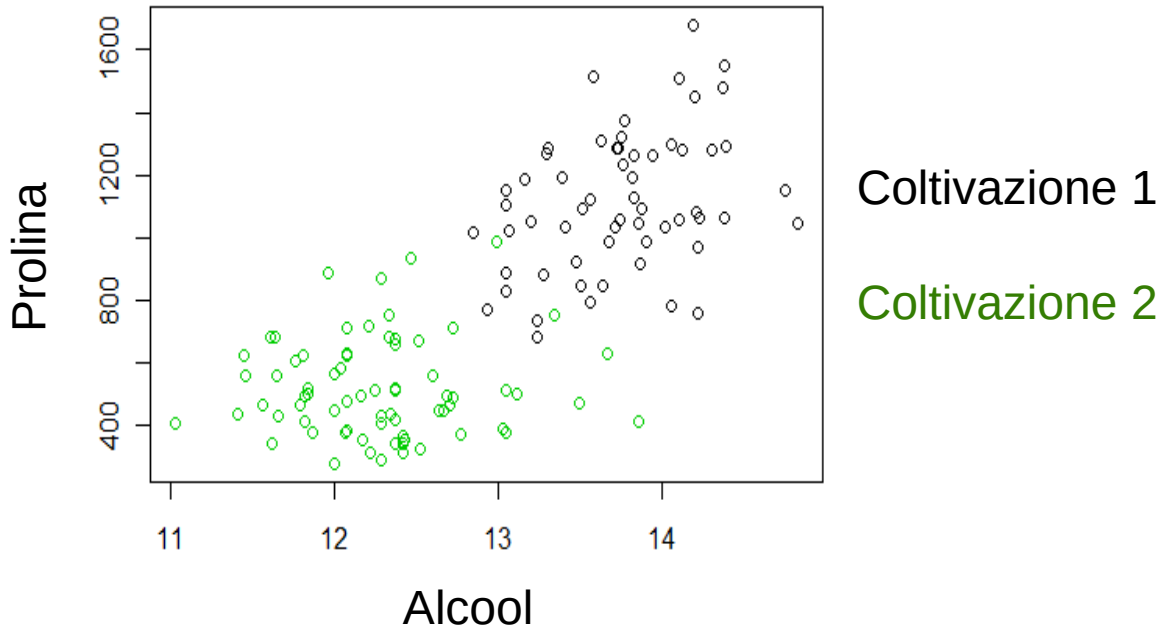


3.1) Esplorazione grafica dei dati e statistica descrittiva

Applicare ai dati gli strumenti di statistica descrittiva più adatti visti nel corso, senza perdere di vista qual è il problema che si vuole affrontare.

Esempio:

Si può ipotizzare un qualche legame tra la concentrazione di alcool e quella di prolina?



3.2) Inferenza statistica

Definire in modo chiaro:

•Quali sono le variabili aleatorie che si stanno considerando

Esempio:

- *A1 = concentrazione di alcool nel vino di una bottiglia della prima coltivazione*
- *A2 = concentrazione di alcool nel vino di una bottiglia della seconda coltivazione*
- *P1 = concentrazione di prolina nel vino di una bottiglia della prima coltivazione*
- *P2 = concentrazione di prolina nel vino di una bottiglia della seconda coltivazione*

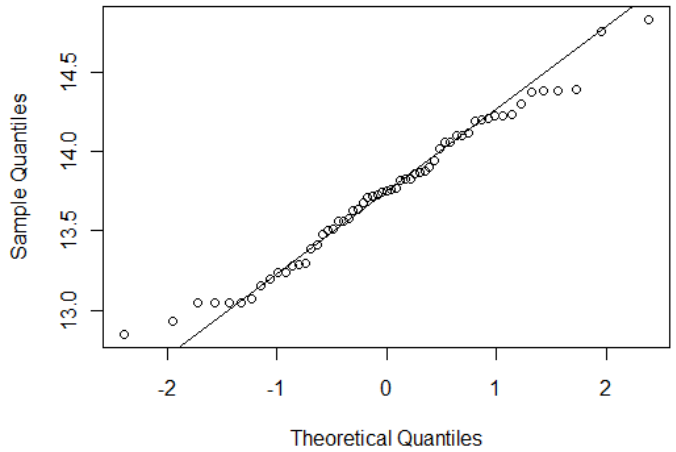
3.2) Inferenza statistica

Definire in modo chiaro:

- Quali sono le variabili aleatorie che si stanno considerando
- Fare le opportune assunzioni sulle leggi delle variabili aleatorie **(solo se servono)**
- Verificare (con strumenti grafici, test) le ipotesi fatte

$$A_1 \sim N(\mu_{A1}, \sigma^2_{A1})$$
$$A_2 \sim N(\mu_{A2}, \sigma^2_{A2})$$
$$P_1 \sim N(\mu_{P1}, \sigma^2_{P1})$$
$$P_2 \sim N(\mu_{P2}, \sigma^2_{P2})$$

Esempio: *QQ-plot di A1*



p-value Shapiro Test = 0.4527

3.2) Inferenza statistica

Tra gli strumenti statistici che vedremo nel corso, **scegliere quelli utili** per rispondere alle domande del problema e applicarli.

Esempio:

La concentrazione di alcool delle due coltivazioni rispetta quella nominale?

RIFORMULARE CORRETTAMENTE LA DOMANDA DA UN PUNTO DI VISTA FORMALE

La concentrazione media di alcool delle due coltivazioni è uguale a quella nominale?

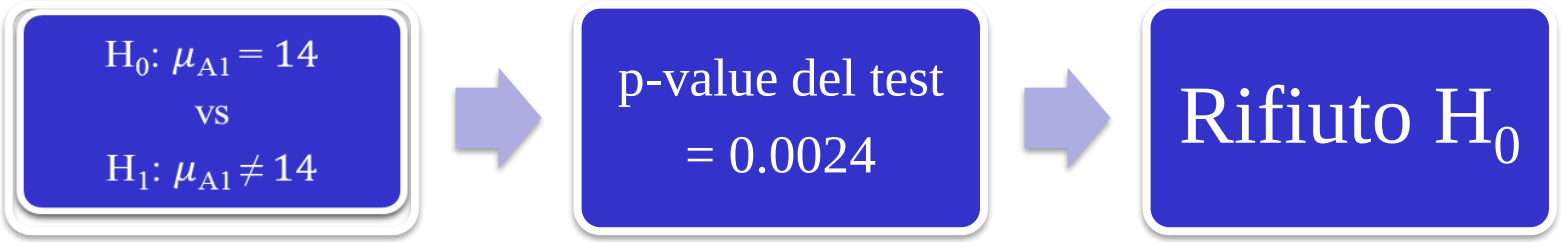
3.2) Inferenza statistica

Tra gli strumenti statistici che vedremo nel corso, **scegliere quelli utili** per rispondere alle domande del problema e applicarli.

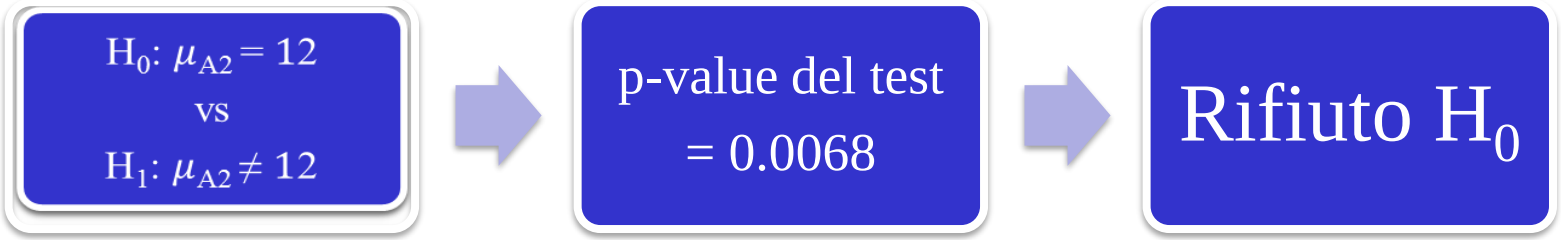
Esempio:

La concentrazione media di alcool delle due coltivazioni è uguale a quella nominale?

- *Test per la media di una popolazione normale, varianza incognita:*



- *Test per la media di una popolazione normale, varianza incognita:*



3.2) Inferenza statistica

Tra gli strumenti statistici che vedremo nel corso, **scegliere quelli utili** per rispondere alle domande del problema e applicarli.

Esempio:

La concentrazione media di alcool delle due coltivazioni è uguale a quella nominale?

Osservazione: se i campioni sono abbastanza numerosi, si può utilizzare il teorema centrale del limite per effettuare questo tipo di test anche in assenza della normalità.

3.2) Inferenza statistica

Tra gli strumenti statistici che vedremo nel corso, **scegliere quelli utili** per rispondere alle domande del problema e applicarli.

Esempio:

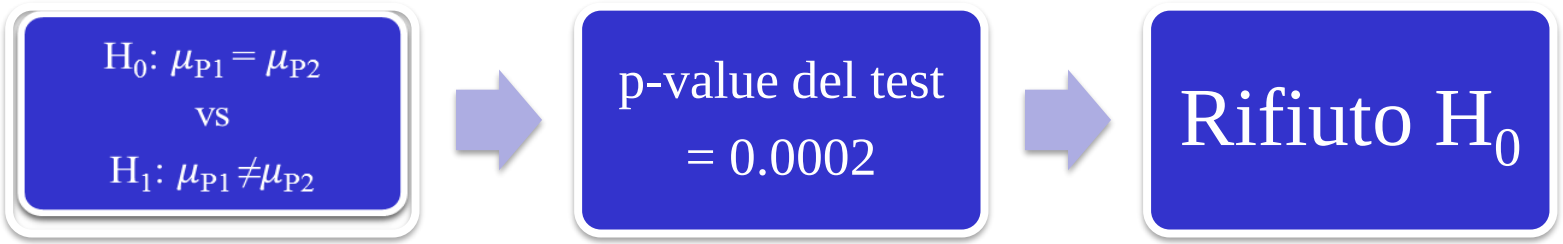
La concentrazione media di prolina è la stessa per le due coltivazioni?

Osservazione:

Per fornire questa risposta, l'ipotesi di normalità è essenziale

↳ *Bisogna verificare l'ipotesi di normalità per P1 e P2 (test di Shapiro)*

Test per il confronto della media di due popolazioni normali:



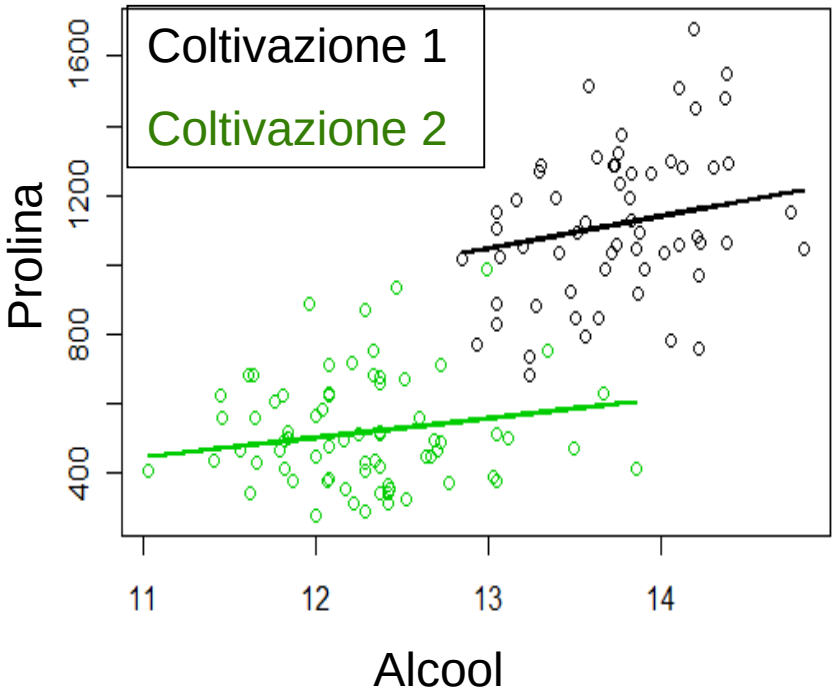
3.2) Inferenza statistica

Tra gli strumenti statistici che vedremo nel corso, **scegliere quelli utili** per rispondere alle domande del problema e applicarli.

Esempio:

Si può ipotizzare un qualche legame tra la concentrazione di alcool e quella di prolina?

A questa domanda si può rispondere stimando un opportuno modello di regressione lineare tra concentrazione di prolina e di alcool e verificandone la significatività (vedremo nelle ultime lezioni del corso). Fra i molti modelli possibili, alcuni che tengono conto dei due gruppi, altri che ignorano la divisione nelle due coltivazioni, scegliamo il modello più significativo.



3.2) Inferenza statistica

Verificare le assunzioni sul modello che sono necessarie per l'analisi mediante grafici o test.

Esempio:

- *verificare la normalità dei residui del modello di regressione (vedremo perché…)*
- *omoschedasticità…*

4) Conclusioni

Fornire al committente (chi ha posto il problema) le risposte ai quesiti posti (possibilmente in un linguaggio il più simile possibile a quello del committente).

Esempio:

- *Le concentrazioni medie di alcool nelle due coltivazioni non corrispondono ai valori nominali.*
- *Le due coltivazioni hanno concentrazioni medie di prolina diverse.*
- *I dati suggeriscono una corrispondenza lineare tra alcool e prolina, che però dipende dalla coltivazione.*

INFORMAZIONI PRATICHE:

Relazione tecnica:

- E' richiesta la stesura di una relazione tecnica di massimo 2 facciate, più stampa della presentazione e/o allegati.
- La relazione deve essere comprensibile al committente dell'analisi e autoesplicativa (con eventuali riferimenti alle slide e/o allegati).
- Dovrà essere consegnata in copia cartacea il giorno della discussione dei progetti (in una o due date da definirsi durante la sessione estiva).

Presentazione:

- I risultati dei lavori saranno presentati oralmente da tutti i membri del gruppo (organizzarsi la divisione delle parti del discorso).
- La presentazione dovrà essere proiettata tramite PC.
- Il tipo di presentazione è libero (pdf, ppt, ecc…) e ne verranno valutati originalità e impatto. Ogni gruppo ha a disposizione 15 minuti per la presentazione, più il tempo per domande e commenti.

INFORMAZIONI PRATICHE:

Il progetto andrà svolto in gruppi da 3 persone (o per particolari necessità da concordare col docente in gruppo da 2 studenti).

La presentazione del progetto sarà discussa indipendentemente dall'esito dello scritto.

Entro il ***data da definirsi*** ogni gruppo dovrà contattare il capoclasse del proprio scaglione e riferire:

- Il nome del responsabile del gruppo completo di indirizzo e-mail
- Il titolo del progetto
- L'elenco dei nomi dei componenti

Ad ogni gruppo verrà assegnato un punteggio (da -3 a 3, congelato per tutto l'A.A.) che si sommerà ai voti che i singoli componenti hanno raggiunto nella prova scritta.