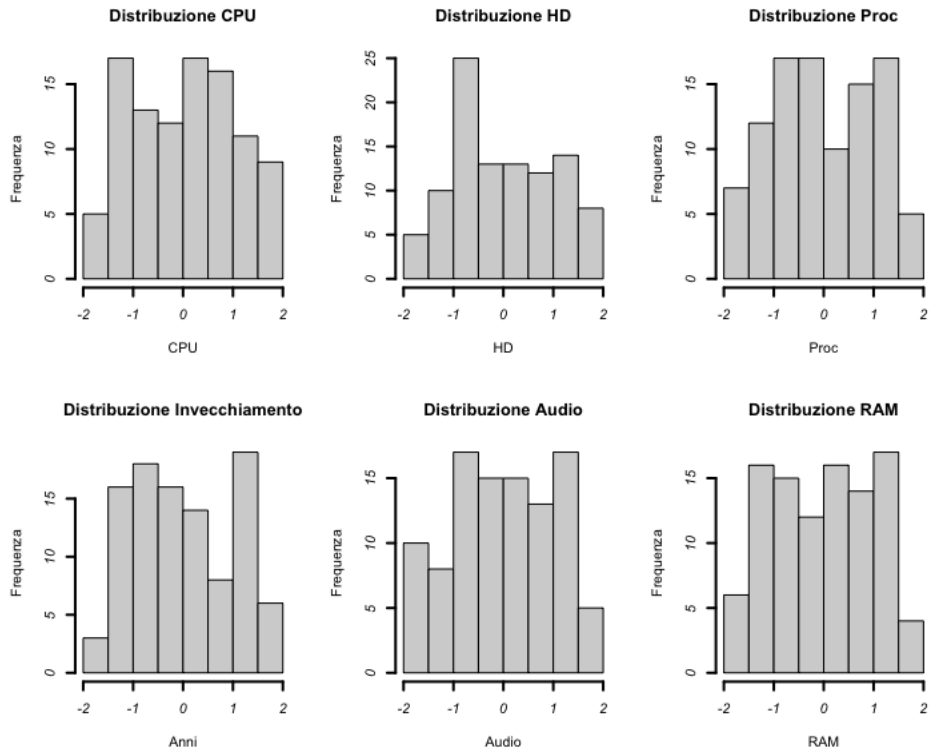
Progetto di Statistica Applicata

(gruppo 9)

Il progetto si articola nell'analisi di un dataset fornito e nella costruzione di un modello di Regressione che possa adattarsi al meglio ai dati forniti.

Si ha a disposizione una variabile dipendente Y, indice delle prestazioni di calcolo del software e sei variabili indipendenti:



X1: Velocità della CPU

X2: Dimensione Hard Disk

X3: Numero di processi software

X4: Indice dell'invecchiamento del Hardware

X5: Prestazioni della scheda audio

X6: Prestazioni della RAM

Analisi preliminare delle variabili

L'istruzione **summary** applicata al dataset ci consente di ottenere valori caratteristici delle variabili in studio, in particolare i quartili (primo e terzo), la media, la mediana(secondo quartile) e valori di massimo e minimo. L'analisi preliminare è stata condotta generando dei grafici tramite il software R in modo da ottenere una visuale molto più chiara e concreta sulla distribuzione dei dati e in particolare delle singole variabili.

Gli istogrammi sono stati realizzati con l'istruzione **hist** utile per vedere il centro, la distribuzione e la forma di un set di dati. I **boxplot,** realizzati tramite il comando **boxplot** mostrano in maniera molto evidente la divisione in quartili (il 25° e il 75° percentile), la mediana, eventuali outlier e valori minimo e massimo.

Analisi di correlazione

L'analisi di correlazione è importante per comprendere se e come le variabili sono correlate tra di loro.

Abbiamo generato degli scatterplot con il comando **plot**, un tipo di grafico cartesiano nel quale sull'asse delle ordinate è stata inserita la variabile dipendente prestazioni, mentre sull'asse delle ascisse le variabili indipendenti Xi.

Abbiamo analizzato l'andamento della nube dei punti in modo visivo confrontandolo con una retta e abbiamo confermato l'aspetto qualitativo calcolando l'indice di correlazione per verificare la forza della relazione lineare (tramite l'istruzione **cor**). Dall'analisi di correlazione si evince che:

- X1: correlazione positiva
- X2: non vi è correlazione
- X3: correlazione positiva
- X4: correlazione negativa
- X5: non vi è correlazione
- X6: correlazione positiva

Per ottenere un'ulteriore conferma è stato realizzato un corrplot, il quale permette di avere una visione globale delle correlazioni tre tutte le variabili e si è inoltre notato che vi è una correlazione tra la variabile CPU e RAM.

Definizione del modello statistico

Per la definizione del modello statistico è stato usato l'algoritmo di backward elimination, che partendo da un modello "completo" elimina in successione, secondo i criteri di arresto impostati, tutte le variabili che apportano uno scarso contributo predittivo, cioè che non dimostrano una forte correlazione con la variabile dipendente. Abbiamo determinato il modello più fedele per descrivere il rapporto tra la variabile e i regressori utilizzando l'istruzione **lm** (che effettua una regressione calcolando tutti i parametri necessari) e abbiamo eliminato dall'analisi i parametri non ritenuti statisticamente rilevanti, connotati da un alto p-value. Inoltre l'istruzione **step**, che consente di trovare il modello migliore è stata un'ulteriore conferma del modello scelto poiché realizza un'analisi basata sul valore di AIC scegliendo il modello con l'AIC minore.

-Il primo modello considerato è un modello di regressione lineare multipla in cui dal modello iniziale, contenente tutte le variabili, sono state eliminate HD e audio, considerate poco rilevanti.

-Il secondo modello tiene in considerazione un fattore curvilineo(variabile proc al quadrato, poiché abbiamo notato che una curva approssima meglio la variabile proc) che ci consente di ottenere un modello che meglio si adatta ai dati, infatti possiede valore più elevato di R squared rispetto al modello prettamente lineare.

Intervalli di confidenza e stima ai minimi quadrati

Abbiamo calcolato gli intervalli di confidenza, utilizzando il comando **confint** che permette di ricavare l'intervallo entro cui è presente il valore desiderato.

Per calcolare i parametri abbiamo utilizzato il comando **coef** fornito da R grazie al quale otteniamo una stima ai minimi quadrati.

Coefficienti di determinazione e grafici diagnostici

Tramite l'istruzione plot abbiamo ottenuti i grafici diagnostici dei nostri modelli presi in analisi, in particolare ci siamo concentrati sul grafico dei residui e il QQplot.

Il coefficiente di determinazione dà un indice della bontà del modello trovato. Se il suo valore è prossimo a uno il modello di regressione lineare si adatta bene ai dati su cui sono stati stimati i parametri(1 miglior adattamento possibile, 0 peggiore).

Per una maggiore sicurezza abbiamo effettuato l'analisi dei residui. Con il comando **plot(___$residuals)**.

Scelta del modello

L'AIC è essenzialmente una misura che stima la qualità di ciascuno dei modelli.

Si può notare che il secondo modello ha un AIC inferiore e quindi ci dà la possibilità di sceglierlo come modello migliore. Questa scelta è stata confermata anche dai grafici diagnostici, poiché quelli del modello 2 risultano avere un migliore adattamento rispetto ai grafici del modello 1.

Abbiamo dunque scelto il modello:

```
Step:  AIC=355.29
y_prestazSWcalc ~ x1_CPU + x3_proc + I(x3_proc^2) + x4_aging +
    x6_RAM

               Df Sum of Sq    RSS    AIC
<none>                      3096.8 355.29
- x6_RAM        1     326.8 3423.6 363.33
- x3_proc       1     790.4 3887.2 376.03
- I(x3_proc^2)  1    2700.8 5797.5 416.00
- x4_aging      1    3255.6 6352.4 425.14
- x1_CPU        1    3764.6 6861.4 432.85
```