

Anno Accademico 21/22

**Progetto di**

**Machine Learning**

Water potability

|  |  |
| --- | --- |
| **Cognome e Nome** | **Matricola** |
| Colombo Andrea | 844557 |
| Sala Mattia | 889888 |

# Indice

1. Dominio di riferimento
2. Obiettivo
3. Scelte di design e assunzioni
4. Pacchetti installati
5. Descrizione del Training Set
6. Principal Component Analisys (PCA)
7. Descrizione dei modelli di Machine Learning
   1. Decision Tree
      1. Train Set e Test Set
      2. Data Exploration
      3. Modello Base
   2. Gradient Boosting
      1. Descrizione dei parametri
      2. Boosting dei parametri
8. Esperimenti
   1. 10 Fold Cross Validation e Stima delle misure di performance
      1. Matrice di Confusione Complessiva
      2. Precision
      3. Recall
      4. F-Measure
      5. ROC e AUC
9. Conclusioni Finali

# **1 – Dominio di Riferimento**

L'accesso all'acqua potabile è un diritto fondamentale per la salute dell’essere umano.

Questo è importante come questione di salute e sviluppo a livello nazionale, regionale e locale.

In alcune regioni, è stato dimostrato che gli investimenti nell'approvvigionamento idrico e nei servizi igienico-sanitari possono produrre un beneficio economico, poiché le riduzioni degli effetti negativi sulla salute e dei costi sanitari superano i costi di realizzazione degli interventi.

Il dataset di riferimento su cui si basa questo progetto è “*Water Quality*”, reperibile al seguente link:

<https://www.kaggle.com/datasets/adityakadiwal/water-potability>

Il dataset contiene 3276 record di diversi campioni d’acqua analizzati.

# **2 - Obiettivo**

L’obiettivo dell’elaborato è di risolvere un problema di classificazione legato alla potabilità di un campione d’acqua. Per farlo, verranno impiegati due algoritmi di apprendimento supervisionato e il dataset sopra citato.

# **3 - Scelte di design e assunzioni**

Il dataset di riferimento contiene i seguenti attributi:

|  |  |
| --- | --- |
| Attributo | Descrizione |
| **pH** | Il pH è un parametro importante per valutare l’acidità o la basicità dell’acqua. La “WHO” considera potabile l’acqua con un pH che varia tra 6,5 e 8,5. Nel dataset considerato, il range varia tra 6.52 e 6.83, che rientra nei parametri definiti dalla WHO. |
| **Hardness** | Hardness is mainly caused by calcium and magnesium salts. These salts are dissolved from geologic deposits through which water travels. The length of time water is in contact with hardness producing material helps determine how much hardness there is in raw water. Hardness was originally defined as the capacity of water to precipitate soap caused by Calcium and Magnesium. |
| **Solids** (Total dissolved solids) | Water has the ability to dissolve a wide range of inorganic and some organic minerals or salts such as potassium, calcium, sodium, bicarbonates, chlorides, magnesium, sulfates etc. These minerals produced un-wanted taste and diluted color in appearance of water. This is the important parameter for the use of water. The water with high TDS value indicates that water is highly mineralized. Desirable limit for TDS is 500 mg/l and maximum limit is 1000 mg/l which prescribed for drinking purpose. |
|  |  |
| **Chloramines** | Chlorine and chloramine are the major disinfectants used in public water systems. Chloramines are most commonly formed when ammonia is added to chlorine to treat drinking water. Chlorine levels up to 4 milligrams per liter (mg/L or 4 parts per million (ppm)) are considered safe in drinking water. |
| **Sulfate** | Sulfates are naturally occurring substances that are found in minerals, soil, and rocks. They are present in ambient air, groundwater, plants, and food. The principal commercial use of sulfate is in the chemical industry. Sulfate concentration in seawater is about 2,700 milligrams per liter (mg/L). It ranges from 3 to 30 mg/L in most freshwater supplies, although much higher concentrations (1000 mg/L) are found in some geographic locations. |
| **Conductivity** | Pure water is not a good conductor of electric current rather’s a good insulator. Increase in ions concentration enhances the electrical conductivity of water. Generally, the amount of dissolved solids in water determines the electrical conductivity. Electrical conductivity (EC) actually measures the ionic process of a solution that enables it to transmit current. According to WHO standards, EC value should not exceeded 400 μS/cm. |
| **Organic\_carbon** | Total Organic Carbon (TOC) in source waters comes from decaying natural organic matter (NOM) as well as synthetic sources. TOC is a measure of the total amount of carbon in organic compounds in pure water. According to US EPA < 2 mg/L as TOC in treated / drinking water, and < 4 mg/Lit in source water which is use for treatment. |
| **Trihalomethanes** | THMs are chemicals which may be found in water treated with chlorine. The concentration of THMs in drinking water varies according to the level of organic material in the water, the amount of chlorine required to treat the water, and the temperature of the water that is being treated. THM levels up to 80 ppm is considered safe in drinking water. |
| **Turbidity** | The turbidity of water depends on the quantity of solid matter present in the suspended state. It is a measure of light emitting properties of water and the test is used to indicate the quality of waste discharge with respect to colloidal matter. The mean turbidity value obtained for Wondo Genet Campus (0.98 NTU) is lower than the WHO recommended value of 5.00 NTU. |
| **Potability** | Variabile target che indica se l’acqua è:   * 0 = Non Potabile * 1 = Potabile |

# **4 - Pacchetti R Installati e Ambiente di sviluppo**

dplyr, ggplot2, skimr, scales, forcats, tidyverse, ggpubr, corrplot, RColorBrewer, varImp, GGally

RStudio

# **5 - Descrizione del Training Set: Analisi Esplorativa**

## Pulizia e formattazione dei dati

Inizialmente si è partito con una lettura dei dati del dataset.

Immagine che contiene tavolo

Descrizione generata automaticamente

Come si può notare, la variabile target è di tipo “double”. Per una migliore operabilità si è deciso di convertirla in “factor”, in quanto può assumere solo due valori (0 e 1).

Qui sotto il codice:

water\_potability$Potability <- as.factor(water\_potability$Potability)

Immagine che contiene tavolo

Descrizione generata automaticamente

Prima di iniziare l’analisi, viene effettuato un controllo e pulizia su eventuali dati mancanti (“NA”).

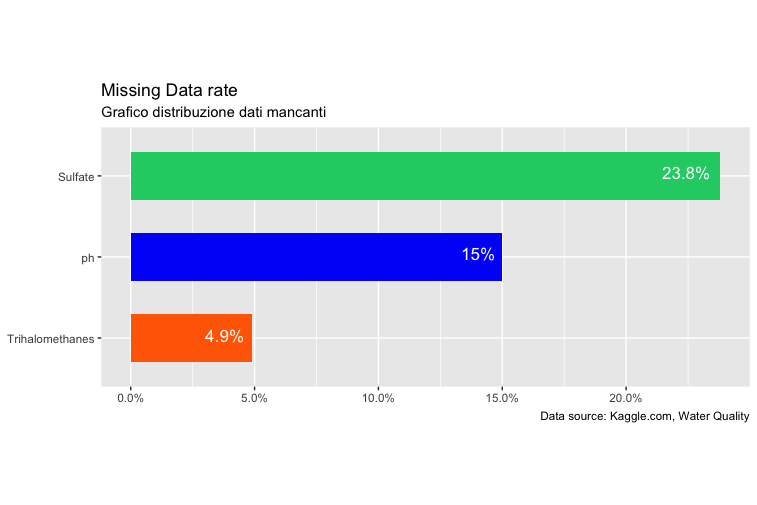
Innanzitutto, viene visualizzata la quantità dei dati mancanti su tutti gli attributi del dataset.

|  |
| --- |
| > water\_potability %>% summarise\_all(~ sum(is.na(.))) |



Su tre attributi manca un discreto numero di dati. Si analizza quindi il missing data rate (rispetto alle 3276 righe del dataset) e viene mostrato un grafico con i valori percentuali. Di seguito viene mostrato il codice:

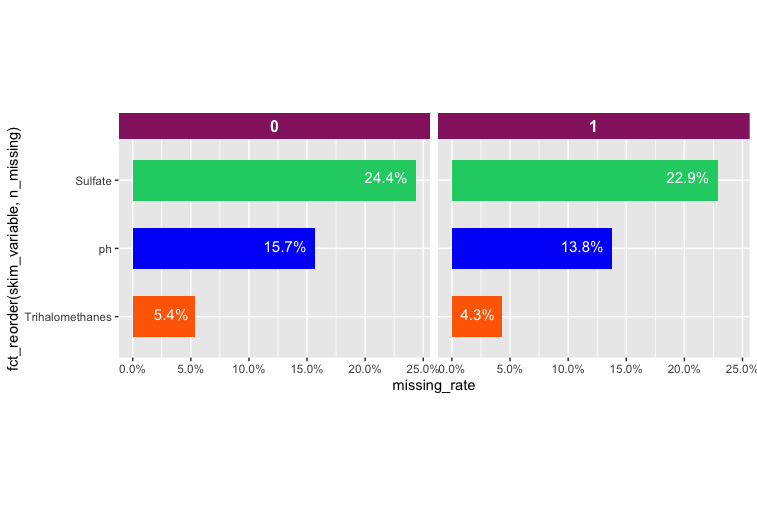
water\_potability %>%  skim() %>%  
    filter(n\_missing != 0) %>%  
    as\_tibble() %>%  
    select(skim\_variable, n\_missing, complete\_rate) %>%  
    mutate(missing\_rate = round(abs(complete\_rate - 1) \* 100, 1)) %>%  
    ggplot(aes(  
        x = fct\_reorder(skim\_variable, n\_missing),  
        y = missing\_rate,  
        fill = skim\_variable,  
        label = paste0(missing\_rate, "%")  
    )) +  
    geom\_col(width = .6) +  
    geom\_text(  
        size = 4.5,  
        hjust = 1.2,  
        vjust = .25,  
        col = "white"  
    ) +  
    coord\_flip() + theme(aspect.ratio = .4) +  
    theme(  
        legend.position = "none"  
    ) +  
    scale\_y\_continuous(label = label\_percent(scale = 1)) +  
    scale\_fill\_manual(values = c("#071ff7",  
                                 "#17cf73",  
                                 "#ff6a00")) +  
    labs(  
        title = "Missing Data rate",  
        subtitle = "Grafico distribuzione dati mancanti",  
        caption = "Data source: Kaggle.com, Water Quality",  
        x = NULL,  
        y = NULL  
    )



Viene ora studiata la distribuzione dei valori nulli rispetto alla variabile target, in modo da capire la diversa influenza che hanno sul risultato:

Vengono valutati il numero di dati mancanti per ciascun attributo, a seconda che la variabile target sia 0 oppure 1:





Come si può notare, la distribuzione è molto simile in entrambi i casi, quindi la presenza di valori nulli non crea evidenti squilibri nella classificazione.

È quindi possibile attuare tre diverse strategie, ovvero:

* **Eliminazione delle colonne contenenti valori nulli superiori ad una soglia limite**
* **Eliminazione record contenenti valori nulli**
* **Sostituzione valori nulli con valori che non inficiano la classificazione**

Si è deciso di scartare la prima opzione in quanto si perderebbero molti dati fondamentali alla classificazione.

Si scarta anche la seconda opzione in quanto verrebbe eliminato un consistente numero di record, utili per la classificazione.

Di conseguenza si decide di usare la terza strategia; ovvero la sostituzione dei valori nulli con i valor medi degli attributi mancanti. In questo modo manteniamo intatto il numero di record del dataset, inserendo un valore neutro per la classificazione.

Di seguito viene presentato un sommario dei valori statistici principali, in particolare:

*media, val. max, val. min, primo quartile, terzo quartile, mediana.*

Immagine che contiene testo, ricevuta

Descrizione generata automaticamente

Ora si procede alla sostituzione dei valori “NA” col corrispettivo valor medio:

|  |
| --- |
| water\_potability <- water\_potability %>%  group\_by(Potability) %>%     mutate(across(where(is.numeric),          ~if\_else(is.na(.),          mean(., na.rm = T),          as.numeric(.)))) %>% ungroup() |

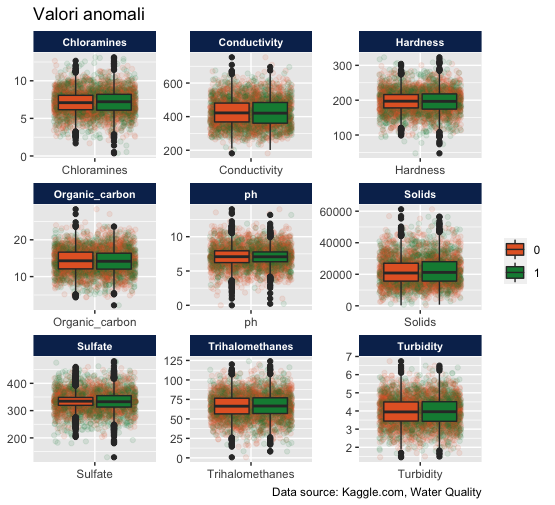
Immagine che contiene tavolo

Descrizione generata automaticamente

Viene ora analizzata la presenza di eventuali valori anomali (Outliers) che potrebbero compromettere l’analisi dei dati in quanto non sono indicativi della reale distribuzione dei dati.

Per verificare la presenza di questi valori viene usato un grafico di tipo boxplot che consente di individuare in modo visuale eventuali datti che si discostano molto dai dati compresi tra 1° e 3° quartile.

|  |
| --- |
| water\_potability %>%     pivot\_longer(cols = -Potability, names\_to = "feature") %>%     ggplot(aes(x = feature, y = value)) +     geom\_jitter(aes(y = value, col = Potability), alpha = 0.1) +     geom\_boxplot(aes(fill = Potability)) +     facet\_wrap(vars(feature), ncol = 3, scales = "free") +     scale\_color\_manual(values = c("#E4652E", "#0E8A41")) +     scale\_fill\_manual(values = c("#E4652E", "#0E8A41")) +     theme(         legend.position = "right",         strip.background = element\_rect(fill = "#0B2D5B"),         strip.text = element\_text(color = "white", face = "bold", size = 8)     ) +     labs(         title = "Valori anomali",         caption = "Data source: Kaggle.com, Water Quality",         x = NULL,         y = NULL,         fill = NULL,         color = NULL     ) |

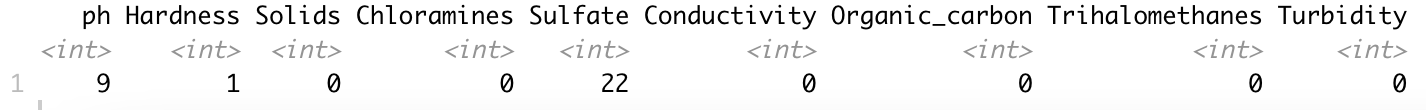


Si nota che potrebbero esserci alcuni outliers, è necessario quindi calcolarne in modo più preciso il numero sfruttandone la definizione stessa:

* Nel lato inferiore, un outlier estremo si trova al di sotto del **Primo Quartile – 3 \* Scarto Interquartile**
* Nel lato superiore, un outlier estremo si trova al di sopra del **Terzo Quartile + 3 \* Scarto Interquartile**

Tramite il codice seguente si sostituiscono con un valore nullo in modo da poter essere contati

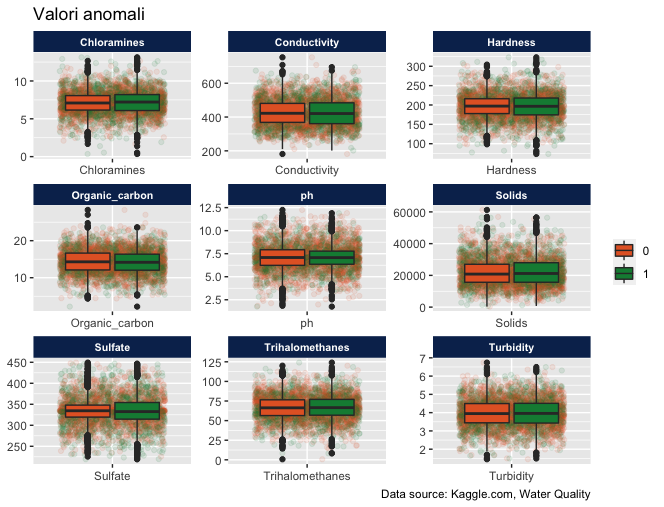
|  |
| --- |
| outliers <- water\_potability %>%    select(-Potability) outliers[] <- lapply(outliers, **function**(x){   qq <- quantile(x, c(0.25, 0.75), na.rm = TRUE)   is.na(x) <-  x < (qq[1] - (3\*(qq[2]-qq[1]))) | x > (qq[2] + (3\*(qq[2]-qq[1])))   x }) outliers %>% summarise\_all(~ sum(is.na(.))) |



Si nota che sono stati trovati (anche se pochi) degli outliers, si procede quindi alla loro sostituzione con la media

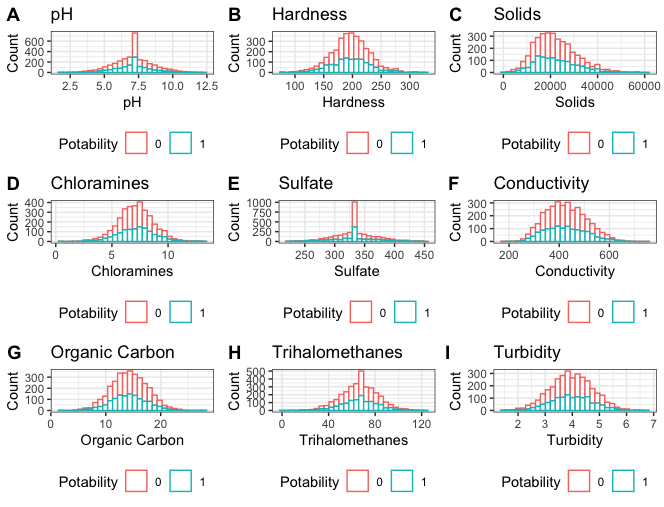
|  |
| --- |
| outliers <- outliers %>%    mutate(across(where(is.numeric),                  ~if\_else(is.na(.),                           mean(., na.rm = T),                             as.numeric(.)))) %>%    ungroup()  outliers$Potability <- water\_potability$Potability water\_potability <- outliers |

Si procede visualizzando nuovamente il boxplot relativo



Viene ora visualizzato un istogramma con la distribuzione dei dati per avere un ulteriore riscontro dell’assenza di valori anomali

|  |
| --- |
| p1 <- ggplot(water\_potability, aes(ph, color = as.factor(Potability)))+     geom\_histogram(bins = 30, fill = "white") +     labs(x = "pH", y = "Count", col = "Potability") +     theme\_bw() +      theme(legend.position = "bottom")+     labs(title = "pH") p2 <- ggplot(water\_potability, aes(Hardness, color = as.factor(Potability)))+     geom\_histogram(bins = 30, fill = "white") +     labs(x = "Hardness", y = "Count", col = "Potability") +     theme\_bw() +      theme(legend.position = "bottom")+      labs(title = "Hardness") p3 <- ggplot(water\_potability, aes(Solids, color = as.factor(Potability)))+     geom\_histogram(bins = 30, fill = "white") +     labs(x = "Solids", y = "Count", col = "Potability") +     theme\_bw() +      theme(legend.position = "bottom") +      labs(title = "Solids") p4 <- ggplot(water\_potability, aes(Chloramines, color = as.factor(Potability)))+     geom\_histogram(bins = 30, fill = "white") +     labs(x = "Chloramines", y = "Count", col = "Potability") +     theme\_bw() +      theme(legend.position = "bottom") +      labs(title = "Chloramines") p5 <- ggplot(water\_potability, aes(Sulfate, color = as.factor(Potability)))+     geom\_histogram(bins = 30, fill = "white") +     labs(x = "Sulfate", y = "Count", col = "Potability") +     theme\_bw() +      theme(legend.position = "bottom") +      labs(title = "Sulfate") p6 <- ggplot(water\_potability, aes(Conductivity, color = as.factor(Potability)))+     geom\_histogram(bins = 30, fill = "white") +     labs(x = "Conductivity", y = "Count", col = "Potability") +     theme\_bw() +      theme(legend.position = "bottom") +      labs(title = "Conductivity") p7 <- ggplot(water\_potability, aes(Organic\_carbon, color = as.factor(Potability)))+     geom\_histogram(bins = 30, fill = "white") +     labs(x = "Organic Carbon", y = "Count", col = "Potability") +     theme\_bw() +      theme(legend.position = "bottom") +      labs(title = "Organic Carbon") p8 <- ggplot(water\_potability, aes(Trihalomethanes, color = as.factor(Potability)))+     geom\_histogram(bins = 30, fill = "white") +     labs(x = "Trihalomethanes", y = "Count", col = "Potability") +     theme\_bw() +      theme(legend.position = "bottom") +      labs(title = "Trihalomethanes") p9 <- ggplot(water\_potability, aes(Turbidity, color = as.factor(Potability)))+     geom\_histogram(bins = 30, fill = "white") +     labs(x = "Turbidity", y = "Count", col = "Potability") +     theme\_bw() +      theme(legend.position = "bottom") +      labs(title = "Turbidity") figure <- ggarrange(p1, p2, p3, p4, p5, p6, p7, p8, p9, nrow = 3, ncol = 3, labels = "AUTO")  figure |

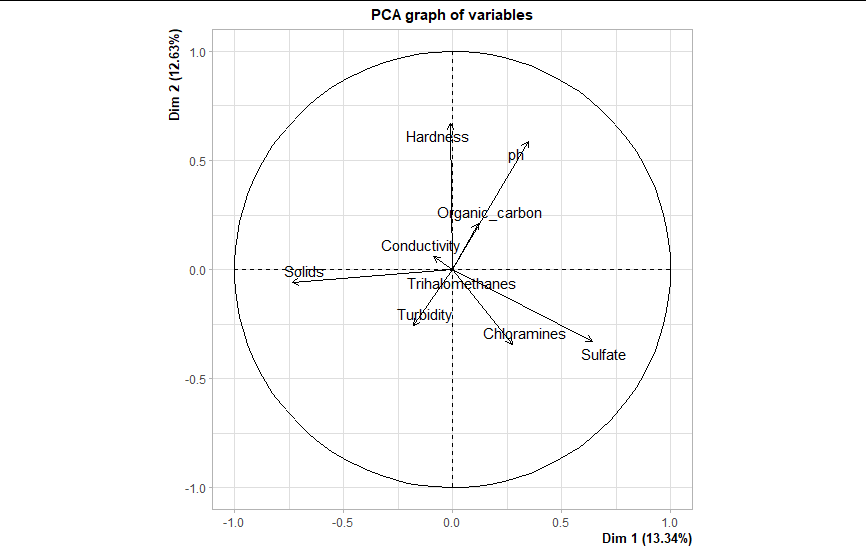


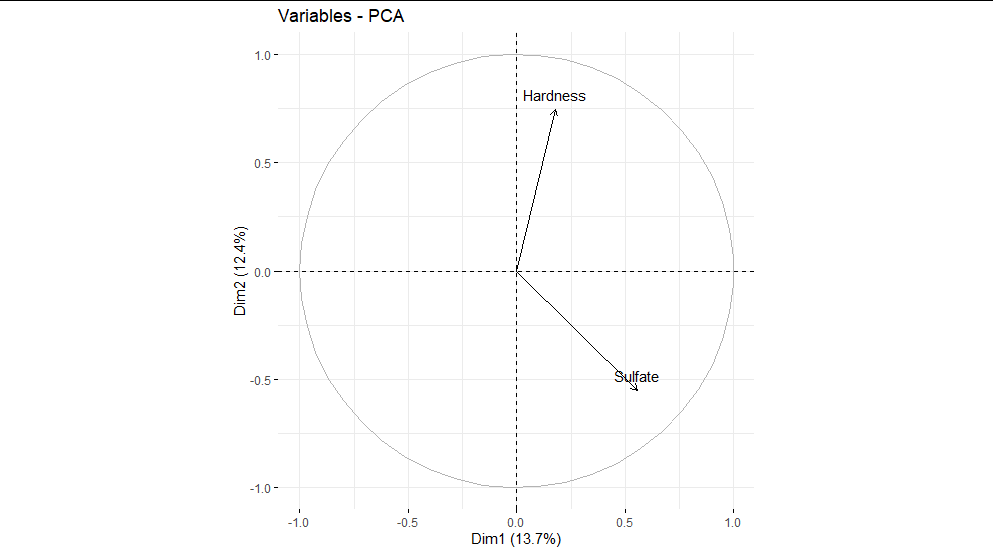
# **6 - Principal Component Analisys (PCA)**

Con la PCA si vuole analizzare quali siano le componenti più incisive del dataset, in modo da utilizzarle in fase di training del modello.

Di seguito viene presentato il codice per computare la PCA. Si prendono le feature “attive” (*tutti gli attributi del dataset ad eccezione della variabile*). Dopodichè si visualizzano i risultati sul grafico delle componenti principali.

|  |
| --- |
| res.pca <- PCA(as.data.frame(water\_potability[1:9]), graph=TRUE) |



Tra tutte le componenti individuate, si considerano solo quelle più importanti, cioè quelle 

Si valutano gli autovalori col comando:

|  |
| --- |
| eig.val <- get\_eigenvalue(res.pca) |

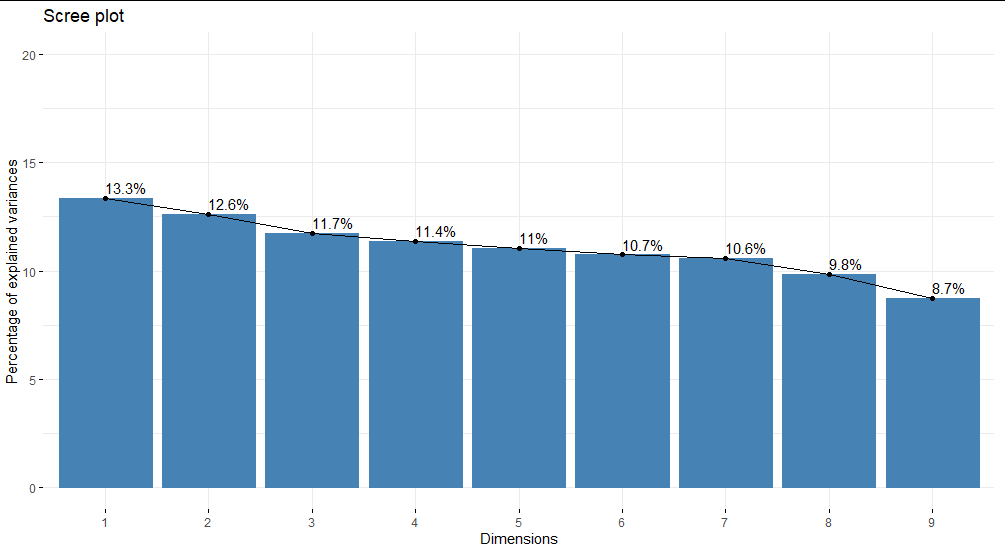
Immagine che contiene testo, monitor, schermo, tabellonesegnapunti

Descrizione generata automaticamente

Il 70% della varianza cumulativa è spiegato dalle prime 6 dimensioni; in particolare le prime 2 dimensioni spiegano soltanto il 25% cumulativo della varianza (*rispettivamente 13% e 12%*).

Viene anche visualizzato lo Scree Plot, ottenibile con il seguente codice:

|  |
| --- |
| fviz\_eig(pca.res, addlabels = TRUE, ylim = c(0, 20)) *#scree plot* |



Dopodichè si vuole osservare che incidenza ha ciascun attributo sulle varie dimensioni:

|  |
| --- |
| *# Contributions to the principal components* var <- get\_pca\_var(pca.res) head(var$contrib) |

Immagine che contiene testo, batteria, targa

Descrizione generata automaticamente  
Come si può vedere, gli attributi con più incidenza sulla “**Dimensione 1**”, sono:

     - **Solids** *44%*

     - **Sulfate** *34%*

     - **pH** *10%*

responsabili del ***88%*** della varianza sulla prima dimensione.

Per quanto riguarda la “**Dimensione 2**” invece, si ha:

- **pH** 30%

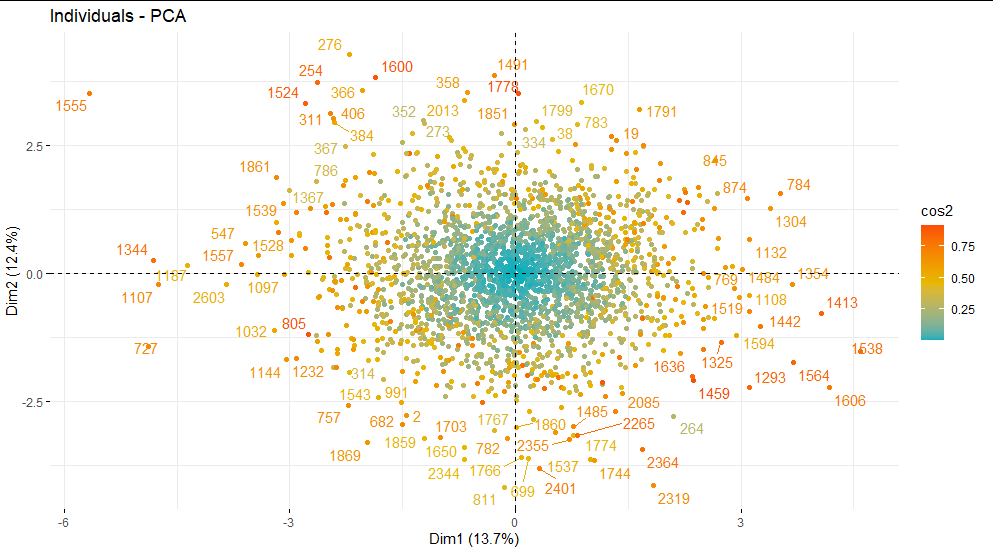
- **Hardness** 39%

- **Chloramines** 10 %

Responsabili del ***79%*** della varianza sulla seconda dimensione.

 Successivamente si visualizzano gli “individui”:

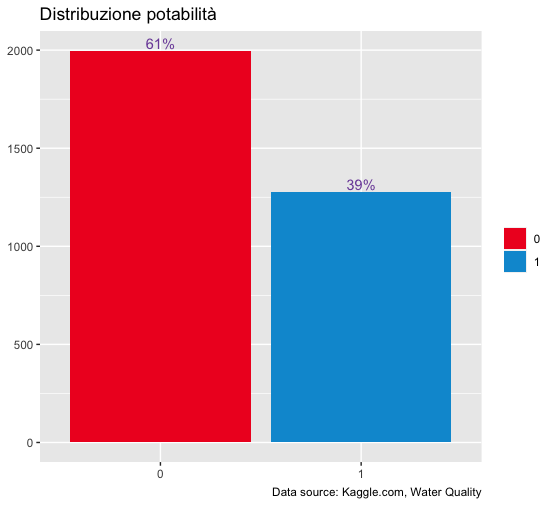
|  |
| --- |
| fviz\_pca\_ind(pca.res, col.ind = "cos2",                      gradient.cols = c("#00AFBB", "#E7B800", "#FC4E07"),                      repel = TRUE *# Avoid text overlapping (slow if many points)* ) |



Come è possibile notare, la densità di individui ben classificati è nettamente inferiore a quelle degli individui non ben classificati.

Distribuzione di potabilità

|  |
| --- |
| water\_potability %>%     select(Potability) %>%     count(Potability) %>% mutate(percent = paste0(round(n / sum(n) \* 100), "%"), 2) %>%     ggplot(aes(         x = Potability,         y = n,         label = percent,         fill = Potability     )) +     geom\_col() +     geom\_text(vjust = -0.2, color = "#7C4EA8") +     scale\_fill\_manual(values = c("#EF1A25", "#0099D5")) +     labs(         title = "Distribuzione potabilità",         caption = "Data source: Kaggle.com, Water Quality",         x = NULL,         y = NULL,         fill = NULL     ) |



# **7 - Descrizione dei modelli di Machine Learning usati**

La scelta di un algoritmo di apprendimento automatico su cui costruire un modello è un’operazione delicata e che richiede uno studio profondo del dominio di riferimento e dei dati a disposizione.  
La prima cosa da fare è categorizzare il problema che si sta affrontando, in modo da escludere e priori algoritmi poco consoni.

Per questo motivo, essendo un problema di classificazione con solo 2 possibili classi, con valori numerici ed un numero relativamente basso di record e features si è deciso di utilizzare come primo algoritmo i Decision Tree.

La distribuzione dei valori delle varie features risulta molto concentrata in un range molto piccolo, condizione che porta all’esclusione di algoritmi che utilizzano un iperpiano per separare le classi.

Si è deciso dunque di utilizzare un algoritmo derivante sempre dai Decision Tree ma che ne consente un incremento sostanziale delle performance, il Gradient Boosting.

## Gradient Boosting

Come per Decision Tree si procede allo split del dataset 70%-30%

Si procede poi a convertire le classi di potabilità in intero in  quanto l’algoritmo richiede delle classi intere per la creazione delle matrici

Essendo che il comando as.integer restituisce un intero nel range [1, n], mentre il comando xgb.DMatrix richede valori in un range [0, n] sottraiamo 1 al valore converitio.

Viene poi rimossa la colonna potability dal trainset e dal testset.

|  |
| --- |
| split.data = **function**(data, p = 0.7, s = 1){   set.seed(s)   index = sample(1:dim(data)[1])   train = data[index[1:floor(dim(data)[1] \* p)], ]   test = data[index[((ceiling(dim(data)[1] \* p)) + 1):dim(data)[1]], ]   **return**(list(train=train, test=test)) }  allset = split.data(water\_potability, p=0.7) trainset = allset$train testset = allset$test  y\_train <- as.integer(trainset$Potability) - 1  y\_test <- as.integer(testset$Potability) - 1 X\_train <- trainset %>% select(-Potability) X\_test <- testset %>% select(-Potability) |

a questo punto creiamo 2 matrici (train e test) che verranno utilizzate dall'algoritmo per il leaning e la verifica

|  |
| --- |
| xgb\_train <- xgb.DMatrix(data = as.matrix(X\_train), label = y\_train) xgb\_test <- xgb.DMatrix(data = as.matrix(X\_test), label = y\_test) |

passiamo ora alla fase di definizione dei parametri di default che studieremo per poi effettuare il tuning migliore possibile per aumentare l’efficienza dell’algoritmo.

|  |
| --- |
| #default parameters xgb\_params <- list(   booster = "gbtree",    objective = "multi:softprob",    eta=0.3,    max\_depth=6,    num\_class = length(levels(water\_potability$Potability)),   )  #printa il valore ideale per nround con  parametri default xgbcv <- xgb.cv(    params = xgb\_params,    data = xgb\_train,    nrounds = 100,    nfold = 5,    showsd = T,    stratified = T,    print\_every\_n = 10,    early\_stopping\_rounds = 20,    maximize = F) |

In particolare I parametri indicano:

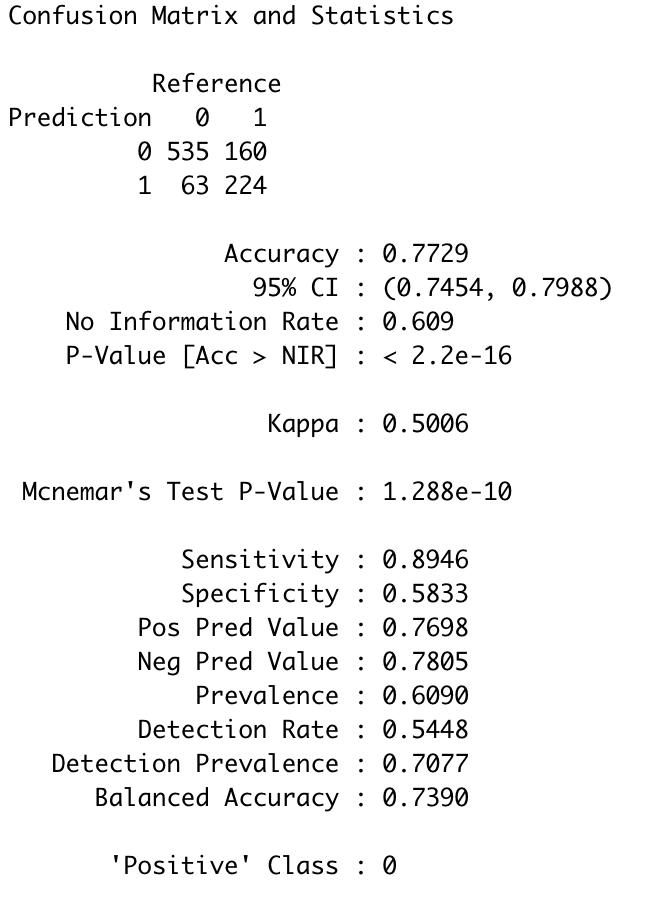
* Booster: tipologia di booster da utilizzare (tree based o funzione lineare)
* Objective: Specificare il learning task e il learning objective corrispondente (in questo caso classificazione muulticlass)
* Eta: riduce i pesi delle funzionalità per rendere il processo di potenziamento più conservativo (più è basso e meglio resiste all’overfitting)
* Max\_depth: altezza massima di un albero
* Num\_class: numero di classificazioni possibili

Tramite la funzione xgb.cv otteniamo il numero ideale di iterazioni per cui l’errore ricavato con il trainSet è minore dell’errore ricavato nel testSet.

In questo caso il numero ideale di iterazioni è 20

Eseguiamo ora il train con i nuovi parametri e mostriamo la confusion matrix finale

|  |
| --- |
| xgb\_model <- xgb.train(   params = xgb\_params,   data = xgb\_train,   nrounds = 20,   verbose = 1, )  xgb\_preds <- predict(xgb\_model, as.matrix(X\_test), reshape = TRUE)  xgb\_preds <- as.data.frame(xgb\_preds)  colnames(xgb\_preds) <- levels(water\_potability$Potability)  xgb\_preds$PredictedClass <- apply(xgb\_preds, 1, **function**(y) colnames(xgb\_preds)[which.max(y)]) xgb\_preds$ActualClass <- levels(water\_potability$Potability)[y\_test + 1]  confusionMatrix(as.factor(xgb\_preds$PredictedClass), as.factor(xgb\_preds$ActualClass)) |



# **8 - Esperimenti**

# **9 - Conclusioni**