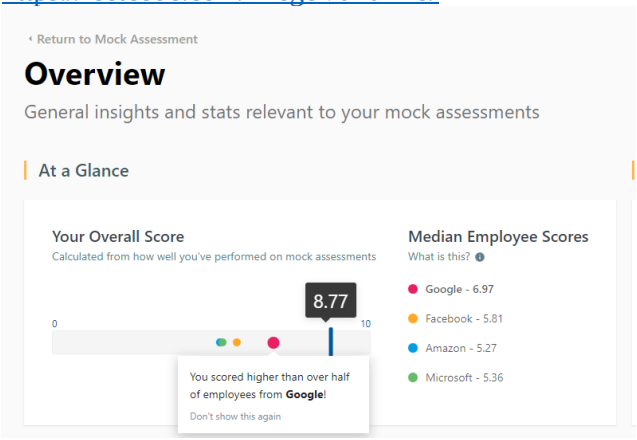
Project developed by Diego Vallarino in the last 5 years.

Github (with some projects)
https://github.com/DiegoVallarino
Leetcode:
https://leetcode.com/DiegoVallarino/



Project 1
Design an algorithm for the design of an investment strategy.

```
library (neuralnet)
library(rpart )
library(GA)
library (caret )
library(psych)
library(quantmod )

# load the data
dji<-getSymbols ("^DJI", from = Sys.Date () - 15*365, to = Sys.Date (), src ="yahoo", adjust = TRUE)
stocks<- getSymbols("Company1", "Company1",…, "Companyn", from = Sys.Date () - 15*365, to = Sys.Date (),
src ="yahoo", adjust = TRUE)
cor(dji,stocks)
Note: A portfolio is created with the companies that have the greatest correlation with DJI. Few companies with a
behavior similar to the index.

## Neural Network
formula <- price ~ volume + news
model <- train(formula, data = data, method = " neuralnet ")

## Trees
formula <- decision ~ price + economic_data + trends
model <- train (formula, data = data, method = " rpart ")

## Genetic Networks
formula <- c(price, allocation_funds ) ~ return + risk
model <- train (formula, data = data, method = " ga ")

How would I define a good strategy (trading algorithm) in reality?
# Calculate daily returns
returns <- diff (log(dji))

# Definition of parameters
SL <-TBD
TP <- TBD
threshold <-TBD
*sorry but I don't show all the cards!! ... 😊

# initialization of variables
n <- length (dji)
positions <- rep(0, n)
positions[ 1] <- 0
buy.price <- rep(0, n)
sell.price <- rep(0, n)
pnl <- rep(0, n)
```

Strategy calculation
It is confidential.

```
# Results of the strategy
PosOps <- sum(pnl > 0)
NegOps <- sum(pnl < 0)
Ops <- PostOps + NegOps
HR <- PostOps / Ops
TR <- cumsum(pnl)
AR <- mean(pnl)
MD <- max(cummax(TR) - TR)
```

Project 2

Design of a default risk analysis model.

```
# Create a list of the models
models2 <- list(
  "glm" = train(dividend ~ ., data = trainset, method = "glm", trControl = trainControl(method = "cv", number = 25)),
  "Neural Network" = train(dividend ~ ., data = trainset, method = "nnet", trControl = trainControl(method = "cv",
number = 25)),
  "Random Forest" = train(dividend ~ ., data = trainset, method = "rf", trControl = trainControl(method = "cv",
number = 25)),
  "SVA" = train(dividend ~ ., data = trainset, method = "svmRadial", trControl = trainControl(method = "cv", number
= 25)),
  "XGBoost" = train(dividend ~ ., data = trainset, method = "xgbTree", trControl = trainControl(method = "cv",
number = 25))
)

results2 <- resamples(models2)
summary(results2)
```

Project 3

Right now, I am working on a paper analysing **different survival models**, where I take parametric, non-parametric and semi-parametric models and compare them with a machine learning model on neural networks (multitask logistic regression (MTLR). This implies that the censoring factor (censored by right) in the ML model is incorporated and provides more information, improving the model. C-index has a major improvement.

A fundamental problem is understanding the **relationship between covariates and the (distribution of) survival times (times to event).** This type of analysis can allow you to understand the time needed until the event, purchase, repurchase, churn, etc.

```
#Train & Test data
set.seed(123)
data.train <- sample_frac(veteran, 0.7)
train_index <- as.numeric(rownames(data.train))
data.test <- veteran [-train_index, ]

surv_obj = Surv(data.test$time, data.test$status)

fit2 <- survreg(Surv(time, status) ~ karno + age + trt, data=data.train)
predictfit2<-predict(fit2, data.test)
metrics_fit2<-Cindex(surv_obj, predicted = predictfit2)

fit3<-survfit(Surv(time, status) ~ 1, data = data.train)
dis_timefit3 = fit3$time
med_indexfit3 = median(1:length(dis_timefit3))
predictfit3<-predictSurvProb(fit3, data.test, dis_timefit3)
metrics_fit3 = Cindex(surv_obj, predicted = predictfit3[, med_indexfit3])

fit4 <- coxph(Surv(time, status) ~ ., data=data.train, x = TRUE)
shapiro.test(fit4$residuals)
anova(fit4)
dis_timefit4 = fit3$time
med_indexfit4 = median(1:length(dis_timefit4))
predictfit4<-predictSurvProb(fit4, data.test, dis_timefit4)
metrics_fit4 = Cindex(surv_obj, predicted = predictfit4[, med_indexfit4])
```

Project 4

An experiment was conducted to compare the a**ccuracy of three mass spectrometers in measuring the proportions of 14N to 15N,** two soil samples were taken from each of three plots treated with 15N and two subsamples of each sample were analysed on each of the machines. The resulting design has spectrometers “crossed” with plots and samples, but the samples are “nested” within the plots.

Project 5

Reduced the bank cost around USD 1 million per month by developed an ANN and MLR algorithm (over 2 million people in the CRC-BCU) to estimate the individual income of the Uruguayan population.

1. Pain Point: need to understand the individual income.

2. Opportunity (new data) The Uruguayan central bank opened some partial information related with the financial market. *ID-contingency-bank*
3. We get that data and enrich this data with our data (from our clients)
4. We use a multi category independent variable (target) to infer the different income ranges (1-10).
5. I use a MLR white box (organizational issues)
6. Process
    a) ETL
    b) data split (data.train, data.test)
    c) lm
    d) anova
    e) shapiro.test (normality of residuals)
    f) leveneTest (homoscedasticity)
    g) VIF (multicolinealidad)
    h) dwt(autocorrelación)
    i) predict
    j) predictROC
    k) performance
    l) performanceAUC

Project 6:
Increased the origination-rate by +10x by lead the development of ANN model to predict the consumer loans propensity.

1. Propensity by Product (personal loans, car loans, credit card) by Channel (ATM, Call Center, and Internet)
2. Different Models in different Channels in different Products
3. We Scored 200K people.
4. Preferences (moment life, moment day, trade-off, etc), Utility Function, Behavioural Economics, Info Economics (asymmetric, biases, etc)
5. Process (credit cards ~ 0.5%)
    a. Unbalance (credit cards) – *Undersampling* (on leads), *Oversampling* (on new clients), *SMOTE* (new dataset with same characteristics but balanced)
    library(DMwR2)

Project 7
My PhD thesis I used the **Dif in Dif model** to analyse the impact of (tax) incentives in investment decisions.
Dif in Dif working by comparing the average change over time in the outcome variable for the treatment group to the average change over time for the control group. Although it is intended to mitigate the effects of extraneous factors and selection bias, depending on how the treatment group is chosen, this method may still be subject to certain biases (e.g., mean regression, reverse causality and omitted variable bias).

Project 8
I have received recognition from the Ministry of Public Health of Uruguay for helping with the modelling of **Covid in Uruguay**. This is the **ARMA model** (autoregressive moving average) with a projection horizon of t=15 days, adjustable daily.