R Notebook

Definición del problema

Crear un modelo de regresión con el objetivo de predecir el ancho del sépalo.

Iris

La base de datos antes mencionada contiene información acerca de la especies de flor de iris existen tres tipos en este datset setosa,virginica y versicolor.

A su ves tiene información acerca del pétalo y el sépalo ambos con sus respectivas medidas de ancho y largo respectivamente.

Importaciones de las librerias

```
library(DataExplorer)
library(ggplot2)
library(dplyr)
```

```
library(caret)
```

```
df<-read.csv("C:\\Users\\amado\\Desktop\\Ciencias de datos\\Bases de datos\\Iris.csv")
```

```
names(df)
```

```
## [1] "Id"            "SepalLengthCm" "SepalWidthCm"  "PetalLengthCm"
## [5] "PetalWidthCm"  "Species"
```

Selección de variables

Seleccionamos las variables que tengan más relación con el grosor del pétalo.

```
df<- df %>%
  select(PetalLengthCm,PetalWidthCm,Species)
```

```
summary(df)
```

```
##  PetalLengthCm    PetalWidthCm     Species
##  Min.   :1.000   Min.   :0.100   Length:150
##  1st Qu.:1.600   1st Qu.:0.300   Class :character
##  Median :4.350   Median :1.300   Mode  :character
```

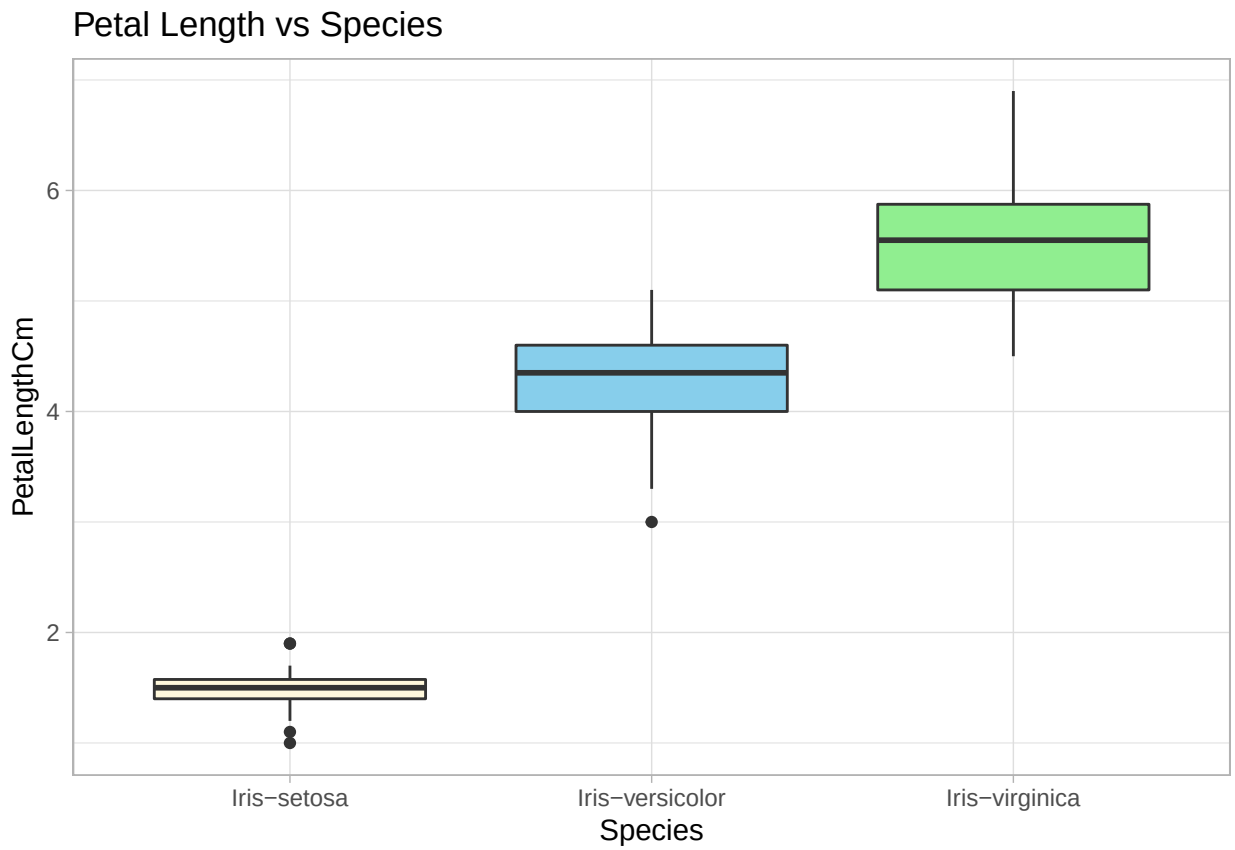
```
##  Mean   :3.759   Mean   :1.199
##  3rd Qu.:5.100   3rd Qu.:1.800
##  Max.   :6.900   Max.   :2.500
```

```
df$Species<-as.factor(df$Species)
```

Análisis Exploratorio

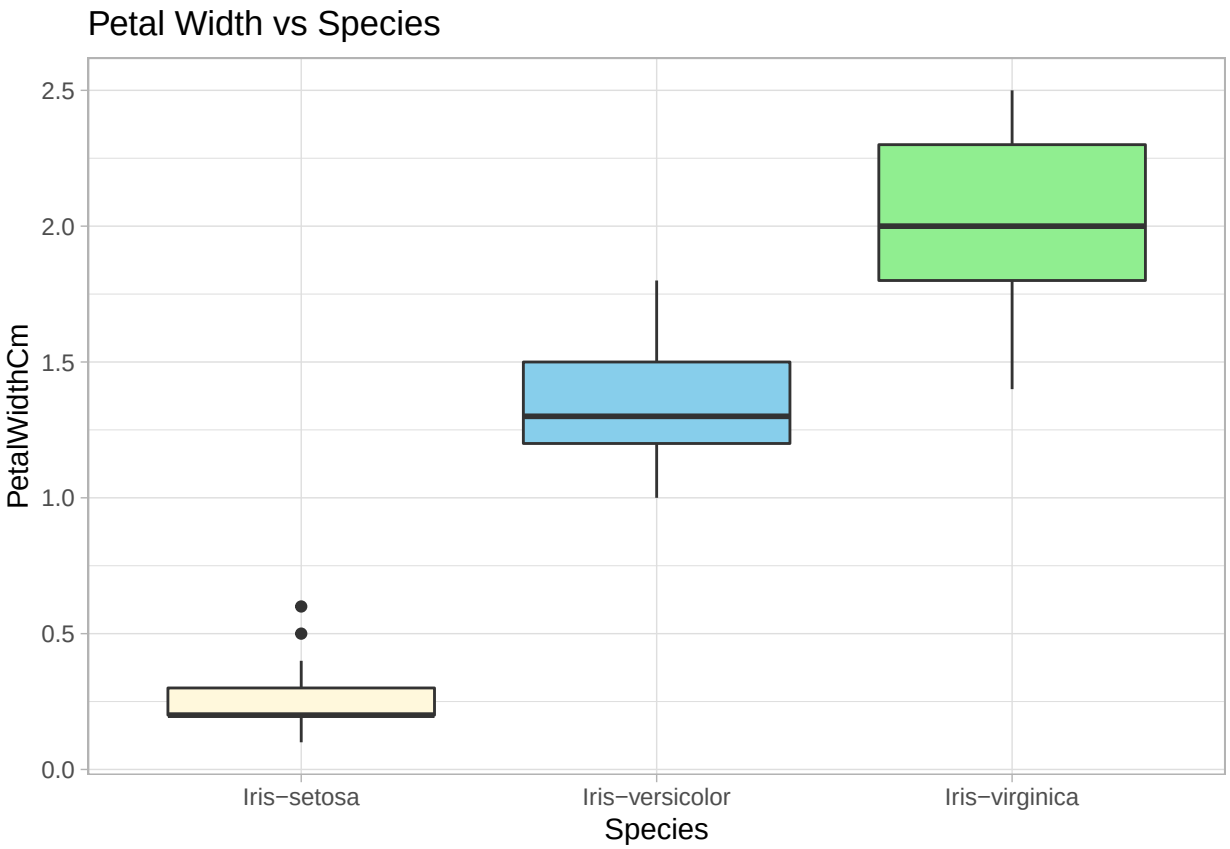
¿Qué especie de iris tiene mayor longitud del péatalo?

```
ggplot(data=df,aes(x=Species,y=PetalLengthCm,fill=Species)) +  geom_boxplot(fill=c('cornsilk','skyblue'
theme(legend.position = 'top') + theme_light() + ggtitle('Petal Length vs Species')
```



¿Qué especie de iris tiene mayor grosor péatalo?

```
ggplot(data=df,aes(x=Species,y=PetalWidthCm,fill=Species)) +  geom_boxplot(fill=c('cornsilk','skyblue',
theme(legend.position = 'top') + theme_light() + ggtitle('Petal Width vs Species')
```

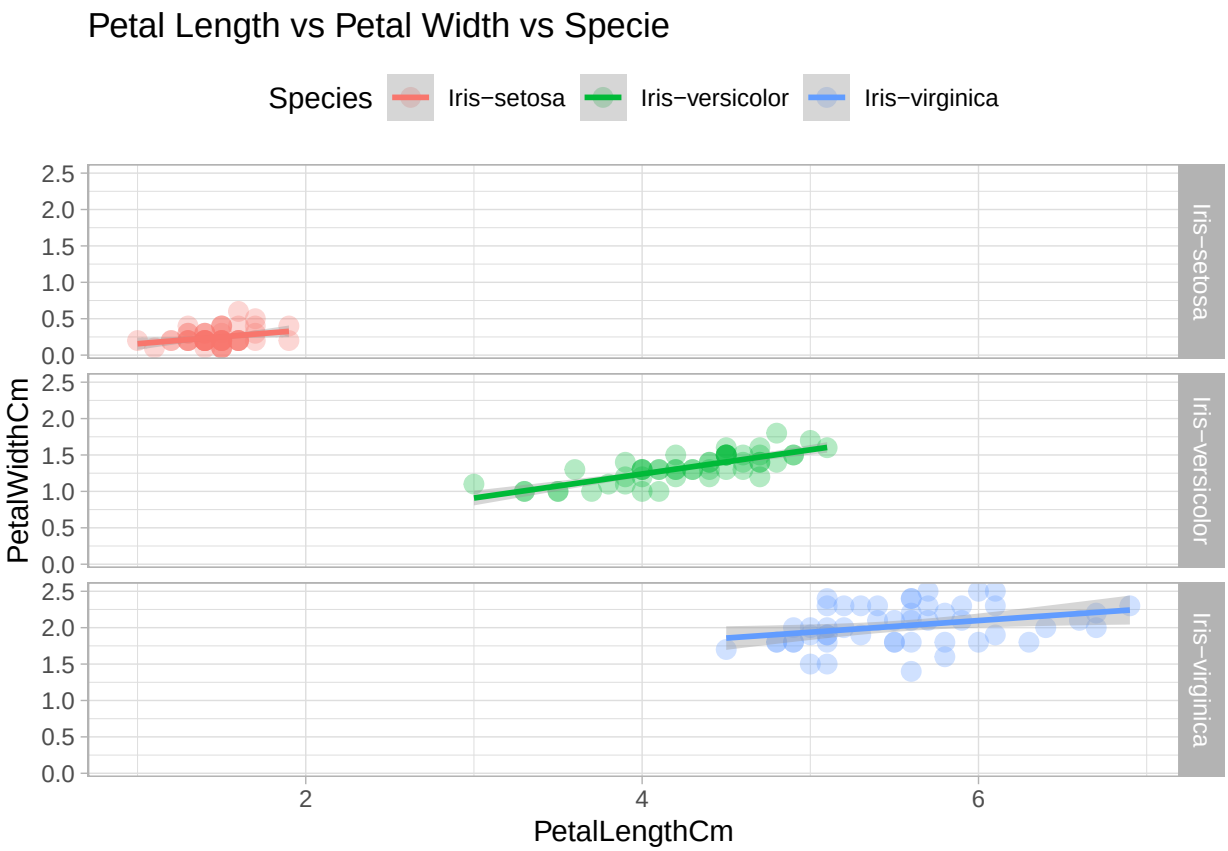


Tanto para la logitud y el grosor del sépalo la especie de flor iris virginica es superior en ambos atributos ya mencionados.

A su vez se aprecia valores fuera de lo normal.

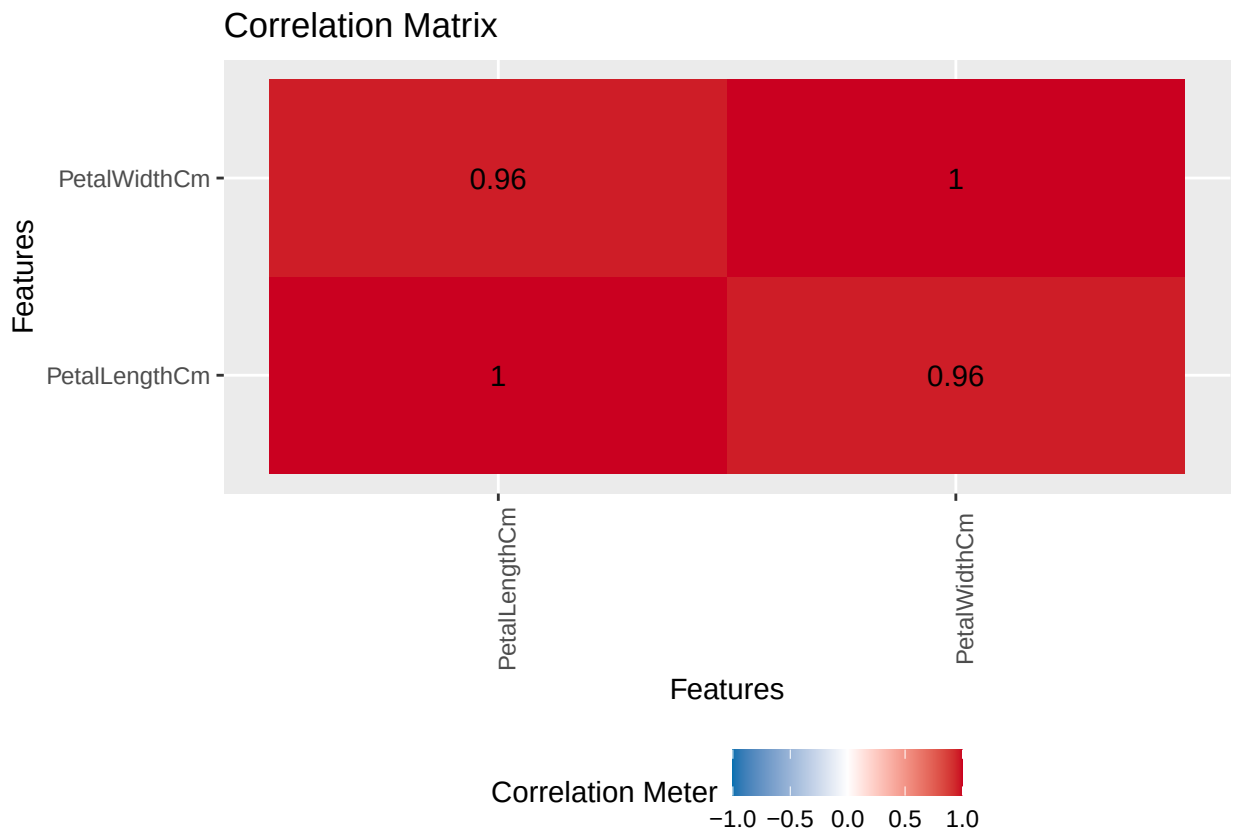
```
ggplot(data=df,aes(x=PetalLengthCm,y=PetalWidthCm,color=Species)) +  geom_point(size=3,alpha=0.3) +
 theme_light() + ggtitle('Petal Length vs Petal Width vs Specie') + facet_grid(Species~.) + geom_smooth
```

```
## `geom_smooth()` using formula 'y ~ x'
```



Hay una clara relación de tendecia lineal entre ambos atributos.

```
plot_correlation(df,type = 'continuous',title = 'Correlation Matrix')
```



Hay un alto grado de coeficiente de relación de Pearson que va de 0 a 1 si es una relación positiva caso contrario de 0 a -1 si.

Ingeniería de carácteristicas

```
iris_splits<-split(df,df$Species)

setosa<-iris_splits[[1]]
versicolor<-iris_splits[[2]]
virginica<-iris_splits[[3]]
```

Creamos pequeños subconjuntos de acuerdo al especie de iris.

```
lower_limit<-function(x){mean(x)-1.5*sd(x)}
upper_limit<-function(x){mean(x)+1.5*sd(x)}
```

Limite inferior y superior longitud del pétalo

```
aggregate(PetalLengthCm~Species,FUN=lower_limit,data=df)
```

```
##           Species PetalLengthCm
```

```
## 1     Iris-setosa      1.203733
## 2 Iris-versicolor      3.555134
## 3  Iris-virginica      4.724158
```

```
aggregate(PetalLengthCm~Species,FUN=upper_limit,data=df)
```

```
##           Species PetalLengthCm
## 1     Iris-setosa      1.724267
## 2 Iris-versicolor      4.964866
## 3  Iris-virginica      6.379842
```

Limite superior grosor del pétalo.

```
aggregate(PetalWidthCm~Species,FUN=upper_limit,data=df)
```

```
##           Species PetalWidthCm
## 1     Iris-setosa    0.4048143
## 2 Iris-versicolor    1.6226290
## 3  Iris-virginica    2.4379751
```

```
min_lim_replace<-function(x,limit){

  return(ifelse(x<limit,sample(x[x>limit],replace = T),x))
}

max_lim_replace<-function(x,limit){


  return(ifelse(x>limit,sample(x[x<limit],replace = T),x))
}
```

Remplazaremos los valores atípicos por valores que estan en un rango normal.

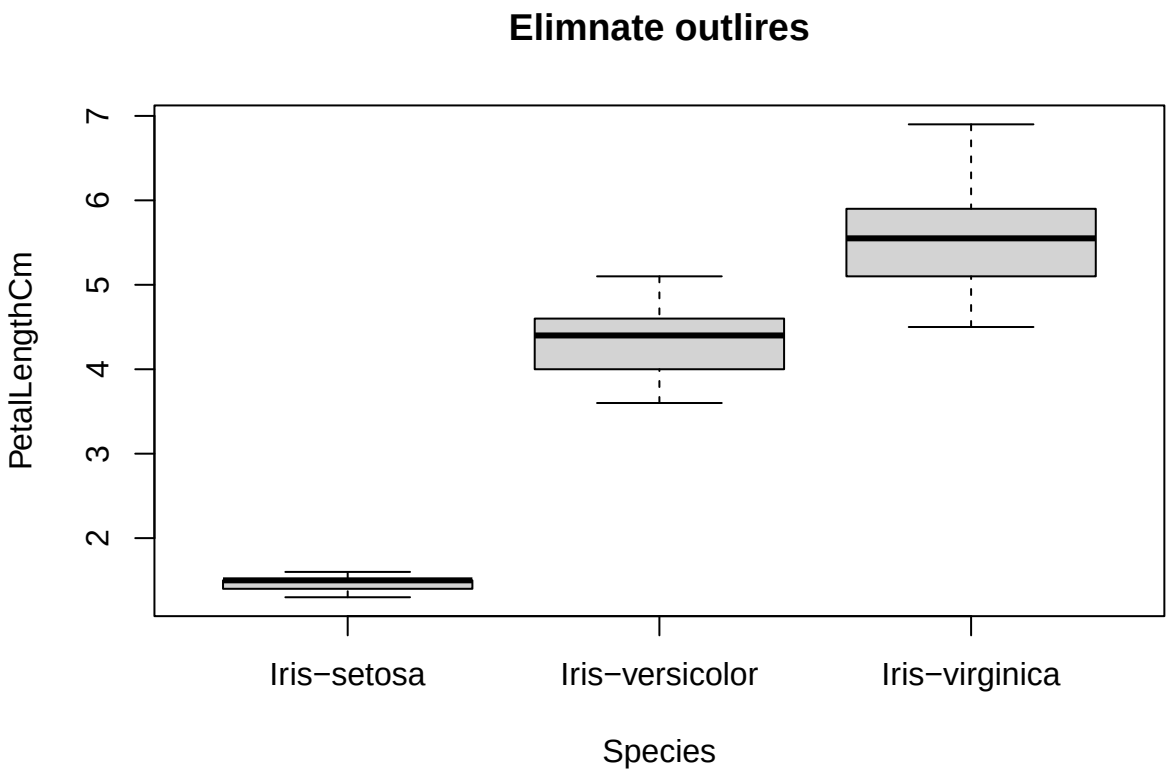
```
setosa$PetalLengthCm<-min_lim_replace(setosa$PetalLengthCm,1.2037)

setosa$PetalLengthCm<-ifelse(setosa$PetalLengthCm>=1.7,
                             sample(setosa$PetalLengthCm[setosa$PetalLengthCm<1.7],
                                    replace = T),
                             setosa$PetalLengthCm)
```

```
versicolor$PetalLengthCm<-min_lim_replace(versicolor$PetalLengthCm,3.555134)
```

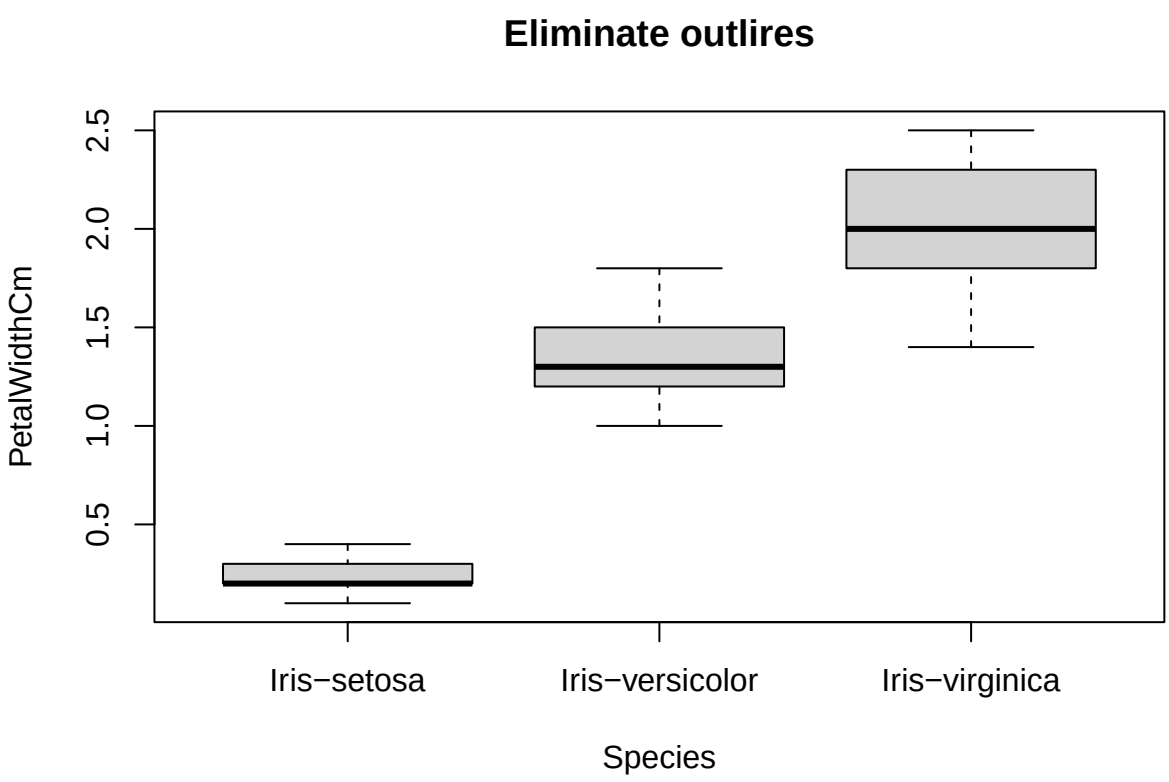
```
setosa$PetalWidthCm<-max_lim_replace(setosa$PetalWidthCm,0.4048143)
```

```
new_df<-rbind.data.frame(setosa,versicolor,virginica)
```

```
boxplot(PetalLengthCm~Species,data=new_df,main='Elimnate outlires')
```




```
boxplot(PetalWidthCm~Species,data=new_df,main='Eliminate outlires')
```



Escalado de los datos

```
new_df$PetalLengthCm<-scale(new_df$PetalLengthCm)
```

Es buena práctica estandarizar los datos con el objetivo de no manejar valores tan grandes pero tampo tan pequeños.

División datos de entrenamiento y validación.

```
set.seed(2018)

training.ids<-createDataPartition(new_df$PetalWidthCm,p=0.7,list=F)

train<-new_df[training.ids,]
test<-new_df[-training.ids,]
```

Creación del modelo.

El mismo modelo crea variables ficticias para las variables tipo factor.

```
lm=lm(PetalWidthCm~.,data=train)

summary(lm)
```

```
##
## Call:
## lm(formula = PetalWidthCm ~ ., data = train)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -0.61611 -0.08357 -0.01088  0.10659  0.47122
##
## Coefficients:
##                          Estimate Std. Error t value Pr(>|t|)
## (Intercept)               0.62925    0.12025   5.233 8.83e-07 ***
## PetalLengthCm             0.30840    0.08809   3.501 0.000687 ***
## SpeciesIris-versicolor    0.59667    0.14995   3.979 0.000129 ***
## SpeciesIris-virginica     1.07165    0.20555   5.214 9.59e-07 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.1857 on 103 degrees of freedom
## Multiple R-squared:  0.941,  Adjusted R-squared:  0.9392
## F-statistic: 547.1 on 3 and 103 DF,  p-value: < 2.2e-16
```

Entre más asteriscos tenga las variables mayor sera el impacto de las variables al valor que tratamos predecir.

```
pred<-(predict(lm,newdata = test))


R2(test$PetalWidthCm,pred)
```

```
## [1] 0.9423872
```

```
test[,c('Predicted values')]<-pred
```

```
test[,c('PetalWidthCm','Predicted values')]
```

```
##     PetalWidthCm Predicted values
## 2            0.2        0.2108804
## 4            0.2        0.2283465
## 8            0.2        0.2283465
## 19           0.3        0.2108804
## 22           0.4        0.2283465
## 27           0.4        0.2458125
## 28           0.2        0.2283465
## 30           0.2        0.2458125
## 33           0.1        0.2283465
## 37           0.2        0.1934143
## 39           0.2        0.1934143
## 43           0.2        0.1934143
## 45           0.4        0.2108804
## 46           0.3        0.2108804
## 47           0.2        0.2458125
## 54           1.3        1.2616705
## 57           1.6        1.3839330
```

```
## 59          1.3        1.3664669
## 61          1.0        1.3490009
## 62          1.5        1.2966027
## 63          1.0        1.2616705
## 68          1.0        1.2791366
## 71          1.8        1.4013991
## 73          1.5        1.4188651
## 75          1.3        1.3140687
## 87          1.5        1.3839330
## 91          1.2        1.3315348
## 92          1.4        1.3664669
## 110         2.5        2.1034364
## 117         1.8        1.9986399
## 118         2.2        2.2082328
## 119         2.3        2.2431649
## 121         2.3        2.0335721
## 123         2.0        2.2082328
## 124         1.8        1.8938435
## 126         1.8        2.0859703
## 127         1.8        1.8763775
## 138         1.8        1.9986399
## 141         2.4        2.0161060
## 142         2.3        1.9287757
## 145         2.5        2.0335721
## 147         1.9        1.9113096
## 148         2.0        1.9462417
```

Hay muy poca variabilidad entre los valores originales y predichos ,cabe resaltar que la unidad con la que se trabajo es centimitros en la mayoria de los casos la diferencia apenas son unos cuantos milimietros.

Guardamos el modelo

```
save(lm,file = 'C:\\Users\\amado\\Desktop\\Blog\\petal_width_lm.Rda')
```