Realizado por:

Juan Camilo Restrepo Velez

Wilder Valencia Ocampo

|  |  |
| --- | --- |
| **Base de datos** | Medicamentos |
| **Descripción del negocio y de los datos** | La rinitis alérgica afecta al 35% de la población en Colombia. Los tratamientos y fármacos son variados por lo que se requiere un estudio para identificar la medicina adecuada para cada persona |
| **Objetivo** | Predecir el medicamento adecuado para 11 personas |

**Tabla de contenido**

[I. (1.0) PREPARACIÓN DE LOS DATOS 4](#_Toc50289194)

[1. Integración de los datos 4](#_Toc50289195)

[2. Eliminar variables irrelevantes y redundantes 4](#_Toc50289196)

[a. ¿Cuáles variables son irrelevantes 4](#_Toc50289197)

[b. ¿Cuáles variables son redundantes 4](#_Toc50289198)

[3. Descripción estadística de los datos 5](#_Toc50289199)

[4. Limpieza de datos 5](#_Toc50289200)

[c. ¿Cuáles son los datos atípicos? 5](#_Toc50289201)

[d. ¿Cuáles son los datos nulos? 6](#_Toc50289202)

[e. Cuál es el valor de la imputación? 7](#_Toc50289203)

[5. Creación de nuevas variables 7](#_Toc50289204)

[6. Análisis de correlaciones 7](#_Toc50289205)

[Matriz de correlaciones 7](#_Toc50289206)

[Correlaciones con la Variable Objetivo 8](#_Toc50289207)

[f. ¿Cuáles variables tienen una alta correlación? 8](#_Toc50289208)

[g. ¿Cuál variable tiene la correlación más alta con la variable objetivo? 8](#_Toc50289209)

[h. ¿Cuál variable tiene la correlación más baja con la variable objetivo? 8](#_Toc50289210)

[7. Reducción de variables 8](#_Toc50289211)

[i. Según el método PCA, ¿cuántos componentes se deben seleccionar para hacer una reducción de variables? 9](#_Toc50289212)

[8. Balanceo de datos 9](#_Toc50289213)

[j. Cuánto es el porcentaje en el balanceo? 9](#_Toc50289214)

[9. Transformación de tipo de datos según el método 10](#_Toc50289215)

[II. (2.0) MODELAMIENTO 12](#_Toc50289216)

[Arboles de decisión 12](#_Toc50289217)

[a. Según el árbol de decisión, ¿cuál es la variable más relevante para hacer la predicción? 12](#_Toc50289218)

[Redes Neuronales 13](#_Toc50289219)

[b. Según la red neuronal, ¿cuántas capas y cuántas neuronas tiene? 14](#_Toc50289220)

[SVM 14](#_Toc50289221)

[c. ¿Cuál kernel tiene el mejor desempeño en la máquina de soporte vectorial? 44](#_Toc50289222)

[Regresión Logística 44](#_Toc50289223)

[d. ¿Cuáles son las ecuaciones de regresión? 44](#_Toc50289224)

[Bayesianos 45](#_Toc50289225)

[e. Según el método NaiveBayes, ¿cuáles son las probabilidades de cada clase? 45](#_Toc50289226)

[KNN 46](#_Toc50289227)

[f. Según el método Knn, ¿con cuántos vecinos se obtiene el mejor resultado? 46](#_Toc50289228)

[III. (1.0) EVALUACIÓN 47](#_Toc50289229)

[Arboles de decisión 47](#_Toc50289230)

[Redes Neuronales 47](#_Toc50289231)

[SVM 48](#_Toc50289232)

[Regresión Logística 52](#_Toc50289233)

[Bayesianos 53](#_Toc50289234)

[KNN 53](#_Toc50289235)

[¿Cuál método obtuvo el mejor resultado? ¿Por qué? 56](#_Toc50289236)

[IV. (1.0) DESPLIEGUE (PREDICCIÓN FUTURA) 57](#_Toc50289237)

[a. ¿Cuál es la predicción de cada método? 57](#_Toc50289238)

[Arboles de decisión 57](#_Toc50289239)

[Redes Neuronales 57](#_Toc50289240)

[SVM 57](#_Toc50289241)

[Regresión Logística 58](#_Toc50289242)

[Bayesianos 59](#_Toc50289243)

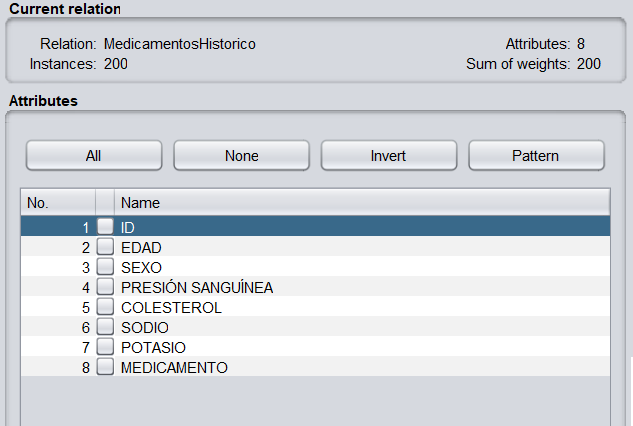
[KNN 59](#_Toc50289244)

[b. ¿Cuál método es más confiable en la predicción? ¿Por qué? 59](#_Toc50289245)

# I. (1.0) PREPARACIÓN DE LOS DATOS

## 1. Integración de los datos

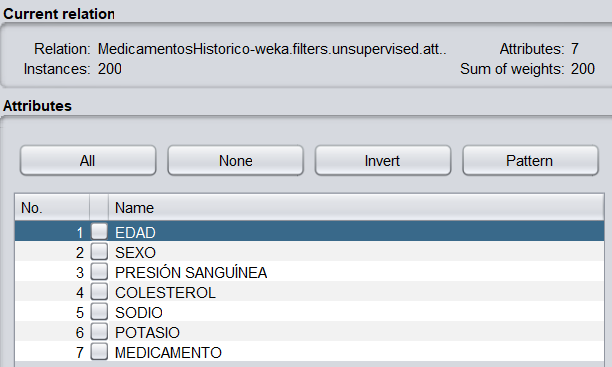
Los datos ya se encuentran integrados en una ‘sabana de datos’



## 2. Eliminar variables irrelevantes y redundantes

### a. ¿Cuáles variables son irrelevantes

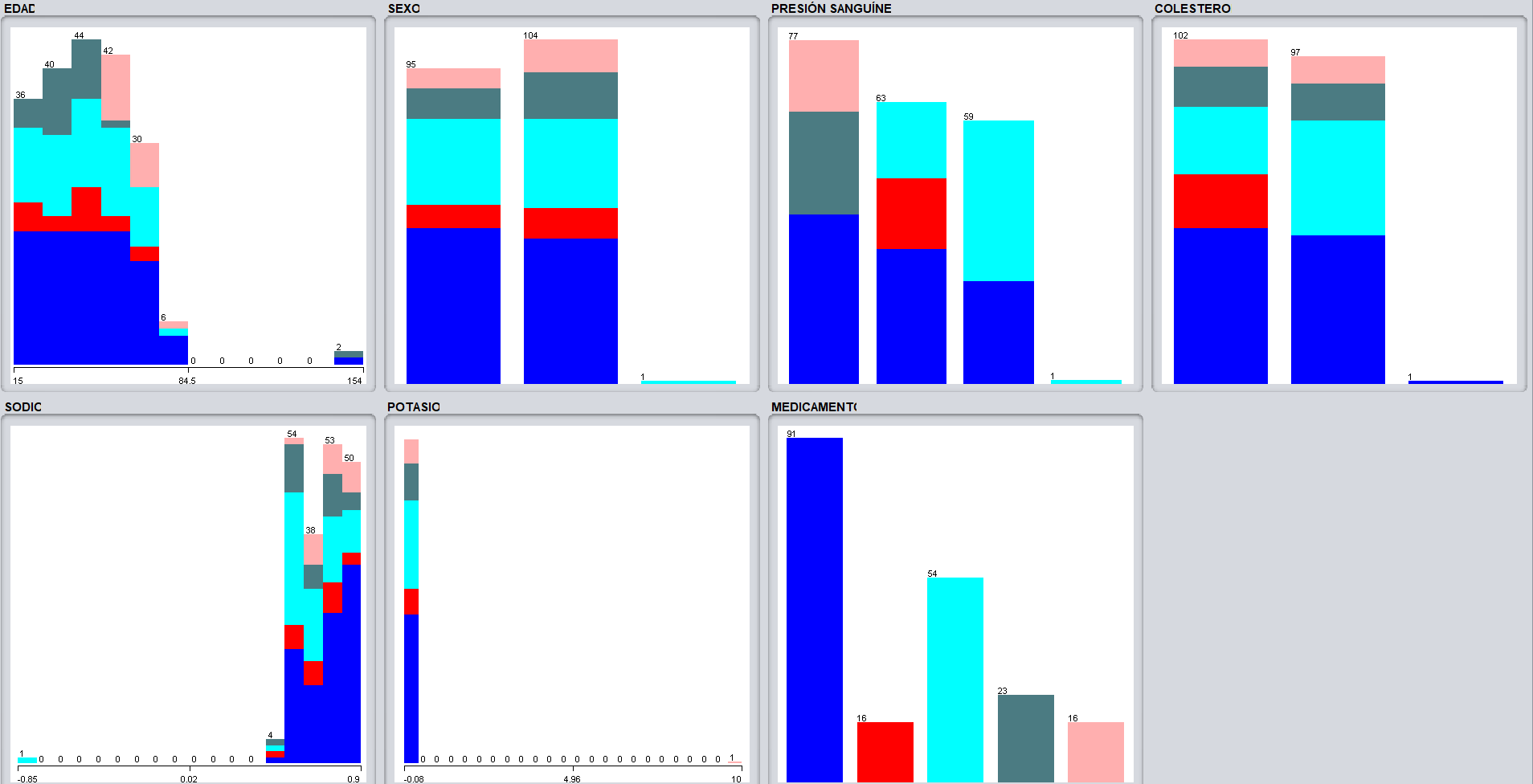
Desde el análisis de negocio y de los datos se concluye que la variable ‘ID’ no es relevante, por lo tanto, se elimina



### b. ¿Cuáles variables son redundantes

No se encuentran variables redundantes a simple vista, esto será comprobado con el paso 6 “Análisis de correlaciones”

## 3. Descripción estadística de los datos



Se encuentran algunas irregularidades en ‘EDAD’, ‘SEXO’, ‘PRESIÓN SANGUÍNEA’, ‘COLESTEROL’, ‘SODIO’, ‘POTASIO’ las cuales se proceden a corregir en el siguiente paso “Limpieza de datos”.

## 4. Limpieza de datos

No se presentan registros duplicados

### c. ¿Cuáles son los datos atípicos?

|  |  |
| --- | --- |
| **Variable** | **Descripción** |
| EDAD | Se encuentran 2 registros con 2 edades que se salen de cualquier registro humano, estos son ‘145’ y ‘154’ años, por lo tanto, se procede a vaciar estos campos (**null\***) |
| SEXO | Existen 2 categorías (con posibilidad de una tercera) y aparecen 3, sin embargo, se observa que es un error de digitación y se corrige. |
| PRESIÓN SANGUÍNEA | Existen 3 categorías (HIGH, LOW, NORMAL), sin embargo, aparece una cuarta (Null), se procede eliminando la categoría ‘Null’ y poniendo dicho valor en vacío(**null\***) |
| COLESTEROL | Existen 2 categorías (HIGH, NORMAL), aparece una tercera denominada ‘hig’, similar a la categoría ‘HIGH’, por lo que se asume es un error de digitación, adicional a ello se toma a esa decisión, pues si bien se pusiera en null, el método “ReplaceMissingValues” procedería a remplazar con la moda que para este caso es “HIGH”. Adicional se elimina esta categoría (hig). |
| SODIO | Un registro presenta un “SODIO” negativo, lo cual no es posible (valores negativos), por lo tanto, este campo se pone como vacío para este registro (**null\***) |
| POTASIO | Dos registros presentan “POTASIO” negativo, lo cual no es posible (valores negativos), por lo tanto, este campo se pone como vacío para estos registros (**null\***).    Otro registro presenta un “POTASIO” muy alto (10), lo cual se aleja demasiado de los datos normales (0,099), por lo tanto, este campo se pone como vacío para este registro (**null\***). |

### d. ¿Cuáles son los datos nulos?

Solo se Presentan datos Nulos como consecuencia del paso anterior “datos atípicos”, marcados con “(null\*)”, a continuación, se describen.

|  |  |  |
| --- | --- | --- |
| **Variable** | **Descripción** | **¿Se Imputa?** |
| EDAD | Se encuentran 2 registros que corresponde al 1% | SI |
| PRESIÓN SANGUÍNEA | Se encuentran 1 registro que corresponde al 1% | SI |
| SODIO | Se encuentran 1 registro que corresponde al 1% | SI |
| POTASIO | Se encuentran 3 registros que corresponde al 2% | SI |

### e. Cuál es el valor de la imputación?

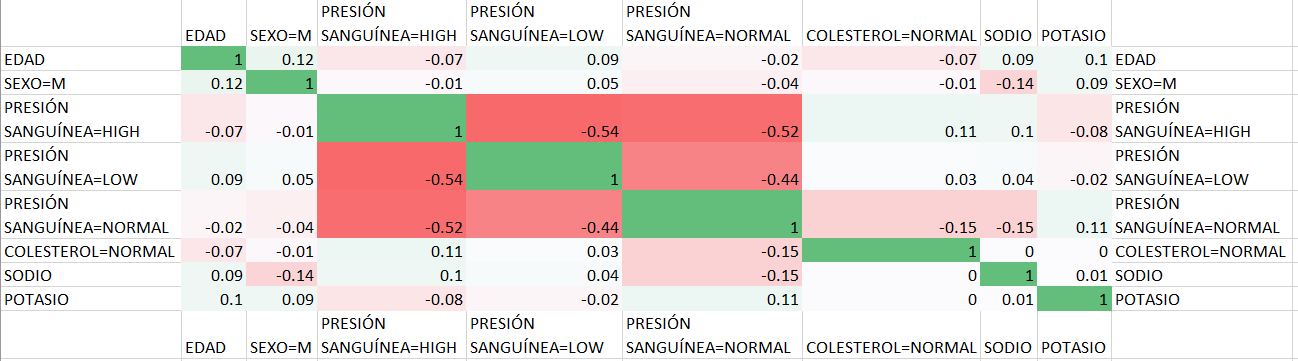
|  |  |  |
| --- | --- | --- |
| **Variable** | **Cantidad Nulos** | **Valor Imputación** |
| EDAD | 2 | 44,535 |
| PRESIÓN SANGUÍNEA | 1 | HIGH |
| SODIO | 1 | 0,696 |
| POTASIO | 3 | 0,05 |

## 5. Creación de nuevas variables

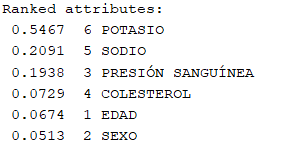
Para este ejercicio no existe la necesidad de crear nuevas variables, pues no hay presentes fechas u otras variables derivadas del negocio.

## 6. Análisis de correlaciones

### Matriz de correlaciones



### Correlaciones con la Variable Objetivo



### f. ¿Cuáles variables tienen una alta correlación?

Las variables que presentan una alta correlación son

* ‘PRESIÓN SANGUÍNEA=HIGH’ con ‘PRESIÓN SANGUÍNEA=LOW’ presentan una correlación de 0,54
* ‘PRESIÓN SANGUÍNEA=HIGH’ con ‘PRESIÓN SANGUÍNEA=NORMAL’ presentan una correlación de 0,52
* ‘PRESIÓN SANGUÍNEA=NORMAL’ con ‘PRESIÓN SANGUÍNEA=LOW’ presentan una correlación de 0,44

Sin embargo, ninguna presenta una correlación que haga pensar que hay presencia de variables redundantes.

### g. ¿Cuál variable tiene la correlación más alta con la variable objetivo?

La variable que presenta mayor correlación con la variable objetivo es ‘POTASIO’ equivalente a 0,5467.

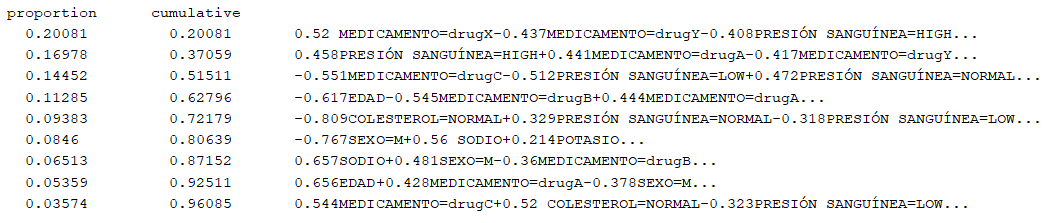
### h. ¿Cuál variable tiene la correlación más baja con la variable objetivo?

La variable que presenta menor correlación con la variable objetivo es ‘SEXO’ equivalente a 0,0513.

Después de este análisis se concluye que **no hay presencia de variables redundantes** debido a que la matriz de correlaciones así lo evidencia. Adicionalmente se observa que si bien algunas variables poseen muy baja correlación con la variable objetivo (SEXO, EDAD, COLESTEROL), estas **no serán eliminadas** debido a que se tienen muy pocas variables y prescindir de alguna de ellas podría conllevar a eliminar información valiosa.

## 7. Reducción de variables

Con una configuración de máximo 3 variables en cada ecuación se obtuvieron las siguientes:

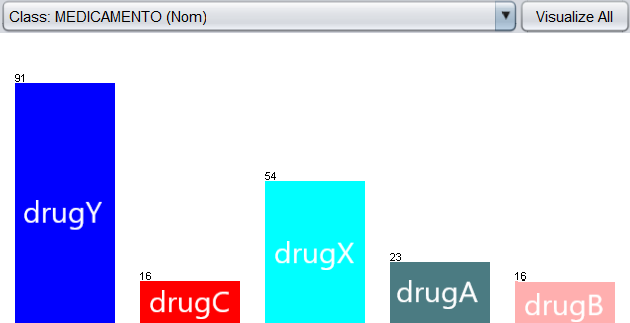


### i. Según el método PCA, ¿cuántos componentes se deben seleccionar para hacer una reducción de variables?

Se deben seleccionar mínimo 6 componentes debido a que se debe cubrir como mínimo el 75% de la varianza de los datos originales, y con estos 6 componentes se cubre el 80% de la varianza.

## 8. Balanceo de datos

El estado Pre-Balanceo es:



### j. Cuánto es el porcentaje en el balanceo?

|  |  |  |  |
| --- | --- | --- | --- |
| **% de Balanceo** | **Categoría Referencia** | **Categoría** | **Aumento** |
| 50% | drugY | drugC | 184,37 % |
| drugX | N/A |
| drugA | 97,82 % |
| drugB | 184,37 % |
| 75% | drugY | drugC | 326,56 % |
| drugX | 26,38 % |
| drugA | 196,73 % |
| drugB | 326,56 % |
| 100% | drugY | drugC | 468,75 % |
| drugX | 68,51 % |
| drugA | 295,65 % |
| drugB | 468,75 % |

Para el desarrollo del ejercicio se toma la decisión de realizar un balanceo al 100%, pues si bien para algunas categorías el % de aumento es alto, este no es mayor a 500% en ninguna de las categorías.

## 9. Transformación de tipo de datos según el método

Teniendo en consideración de la clasificación natural de las variables es la siguiente:

* EDAD à variable Numérica
* SEXO à variable Categórica
* PRESIÓN SANGUÍNEA à variable Categórica
* COLESTEROL à variable Categórica
* SODIO à variable Numérica
* POTASIO à variable Numérica
* MEDICAMENTO (variable objetivo) à variable Categórica

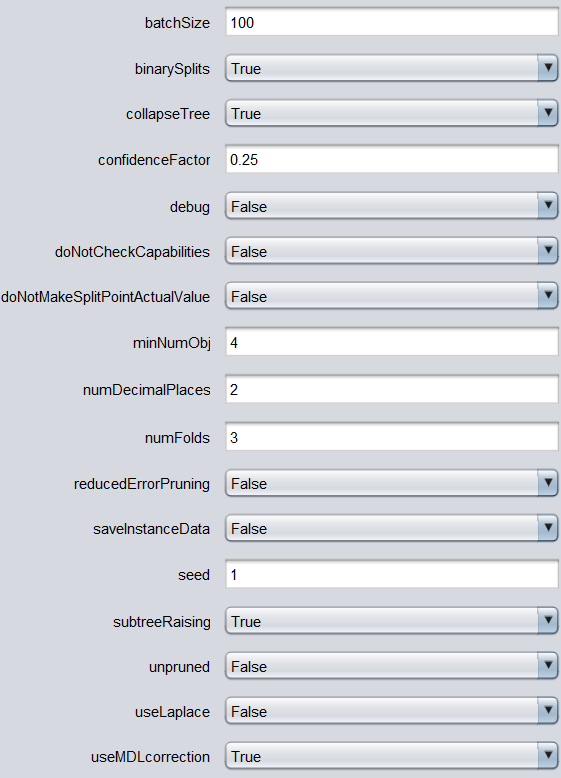
Según el método las transformaciones necesarias, serian:

|  |  |  |
| --- | --- | --- |
| **Método** | **Variables que transformar a categóricas** | **Variables que transformar a numéricas** |
| Arboles de decisión | EDAD, SODIO, POTASIO | N/A |
| Redes Neuronales | N/A | SEXO, PRESIÓN SANGUÍNEA, COLESTEROL |
| SVM | N/A | SEXO, PRESIÓN SANGUÍNEA, COLESTEROL |
| Regresión Logística | N/A | SEXO, PRESIÓN SANGUÍNEA, COLESTEROL |
| Bayesianos | EDAD, SODIO, POTASIO | N/A |
| KNN | N/A | SEXO, PRESIÓN SANGUÍNEA, COLESTEROL |

# II. (2.0) MODELAMIENTO

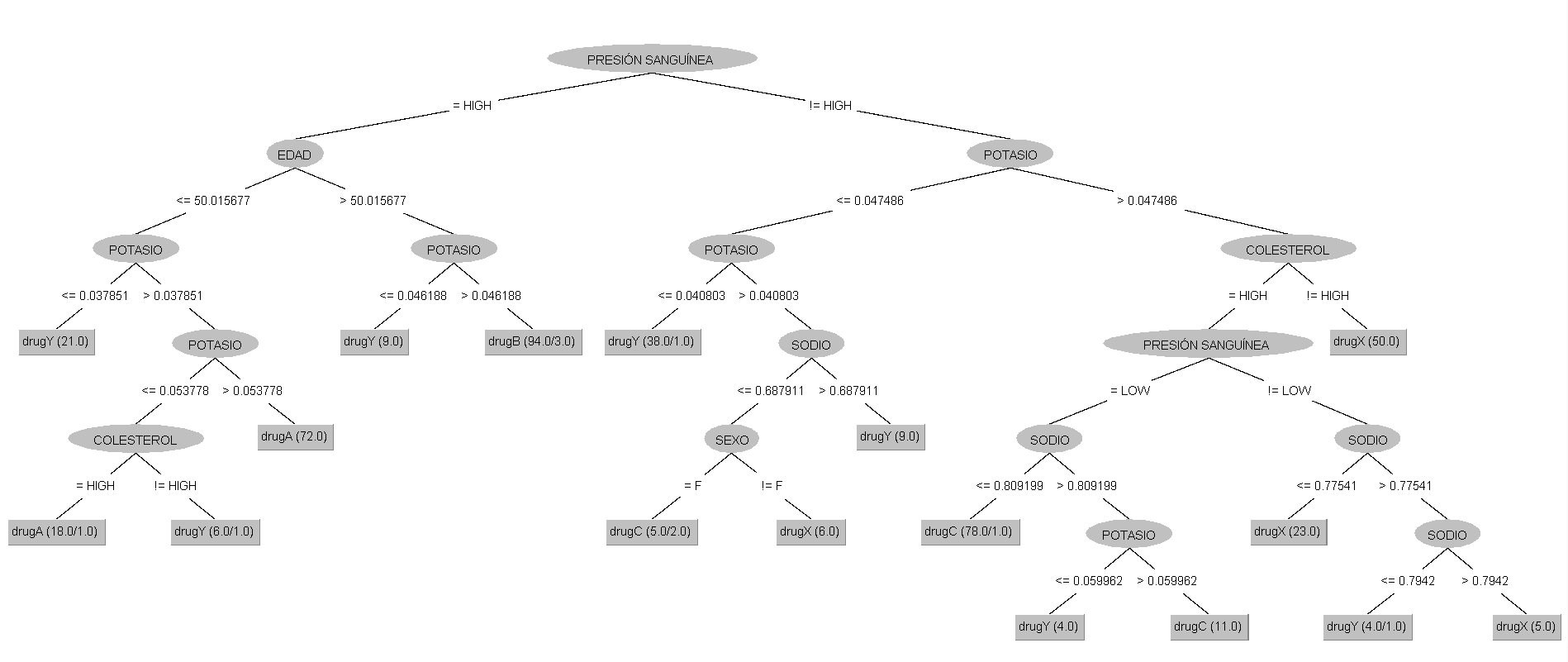
## Arboles de decisión

Se configura un árbol binario, con un ‘minNumObj’ de 4, habilitada la poda y demás configuraciones como se muestra a continuación.



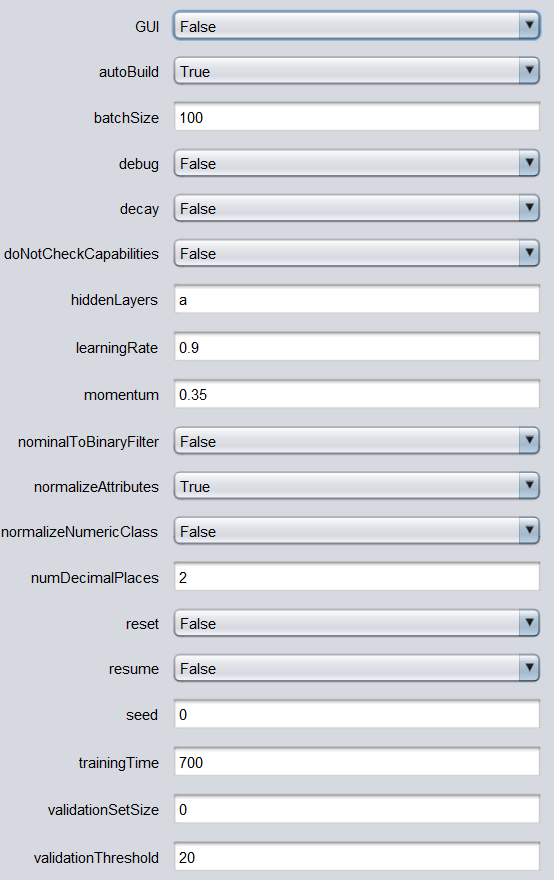
### a. Según el árbol de decisión, ¿cuál es la variable más relevante para hacer la predicción?

Se evidencia por medio de la vista grafica del árbol que la variable más relevante es ‘**PRESIÓN SANGUÍNEA**’ para el desarrollo de la predicción.



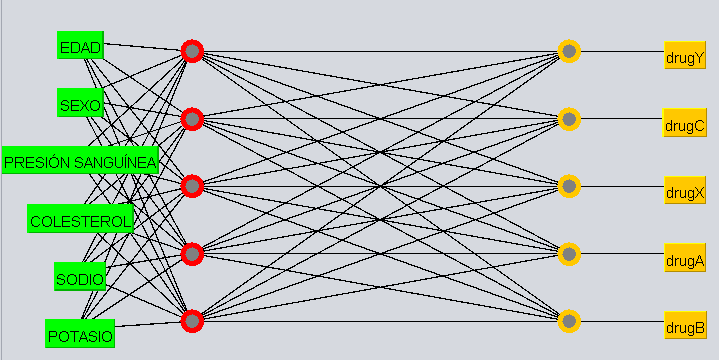
## Redes Neuronales

Se realiza una configuración de MLP con una capa oculta que tiene 5 neuronas (a), con una tasa de aprendizaje de 0,9 , momentum de 0,35 , 700 iteraciones para que no sufra sobre entrenamiento y demás configuraciones a continuación:



### b. Según la red neuronal, ¿cuántas capas y cuántas neuronas tiene?

La red neuronal configurada cuenta con una capa oculta y en esta 5 neuronas, como se observa a continuación



## SVM

* **Normalized Poly**: Se obtuvieron las siguientes ecuaciones de los hiperplanos que separan a dos clases

|  |  |
| --- | --- |
| **Clases** | **Ecuación** |
| drugY, drugC |  |
| drugY, drugX |  |
| drugY, drugA |  |
| drugY, drugB |  |
| drugC, drugX |  |
| drugC, drugA |  |
| drugC, drugB |  |
| drugX, drugA |  |
| drugX, drugB |  |
| drugA, drugB |  |

* **Poly**: Se obtuvieron las siguientes ecuaciones de los hiperplanos que separan a dos clases

|  |  |
| --- | --- |
| **Clases** | **Ecuación** |
| drugY, drugC |  |
| drugY, drugX |  |
| drugY, drugA |  |
| drugY, drugB |  |
| drugC, drugX |  |
| drugC, drugA |  |
| drugC, drugB |  |
| drugX, drugA |  |
| drugX, drugB |  |
| drugA, drugB |  |

* **Puk**: Se obtuvieron las siguientes ecuaciones de los hiperplanos que separan a dos clases

|  |  |
| --- | --- |
| **Clases** | **Ecuación** |
| drugY, drugC |  |
| drugY, drugX |  |
| drugY, drugA |  |
| drugY, drugB |  |
| drugC, drugX |  |
| drugC, drugA |  |
| drugC, drugB |  |
| drugX, drugA |  |
| drugX, drugB |  |
| drugA, drugB |  |

* **RBF**: Se obtuvieron las siguientes ecuaciones de los hiperplanos que separan a dos clases

|  |  |
| --- | --- |
| **Clases** | **Ecuación** |
| drugY, drugC |  |
| drugY, drugX |  |
| drugY, drugA |  |
| drugY, drugB |  |
| drugC, drugX |  |
| drugC, drugA |  |
| drugC, drugB |  |
| drugX, drugA |  |
| drugX, drugB |  |
| drugA, drugB |  |

### c. ¿Cuál kernel tiene el mejor desempeño en la máquina de soporte vectorial?

Según lo analizado en el apartado siguiente (Evaluación), se determinó que el kernel Poly presenta un mejor desempeño que todos los demás.

## Regresión Logística

### d. ¿Cuáles son las ecuaciones de regresión?

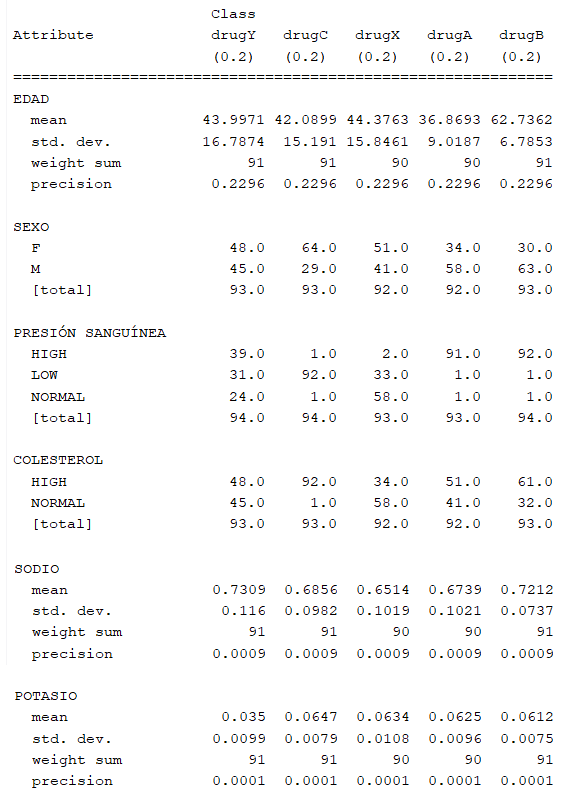
Se realiza el modelamiento por medio de división 70-30 y validación cruzada, donde se crea el modelo con siguientes ecuaciones por clase.

|  |  |
| --- | --- |
| **Clase** | **Ecuación** |
| drugY |  |
| drugC |  |
| drugX |  |
| drugA |  |
| drugB |  |

## Bayesianos

### e. Según el método NaiveBayes, ¿cuáles son las probabilidades de cada clase?

Según el modelo creado, cada clase tiene un 20% de probabilidad



## KNN

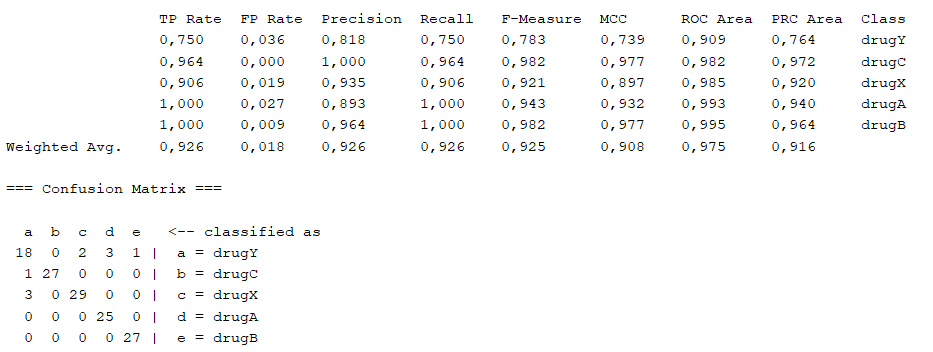
### f. Según el método Knn, ¿con cuántos vecinos se obtiene el mejor resultado?

Según lo analizado en el apartado siguiente (Evaluación), se determinó que con 3 vecinos se obtiene mejor resultado en las medidas de evaluación.

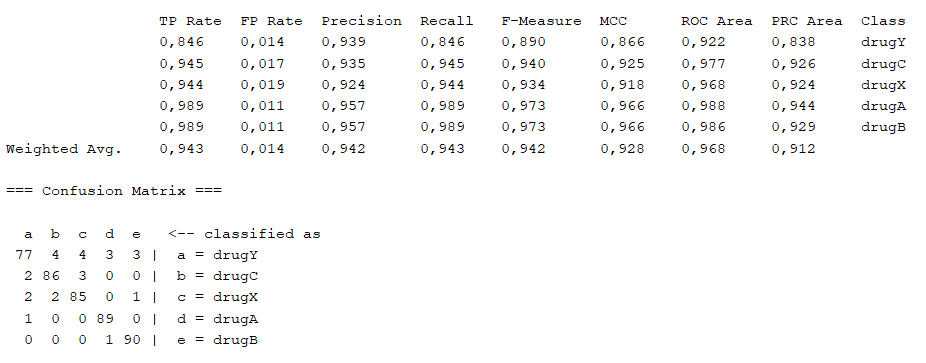
# III. (1.0) EVALUACIÓN

## Arboles de decisión

* **División 70-30**



* **Validación cruzada**



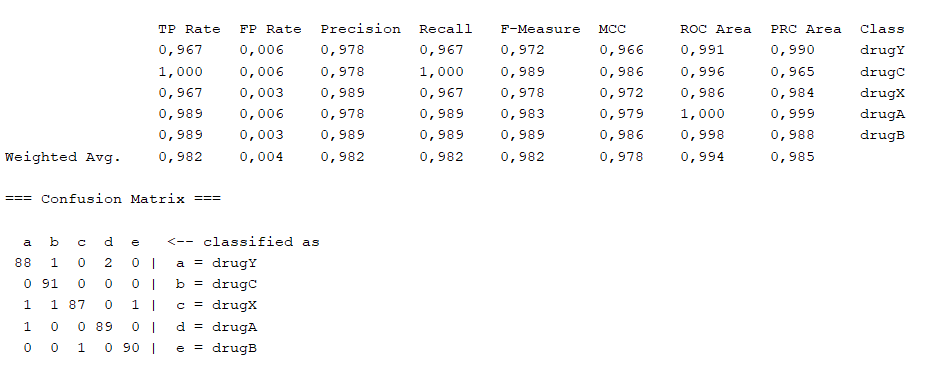
Al mantenerse un área ROC buena por medio de la validación cruzada, que cumple con el diseño de experimentos, se considera que es un modelo aceptable.

## Redes Neuronales

* **División 70-30**

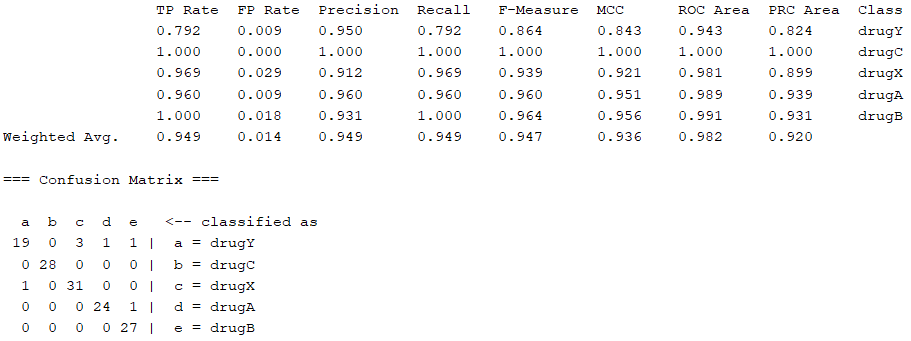
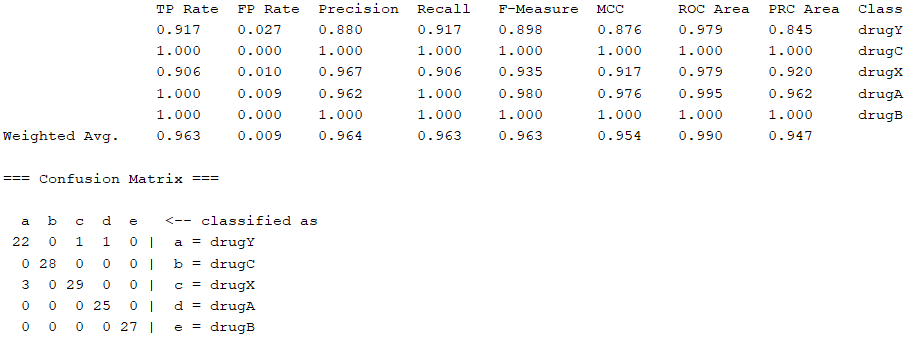
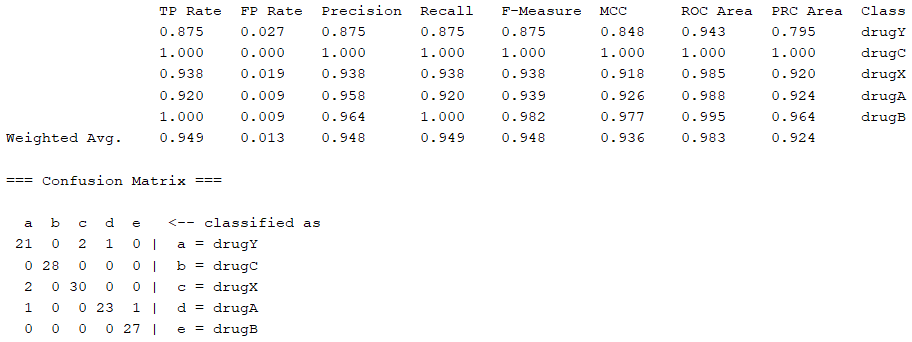
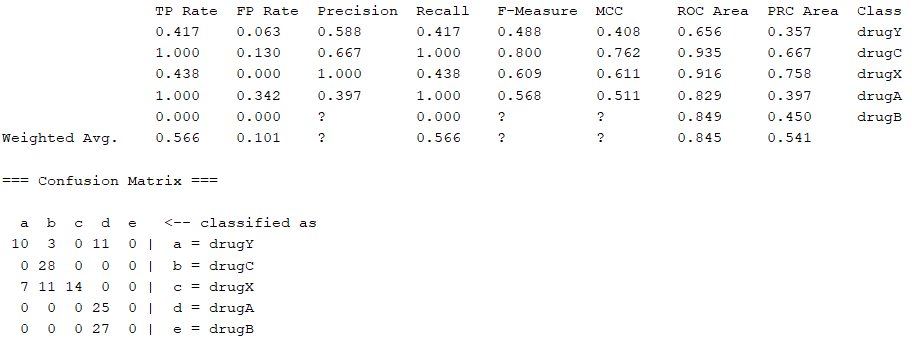
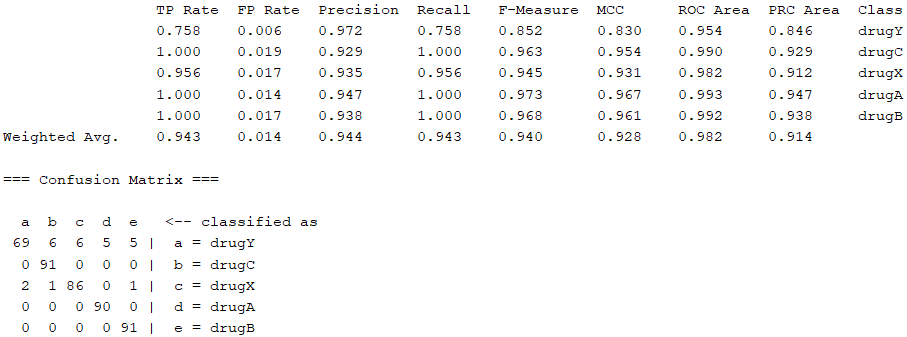
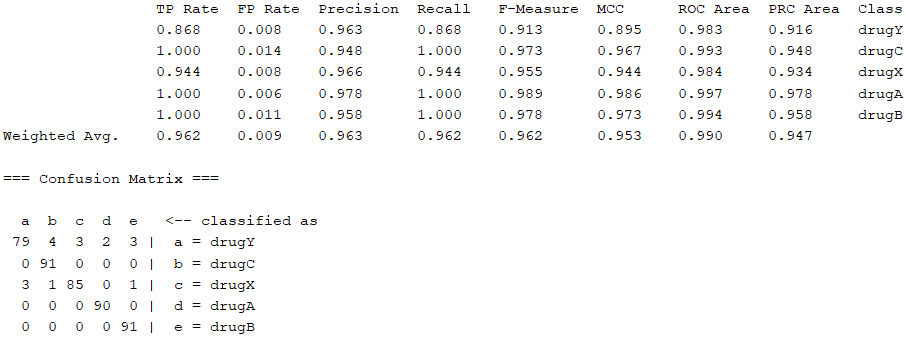
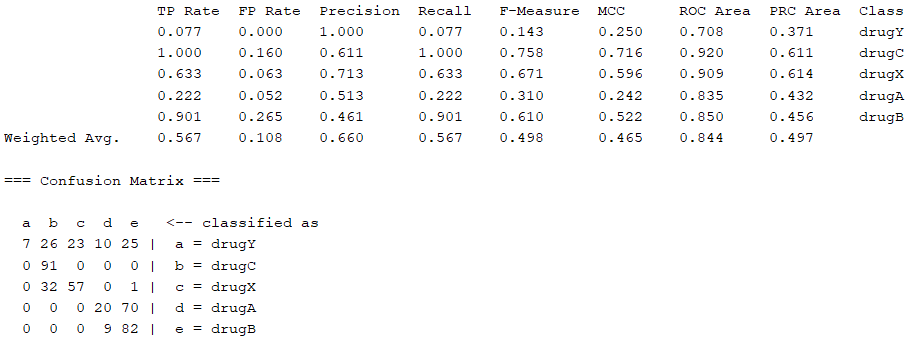


* **Validación cruzada**



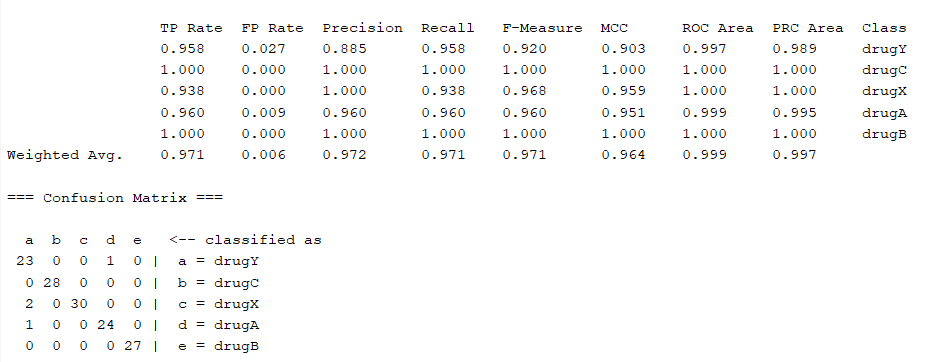
Al mantenerse un área ROC buena por medio de la validación cruzada, que cumple con el diseño de experimentos, se considera que es un modelo aceptable. Sin embargo con división 70-30 se logra un área ROC ideal.

## SVM

* **División 70-30**  
  *Normalized Poly*  
    
  *Poly*  
    
  *Puk*  
    
  *RBF*  
  
* **Validación cruzada**  
  *Normalized Poly*  
    
  *Poly*  
    
  *Puk*  
    
  *RBF*  
  

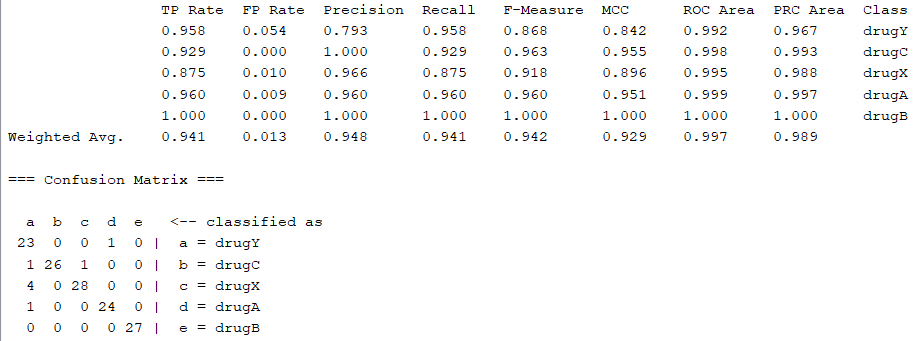
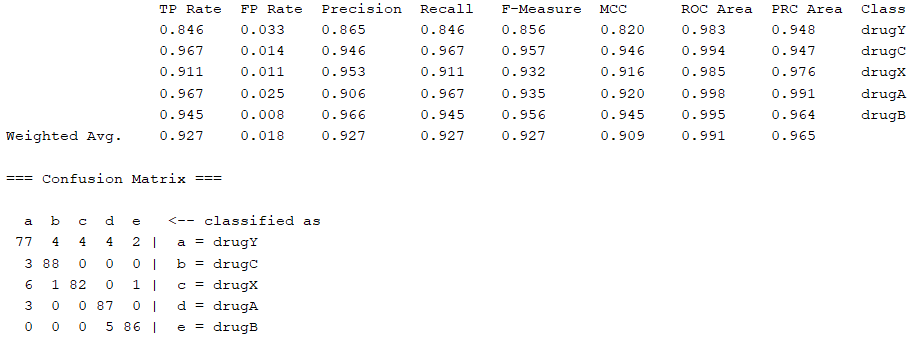
Según las medidas de evaluación que se obtuvieron de los diferentes kernels por medio de las dos divisiones de datos, en ambas el kernel *Poly* presentó el mejor desempeño con un área ROC de 0.99, por lo que dicho kernel se dejará para la predicción futura.

## Regresión Logística

* **División 70-30**  
  
* **Validación cruzada**  
  

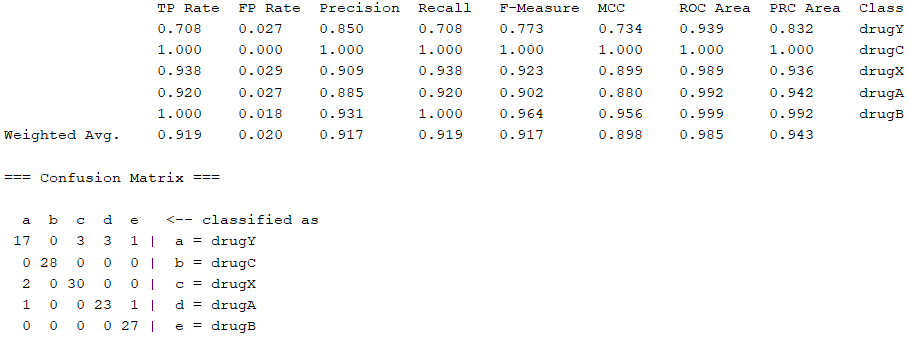
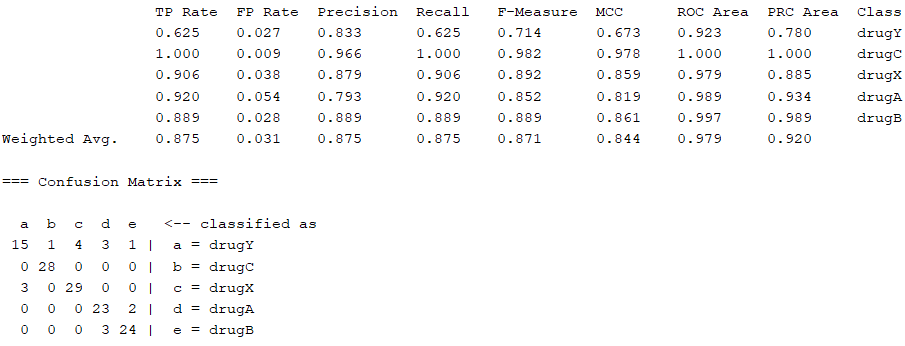
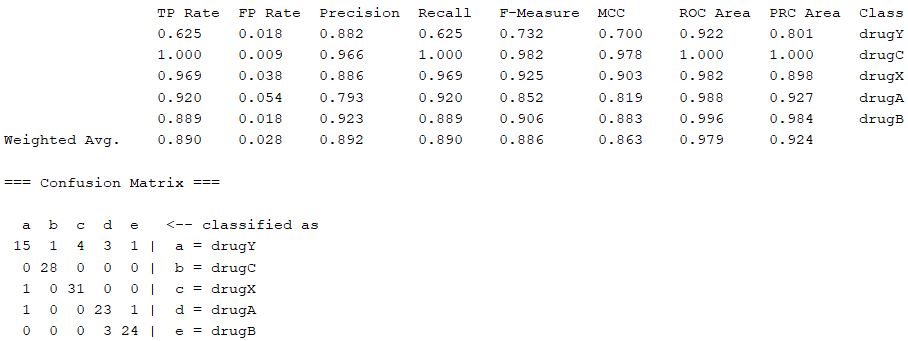
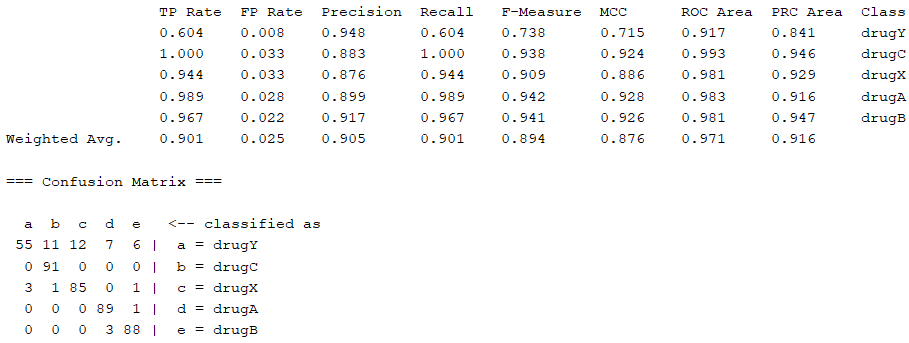
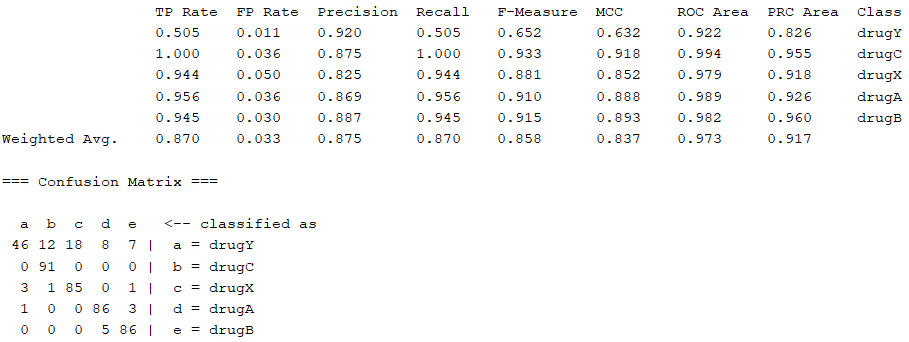
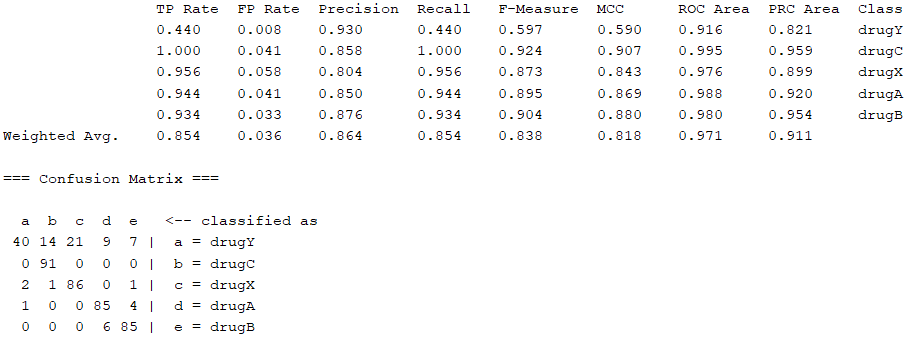
Al mantenerse un área ROC buena por medio de la validación cruzada, que cumple con el diseño de experimentos, se considera que es un modelo aceptable.

## Bayesianos

* **División 70-30**  
  
* **Validación cruzada**  
  

Al mantenerse un área ROC buena por medio de la validación cruzada, que cumple con el diseño de experimentos, se considera que es un modelo aceptable.

## KNN

* **División 70-30**  
  *3 vecinos*  
    
  *5 vecinos*  
    
  *7 vecinos*  
  
* **Validación cruzada**  
  *3 vecinos*  
    
  *5 vecinos*  
    
  *7 vecinos*  
  

Al analizar las diferentes evaluaciones se decide dejar el modelo con 3 vecinos, debido a que con la división 70 – 30 fue el que mejor resultado dio y en cuanto a la validación cruzada fue despreciablemente inferior en el área ROC, pero en las demás medidas dio mejor resultado.

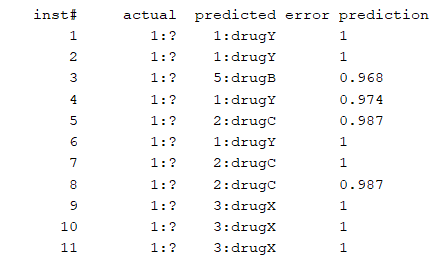
## ¿Cuál método obtuvo el mejor resultado? ¿Por qué?

A través del análisis realizado, se observa que el modelo obtenido con redes Neuronales obtiene un área ROC de 1 con división 70-30, además de eso, valores de ‘Precision’, ‘Recall’ ‘TP rate’ mayores a 0,97 e incluso se mantienen unas buenas medidas por medio de la validación cruzada, por lo tanto, es el método que mejor resultado presenta.

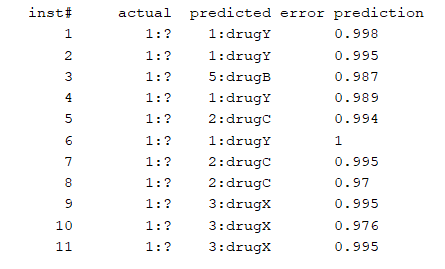
# IV. (1.0) DESPLIEGUE (PREDICCIÓN FUTURA)

## a. ¿Cuál es la predicción de cada método?

### Arboles de decisión

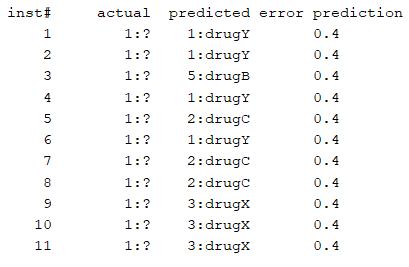


### Redes Neuronales

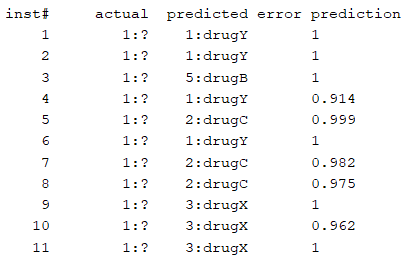


### SVM

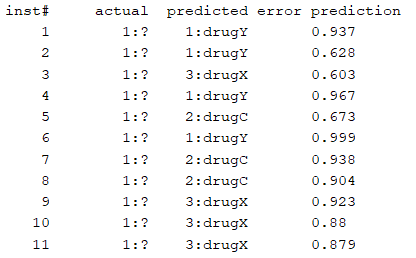
Se realiza la predicción por medio del kernel Poly



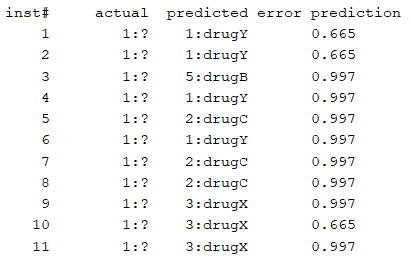
### Regresión Logística



### Bayesianos



### KNN



## b. ¿Cuál método es más confiable en la predicción? ¿Por qué?

En primer lugar, es valioso recopilar/resumir la información obtenida en los pasos desarrollados

|  |  |  |  |
| --- | --- | --- | --- |
| Método | ROC à 70-30 | ROC à VC | Predicción à min(‘prediction’) |
| Arboles de decisión | 0,975 | 0,968 | 0,968 |
| Redes Neuronales | 1 | 0,994 | 0,970 |
| SVM (*poly*) | 0,990 | 0,990 | 0,400 |
| Regresión Logística | 0,999 | 0,996 | 0,914 |
| Bayesianos | 0,997 | 0,991 | 0,603 |
| KNN (*3 vecinos*) | 0,985 | 0,971 | 0,665 |

Con esto es posible afirmar que con los 6 métodos se llegaron a modelos que cumplen con el requisito de área ROC para el área de la salud (>95%). Sin embargo, en la predicción se observan resultados muy interesantes pues varios de estos modelos presentan una confianza de predicción (‘prediction’) bajas o muy bajas en alguna o varas de sus predicciones como: SVM (0,400), Bayesianos (0,603), KNN (0,665) y regresión logística (0,914). En cuanto a la predicción futura estos 4 modelos presentan los mismos resultados excepto en los resultados del método de Bayes, donde para el tercer registro se predijo la droga “drugX” mientras que en los otros fue la droga “drugB”.

Por otro lado, se hallan dos modelos que presentan muy buenos resultados tanto en sus medidas ROC y sus predicciones. El modelo obtenido con árbol de decisión presenta una ROC de 0,975 y su ‘predition’ más bajo es de 0,968. Por su parte el modelo obtenido con redes neuronales (MLP) es de 1 para el área ROC y el ‘predition’ más bajo es de 0,970. Respecto a las predicciones ambos predicen las mismas drogas para los mismos pacientes con confianza de predicción altas (>95%), sin embargo, el modelo de redes neuronales presenta una mayor confianza, por lo tanto, es más confiable y seguro.

En conclusión, si bien tanto el modelo obtenido con árboles y el modelo obtenido con redes neuronales cumplen con las características necesarias para el área de la salud, el modelo que se obtuvo con **Redes neuronales** presenta mejores resultados y por lo tanto es el modelo mas confiable en la predicción.