MODELADO PREDICTIVO DE LA CLASIFICACIÓN DEL CÁNCER DE MAMA: UN ANÁLISIS COMPARATIVO DE TÉCNICAS DE APRENDIZAJE AUTOMÁTICO EN R STUDIO

Integrante: María Paz Santibáñez

Planteamiento del Problema, Objetivo General y Objetivos Específicos

El desafío que se enfrenta es un problema de clasificación binaria, donde el objetivo consiste en predecir si un tumor mamario es benigno o maligno, fundamentándose en diversas características del núcleo celular. El objetivo principal de la investigación es construir y comparar modelos predictivos utilizando varios algoritmos de aprendizaje automático. Los objetivos específicos incluyen el desarrollo de modelos predictivos mediante regresión logística, k-vecinos más cercanos (KNN), árboles de decisión, bosques aleatorios y máquinas de vectores de soporte (SVM). Posteriormente, se procederá a evaluar y comparar el rendimiento de estos modelos utilizando métricas de clasificación y curvas ROC.

Descripción de los Métodos Estadísticos

Se llevará a cabo un análisis descriptivo de las variables seleccionadas, incluyendo medidas de tendencia central y dispersión. A partir de los datos de entrenamiento, se construirán modelos predictivos utilizando varios algoritmos de aprendizaje automático, tales como regresión logística, KNN, árboles de decisión, bosques aleatorios y SVM. Con el fin de garantizar la robustez de los modelos, se implementará una validación cruzada. Cada modelo será evaluado en función de su capacidad para clasificar adecuadamente los casos de cáncer, utilizando métricas de clasificación como exactitud, precisión, sensibilidad y F1 score. Finalmente, se seleccionará el modelo con el mejor rendimiento como el predictor óptimo, basándonos en métricas de clasificación y curvas ROC.

Descripción De Los Datos

La variable dependiente es "Diagnóstico", la cual indica si el cáncer es maligno (M) o benigno (B). Las variables independientes comprenden diez características calculadas para cada núcleo celular, abarcando radio, textura, perímetro, área, suavidad, compacidad, concavidad, puntos cóncavos, simetría y dimensión fractal. Cada característica se presenta en tres dimensiones: media, error estándar y peor.

Los datos consisten en 569 observaciones y 31 variables, generando un total de 17.639 observaciones. Estos datos fueron obtenidos del repositorio de aprendizaje automático de UCI y también están disponibles en Kaggle.

Descripción, Interpretación y Análisis de los Resultados

A continuación, se describen y se interpretan los resultados obtenidos en el proceso de modelado y análisis de los datos utilizando las herramientas previamente mencionadas. En cada caso, se están creando muestras de entrenamiento y validación, de manera que los clasificadores se construyen con los datos de entrenamiento. Posteriormente, se evalúa el poder predictivo de cada uno con los datos de validación.

Preprocesamiento de Datos

No existen datos nulos en la información proporcionada. Las variables ID y X no son consideradas en el análisis. La variable ID, que es una variable entera que identifica a los pacientes, se omite para los fines de nuestra investigación. Además, la variable X, una variable lógica incompatible con los procedimientos a realizar a continuación, se excluye.

Los datos se estandarizan mediante la función 'scale()'. Además, la variable dependiente 'diagnosis' se convierte en un factor utilizando la función 'as.factor'. Ambos procedimientos se combinan en un nuevo marco de datos que incluye la variable dependiente como factor y las variables estandarizadas.

KNN

A medida que aumenta el valor de K, el límite de decisión del clasificador KNN se vuelve más suave y menos sensible al ruido en los datos de entrenamiento. Esto se debe a que el clasificador promedia más instancias vecinas, lo que reduce la varianza de las predicciones, haciendo que el modelo sea más estable. Sin embargo, este aumento en la estabilidad también conlleva un mayor riesgo de sesgo, es decir, existe la posibilidad de desajuste, ya que el modelo puede volverse demasiado simple para capturar los patrones subyacentes en los datos.

Si el error comienza a aumentar considerablemente después de 1/9, sugiere que el modelo podría estar volviéndose demasiado sesgado y no lo suficientemente flexible para capturar la complejidad de los datos. En otras palabras, un valor K de alrededor de 9 parece proporcionar un buen equilibrio entre el sesgo (desajuste) y la varianza (sobreajuste) para el conjunto de datos específico.

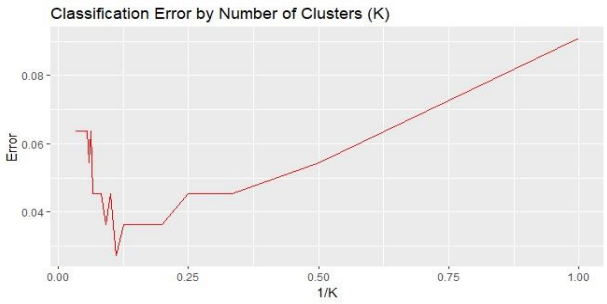


Ilustración 1: Gráfico errores vs 1/K

Sensibilidad (Recall o True Positive Rate): $\frac{TP}{TP+FN} = 1{,}0000$ Esto significa que el modelo identificó correctamente el 100% de los casos positivos.

Especificidad (True Negative Rate): $\frac{TN}{TN+FP} = 0{,}9318$. Esto significa que el modelo identificó correctamente el 93.18% de los casos negativos. **Valor Predictivo Positivo** (Precisión): $\frac{TP}{TP+FP} = 0{,}9565$.

Esto significa que el 95.65% de los casos que el modelo predijo como positivos eran realmente positivos.

Valor Predictivo Negativo: $\frac{TN}{TN+FN} = 1{,}0000$. Esto significa que el 100% de los casos que el modelo predijo como negativos eran realmente negativos.

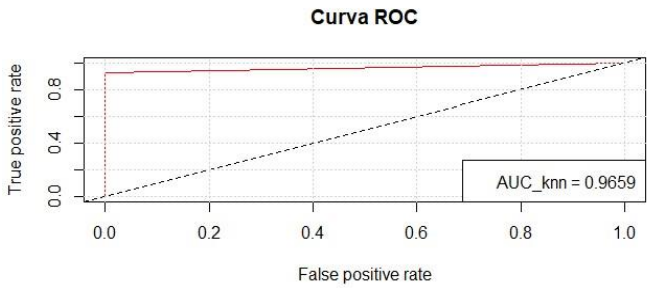
Exactitud (Accuracy): $\frac{TP+TN}{TP+FP+TN+FN} = 0{,}9727$. Esto significa que el modelo hizo predicciones correctas en el 97.27% de los casos.

F1 Score: $\frac{2*Precisión*Recall}{Precisión+Recall} = 0{,}9778$. Esta es una medida ponderada de Precisión y Recall que varía entre 0 (peor) y 1 (mejor).

```
Confusion Matrix and Statistics

          Reference
Prediction  B  M
         B 66  3
         M  0 41

               Accuracy : 0.9727
                 95% CI : (0.9224, 0.9943)
    No Information Rate : 0.6
    P-Value [Acc > NIR] : <2e-16

                  Kappa : 0.9425

 Mcnemar's Test P-Value : 0.2482

            Sensitivity : 1.0000
            Specificity : 0.9318
         Pos Pred Value : 0.9565
         Neg Pred Value : 1.0000
              Precision : 0.9565
                 Recall : 1.0000
                     F1 : 0.9778
             Prevalence : 0.6000
         Detection Rate : 0.6000
   Detection Prevalence : 0.6273
      Balanced Accuracy : 0.9659

       'Positive' Class : B
```

Ilustración 2: Matriz KNN

La Curva ROC es una gráfica que ilustra el rendimiento de un modelo de clasificación en diversos umbrales de clasificación. El Área Bajo la Curva (AUC) se presenta como una métrica que evalúa el desempeño global del modelo, independientemente del umbral de clasificación. Con un AUC de 0.9659, el modelo exhibe una capacidad destacada para discriminar entre clases positivas y negativas. Este resultado se considera excelente. Cabe destacar que un AUC de 1.0 representaría un clasificador perfecto, mientras que un AUC de 0.5 indicaría un clasificador aleatorio.



El modelo KNN con K=9 ha demostrado un rendimiento excepcional. Las métricas de clasificación indican que el modelo es altamente sensible (1.0000) y específico (0.9318), lo que significa que es eficaz para identificar tanto los casos positivos como los negativos. Además, la precisión del modelo es alta (0.9565 para casos positivos y 1.0000 para casos negativos), lo que indica que las predicciones del modelo suelen ser correctas. La exactitud general del modelo es de 0.9727, lo que significa que el modelo hizo predicciones correctas en el 97.27% de los casos. El puntaje F1 de 0.9778 sugiere un equilibrio sólido entre precisión y recuperación. En resumen, el modelo KNN con K=9 es robusto y preciso para este conjunto de datos. Sin embargo, es importante validar estos resultados con nuevos datos para confirmar la capacidad de generalización del modelo.

REGRESIÓN LOGÍSTICA

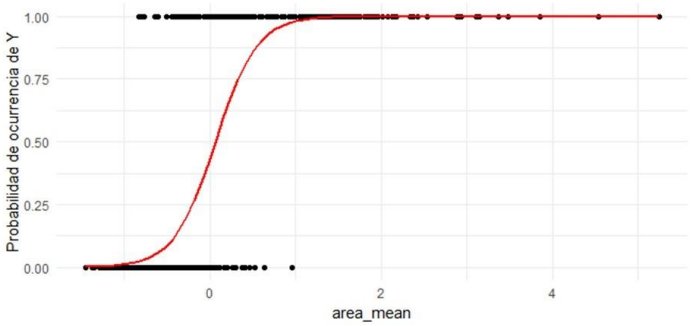


Ilustración 3: Gráfico variable dependiente vs variable independiente

Matriz de Confusión: La matriz es una tabla de 2x2 que contiene 4 resultados producidos por el clasificador binario. En este caso, las clases son 'B' y 'M'. La matriz se ve así:

Precisión: Este es el cociente entre el número total de predicciones correctas y el número total de predicciones. En

este caso, la precisión es 0.9646, lo que significa que el modelo es correcto el 96.46% del tiempo.

Sensibilidad (Tasa de Verdaderos Positivos): Este es el cociente entre las predicciones verdaderamente positivas (B's correctamente predichos como B) y todos los positivos reales (todos los B's). El modelo tiene una sensibilidad de 0.9577, lo que significa que identifica correctamente el 95.77% de todas las instancias positivas reales.

Especificidad (Tasa de Verdaderos Negativos): Este es el cociente entre las predicciones verdaderamente negativas (M's correctamente predichos como M) y todos los negativos reales (todos los M's). El modelo tiene una especificidad de 0.9762, lo que significa que identifica correctamente el 97.62% de todas las instancias negativas reales.

Valor Predictivo Positivo (Precisión): El modelo tiene un VPP de 0.9855, lo que significa que cuando predice que una instancia es positiva, acierta el 98.55% del tiempo.

Valor Predictivo Negativo: El modelo tiene un VPN de 0.9318, lo que significa que cuando predice que una instancia es negativa, acierta el 93.18% del tiempo.

Precisión Equilibrada: Es el promedio de sensibilidad y especificidad, y es útil cuando se trata con conjuntos de datos desequilibrados. El modelo tiene una precisión equilibrada de 0.9670.

Puntuación F1: Es la media armónica de la precisión y la sensibilidad, y proporciona una medida más precisa de los casos clasificados incorrectamente que la Métrica de Precisión. Sin embargo, la puntuación F1 no se proporciona en los resultados.

Kappa: Esta es una estadística que compara una Precisión Observada con una Precisión Esperada (probabilidad aleatoria). El modelo tiene un kappa de 0.9249, que es muy alto, indicando que el modelo está haciendo un buen trabajo en la clasificación.

Prevalencia: Esta es la tasa real de ocurrencia de la clase positiva en los datos. En este caso, la prevalencia es 0.6283, lo que significa que el 62.83% de las instancias son positivas en realidad.

En general, el modelo parece estar funcionando bastante bien, con alta precisión, sensibilidad y especificidad. Está haciendo un buen trabajo clasificando instancias y es correcto la mayor parte del tiempo.

```
Confusion Matrix and Statistics

          Reference
Prediction  B  M
         B 68  1
         M  3 41

               Accuracy : 0.9646
                 95% CI : (0.9118, 0.9903)
    No Information Rate : 0.6283
    P-Value [Acc > NIR] : <2e-16

                  Kappa : 0.9249

 Mcnemar's Test P-Value : 0.6171

            Sensitivity : 0.9577
            Specificity : 0.9762
         Pos Pred Value : 0.9855
         Neg Pred Value : 0.9318
             Prevalence : 0.6283
         Detection Rate : 0.6018
   Detection Prevalence : 0.6106
      Balanced Accuracy : 0.9670

       'Positive' Class : B
```

Ilustración 5: Métricas Modelo Logístico

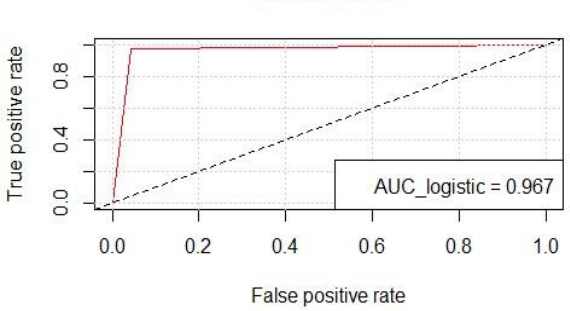


Ilustración 6: ROC Regresión Logística

En el modelo de regresión logística ajustado, se han identificado 'texture_mean' y 'area_mean' como las

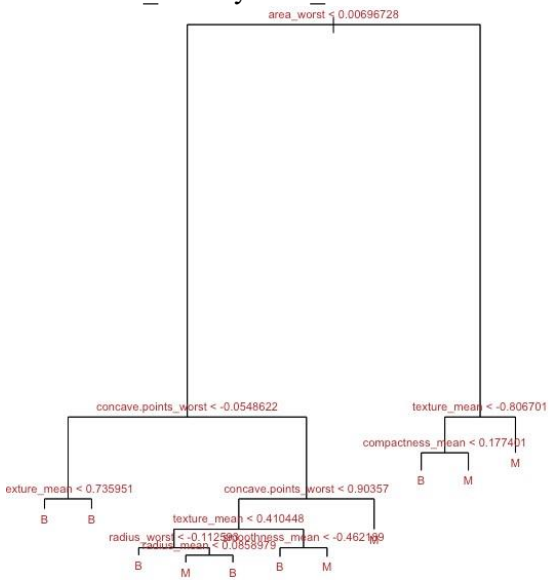


Ilustración 4: Árbol profundo variables más significativas. Sus coeficientes positivos sugieren que a medida que estos valores aumentan, también lo hace la probabilidad del evento.

El coeficiente para texture_mean es 3.017207. Esto significa que, manteniendo todas las demás variables constantes, un aumento de una unidad en texture_mean multiplicará las odds de la variable dependiente por 3.017207.

Se observa una disminución considerable en la desviación, pasando de 710.4 en el modelo nulo a 110.7 en nuestro modelo, indicando un mejor ajuste con las variables predictoras.

Las métricas de clasificación muestran que el modelo tiene una precisión del 96.46%, una sensibilidad del 95.77% y una especificidad del 97.62%. Además, el valor AUC de 0.967 indica una alta capacidad para distinguir entre clases.

En resumen, el modelo de regresión logística exhibe un rendimiento sólido con variables predictoras significativas y fuertes métricas de clasificación. No obstante, es crucial validar estos resultados con nuevos datos para confirmar la capacidad de generalización del modelo.

Para identificar las variables independientes que mejor explican la variable dependiente en un modelo de regresión logística, puede observar los **odds ratios** y los **valores p** del análisis de varianza (ANOVA).Para identificar las variables independientes que mejor explican la variable dependiente en un modelo de regresión logística, se puede observar los "**odds ratios**" y los "**valores p**" del análisis de varianza (ANOVA).

ÁRBOLES DE DECISIÓN

En el modelado de los árboles de decisión, se han creado dos conjuntos de datos: uno de entrenamiento que abarca el 80% de los datos, y otro de clasificación que comprende el 20% restante. Con el objetivo de garantizar la reproducibilidad de los resultados, se ha establecido una semilla, asegurando que al ejecutar el código programado se obtengan resultados idénticos.

Se ha construido un árbol de decisión mediante el uso de la función 'tree()', modelando la variable de diagnóstico en función de todas las variables independientes previamente descritas, empleando el conjunto de datos de entrenamiento. El gráfico resultante presenta nodos, en los cuales se establecen condiciones basadas en las variables predictoras. El proceso de toma de decisiones en el árbol depende del valor de un dato específico, que se compara con la condición establecida en cada nodo, determinando así el camino a seguir.

Las variables que aparecen en el árbol se pueden considerar como aquellas que poseen un mayor grado de influencia sobre la variable dependiente.
Este árbol se clasifica como profundo, ya que no se aplican restricciones o instrucciones que limiten su crecimiento. Bajo esta perspectiva, se hace necesario evaluar el poder predictivo del modelo utilizando, en esta fase, los datos de validación. Para los árboles de decisión, se realiza una validación cruzada con un parámetro k=5, que produce distintos tamaños de árbol y las tasas de error correspondientes a cada uno.

Con esta información y mediante el uso de la función 'cv.tree()', se obtiene el tamaño óptimo del árbol, identificado como aquel con la menor tasa de error en el proceso de validación cruzada. En este caso, dicho valor es igual a 8. Utilizando este valor, se poda el árbol a través de la función 'prune.misclass()', resultando en el árbol podado.

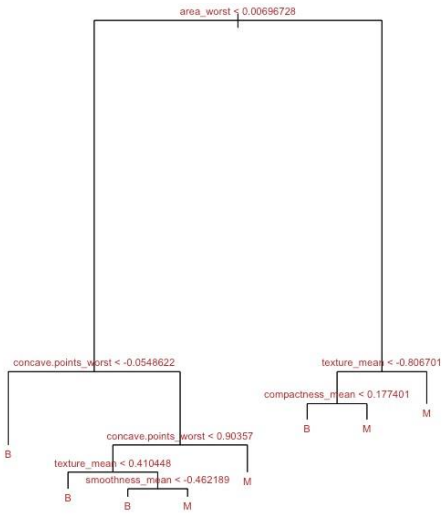


Ilustración 7: Árbol podado

Es importante señalar que en el árbol profundo inicialmente se contaba con 10 nodos, que incluían a las variables 'area_worst', 'concave.points_worst', 'texture_mean', 'radius_worst', 'radius_mean', 'smoothness_mean' y 'compactness_mean'. Tras la poda del árbol, se redujo el número de nodos a 7, y ya no se incluyen las variables 'radius_worst' y 'radius_mean', indicando una disminución en la cantidad de variables explicativas que tienen una influencia significativa en la variable dependiente.

Con todos estos datos, se realizan predicciones en el conjunto de validación tanto para el árbol profundo como para el árbol podado, con el objetivo de comparar sus resultados. El accuracy en la matriz de confusión para el árbol profundo es igual a 0.9292, mientras que, en el árbol podado, el accuracy es de 0.9204. Este ligero decremento en el accuracy puede atribuirse a un trade-off entre la precisión y el sobreajuste de datos que podría estar presente en un árbol profundo.

A continuación, se calcula el Área Bajo la Curva (AUC) y la curva ROC como métodos adicionales para evaluar la capacidad predictiva de ambos modelos. Los resultados se presentan en las gráficas siguientes.

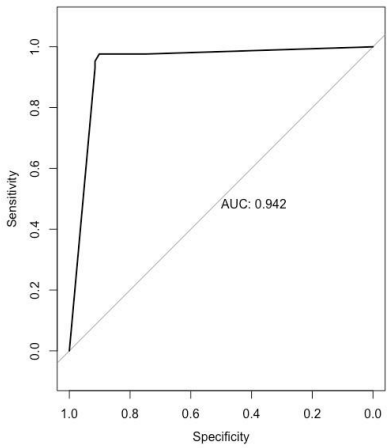


Ilustración 8: Curva ROC árbol profundo

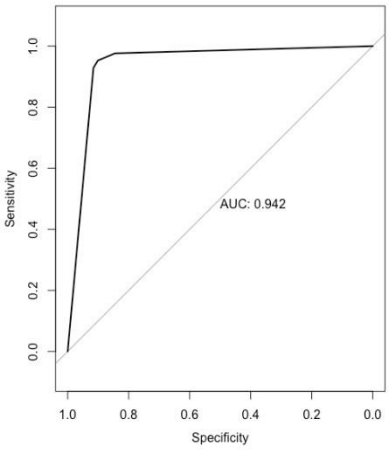


Ilustración 9: Curva ROC árbol podado

En el caso del árbol profundo, se observa un AUC igual a 0.9416, mientras que, para el árbol podado, el AUC alcanza el valor de 0.9418. No se evidencian diferencias significativas entre ambas métricas; sin embargo, dado que el AUC del árbol podado se acerca más a 1, se puede considerar que este modelo exhibe un rendimiento ligeramente superior.

A continuación, se revisan las mismas métricas para los datos de entrenamiento:

En el caso de los datos de entrenamiento, se observa que el árbol de decisión profundo presenta un valor cercano a 1. Aunque la diferencia no resulta significativa, se destaca como un clasificador ligeramente superior, incluso en comparación con los árboles de decisión evaluados con datos de validación.

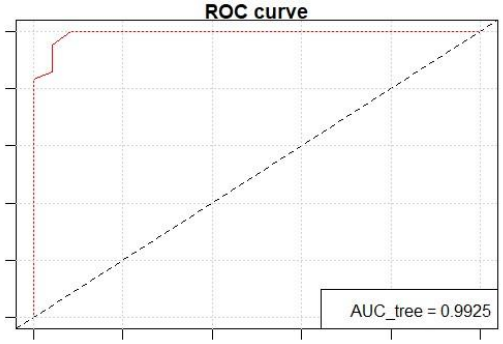


Ilustración 10: Curva ROC Árbol de Decisión

Random Forest

```
> # Imprime los mejores parámetros
> print(paste("Best number of trees: ", best
[1] "Best number of trees:  200"
> print(paste("Best minimum node size: ", be
[1] "Best minimum node size:  3"
> print(paste("Best accuracy: ", best_accura
[1] "Best accuracy:  0.96830985915493"
>
```

Ilustración 11: Fragmento código Random Forest

En el modelo de Random Forest que se ajustó, se empleó la validación cruzada para determinar el tamaño óptimo de los árboles y los nodos, resultando en un mínimo de 3 nodos y 200 árboles. Este equilibrio entre el número de nodos y árboles es crucial para evitar el sobreajuste y garantizar que el modelo pueda generalizar bien a nuevos datos.

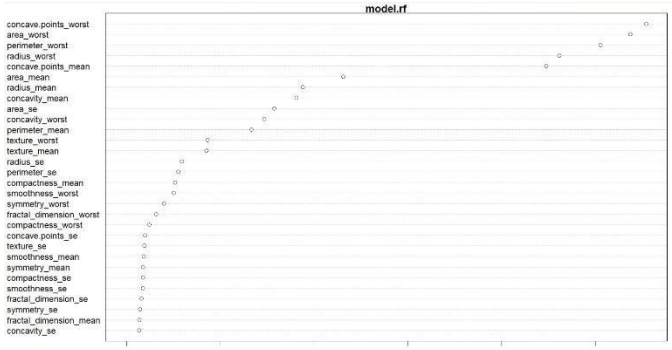


Ilustración 12: Gráfico resultados Random Forest

```
Confusion Matrix and Statistics

          Reference
Prediction  B  M
         B 68  0
         M  3 42

               Accuracy : 0.9735
                 95% CI : (0.9244, 0.9945)
    No Information Rate : 0.6283
    P-Value [Acc > NIR] : <2e-16

                  Kappa : 0.944

 Mcnemar's Test P-Value : 0.2482

            Sensitivity : 0.9577
            Specificity : 1.0000
         Pos Pred Value : 1.0000
         Neg Pred Value : 0.9333
             Prevalence : 0.6283
         Detection Rate : 0.6018
   Detection Prevalence : 0.6018
      Balanced Accuracy : 0.9789

       'Positive' Class : B
```

Ilustración 13: Métricas RF

Las métricas de clasificación revelan que el modelo tiene una precisión del 97.35%, indicando que realiza

predicciones correctas la mayoría de las veces. Tanto la sensibilidad como la especificidad del modelo son altas (95.77% y 100% respectivamente), sugiriendo que el modelo es eficaz para identificar tanto los casos positivos como los negativos. Además, el valor AUC de 0.9993 indica que el modelo tiene una alta capacidad para distinguir entre clases positivas y negativas.

Support Vector Machine

Al igual que en los modelos anteriores, se forman dos conjuntos: uno de entrenamiento, que abarca el 80% de los datos, y otro de validación, que comprende el 20% restante. Previo al entrenamiento de los modelos con diferentes kernels, se llevan a cabo dos acciones necesarias: se definen 10 pliegues para una futura validación cruzada como objeto de control y se establecen parámetros de preprocesamiento. A través de la fórmula 'train()' y utilizando los métodos 'SVMLinear', 'SVMRadial' y 'SVMPoly', se entrenarán tres modelos de Support Vector Machine con kernels lineal, radial y polinómico, respectivamente.

Los resultados para el kernel lineal son los siguientes:

Accuracy	Kappa
0.9560848	0.9023754

Ilustración 14: Kernel lineal

Los resultados para el kernel radial corresponden a:

C	Accuracy	Kappa
0.25	0.9537593	0.8964768
0.50	0.9691238	0.9307744
1.00	0.9712978	0.9365554

Ilustración 15: Kernel radial Los resultados para el kernel polinomial corresponden a:

degree	scale	C	Accuracy	Kappa
1	0.001	0.25	0.8021256	0.4970225
1	0.001	0.50	0.9010628	0.7681906
1	0.001	1.00	0.9341546	0.8495534
1	0.010	0.25	0.9495652	0.8846418
1	0.010	0.50	0.9495652	0.8849254
1	0.010	1.00	0.9671014	0.9258248
1	0.100	0.25	0.9692271	0.9316569
1	0.100	0.50	0.9735266	0.9418887
1	0.100	1.00	0.9757488	0.9468011
2	0.001	0.25	0.8988889	0.7625264
2	0.001	0.50	0.9341546	0.8495534
2	0.001	1.00	0.9495169	0.8851869
2	0.010	0.25	0.9495169	0.8851869
2	0.010	0.50	0.9693237	0.9307428
2	0.010	1.00	0.9671014	0.9261293
2	0.100	0.25	0.9692271	0.9325483
2	0.100	0.50	0.9735749	0.9421838
2	0.100	1.00	0.9713527	0.9379555
3	0.001	0.25	0.9209662	0.8179520
3	0.001	0.50	0.9385507	0.8598989
3	0.001	1.00	0.9517391	0.8899631
3	0.010	0.25	0.9605797	0.9101512
3	0.010	0.50	0.9715459	0.9358275
3	0.010	1.00	0.9670531	0.9269181
3	0.100	0.25	0.9714493	0.9370026
3	0.100	0.50	0.9714493	0.9372742
3	0.100	1.00	0.9735749	0.9420897

Ilustración 16: Kernel polinomial Una vez construidos los clasificadores, se evalúa el poder predictivo de los modelos, para esto, se realizan predicciones utilizando el conjunto de validación y se calculan las matrices de confusión:

Para el SVM Lineal los resultados fueron:

```
          Reference
Prediction  B  M
         B 61  1
         M  0 52

               Accuracy : 0.9912
                 95% CI : (0.9521, 0.9998)
    No Information Rate : 0.5351
    P-Value [Acc > NIR] : <2e-16

                  Kappa : 0.9823

 Mcnemar's Test P-Value : 1

            Sensitivity : 1.0000
            Specificity : 0.9811
         Pos Pred Value : 0.9839
         Neg Pred Value : 1.0000
             Prevalence : 0.5351
         Detection Rate : 0.5351
   Detection Prevalence : 0.5439
      Balanced Accuracy : 0.9906

       'Positive' Class : B
```

Ilustración 17: Métricas SVM

Para el SVM Radial los resultados fueron:

```
          Reference
Prediction  B  M
         B 61  3
         M  0 50

               Accuracy : 0.9737
                 95% CI : (0.925, 0.9945)
    No Information Rate : 0.5351
    P-Value [Acc > NIR] : <2e-16

                  Kappa : 0.9469

 Mcnemar's Test P-Value : 0.2482

            Sensitivity : 1.0000
            Specificity : 0.9434
         Pos Pred Value : 0.9531
         Neg Pred Value : 1.0000
             Prevalence : 0.5351
         Detection Rate : 0.5351
   Detection Prevalence : 0.5614
      Balanced Accuracy : 0.9717

       'Positive' Class : B
```

Ilustración 18: Métrica SVM Radial

Para el SVM Polinomial los resultados fueron:

```
          Reference
Prediction  B  M
         B 61  2
         M  0 51

               Accuracy : 0.9825
                 95% CI : (0.9381, 0.9979)
    No Information Rate : 0.5351
    P-Value [Acc > NIR] : <2e-16

                  Kappa : 0.9647

 Mcnemar's Test P-Value : 0.4795

            Sensitivity : 1.0000
            Specificity : 0.9623
         Pos Pred Value : 0.9683
         Neg Pred Value : 1.0000
             Prevalence : 0.5351
         Detection Rate : 0.5351
   Detection Prevalence : 0.5526
      Balanced Accuracy : 0.9811

       'Positive' Class : B
```

Ilustración 19: Métricas SVM Polinomial El accuracy del SVM Lineal es el mejor clasificador al ser el más cercano a 1. Para un análisis más profundo se realiza el cálculo de la curva ROC y el AUC.

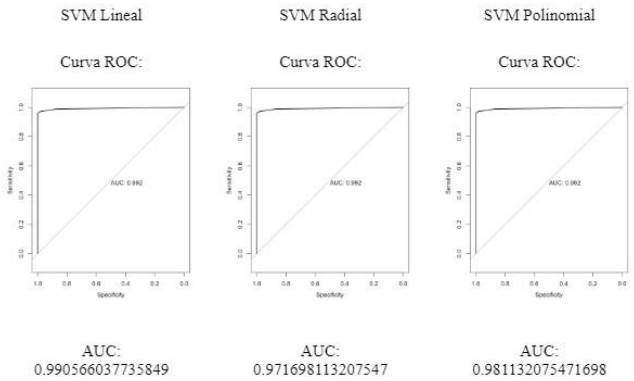


Ilustración 20: Curvas ROC SVM

Los resultados permiten identificar varias situaciones destacadas. Si consideramos M como el margen del que se habla en SVC, este está influenciado por C, ya que un valor de C más alto estrecha el margen y funciona como un parámetro para el trade-off entre precisión y sobreajuste. Penalizar los errores de clasificación permite que, a mayor C, haya mayor varianza, pero reducir C conlleva a un sesgo más alto. Lo anterior se puede apreciar en los resultados, influyendo enormemente en el rendimiento de los clasificadores.

Bajo este contexto, el SVM Lineal es el que posee un mejor rendimiento dentro de los modelos SVM estudiados.

Comparación de modelos

Se procede a realizar el estudio de los modelos en forma simultánea, obteniéndose la siguiente gráfica.

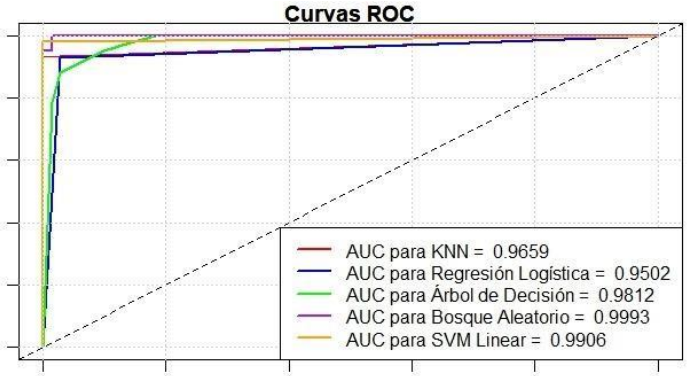


Ilustración 21: Comparación curvas ROC

Conclusiones

kNN (k-Nearest Neighbors): Este modelo es intuitivo y fácil de entender, pero no tan interpretable en términos de entender la contribución de las características individuales.

Regresión logística: Este es uno de los modelos más interpretables. Los coeficientes del modelo representan el cambio en el logaritmo de odds para cada unidad de cambio en las variables predictoras.

Árboles de decisión: Son altamente interpretables y pueden visualizarse fácilmente, pero pueden volverse muy complejos a medida que aumenta la profundidad del árbol.

Random Forest: Este es un modelo más complejo que combina múltiples árboles de decisión. Aunque los árboles individuales son interpretables, un bosque aleatorio puede ser difícil de interpretar debido a la cantidad de árboles.

SVM (Support Vector Machines): Este es un modelo complejo, especialmente cuando se utilizan kernels para mapear los datos a un espacio de mayor dimensión. Los SVM pueden ser difíciles de interpretar.

Todos los modelos excepto regresión logística consideran las 30 variables predictoras, ya que un problema que tiene la regresión logística es que no puede lidiar con la multicolinealidad, por lo que para este modelo se eliminaron las variables altamente correlacionadas. Por otro lado, los modelos de árboles de decisión y random forest consideran las variables area_worst, concave.points_worst, texture_mean, smoothness_mean y compactness_mean como las más importantes, en cambio, el modelo de regresion logistica considera texture mean y área mean.

Los valores AUC para los cinco modelos varían desde 0.9502 para la regresión logística hasta 0.9993 para el bosque aleatorio, como se observa en el gráfico de curvas ROC. Estos valores indican que todos los modelos tienen una alta capacidad para distinguir entre clases positivas y negativas. Sin embargo, el bosque aleatorio parece tener el mejor rendimiento según esta métrica.