Métodos supervisados en aprendizaje automático (proyecto final)

Realizado por: Julian Abello - Julian González - Juan Ortiz

Introducción

El aprendizaje supervisado es esencial en el campo del aprendizaje automático, entrenando algoritmos en datos etiquetados para reconocer patrones y realizar predicciones o clasificaciones. Su importancia abarca distintos campos como la ciencia de datos y la ingeniería de aprendizaje, siendo clave en finanzas, salud y ventas. Desde distinguir spam hasta predecir precios de viviendas, estos modelos automatizan percepciones en áreas como el reconocimiento de imágenes y el procesamiento de lenguaje natural, ofreciendo información valiosa en diversas industrias. Al extraer datos relevantes de información etiquetada, permite predicciones y clasificaciones precisas, fundamentales en la toma de decisiones en distintos sectores. Su aplicación va desde la predicción hasta la automatización de decisiones, siendo crucial en escenarios del mundo real.

Conceptos y técnicas

Tipos de problemas:

Regresión y Clasificación.

Algoritmos:

Regresión lineal, regresión logística, árboles de decisión, bosques aleatorios, máquinas de vectores de soporte.

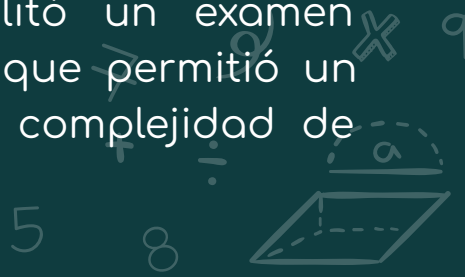
Modelos de entrenamiento y prueba:

Capacitación y pruebas.
Sobreajuste y desajuste.
Validación cruzada.
Métricas de rendimiento.

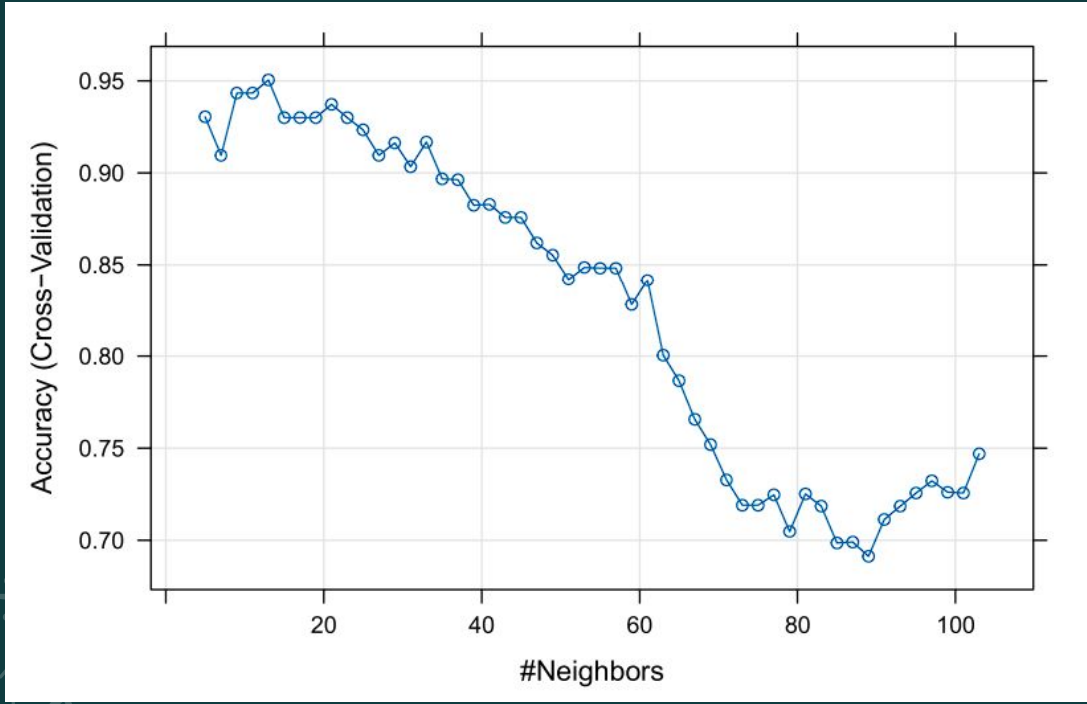
Metodología

El enfoque principal de este estudio radica en utilizar el paquete dinámico de controladores de cáncer para realizar un preprocesamiento integral de datos. Este paquete se empleó para realizar análisis y refinamiento exhaustivos de los datos, preparando de manera efectiva el conjunto de datos original para análisis posteriores.

Además, se crearon subconjuntos de datos a partir del conjunto de datos original para aplicar varios modelos de aprendizaje automático. Estos subconjuntos se diseñaron y adaptaron específicamente para abordar la clasificación de diferentes tipos de cáncer y predecir resultados clínicos relevantes. La diversidad de modelos utilizados facilitó un examen detallado de múltiples facetas dentro de los datos, lo que permitió un enfoque más preciso y detallado para comprender la complejidad de varios tipos de cáncer y sus posibles resultados clínicos.



Resultados



Gráfica en la que se evidencia la precisión del número de vecinos (K), la cual nos ayuda a encontrar la cantidad de datos que vamos a tomar para nuestro algoritmo.

Resultados

```
## Confusion Matrix and Statistics
##
##                          Reference
## Prediction               Primary Tumor Solid Tissue Normal
##   Primary Tumor                     34                    0
##   Solid Tissue Normal               16                   49
##
##                Accuracy : 0.8384
##                  95% CI : (0.7509, 0.9047)
##     No Information Rate : 0.5051
##     P-Value [Acc > NIR] : 4.317e-12
##
##                   Kappa : 0.6778
##
##  Mcnemar's Test P-Value : 0.0001768
```

Matriz de confusión, que representa los valores predictivos al momento de aplicar el modelo knn.

Resultados

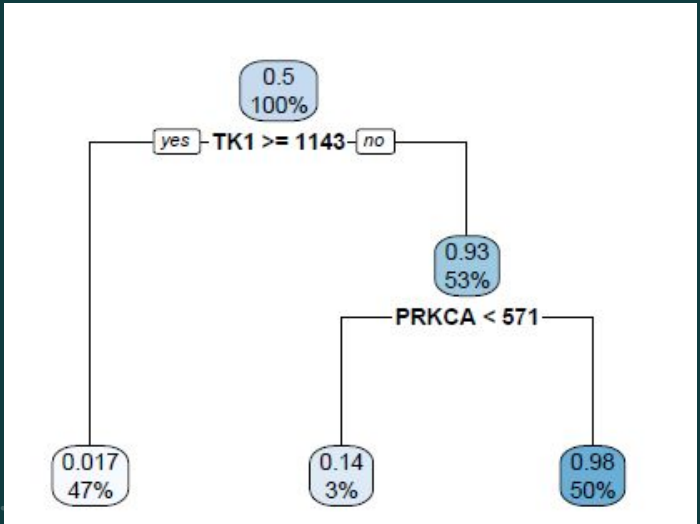
Regresión lineal

```
##
## Call:
## lm(formula = sample_type ~ ., data = train.data)
##
## Residuals:
##       Min         1Q     Median         3Q        Max
## -0.185117 -0.035942  0.001759  0.037991  0.207019
##
## Coefficients:
##               Estimate Std. Error t value Pr(>|t|)
## (Intercept)  7.342e-01  1.139e-01   6.444 6.24e-08 ***
## TP53         1.436e-04  3.041e-05   4.722 2.22e-05 ***
## CREBBP      -1.951e-05  2.224e-05  -0.877 0.384812
## EP300        2.466e-05  3.022e-05   0.816 0.418698
## YWHAG       -2.062e-05  1.361e-05  -1.515 0.136734
## SMAD3       -3.005e-05  1.932e-05  -1.555 0.126806
## GRB2         2.115e-05  1.138e-05   1.858 0.069578 .
```

```
## Linear Regression
##
## 147 samples
## 100 predictors
##
## No pre-processing
## Resampling: Cross-Validated (10 fold)
## Summary of sample sizes: 132, 133, 133, 132, 132, 132, ...
## Resampling results:
##
##   RMSE       Rsquared   MAE
##   0.9961332  0.3216684  0.6099457
##
## Tuning parameter 'intercept' was held constant at a value of TRUE
```

Este es un resultado de un modelo de regresión lineal, donde se evalúan las relaciones entre las variables de entrada (TP53, CREBBP, EP300, etc.) y el tipo de muestra. Los coeficientes muestran cómo cada variable contribuye al resultado final y si esa contribución es significativa. Los valores de Pr(>|t|) indican la significancia estadística de cada coeficiente. El "Residual standard error" indica la diferencia entre los valores observados y los valores predichos por el modelo. El valor de R-cuadrado indica qué tan bien el modelo se ajusta a los datos, siendo cercano a 1 un buen ajuste. En resumen, aquí se proporciona información sobre la influencia de cada variable en la predicción del tipo de muestra y la calidad general del modelo para predecir esos tipos.

Árboles de decisión



Los resultados presentados sugieren el análisis de un modelo de árbol de decisión construido a partir de un conjunto de datos de 246 observaciones. Este modelo divide los datos en nodos basados en las variables TK1 y PRKCA. Por ejemplo, cuando TK1 es mayor o igual a 1142.5, el nodo terminal predice una clase con una proporción de 0.01724138. Por otro lado, cuando TK1 es menor que 1142.5 y PRKCA es menor que 570.5, el nodo terminal predice otra clase con una proporción de 0.14285710, mientras que cuando TK1 es menor que 1142.5 y PRKCA es mayor o igual a 570.5, el nodo terminal predice una clase con una proporción de 0.97560980. Estos nodos terminales representan las predicciones resultantes del modelo para diferentes conjuntos de condiciones en las variables analizadas.

Máquinas de soporte vectorial

```
## Confusion Matrix and Statistics
##
##           Reference
## Prediction  0  1
##          0 30  1
##          1  1 42
##
##                Accuracy : 0.973
##                  95% CI : (0.9058, 0.9967)
##     No Information Rate : 0.5811
##     P-Value [Acc > NIR] : 5.216e-15
##
##                   Kappa : 0.9445
```

Los resultados presentan la evaluación de un modelo de máquinas de soporte vectorial (SVM) mediante una matriz de confusión y estadísticas asociadas. La precisión general del modelo es del 97.3%, con una alta sensibilidad del 96.77% para la clase 0 y una especificidad del 97.67% para la clase 1. El coeficiente Kappa, que indica la concordancia entre las predicciones y los valores reales, es de 0.9445, lo que muestra un nivel significativo de consistencia en las predicciones. Además, tanto el valor predictivo positivo como el negativo son altos (96.77% y 97.67%, respectivamente), indicando una precisión sólida en las predicciones de ambas clases. En resumen, el modelo SVM exhibe un desempeño sobresaliente, con altas tasas de precisión y concordancia en la clasificación de las clases.

Resumen y conclusión

En resumen, los hallazgos de los modelos de aprendizaje supervisado destacan la importancia del preprocesamiento cuidadoso de los datos, incluyendo la filtración de columnas, el manejo de datos faltantes y el procesamiento de información de dominio específico. Se subraya la necesidad de comprender el dominio de los datos y adaptar las técnicas de preprocesamiento en consecuencia.

Las implementaciones de diversos algoritmos, como KNN, regresión lineal, árboles de decisión, bosques aleatorios y SVM, resaltan la versatilidad de éstos en el manejo de diferentes tipos de datos y tareas. Cada modelo presenta fortalezas y debilidades, destacando la importancia de elegir modelos adaptados a las características específicas de los datos y los problemas.
Los desafíos encontrados, como el manejo de datos faltantes y la optimización de hiperparámetros, enfatizan el pensamiento crítico y las habilidades de resolución de problemas necesarias en el ámbito de la ciencia de datos y el aprendizaje automático. Estos requieren creatividad, conocimiento del dominio y una comprensión profunda del comportamiento del modelo para tomar decisiones informadas y mejorar su rendimiento.

La evaluación del modelo mediante diversas métricas de desempeño destaca la importancia de metodologías de evaluación sólidas, como la exactitud, precisión, recuperación y el área bajo la curva ROC. Estas métricas son cruciales para medir la efectividad del modelo y tomar decisiones informadas sobre su despliegue. En conclusión, los hallazgos resaltan la necesidad de habilidades multidimensionales en roles de ciencia de datos y aprendizaje automático, que van desde la implementación técnica de algoritmos y preprocesamiento de datos hasta la comprensión del dominio y una sólida evaluación métrica; la capacidad para enfrentar desafíos y tomar decisiones basadas en datos son esenciales para el éxito en estos roles.