APLICACIÓN DEL APRENDIZAJE DE MÁQUINAS EN LA INTELIGENCIA ARTIFICIAL: UNA APROXIMACIÓN AL CASO DE ESTUDIO DE CÁNCER EN WISCONSIN

APPLICATION OF MACHINE LEARNING IN ARTIFICIAL INTELLIGENCE: A CANCER WISCONSIN'S CASE STUDY APPROACH

Leonardo Hernán Talero Sarmiento[1] Yuly Andrea Ramirez Sierra[2]

RESUMEN

El aprendizaje de máquinas es una metodología de Inteligencia Artificial la cual se enfoca en el entrenamiento de algoritmos para clasificación y pronóstico para la resolución de problemas en múltiples áreas del conocimiento. Durante el presente trabajo se realizó un estudio comparativo de 7 técnicas de aprendizaje con el fin de evaluar cuál lograba pronosticar de manera más acertada el tipo de tumor presentado por diversos pacientes en Winsconsin, para ello se aplicaron de manera simplificada algoritmos pre-programados en el lenguaje Python a un dataset de 30 variables y 569 registros sin faltantes encontrando que los algoritmos de mejora para las técnicas de fraccionamiento presentaron mejor grado de acierto: Random Forest y Adaboost respectivamente.

PALABRAS CLAVE

Aprendizaje Automático, Breast Cancer, Comparación, Inteligencia Artificial, Python,

1. Introducción

La expresión Inteligencia Artificial (Artificial Intelligence, AI) puede estar relacionada en el consiente colectivo con las grandes obras de Asimov o Lucas, quienes de manera imaginativa y literaria proponen el sofisticado y perfecto desarrollo de las interacciones hombre-máquina. De manera análoga, (según exponen (Palma Méndez & Morales, 2008)) los científicos Marvin Minsky, John McCarty, Nathan Rochester y Claude Shannon en 1955 se congregaron con el fin de analizar y proponer la futura interacción hombre-máquina (sea esta última un robot, un computador, un algoritmo, etc.) en el marco de una discusión puntual: *¿la conjetura de que todos los aspectos del aprendizaje o de cualquier otra característica de la inteligencia pueden, en principio, ser descritos de modo tan preciso que se pueda construir una máquina capaz de simularlos?*

Dicha inquietud relacionada con el aprendizaje humano y su replicación perfecta en máquinas sienta las bases de la Inteligencia Artificial (IA) como Ciencia (IAC) e Ingeniería (IAI). La primera al estudiar todos los procesos cognitivos del ser humano y la segunda al desarrollar tareas y métodos de solución generando así un ciclo entre la Ingeniería Bio-inspirada y la Neurociencia Computacional. A partir de dichas bases diversos estudios se han desarrollado en el área de la Inteligencia Artificial, haciendo de ésta uno de los campos más prolíferos (Palma Méndez & Morales, 2008) abarcando estudios comportamentales, sistemas de navegación, controladores

[1] Ingeniero Industrial y estudiante de Maestría en Ingeniería Industrial, Universidad Industrial de Santander, Bucaramanga, Colombia. E-mail: leonardo.talero@correo.uis.edu.co

[2] Ingeniera Industrial y estudiante de Maestría en Ingeniería Industrial, Universidad Industrial de Santander, Bucaramanga, Colombia. E-mail: yuly2169087@correo.uis.edu.co

inteligentes y demás; sin embargo, los agentes perfectamente autónomos, con una inteligencia tan completa y tan compleja como la de los seres humanos no han sido desarrollados aún (Ponce, 2010).

En pro de dicho avance Palma y otros (2008) resaltan la síntesis de Mira (1995) quien expone como propósito de la Inteligencia Artificial el desarrollar: (1) Modelos conceptuales, (2) procedimientos de reescritura formal de esos modelos y (3) estrategias de programación y máquinas físicas para reproducir de la forma más eficiente y completa posible las tareas cognitivas y científico-técnicas más genuinas de los sistemas biológicos a los que hemos etiquetado de inteligentes. Discurriendo en la importancia del modelamiento ya que según Shapiro: cada aspecto del aprendizaje o cualquier otra representación de inteligencia que pueda en principio ser precisamente descrita, puede ser simulado por una máquina. (Shapiro, 1992).

Para ello es necesario transformar cualquier actividad e interacción (áreas estudiadas por la IAC) como ver, oír, interpretar, manipular, predecir, configurar, etc. En una serie de tareas sub-divisibles hasta el punto de llegar al nivel de inferencias primitivas tales como seleccionar, comparar, etc. De estas últimas inferencias se despliega la Inteligencias Artificial en Ingeniería la cual de manera general trabaja en pro de: (1) Generar soluciones, (2) Sistemas expertos, (3) Procesamiento del lenguaje natural, (4) Reconocimiento de modelos, (5) Robótica, (6) Aprendizaje de las máquinas, (7) Lógica y (8) Incertidumbre y lógica difusa. Y se pueden clasificar como: (1) Sistemas que piensan como humanos, (2) Sistemas que actúan como humanos, (3) Sistemas que piensan racionalmente y (4) Sistemas que actúan racionalmente (Ponce, 2010).

El presente trabajo se enfoca en realizar una aproximación al aprendizaje de máquinas (Machine Learning) en el ámbito médico como una metodología de la IA, debido a la facilidad de la misma para la identificación de patrones y en extensión la adaptación a nuevas circunstancias (pronósticos) (Talwar & Kumar, 2013). Para ello se desarrolla en el lenguaje de programación Python 3.5 diversos análisis sobre una base de datos de pacientes con cáncer la cual fue construida por Wolberg & otros (1995). La estructura del presente documento consta de una primera sección donde se hace un pequeño análisis sobre tendencias de Machine Learning (ML) en medicina, seguida de una segunda sección enfocada a la metodología aplicada para el tratamiento de datos y análisis de los mismos. Posteriormente se desarrolla un apartado de Resultados, en el cual se consignan los valores de ajuste y varianza obtenidos durante el presente trabajo. Seguido se encuentran las secciones de Discusiones en la cual se exponen las dificultades y limitantes de la investigación, Conclusiones donde se realiza un breve análisis de los resultados y su semejanza con los valores esperados y, finalmente, Bibliografía y Anexos.

2. Antecedentes

Los primeros estudios realizados a los pacientes fueron desarrollados por Wolberg y Mangasarian en 1993 y 1994 respectivamente. El primer estudio se enfocó en la extracción de características (clasificación) de los núcleos celulares (Street, Wolberg, & Mangasarian, 1993) mientras el segundo se enfocó en la realización de un pronóstico sobre cáncer mediante la matematización de un modelo lineal.(W H Wolberg, Street, & Mangasarian, 1994). Posteriormente los autores realizan

diversos estudian de la mano de William buscando desarrollar de manera computacional técnicas de clasificación y pronóstico a partir del aprendizaje de máquinas (Mangasarian, Street, & Wolberg, 1995; W. Wolberg & Street, 1995; W H Wolberg, Street, Heisey, & Mangasarian, 1995; William H. Wolberg, Street, Heisey, & Mangasarian, 1995).Dentro de los estudios relacionados con el Dataset, se estiman alrededor de 40 publicaciones relacionadas con la aplicación de máquinas de aprendizaje, mejoramiento de algoritmos, prueba de *kernels* medidas de eficiencia, programación evolutiva y diversos afines

3. Metodología

3.1. Datos

El *Breast Cancer Wisconsin (Diagnostic) Data Set* es una recopilación de información construída a partir del análisis gráfico de tumores realizado por Bennet (1992) mediante árboles de decisión; dicho análisis describe las características de los núcleos celulares con presencia de anomalías. En el dataset se encuentra 9 variables *(Radius, texture, perimeter, área, smoothness, compactness, concavity, concave points* y *symmetry)* registradas en 3 parámetros (Media "*mean",* Desviación estándar "*se"* y valor mínimo *"worst"*) además de la clasificación del tumor (*diagnosis)* en Benigno o Maligno, el identificador del paciente y una columna vacía.

La evaluación de los 9 parámetros es realizada a 569 imágenes cada una correspondiente a un único paciente generando así un dataset de 30 columnas y 569 registros sin ningún campo vacío. La descripción de las variables se registra en la Tabla 1.

Variable	*Registros*	*Tipo*	*Variable*	*Registros*	*Tipo*	*Variable*	*Registros*	*Tipo*
id	569 non-null	int64	fractal_dimension_mean	569 non-null	float64	radius_worst	569 non-null	float64
diagnosis	569 non-null	object	radius_se	569 non-null	float64	texture_worst	569 non-null	float64
radius_mean	569 non-null	float64	texture_se	569 non-null	float64	perimeter_worst	569 non-null	float64
texture_mean	569 non-null	float64	perimeter_se	569 non-null	float64	area_worst	569 non-null	float64
perimeter_mean	569 non-null	float64	area_se	569 non-null	float64	smoothness_worst	569 non-null	float64
area_mean	569 non-null	float64	smoothness_se	569 non-null	float64	compactness_worst	569 non-null	float64
smoothness_mean	569 non-null	float64	compactness_se	569 non-null	float64	concavity_worst	569 non-null	float64
compactness_mean	569 non-null	float64	concavity_se	569 non-null	float64	concave points_worst	569 non-null	float64
concavity_mean	569 non-null	float64	concave points_se	569 non-null	float64	symmetry_worst	569 non-null	float64
concave points_mean	569 non-null	float64	symmetry_se	569 non-null	float64	fractal_dimension_worst	569 non-null	float64
symmetry_mean	569 non-null	float64	fractal_dimension_se	569 non-null	float64	Unnamed: 32	0	float64

Tabla 1 Descripción de las variables del dataset original

3.2. Análisis descriptivo del instrumento

Con el fin de realizar una aproximación al comportamiento de las 9 variables durante el presente trabajo se realiza un enfoque a los valores promedios puesto que estos indican la tendencia central de los datos. (Los valores son registrados en la Tabla 2). Se puede observar que de manera general en las 9 variables existe una distribución con tendencia a valores por debajo de la media y presencia de datos atípicos en el límite superior.

	radius mean	*texture mean*	*perimeter mean*	*area mean*	*smoothness mean*	*compactness mean*	*concavity mean*	*concave points mean*	*symmetry mean*	*fractal dimension mean*
Conteo	569,00	569,00	569,00	569,00	569,00	569,00	569,00	569,00	569,00	569,00
Promedio	14,13	19,29	91,97	654,89	0,10	0,10	0,09	0,05	0,18	0,06
Desviación st	3,52	4,30	24,30	351,91	0,01	0,05	0,08	0,04	0,03	0,01
min	6,98	9,71	43,79	143,50	0,05	0,02	0,00	0,00	0,11	0,05
25%	11,70	16,17	75,17	420,30	0,09	0,06	0,03	0,02	0,16	0,06
50%	13,37	18,84	86,24	551,10	0,10	0,09	0,06	0,03	0,18	0,06
75%	15,78	21,80	104,10	782,70	0,11	0,13	0,13	0,07	0,20	0,07
max	28,11	39,28	188,50	2.501,00	0,16	0,35	0,43	0,20	0,30	0,10

Tabla 2 descripción de las 9 variables de interés

3.3. Reducción dimensional del instrumento

Partiendo de la estructura del dataset (compuesto por 9 variables estudiadas desde tres perspectivas (ver sección 3.1)) y la relación interna entre variables; en el presente trabajo se propone sintetizar dicha información mediante un Análisis de Componentes Principales (ACP) el cual se enfoca en la reducción dimensional a partir de la transformación lineal de variables (Pearson, 1901) la cual construye un nuevo sistema de coordenadas para el conjunto original de datos manteniendo la varianza en el nuevo sistema (si tiene igual número de variables al sistema original) y en el que la variabilidad del mayor tamaño del conjunto de datos es capturada en el primer eje (llamado el Primer Componente Principal), la segunda varianza más grande es el segundo eje, y así sucesivamente. Lo anterior sujeto a que la transformación parte de la matriz de covarianza o matriz de coeficientes de correlación, ésta al ser simétrica y definida positiva, posee una base completa de vectores propios. Teniendo en cuenta que el objetivo del presente trabajo es realizar una aproximación al aprendizaje de máquinas no se explica en detalle la transformación y se resalta que lo buscado con la misma es proyectar los datos en un plano (R^2) con el fin de realizar discriminaciones.

3.4. Estudios de clasificación en componentes principales

Una vez proyectado los datos en R^2 se realiza un análisis discriminante; para ello y teniendo en cuenta la variable predicción, se desarrollan dos modelos de clasificación: (1) Regresión logística y (2) Máquina de vector de soporte con separación lineal. Los modelos nombrados anteriormente son comparados mediante su nivel de ajuste o eficacia (pronósticos acertados) mediante la separación del dataset en una sección de entrenamiento y otra de prueba.

3.5. Estudios de aprendizaje de máquinas

A partir de las consideraciones de la sección 3.2 Análisis descriptivo del instrumento, se realiza un estudio referente a la relación entre variables con el fin de evitar la multicolinealidad, para ello se construye la matriz de correlaciones y se simplifican o eliminan aquellas variables con un alto valor de correlación (durante el presente trabajo se toma el criterio de $correlación > 0.9$. Un ejemplo de la correlación de las variables se exponen en el mapa de calor de la Ilustración 1 en la cual se

evidencia una fuerte correlación entre el radio promedio (*radius_mean*), el perímetro promedio (*perimeter_mean*) y el área promedio (*area_mean*). Por tanto para los análisis del presente trabajo se simplificará a una sola variable. Por otra parte, la compactación media (*compactness_mean*), la concavidad media (*concavity_mean* y el punto de concavidad medio (*concavepoint_mean*) también presentan un alto grado de correlación realizando un proceso de simplificación similar al descrito anteriormente. En los anexos se consignan los mapas de calor para la desviación estándar y los valores mínimos. (Ilustración 6 e Ilustración 7 respectivamente).

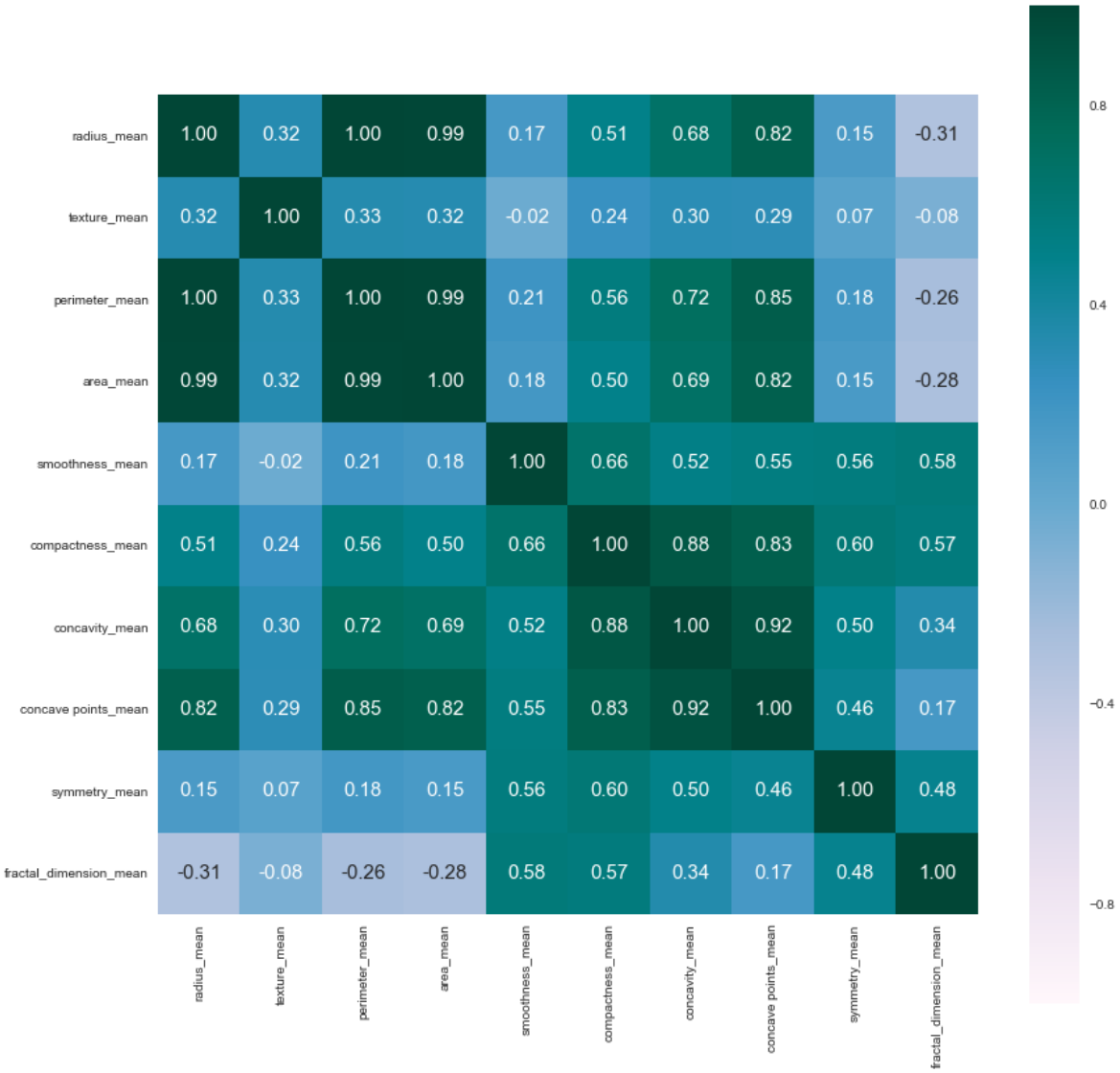


Ilustración 1 Mapa de calor para la correlación entre variables medias

Una vez modificado el dataset, las variables son analizadas con el fin de observar relaciones y tendencias puesto que en la Ilustración 1 se encuentran valores de correlación significativamente altos ($correlación < 0.8$).Dicho comportamiento puede ser contrastado con los diagramas de dispersión de la Ilustración 2 los cuales indican que si bien existe tendencias positivas o negativas,

la variabilidad de las mismas juegan un rol importante ya que aparentemente existe mayor dispersión en los datos correspondientes a tumores malignos.

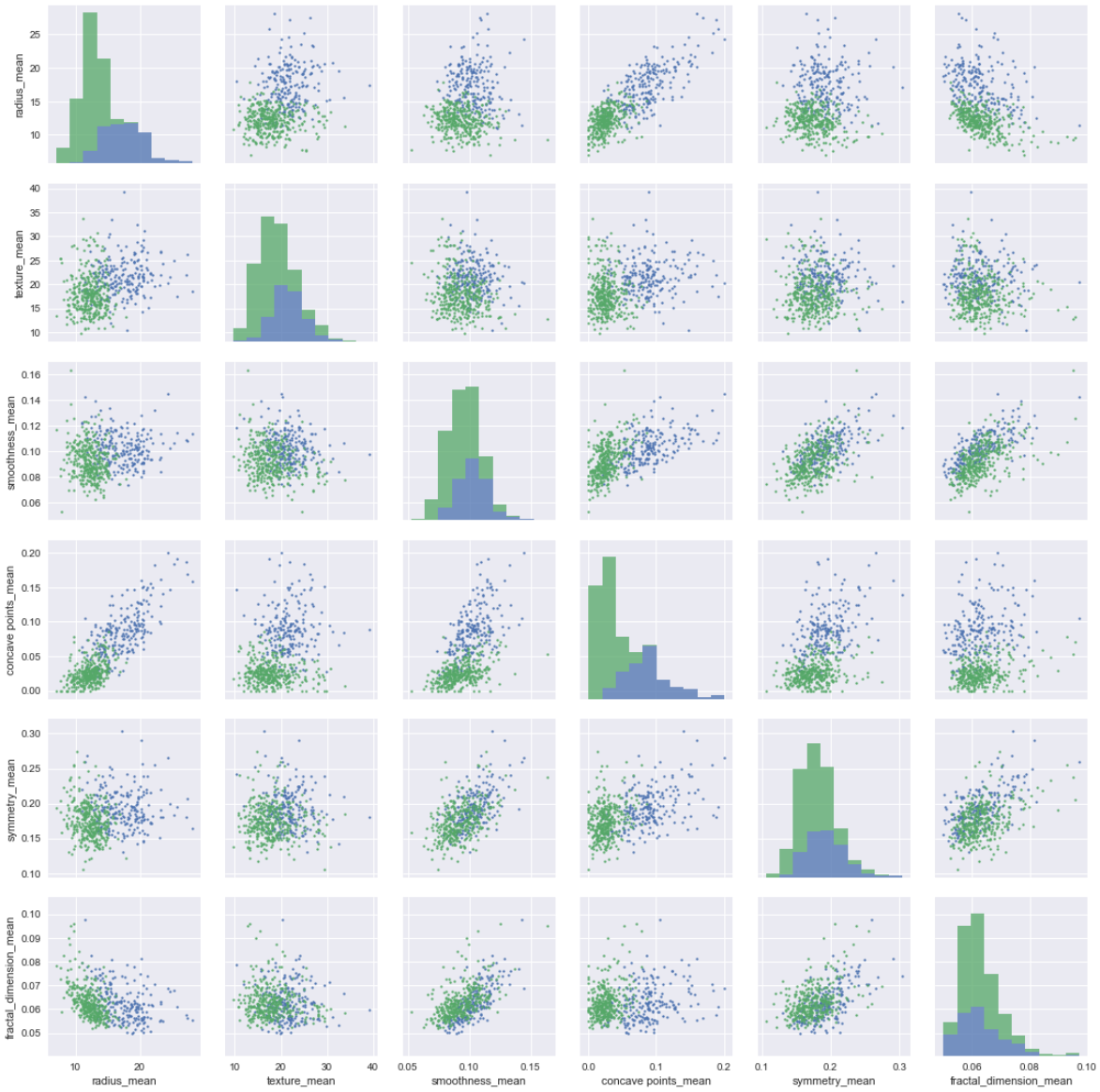


Ilustración 2 Diagrama cruzado de dispersión y distribución

Ahora bien, al contrastar la dispersión de la Ilustración 2 (tanto para las relaciones cruzadas como la correspondiente a cada variable) con el conteo de casos benignos (B) y malignos (M) en la Ilustración 3, se genera una interrogante: y es si la dispersión se relaciona con el número de casos reportados para las categorías.. Por otra parte, radio promedio y puntos de concavidad promedios presentan una mayor diferenciación entre los valores promedios de cada diagnóstico.

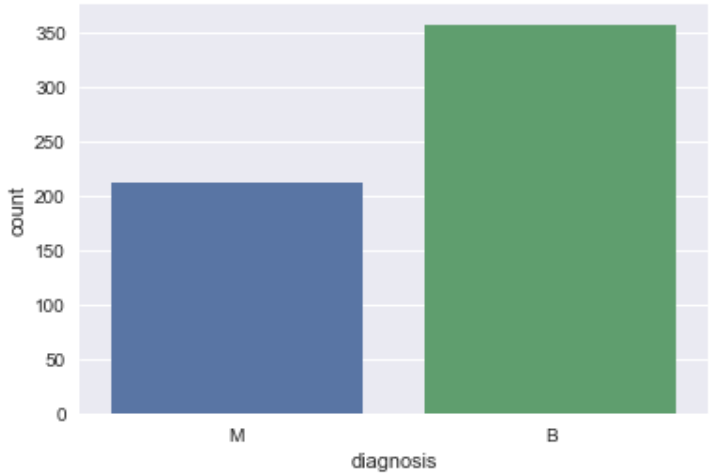


Ilustración 3Conteo de diagnósticos malignos y benignos.

Teniendo en cuenta el comportamiento observados de los datos, durante el presente trabajo se propone aplicar 7 clasificadores a las variables *radius_mean, texture_mean, smoothness_mean, concave points_mean, symmetry_mean, fractal_dimension_mean.*Y se divide el dataset de manera aleatoria en un 70% de datos de entrenamiento y 30% para evaluación.

3.5.1. Clasificador Árboles de Decisiones

Árboles de decisiones *(Decision Tree)* es un método comúnmente utilizado en la minería de datos (Rokach & Maimon, 2007) para la clasificación supervisada; sin embargo, puede ser usado para el pronóstico mediante la división del dataset en datos de entrenamiento y de evaluación. El objetivo es crear un modelo que predice el valor de una variable de destino en función de las variables predictoras mediante la construcción de nodos y la relación entre estos. Se caracteriza por la facilidad de interpretación y de tratamiento de datos aunque para el caso de árboles muy complejos (cantidad de variables predictoras y sus relaciones) suelen tener inconvenientes para generar buenos pronósticos debido al sobre ajuste(Strobl, Malley, & Tutz, 2009).

3.5.2. Clasificador Bosque Aleatorio

Clasificador de bosque aleatorio o selva aleatoria *(Random forest)* es una combinación de árboles predictores tal que cada árbol depende de los valores de un vector aleatorio probado independientemente y con la misma distribución para cada uno de estos (Breiman, 2001). Es decir, a diferencia del clasificador de Árbol de decisiones, la selva aleatoria genera múltiples árboles los cuales generan el pronóstico de manera simultánea a partir de *sub-datasets,* luego los modelos obtenidos son comparados entre sí (con los datos de evaluación) con el fin de seleccionar los mejores y generar así un modelo más robusto y eficaz.

3.5.1. Clasificador Adaboost

Adaboost o *Adaptative boosting*, es un algoritmo cuya finalidad es generar un clasificador fuerte o robusto a partir de la combinación de clasificadores más débiles (Freund & Schapire, 1995), para ello genera múltiples iteraciones y en cada una analiza un clasificador débil el cual según su nivel de acierto se le asigna un peso. A medida que se ejecutan las iteraciones se van generando nuevos clasificadores los cuales finalmente son combinados mediante una suma ponderada y se espera que su combinación (clasificación de los datos) obtenga un mejor desempeño que los clasificadores débiles obtenidos. Para el presente trabajo se construyen los clasificadores débiles mediante una selva aleatoria.

3.5.2. Clasificador Máquina Vector de Soporte

Support Vector Machines, (SVMs) son algoritmos de aprendizaje supervisado relacionados estrechamente con los problemas de clasificación y regresión con los cuales se busca generar espacios en los cuales las variables de predicción puedan estar lo más separado posible con el fin de ser divididas mediante un hiperplano (Cortes & Vapnik, 1995) el cual maximiza la distancia o proyección de los individuos de las diferentes categorías más cercanos entre sí. Dentro de las variantes de los SVMs existen algunas en las cuales se generan dimensiones extras con el fin de separarlos de manera lineal, polinomial, perceptón, de base radial gausseana etc; durante el presente trabajo y teniendo en cuenta el análisis descriptivo de la sección 3.5 se utiliza una máquina de separación lineal.

3.5.3. Clasificador k Vecinos Más Cercano

Es un método de clasificación supervisada no paramétrico que estima el valor de la función de densidad de probabilidad *a posteriori* de que un elemento x pertenezca a la clase C_j (Fix & Hodges, 1951) a partir de la información de una base de entrenamiento para ello no se realizan suposiciones sobre el comportamiento de las variables predictoras.

3.5.4. Clasificador de Regresión Logística

Esta metodología se enmarca dentro del conjunto de Modelos Lineales Generalizados (MLG) (Montgomery, Peck, & Vining, 2012) que usa como función de enlace la función logit. Las probabilidades que describen el posible resultado de un único ensayo se modelan, como una función de variables explicativas, utilizando una función logística. La estimación de los coeficientes se realiza mediante la técnica de máxima verosimilitud de la función de probabilidad Binomial ya que la variable predicción con cumple con el supuesto de homocedasticidad.

3.5.5. Clasificador de Bayesiano Ingenuo

Es un clasificador probabilístico fundamentado en la teoría de Bayes y algunas hipótesis simplificadoras que se suelen resumir en la hipótesis de independencia entre las variables

predictoras, que recibe el apelativo de ingenuo. Entre la academia aún se discute el momento en el que se formuló la aplicación siendo posiblemente Nicholas Saunderson antes de la muerte de Bayes (Stigler, 1983). Si bien en la sección 3.5 se encuentra altos grados de relación entre los datos, durante el presente trabajo se supondrá que son independientes las variables predictoras. Con el fin de evitar el sobreajuste de los modelos se desarrolla un análisis de validación cruzada.

3.6. Validación cruzada

Teniendo en cuenta las características de las técnicas de aprendizajes anteriormente descritas existe la posibilidad de generar modelos robustos con un mínimo error de calibración que explican con altísima precisión el comportamiento de los datos de entrenamiento, el inconveniente surge cuando el ajuste es tan alto que para efectos prácticos se desarrolla un modelo exacto y por tanto deja de ser un modelo estadístico impidiendo la predicción de los datos de evaluación.

Para evitar dicho problema durante el presente trabajo se utilizará un análisis de validación cruzada tipo *k-folds* el cual subdivide de manera iterativa los datos de entrenamiento en *k* partes las cuales se entrenan y estiman su contra subdivisión. De esta manera se obtienen réplicas de ajuste para k-1 *subdataset* de entrenamiento, permitiendo así ignorar un posible primer modelo sobre ajustado.

4. Resultados

4.1. Reducción dimensional

Una vez aplicada la reducción dimensional se obtiene un plano en el cual se distribuyen las características (Benigno-Maligno) con una notable diferencia (ver Ilustración 4). La proyección obtenida logra reflejar aproximadamente el 63,24% de la varianza original de los datos.

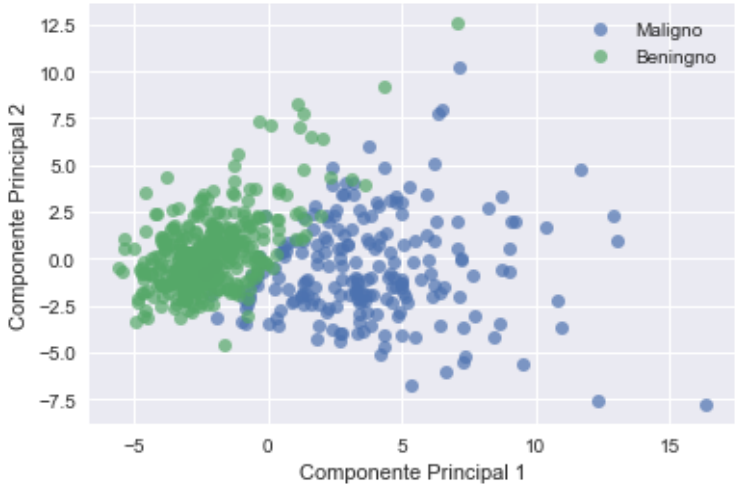


Ilustración 4 Análisis de Componentes Principales

El eje principal explica el 44,27% de la varianza mientras el segundo eje el 18,97%, donde los tumores malignos se distribuyen aparentemente en el eje positivo del factor principal y ambos de

manera parcialmente homogénea en el segundo componente. Al tener en cuenta la distribución espacial en el plano se desarrolla una clasificación lineal mediante una Máquina de Vector de Soporte Lineal (MVSL) y un Análisis de Regresión Logística (ARL).

4.1. Comparación lineal bi-dimensional

Una vez desarrollados los modelos de aprendizaje de máquinas se evidencia visualmente la similitud de pronóstico entre éstos (ver Ilustración 5), de hecho, los valores de ajuste obtenidos por la MVSL y la ARL son de 0,94 y 0,95 respectivamente lo cual indica un buen desempeño; sin embargo, es necesario tener en cuenta que la disposición espacial de los datos representa el 63,24% de la información original (medida en la dispersión de éstos) por tanto, en la siguiente sección se realizará un análisis a las variables esperadas (categoría *mean*) depuradas luego del análisis de correlación.

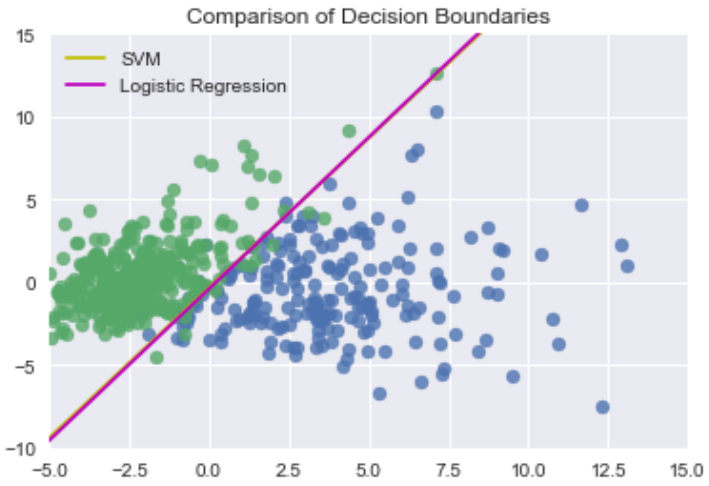


Ilustración 5 Clasificación lineal de los tumores en el plano

4.1. Desempeño de los algoritmos

Durante el presente trabajo se validan los 7 modelos de aprendizaje mediante una validación cruzada k-folds con $k = 5$ donde los resultados se consignan en la Tabla 3.

	Puntaje de validación cruzada					
Clasificador	*k=1*	*k=2*	*k=3*	*k=4*	*k=5*	*Promedio*
Árboles de decisión	86,84%	87,28%	89,18%	90,35%	90,16%	88,76%
Bosque aleatorio	88,60%	89,91%	91,52%	92,33%	92,80%	91,03%
Adaboost	82,11%	82,11%	93,86%	93,64%	92,79%	88,90%
Vectores de Soporte	77,19%	83,33%	85,67%	88,16%	87,69%	84,41%
K vecinos cercanos	75,44%	80,70%	84,50%	86,62%	86,47%	82,75%
Regresión logística	71,05%	77,63%	84,21%	86,84%	87,53%	81,45%
Bayesiano ingenuo	85,09%	88,16%	90,64%	91,67%	91,74%	89,46%

Tabla 3 Puntaje de exactitud en la clasificación de las máquinas de aprendizaje

Una vez realizadas las comparaciones se puede apreciar que aparentemente la técnica de Bosque Aleatorio generó el mejor modelo de predicción.

5. Discusión

Teniendo en cuenta el objetivo del presente trabajo se desarrollaron diversos tipos de análisis los cuales parten de supuestos bastante fuertes como en el caso del Bayesiano Ingenuo donde a pesar de encontrar correlaciones con valores superiores al 0.8 se asumió que eran espurias y por tanto las variables independientes o, que de manera n-dimensional los datos se podían separar linealmente y por tanto la aplicación de las máquinas de vectores se realizaba sin ningún análisis de mayor profundidad.

Otras características que no se abordaron con el rigor científico que requiere un análisis comparativo (y no una aproximación somera como en este caso) son la identificación de parámetros requeridos por los métodos de aprendizaje y su respectiva comparación como indicar el valor de *k* para el análisis de correspondencia, la cantidad de árboles en *random forest,* la cantidad de modelso débiles en Adaboos y la respectiva manera en la cual se generan (durante el presente trabajo simplemente se utilizó *random forest*) entre otras cosas. Lo cual recalca la necesidad de desarrollar estudios más profundos con el fin de identificar relaciones y distribuciones de los datos así mismos, evaluar de manera alternativa las diferentes máquinas de aprendizaje y mejorar los métodos de contraste.

6. Conclusiones

A partir de los desempeños estimados en la sección 4.1 se identifica que los algoritmos enfocados en la separación de datos mediante distancias presentaron peor desempeño, lo anterior puede indicar que en el espacio existen diversos traslapes entre las categorías malignas y benignas y que éstas no son trivialmente separables al menos de manera lineal. Además, se encuentra que los algoritmos de mejora como Bosque Aleatorio y Adaboost impactaron positivamente el desempeño del Árbol de decisiones el cual tiene la limitante de generar sesgos hacia las categorías con más niveles a la par de poseer una tendencia al sobre ajuste, de hecho, durante el presente trabajo se identificó que sin la aplicación del análisis cruzado los modelos de bosque Aleatorio y Adaboost presentaban ajustes del 100% en los datos de entrenamiento y eran incapaces de realizar pronósticos con una calidad similar a las otras máquinas estudiadas.

7. Agradecimiento

El presente trabajo se basó en los aportes realizados por Manish Kumar[3], DrGuillermo[4] y Luis Bronchal[5] de la comunidad Kaggle[6]

8. Bibliografía

Bennett, K. (1992). Decision tree construction via linear programming. Retrieved from http://www.rpi.edu/~bennek/papers/DecisionTree.pdf

Breiman, L. (2001). Random Forests. *Machine Learning*, *45*(1), 5–32. https://doi.org/10.1023/A:1010933404324

Cortes, C., & Vapnik, V. (1995). Support-Vector Networks. *Machine Learning*, *20*(3), 273–297. https://doi.org/10.1023/A:1022627411411

Fix, E., & Hodges, J. L. (1951). An Important Contribution to Nonparametric Discriminant Analysis and Density Estimation. *International Statistical Review*, *3*(57), 233–238. Retrieved from https://scholar.google.com/scholar?q=An+Important+Contribution+to+Nonparametric+Discriminant+Analysis+and+Density+Estimation&btnG=&hl=es&as_sdt=0%2C5

Freund, Y., & Schapire, R. (1995). A desicion-theoretic generalization of on-line learning and an application to boosting. *Computational Learning Theory*, *55*, 119–139. https://doi.org/10.1006/jcss.1997.1504

Mangasarian, O. L., Street, W. N., & Wolberg, W. H. (1995). Breast Cancer Diagnosis and Prognosis Via Linear Programming. *Operations Research*, *43*(4), 570–577. https://doi.org/10.1287/opre.43.4.570

Mira, J. (1995). Aspectos básicos de la Inteligencia Artificial. Retrieved from https://dialnet.unirioja.es/servlet/libro?codigo=369975

Montgomery, D. C., Peck, G. E. A., & Vining, G. (2012). *Introduction to linear regression analysis*. Retrieved from https://books.google.com/books?hl=es&lr=&id=27kOCgAAQBAJ&oi=fnd&pg=PP1&dq=montgomery+introduction+to+linear+regression+montgomery&ots=hQYtiSr5sC&sig=VU-3eAlmyvDMvDGAQygY8Fk7cyw

Palma Méndez, J. T., & Morales, R. M. (2008). *Inteligencia artificial – Métodos y aplicaciones*. Madrid: McGraw-Hill. Retrieved from https://books.google.com.co/books/about/Inteligencia_artificial_Técnicas_métod.html?id=cB8PPwAACAAJ&redir_esc=y

Pearson, K. (1901). On lines and planes of closest fit to systems of points in space. *The London, Edinburgh, and Dublin Philosophical Magazine and Journal of Science*, *2*(1), 559–572. https://doi.org/10.1080/14786440109462720

Ponce, P. (2010). *Inteligencia artificial con aplicaciones a la ingeniería*. *ISBN 978-84-267-1706-1*. Mexico: Alpha Omega.

Rokach, L., & Maimon, O. (2007). *Data mining with decision trees: theory and applications*. (Springer, Ed.), *Data Mining and Knowledge Discovery*. Retrieved from https://books.google.com/books?hl=es&lr=&id=OVYCCwAAQBAJ&oi=fnd&pg=PR6&dq=+Data+mining+with+decision+trees:+theory+and+applications&ots=tIp7f-5fUR&sig=6aofERY1BVDvjTngTlNAUrdYAEg

Shapiro, S. C. (1992). *Encyclopedia of artificial intelligence*. Wiley. Retrieved from https://books.google.com.co/books/about/Encyclopedia_of_artificial_intelligence.html?id=fKURAQAAMAAJ&redir_esc=y

Stigler, S. M. (1983). Who Discovered Bayes's Theorem? *The American Statistician*, *37*(4a), 290–296. https://doi.org/10.1080/00031305.1983.10483122

Street, W., Wolberg, W. H., & Mangasarian, O. L. (1993). Nuclear feature extraction for breast tumor diagnosis. *ISandT/SPIE International Symposium on Electronic Imaging: Science and Technology*,

[3] https://www.kaggle.com/gargmanish/d/uciml/breast-cancer-wisconsin-data/basic-machine-learning-with-cancer

[4] https://www.kaggle.com/drgilermo/d/uciml/breast-cancer-wisconsin-data/exploration-of-svm

[5] https://www.kaggle.com/lbronchal/d/uciml/breast-cancer-wisconsin-data/breast-cancer-dataset-analysis

[6] https://www.kaggle.com/uciml/breast-cancer-wisconsin-data/kernels

1905, 861–870. https://doi.org/10.1117/12.148698

Strobl, C., Malley, J., & Tutz, G. (2009). An introduction to recursive partitioning: Rationale, application, and characteristics of classification and regression trees, bagging, and random forests. *Psychological Methods*, *14*(4), 323–348. https://doi.org/10.1037/a0016973

Talwar, A., & Kumar, Y. (2013). Machine Learning: An artificial intelligence methodology. *International Journal of Engineering and Computer*. Retrieved from http://ijecs.in/issue/v2-i12/11 ijecs.pdf

Wolberg, W. H., Street, N., & Mangasarian, O. L. (1995). Breast Cancer Wisconsin (Diagnostic) Data Set. Wisconsin: UCI Machine Learning Repository. Retrieved from https://archive.ics.uci.edu/ml/datasets/Breast+Cancer+Wisconsin+(Diagnostic)

Wolberg, W. H., Street, W. N., Heisey, D. M., & Mangasarian, O. L. (1995). Computer-derived nuclear features distinguish malignant from benign breast cytology. *Human Pathology*, *26*(7), 792–796. https://doi.org/10.1016/0046-8177(95)90229-5

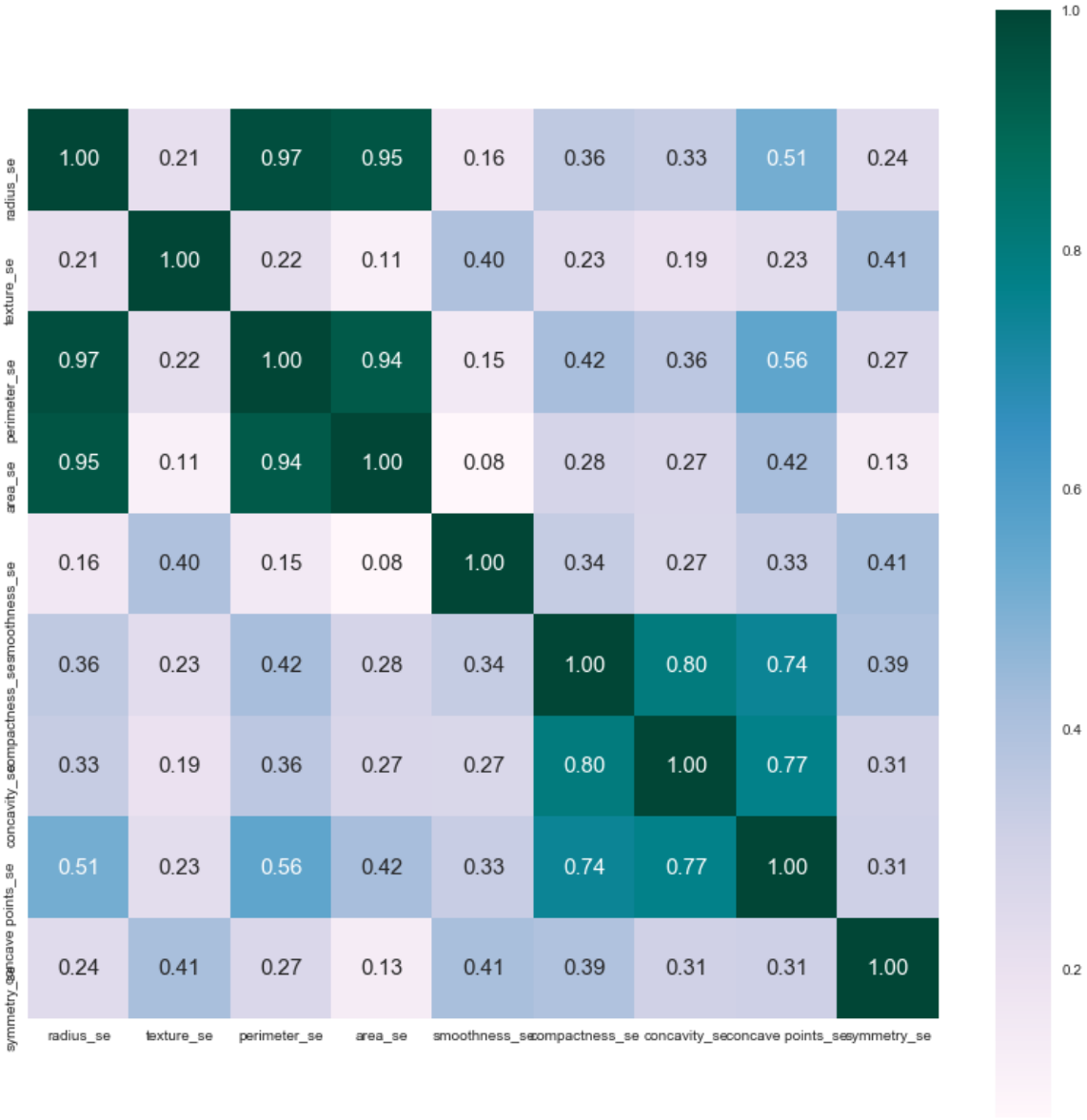
Wolberg, W. H., Street, W. N., Heisey, D. M., & Mangasarian, O. L. (1995). Computerized breast cancer diagnosis and prognosis from fine-needle aspirates. *Arch Surg*, *130*(5), 511–516. https://doi.org/10.1001/archsurg.1995.01430050061010

Wolberg, W. H., Street, W. N., & Mangasarian, O. L. (1994). MACHINE LEARNING TECHNIQUES TO DIAGNOSE BREAST-CANCER FROM IMAGE-PROCESSED NUCLEAR FEATURES OF FINE-NEEDLE ASPIRATES. *Cancer Letters*, *77*(2–3), 163–171. https://doi.org/10.1016/0304-3835(94)90099-x

Wolberg, W., & Street, W. (1995). Image analysis and machine learning applied to breast cancer diagnosis and prognosis. *And Quantitative Cytology …*. Retrieved from https://www.researchgate.net/profile/Nick_Street/publication/15587278_Image_analysis_and_machine_learning_applied_to_breast_cancer_diagnosis_and_prognosis/links/549f75a80cf267bdb8fdbdd8.pdf

9. Anexos

Ilustración 6 Mapa de calor para la correlación entre variables de desviación estándar



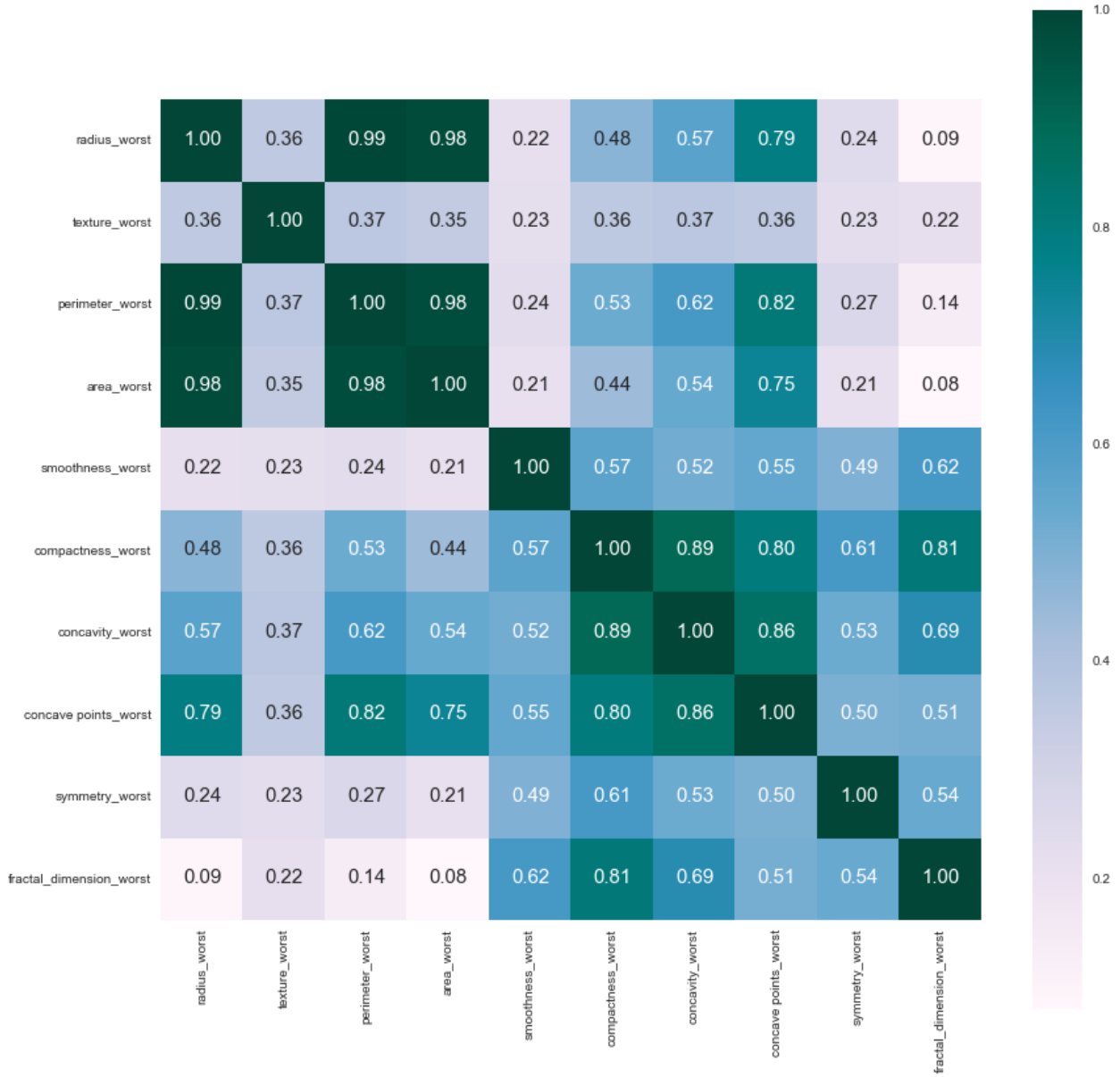


Ilustración 7 Mapa de calor para la correlación entre los valores mínimos de las variables