Machine learning: Cancer

Diego Alejandro Ríos Pérez[1]
Universidad de antioquia
Facultad de ciencias exactas y naturales
17 de diciembre de 2021

Resumen

El Machine Learning es una técnica computacional comúnmente usada para desarrollar entrenamiento de modelos y predicciones de eventos en diferentes campos de la ciencia. Particularmente, este trabajo intenta desarrollar una predicción relativa al estado de la diagnosis de un tumor (maligno o benigno) que se ha detectado en ciertos pacientes, lo cual puede contribuir al tratamiento prematuro de la causa del tumor. Se efectúa un proceso de predicción de diagnosis, que finaliza en un análisis estadístico relativo a lo que arrojan los resultados computacionales.

1. Introducción

El estudio que aquí se presenta, toma conjunto de datos de 569 pacientes que han sido diagnosticados con un tumor de cierto tipo y de los cuales, se han recopilado diferentes características asociadas a la geometría de las células que este contiene:

- Radio.
- Texture.
- Perímetro.
- Área.
- Suavidad.
- Compacidad.
- Concavidad.
- Puntos cóncavos.
- Simetría.
- Dimensión fractal.

Se pretente desarrollar un programa con Machine learning que, basado en los diferentes integrantes del conjunto, se puedan hacer predicciones con respecto de los pacientes implicados en general.

Se iniciará efectuando un análisis estadístico de la muestra, para posteriormente desarrollar árboles de decisión e implementar las predicciones bajo móludos diferentes en `python 3`.

2. Aplicación

Inicialmente, se desea desarrollar el análisis de la muestra de manera estadística. Un total de 63 % de los datos, representan tumores benignos, mientras que el 37 % restante se asocia a malignos. Para iniciar el estudio, es importante que todas las estradas de la matriz `DataFrame` contengan entradas tipo `float`, para ello, se fijó a los valores "M" y "B" de la diagnosis (maligno y benigno), los valores 1 y 0 respectivamente.

Los datos poseen tres grandes clasificaciones:

- **Valores promedios:** Se refiere al valor medio de las características geométricas de todas las células.

[1] diego.riosp@udea.edu.co

- **Valores peores:** Es el valor peor encontrado para cada característica, por ejemplo, en el caso del radio de la célula, es el radio más grande.
- **Error estándar:** Es la parte del conjunto de datos que recopila el error estandar en la medición de una característica dada

La idea consiste entonces en iniciar la predicción de la diagnosis del estado del cáncer de acuerdo a los valores medios reportados en el conjunto. La tabla 1 muestra una generalidad de la muestra en estudio, respecto de algunas de las características celulares mencionadas.

	Diagnosis	**Radio**	**Textura**	**Perímetro**	**Área**	**Suavidad**
Media	0.37	14.13	19.29	91.97	654.89	0.10
Desviación	0.48	3.52	4.30	24.30	351.91	0.01
Mínimo	0.00	6.98	9.71	43.79	143.50	0.05
25 %	0.00	11.70	16.17	75.17	420.30	0.09
50 %	0.00	13.37	18.84	86.24	551.10	0.10
75 %	1.00	15.78	21.80	104.10	782.70	0.11
Máximo	1.00	28.11	39.28	188.50	2501.00	0.16

Cuadro 1: Valores promediados sobre todas las células

Nótese que el área posee una gran desviación estándar, posiblemente asociada a la dificultad de su medición directa. Por otro lado, el radio tiene buena precisión en la medida. Además, es de esperarse que los valores grandes de radios celulares, impliquen que el cancer sea maligno, tal como se aprecia para valores por encima del tercer cuartil, done el estado de diagnosis es 1.

Es importante revisar las correlaciones entre variables, puesto que nos interesa observar los parámetros ideales para desarrollar predicciones. Para ello, se desarrolla una regresión anotando los valores del grado de correlación entre las variables tal como lo muestra la figura 1. Por supuesto, la diagonal es unitaria por la repetición de variables. Se aprecia una estrecha relación entre las propiedades geométricas como radio y perímetro, lo cual es evidente. Asún asó, surgen relaciones interantes como área y concavidad con altos coeficientes de correlación. La suavidad y la textura, por ejemplo, no tienen nada que ver por su valor -0.02 en la correlación.

Posterior al análisis de correlación, se inicia la modelación para la predicción del estado del cancer. Usando la función de *train-test*, se escoge una lista aleatoria para `train` (70 % de los datos) y otra para `test` (30 %). Lo anterior, con el propósito de que el programa aprenda con los valores contenidos en `train`, y logre predecir la diagnosis respecto de las características dadas en `ttest`.

La generación de árboles de decisión se efectuó con dos módulos diferentes:

- `RandomForestClassifier`
- `SVM`

El uso de ambos módulos, se hace para comparar la eficiencia de cada uno y optar por usar el que mejor muestra de exactitud imprima. Así, la implementación de ambos módulos arrojó una exactitud del 94.15 % para el caso de `RandomForestClassifier` y 91.81 % tratándose de `SVM`, con respecto de la muestra real. La comparación predicción-realidad, se muestra en las figuras 2 y 3. Allí se observa

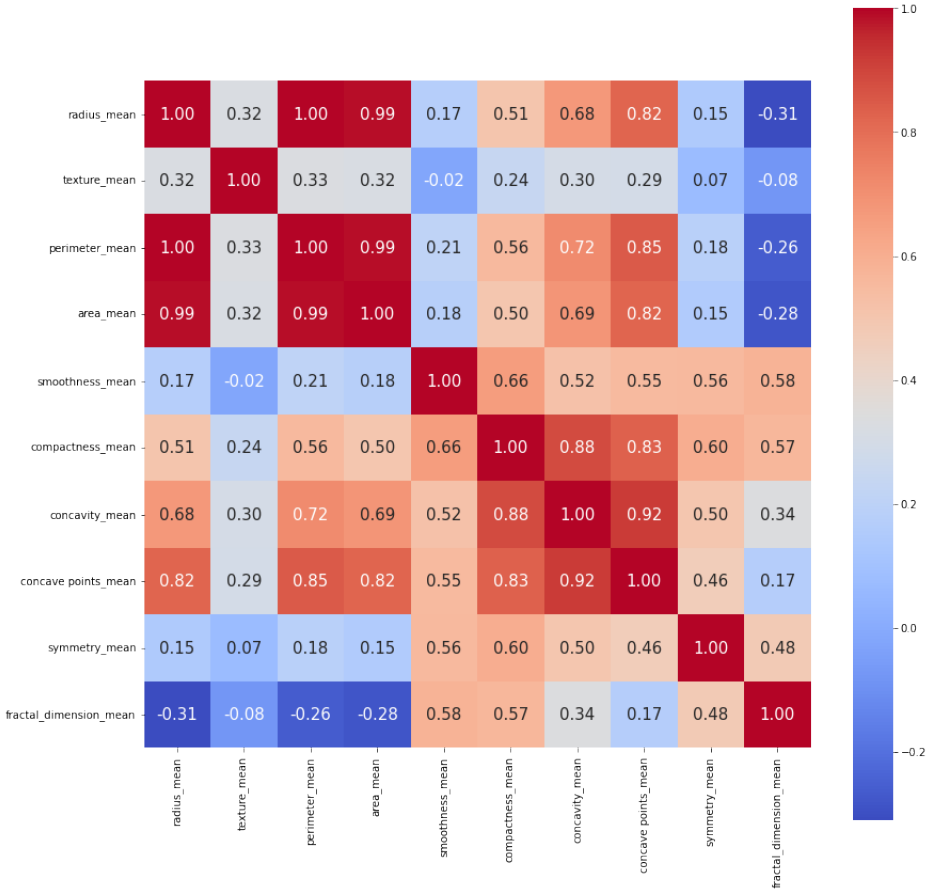

Figura 1: Correlación entre variables

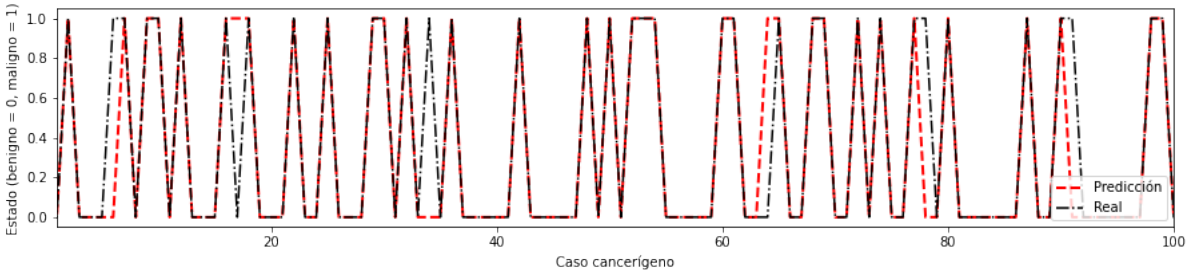

Figura 2: Relación predicción-realidad bajo `RandomForestClassifier`

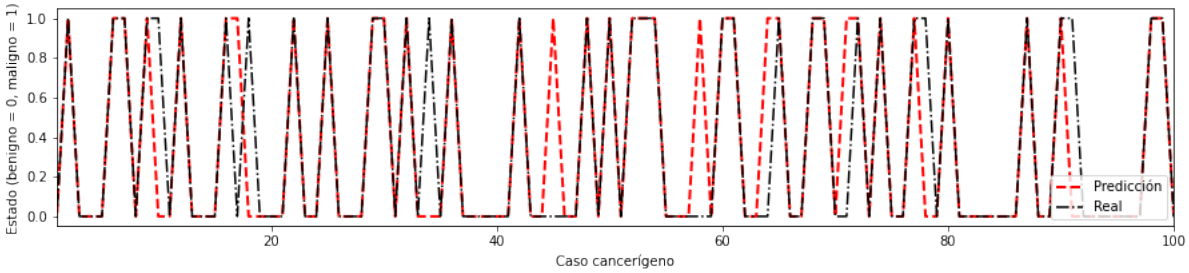

Figura 3: Relación predicción-realidad bajo `SVM`

que la comparación entre módulos es innecesaria, puesto que ambos predicen muy bien los casos. Sin embargo, si de exactitud se tratase, realmente vale la pena escoger `RandomForestClassifier` por su mayor porcentaje de acertación.

Es importante también observar las implicaciones de las diferentes características celulares, es decir, aquella que mayor relevancia tienen para la predicción. Lo anterior, puede conocerse con el simple uso de la función `sort_values` sobre el conjunto de datos para la predicción. Esto, arrojó la siguiente caracterización de variables de acuerdo a su nivel de importancia (tabla 2):

Variable	**Impacto sobre la predicción**
Perímetro	0.23
Radio	0.22
Área	0.18
Puntos de concavidad	0.17
Concavidad	0.07
Suavidad	0.03
Textura	0.03
Compacidad	0.03
Simetría	0.02
Dimensión fractal	0.02

Cuadro 2: Variables con mayor relevancia para la predicción de la diagnosis

La anterior tabla, se sustenta en el desarrollo de la predicción tomando valores críticos del conjunto, es decir, aquellos que están en el mencionado conjunto de los peores.

Basados en la imformación suministrada por las variables relevantes, se optó por hacer el análisis y la predicción usando solo las 5 primeras variables importante, para observar cómo aumenta o disminuye la exactitud de predicción. El resultado, es similar pero empeoró. `RandomForestClassifier` 93.57 % y `VSM` 89.47 %. La continuación de pruebas con los valores peores y los medios, implicó en todos los casos una sugerencia de escoger el módulo `RandomForestClassifier` para las predicciones e incluir todas las variables, aunque existan algunas no tan relacionadas a la predicción.

Finalmente, se muestran los diagramas de dispersión asociados a las variables promediadas. Se observan en ellos relaciones evidentes y principalmente, que el tumor maligno implica un general incremento en todas las variables, puesto que los puntos rojos en la dispersión están en general a la derecha de los azules y por encima.

3. Conclusiones

El machine learnign, con su entrenamiento de modelos funciona bastante bien en la predicción de posibles estados diagnósticos del cancer. Usando el mósulo `RandomForestClassifier` se observan predicciones aceptables por encima del 95 %, lo cual da un nivel de confianza discutible si se tratase del campo real en la medicina.

La diagnosis de tumores es fuertemente dependiente de las características geométricas que lo identifican, lo cual se observó en los escán de dispersión. Además, las variables como la simetría y la dimensión fractal de las células asociadas, tienen un impacto casi nulo en la estimación de la diagnosis de un tumor.

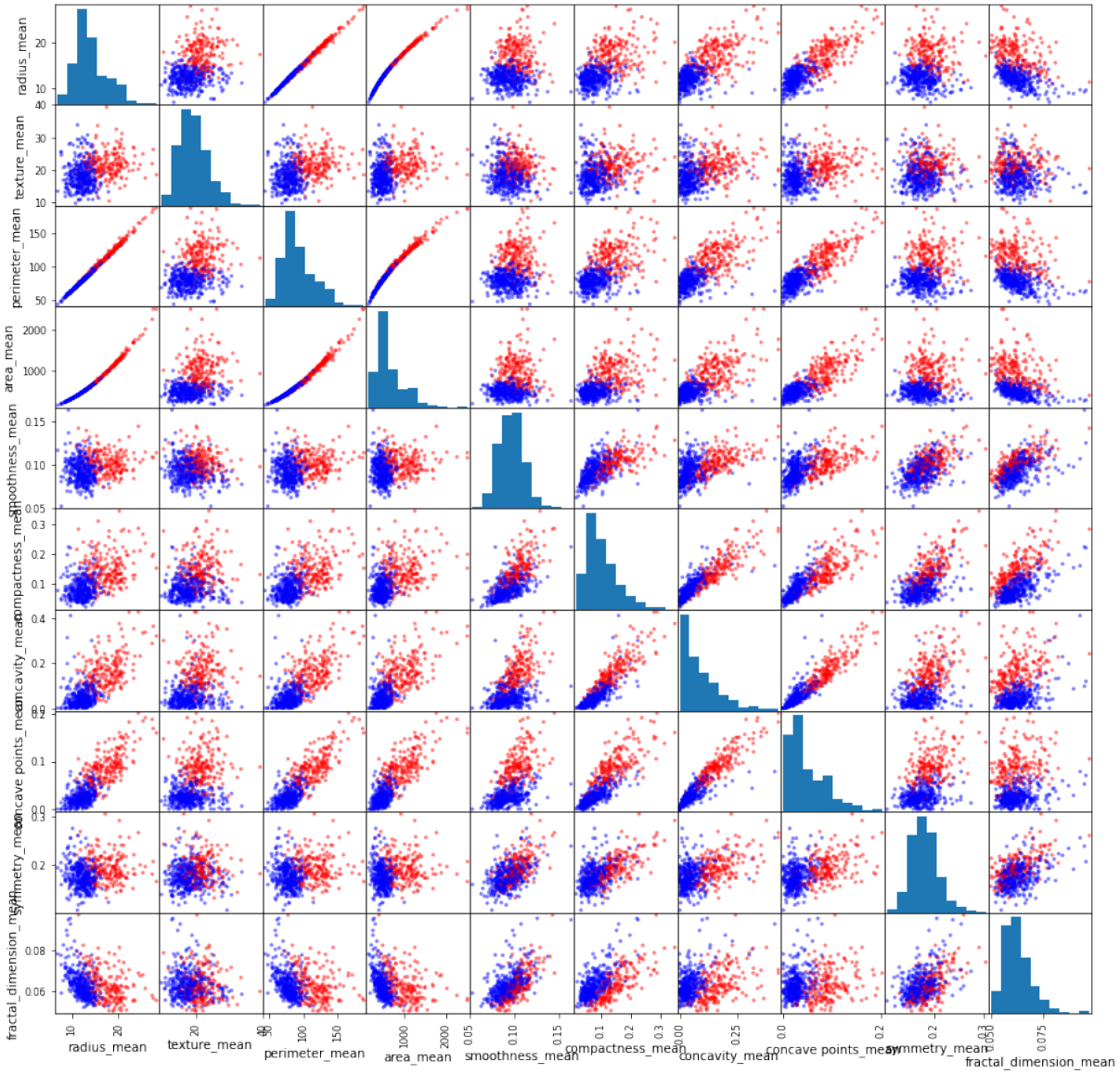

Figura 4: Dispersión respecto de las variables imlicadas