INSTITUTO TECNOLÓGICO AUTÓNOMO DE MÉXICO

Proyecto Final

Comparación de desempeño de modelos aplicados a un problema real

Carlos Marcelo Barrera Nolasco
Machine Learning

15 de diciembre de 2017

Abstract

Las técnicas de Aprendizaje de Máquina (ML por sus siglas en inglés) que estudiamos nos ofrecen la posibilidad de analizar data sets con el fin de encontrar patrones en ellos y, hacer predicciones sobre datos futuros. Este proyecto tiene como propósito el aplicar en un problema de la vida real algunos de los modelos de ML que estudiamos durante el curso; asimismo se hará una comparación de su desempeño con el fin de escoger el modelo que mejor se adapte a las necesidades particulares del problema.

Nota preliminar

La descripción del proyecto pedía un data set con al menos 5000 entradas y 10 atributos, sin embargo, no pude encontrar alguno que satisficiera ambos criterios. El data set de Breast Cancer Wisconsin (Prognostic) me pareció muy interesante de analizar pero tiene solamente 569 observaciones. Debido a lo anterior, haré ambos análisis, tanto del data set "interesante" como de uno que cumpla con los criterios pedidos.

Parte 1 Breast Cancer Wisconsin (Prognostic)

1 Introducción

El cáncer de mama (también conocido como carcinoma de mama) es el tipo de cáncer más común en mujeres en todo el mundo y, a nivel nacional, es la principal causa de muerte en mujeres mayores de 25 años (Centro Nacional de Equidad de Género y Salud Reproductiva).

Como ocurre con todos los tipos de cáncer, la detección temprana (es decir, su detección antes de que empiecen los síntomas) es un elemento clave para la lucha contra esta enfermedad elevando las posibilidades de curación. (IMSS)

Uno de los métodos utilizados para la detección de cáncer es obtener una biopsia de tejido de mama y analizar visualmente las características de los núcleos de las células de dicha biopsia. Es precisamente en este método donde podemos apoyarnos en las técnicas de ML que cada vez son más usados en el campo del diagnóstico médico debido a su efectividad en procesos de clasificación.

2 Descripción del data set

El data set fue descargado desde la página de Kaggle (Kaggle, *Breast Cancer Wisconsin (Prognostic) Data Set*) cuenta con 569 observaciones y 32 atributos.

Descripción de los atributos

Como se dijo, el data set cuenta con 32 atributos que se distribuyen de la siguiente manera:

1	ID number
2	Variable target: diagnóstico (M= maligno, B= benigno)
3 – 32	30 atributos explicativos; promedio, error estándar y el valor más extremo dentro de la imagen de cada uno de los siguientes parámetros:

Radio, textura, perímetro, área, suavidad (smoothness), compacidad (compactness), concavidad, puntos cóncavos, simetría y dimensión fractal.

Los atributos explicativos son números reales redondeados a 4 decimales.

El data set no tiene valores faltantes.

	id	diagnosis	radius_mean	texture_mean	perimeter_mean	area_mean	smoothness_mean	compactness_mean	concavity_mean	concave points_mean	...	te
0	842302	M	17.99	10.38	122.80	1001.0	0.11840	0.27760	0.3001	0.14710	...	
1	842517	M	20.57	17.77	132.90	1326.0	0.08474	0.07864	0.0869	0.07017	...	
2	84300903	M	19.69	21.25	130.00	1203.0	0.10960	0.15990	0.1974	0.12790	...	
3	84348301	M	11.42	20.38	77.58	386.1	0.14250	0.28390	0.2414	0.10520	...	
4	84358402	M	20.29	14.34	135.10	1297.0	0.10030	0.13280	0.1980	0.10430	...	

Fig 1. Head del data set

3 Metodología

Realizaremos los siguientes pasos para el análisis de este data set.

3.1 Análisis Eploratorio

A continuación haremos un análisis exploratorio gráfico (GEDA) de los atributos explicativos contra la variable target, al ser números reales todas las explicativas, se usaron boxplots para graficar la relación. Se muestran solamente algunas de las variables con su explicación. (El resto de las gráficas se encuentran en el Anexo 1)

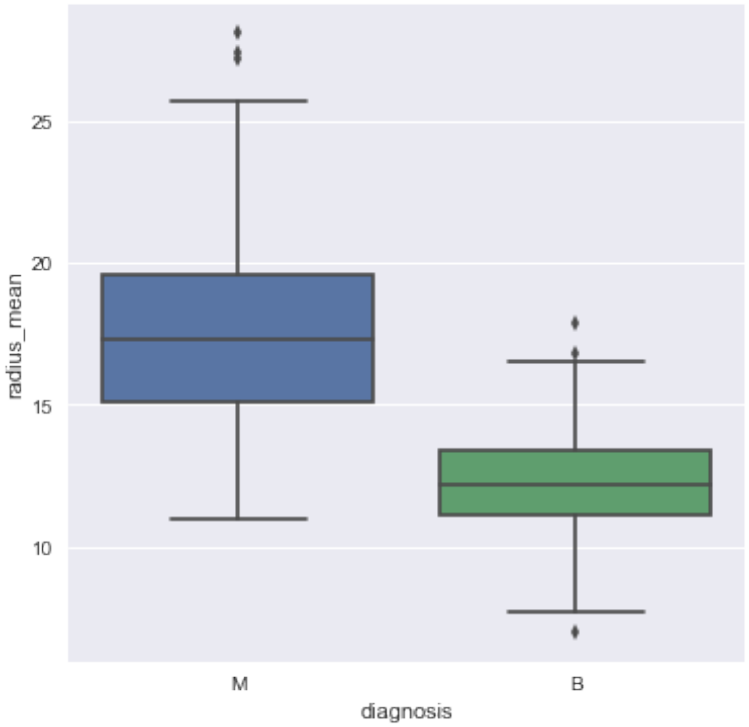


Fig 2. radius_mean vs diagnosis GEDA

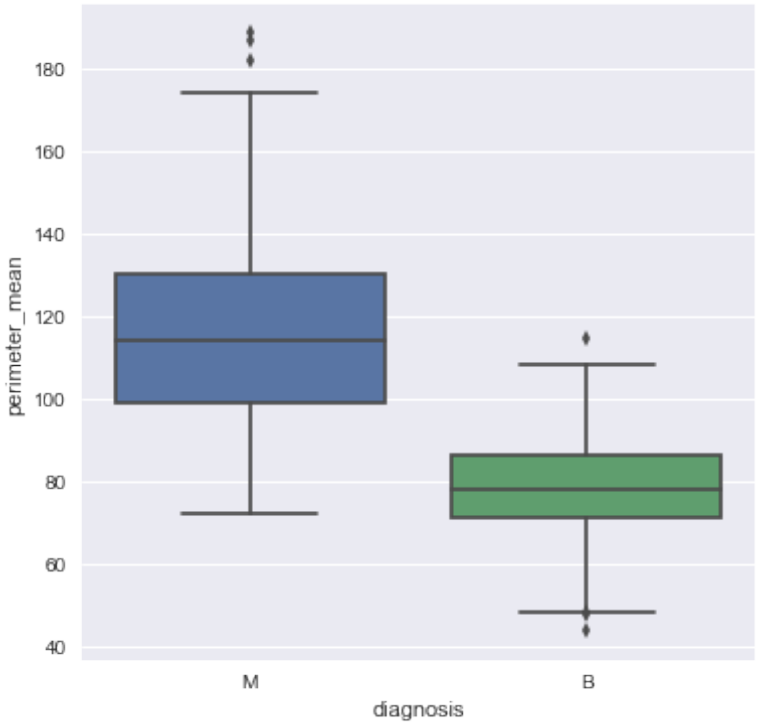


Fig 3. perimeter_mean vs diagnosis GEDA

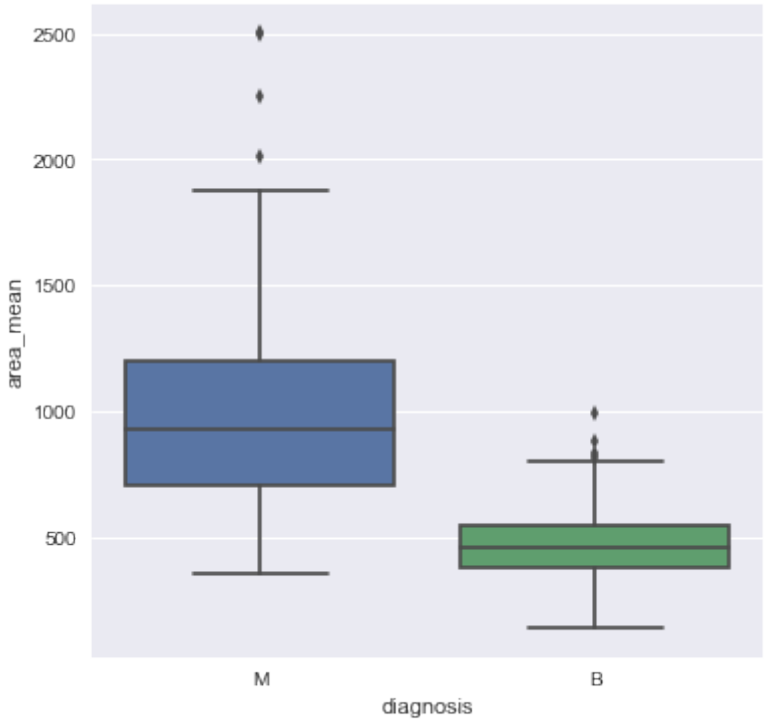


Fig 4. area_mean vs diagnosis GEDA

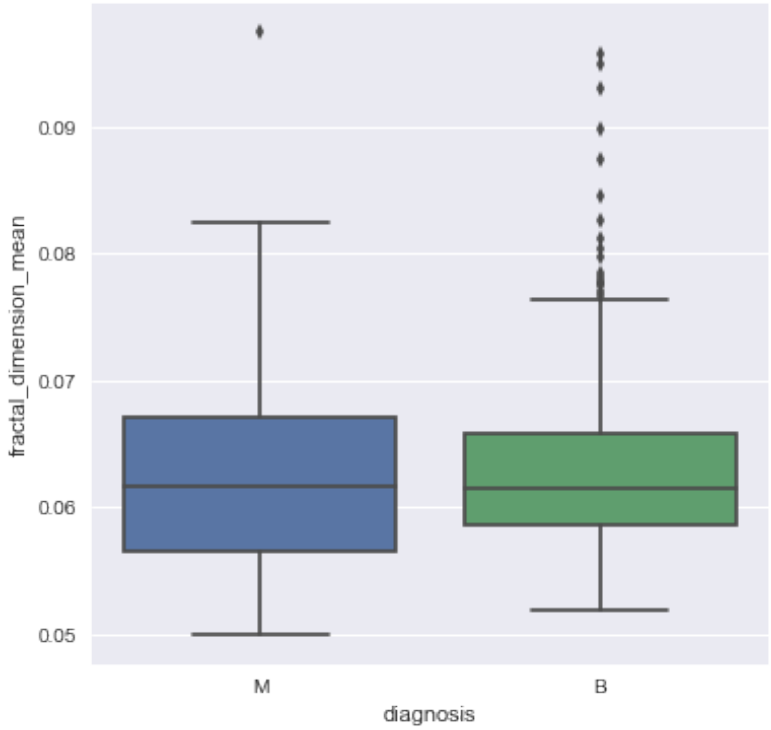


Fig 5. fractal_dimension_mean vs diagnosis GEDA

Como podemos observar a partir de las figuras 2, 3 y 4, los atributos directamente relacionados con las dimensiones como el perímetro el radio y el área de los núcleos de las células pueden separar con bastante confianza las clases “maligno” y “benigno” tendiendo a ser mas “grandes” (mayor radio, mayor perímetro, mayor área) los núcleos pertenecientes a observaciones benignas.

Por otro lado una medida menos convencional como lo es la dimensión fractal (la aproximación del perímetro de una superficie con irregularidades en su borde) arroja resultados muy similares para ambos grupos, aportando poca información a nuestro proceso de clasificación.

3.2 Separación del data set en entrenamiento y test

Para el entrenamiento y test de nuestros modelos y, debido a que contamos con pocos datos, realizaremos una partición 70-30.

3.3 Normalización de los datos

Antes de ejecutar los modelos, normalizaremos los datos haciendo uso del StandardScaler de scikit learn (haremos fit del Scaler únicamente con los datos de entrenamiento).

3.4 Aplicación de varios modelos

En todos los casos, utilizaremos un k-fold cross validation con un k=5 debido al reducido número de observaciones para la comparación de resultados con distintos hiperparámetros.

3.4.1 Máquinas de Soporte Vectorial (SVM)

Evaluaremos los siguientes hiperparámetros:

kernel = {linear, rbf}
C = {1,100}

3.4.2 Decision Tree

Evaluaremos los siguientes hiperparámetros

max_features = {auto, log2, None}

3.4.3 Random Forest

Evaluaremos los siguientes hiperparámetros

max_features = {auto, log2}
n_estimators {10, 100}

3.4.4 Artificial Neural Networks

Evaluaremos una red neuronal con una capa oculta de 10 neuronas

4 Resultados

A continuación se mostrará una tabla con los resultados de la matriz de confusión para cada uno de los modelos que se mencionan en el apartado anterior. Utilizaremos la matriz de confusión para comparar el desempeño ya que la variable que más nos debe importar son los falsos negativos.

En el contexto de este problema, lo más importante es minimizar esta métrica ya que las consecuencias de un falso positivo son menos dañinas, profundizaremos sobre esta elección en el punto 5.

Modelo	Hiperparámetros	TP	TN	FP	FN
SVM	C=1, kernel = linear	115	52	3	1
	C=1, kernel = rbf	115	52	3	1
	C=100, kernel = linear	114	53	2	2
	C=100, kernel = rbf	110	52	3	6
Decision Trees	max_features = auto	108	53	2	8
	max_features = log2	106	48	7	10
	max_features = None	105	52	3	11
Random Forest	max_features = auto, n_estimators=10	114	53	2	2
	max_features = auto, n_estimators=100	109	53	2	7
	max_features = log2, n_estimators=10	112	51	4	4
	max_features = log2, n_estimators=100	112	54	1	4
ANN	1 capa oculta de 10 neuronas	114	51	4	2

Tabla 1. Resultados de matriz de confusión de los modelos revisados

5 Conclusiones

Como se mencionó en el punto anterior, se escogió el valor de FN como la métrica más importante en el contexto de este problema, esto obedece al siguiente razonamiento:

Pensemos en los falsos positivos; en este caso, se le comunicaría a una persona sana que se detectó tejido maligno en su biopsia, esto podría generar angustia en el paciente y posiblemente gastos en estudios y tratamientos innecesarios. Por otra parte pensemos en los falsos negativos; en este caso se le comunicaría a una persona con cáncer de mama que está sana, lo cual ocasionaría que el paciente no se hará nuevos estudios ni recibirá tratamiento adecuado a tiempo, dando al cáncer la posibilidad de expandirse llevando al paciente a un estado que requerirá tratamientos más agresivos y, posiblemente, a la muerte.

En la Tabla 1 podemos observar que, para este problema, las máquinas de soporte vectorial se comportan bastante bien con un valor de C=1 sin importar si se usa el kernel lineal o el rbf; en ambos casos el desempeño es el mismo y se minimiza el valor de los FN. En ambos casos, solamente se presenta un valor en esta categoría.

El siguiente paso es trazar la curva ROC de estos dos modelos y calcularemos el área bajo la curva (AUC) para tratar de hacer un desempate.

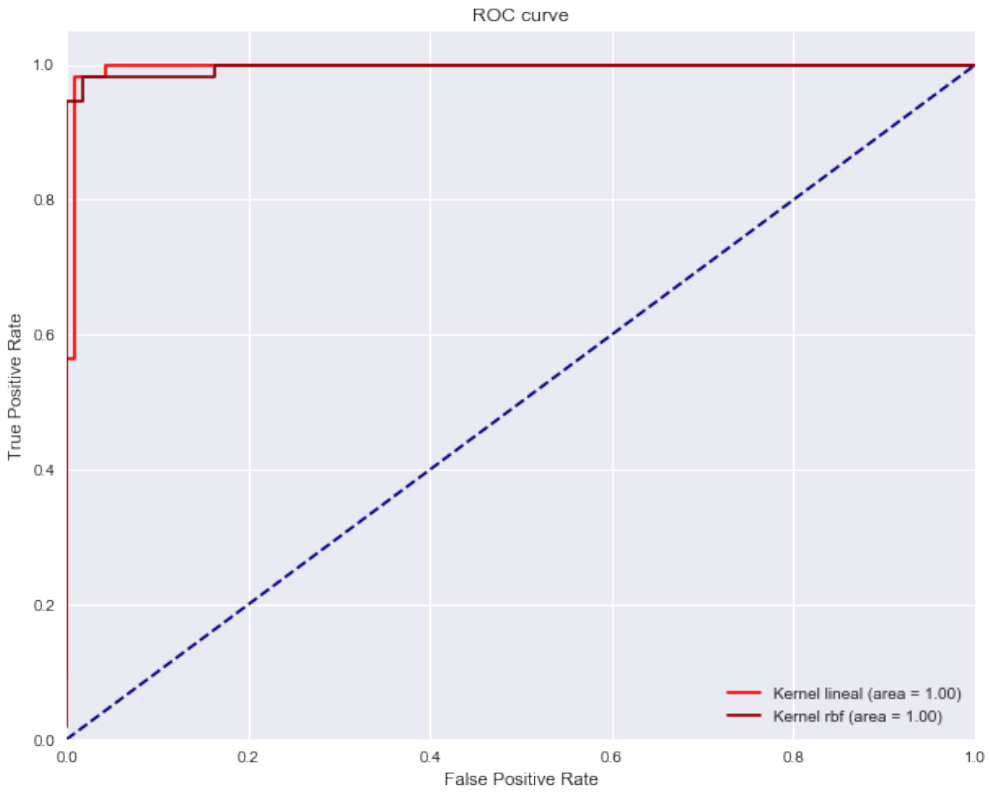


Fig 6. Curva ROC para los modelos SVM con C=1

En la figura 6, podemos observar que las curvas son muy similares e, incluso, las etiquetas del AUC son iguales (observamos que tienen valor 1, pero eso se debe al redondeo). Los valores (no redondeados) de AUC son:

Kernel lineal: 0.99561128526645759
Kernel rbf: 0.99639498432601892

Notamos que efectivamente ambos valores son muy cercanos a 1, sin embargo, el modelo **SVM con C=1 y kernel=rbf** es ligeramente mejor, de tal modo que será el seleccionado.

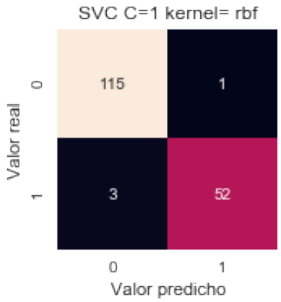


Fig 7. Matriz de confusión del modelo seleccionado

A futuro sugeriría que se digitalizaran cada vez más imágenes para poder hacer modelos más robustos, asimismo sería importante dotar a los oncólogos con acceso a un producto de datos que recibiera como input la imagen y diera como output el diagnóstico encontrado por dicho modelo. Debido a las implicaciones de un diagnóstico equivocado, sugeriría que el resultado de este modelo fuera utilizado solamente como una herramienta más para el doctor y no como un sustituto de su juicio ya que el modelo podría omitir información relevante que un humano con experiencia podría notar.

Parte 2 Adult Data Set

1 Introducción

El ingreso anual de una persona es un dato que puede tener gran importancia en una gran variedad de aplicaciones, desde las más obvias como la mercadotecnia o el sector financiero, hasta otras no tan evidentes como la política, ya que los candidatos a un cargo público pueden segmentar a su audiencia, generando discursos adaptados específicamente a un sector socioeconómico.

Dado lo anterior, por razones de seguridad, este dato es sensible y no es común que lo tengamos, por lo cual, sería de gran utilidad el poder predecirlo con base en otros datos más sencillos de obtener como su edad, sexo, raza o estado civil.

2 Descripción de los datos

El data set fue descargado desde la página de Kaggle (Kaggle, *Adult Income DataSet*) cuenta con 32561 observaciones y 15 atributos.

Descripción de los atributos

Como se dijo, el data set cuenta con 15 atributos, 14 explicativos y la clase target.

Atributos explicativos:

age	numérico
workclass	categórico
fnlwgt:	numérico
education	categórico
education-num	numérico
marital-status	categórico
occupation	categórico

relationship	categórico
race	categórico
sex	categórico
capital-gain	numérico
capital-loss	numérico
hours-per-week	numérico
native-country	categórico

Clase target:

Income	categórico (>50k, <=50k)

	age	workclass	fnlwgt	education	education_num	marital_status	ocupation	relationship	race	sex	capital_gain	capital_loss	hours_per_week	native
0	39	State-gov	77516	Bachelors	13	Never-married	Adm-clerical	Not-in-family	White	Male	2174	0	40	Unit
1	50	Self-emp-not-inc	83311	Bachelors	13	Married-civ-spouse	Exec-managerial	Husband	White	Male	0	0	13	Unit
2	38	Private	215646	HS-grad	9	Divorced	Handlers-cleaners	Not-in-family	White	Male	0	0	40	Unit
3	53	Private	234721	11th	7	Married-civ-spouse	Handlers-cleaners	Husband	Black	Male	0	0	40	Unit
4	28	Private	338409	Bachelors	13	Married-civ-spouse	Prof-specialty	Wife	Black	Female	0	0	40	

Fig 8. Head del data set Adult

El objetivo es tratar de determinar si el ingreso anual de una persona es mayor a 50,000 dólares.

3 Metodología

Utilizaremos la misma metodología que usamos en la Parte 1

- 3.1 Análisis Exploratorio
- 3.2 Separación de los datos en entrenamiento y test
- 3.3 Normalización de los datos
- 3.4 Aplicación de varios modelos
  - 3.4.1 SVM
  - 3.4.2 Decision Tree
  - 3.4.3 Random Forest
  - 3.4.4 Artificial Neural Network

3.1 Análisis Eploratorio

Se realizó un análisis gráfico de las variables explicativas en contra de la target, esto nos ayuda a encontrar información interesante (solo se muestran cuatro gráficas, las demás se encuentran en el anexo 2)

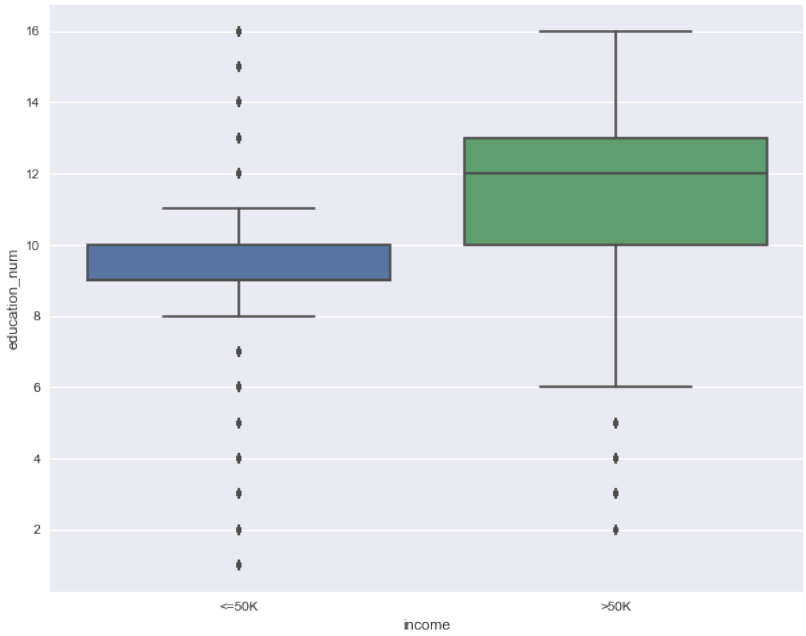


Fig 9. education_num vs income

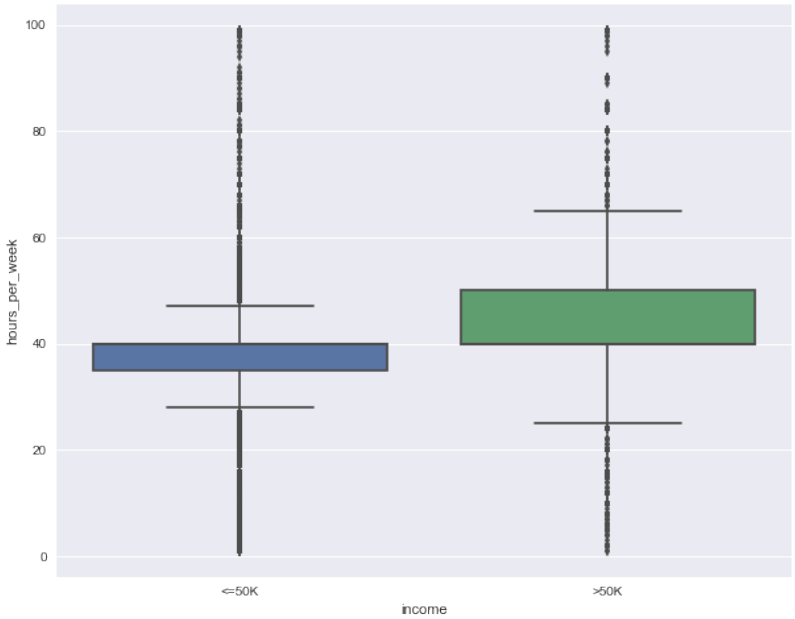


Fig 10. hours_per_week vs income

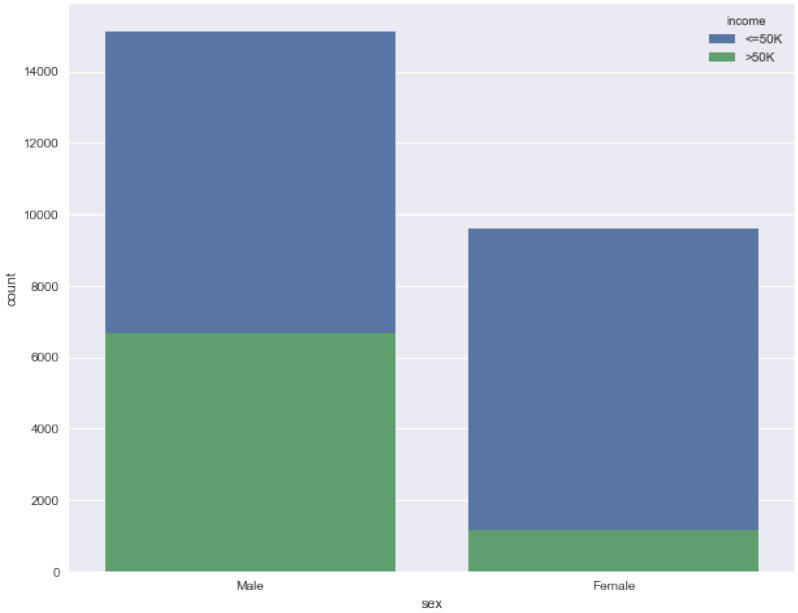


Fig 11. sex vs income

Estas gráficas nos hacen ver la relación de los atributos explicativos con la variable target y encontramos resultados interesantes. Por ejemplo, se confirman relaciones que podríamos anticipar como lo es el número de años de educación con respecto al ingreso anual, de la gráfica podemos ver que los individuos que ganan más de 50000 dólares al año tienen en promedio 12 años de estudios mientras que aquellos que ganan una cantidad menor, tienen una media de estudios de 9 años. También podemos observar que la gente con ingresos superiores trabaja, en promedio, más horas a la semana que la gente con ingresos inferiores.

Otro detalle interesante es que el sexo sigue siendo un factor muy importante ya que el número total de mujeres que tienen ingresos por arriba de los 50k es muy pequeño y, aunque hubo menos observaciones de mujeres, si nos enfocamos solamente en ellas, la proporción de observaciones con ingresos altos con respecto a aquellas con ingresos bajos es menor que la misma proporción observada en el sector masculino.

3.2 Separación del data set en entrenamiento y test

Para el entrenamiento y test de nuestros modelos y, debido a que contamos con una buena cantidad de datos, realizaremos una partición 80-20.

3.3 Transformación de los datos

Antes de ejecutar los modelos, convertiremos las variables categóricas en numéricas aplicando el método de One-Hot encoding y posteriormente normalizaremos los datos haciendo uso del StandardScaler de scikit learn (haremos fit del Scaler únicamente con los datos de entrenamiento).

3.4 Aplicación de varios modelos

En todos los casos, utilizaremos un k-fold cross validation con un k=10 ya que tenemos un número suficiente de observaciones para la comparación de resultados con distintos hiperparámetros.

3.4.1 Máquinas de Soporte Vectorial (SVM)

Evaluaremos los siguientes hiperparámetros:

kernel = {linear, rbf}
C = {1,100}

3.4.2 Decision Tree

Evaluaremos los siguientes hiperparámetros

max_features = {auto, log2, None}

3.4.3 Random Forest

Evaluaremos los siguientes hiperparámetros

max_features = {auto, log2}
n_estimators {10, 100}

3.4.4 Artificial Neural Networks

Evaluaremos una red neuronal con una capa oculta de 30 neuronas

4 Resultados

A continuación se mostrará una tabla con los resultados del *clasification report* generado por sklearn para cada uno de los modelos que se mencionan en el apartado anterior.

Modelo	Hiperparámetros	Precision	Recall	F1 Score
SVM	C=1, kernel = linear	0.84	0.85	0.84
	C=1, kernel = rbf	0.85	0.85	0.85
	C=10, kernel = linear	0.84	0.85	0.84
	C=10, kernel = rbf	0.84	0.85	0.84
Decision Trees	max_features = auto	0.81	0.81	0.81
	max_features = log2	0.80	0.80	0.80
	max_features = None	0.81	0.81	0.81
Random Forest	max_features = auto, n_estimators=10	0.83	0.84	0.83
	max_features = auto, n_estimators=100	0.84	0.85	0.85
	max_features = log2, n_estimators=10	0.83	0.84	0.84
	max_features = log2, n_estimators=100	0.84	0.85	0.84
ANN	1 capa oculta de 30 neuronas	0.58	0.76	0.66

Tabla 2 Resultados de los reportes de clasificación (avg) de los modelos

A continuación se compararán las curvas ROC de los modelos y mostraremos sus valor de Área Bajo la Curva (AUC)

Modelo	Hiperparámetros	AUC
SVM	C=1, kernel = linear	0.900885
	C=1, kernel = rbf	0.894735
	C=10, kernel = linear	0.900877
	C=10, kernel = rbf	0.886285
Decision Trees	max_features = auto	0.739175
	max_features = log2	0.731871
	max_features = None	0.740591
Random Forest	max_features = auto, n_estimators=10	0.876251
	max_features = auto, n_estimators=100	0.900210
	max_features = log2, n_estimators=10	0.875611
	max_features = log2, n_estimators=100	0.897853
ANN	1 capa oculta de 30 neuronas	0.500000

Tabla 3 AUC de los modelos

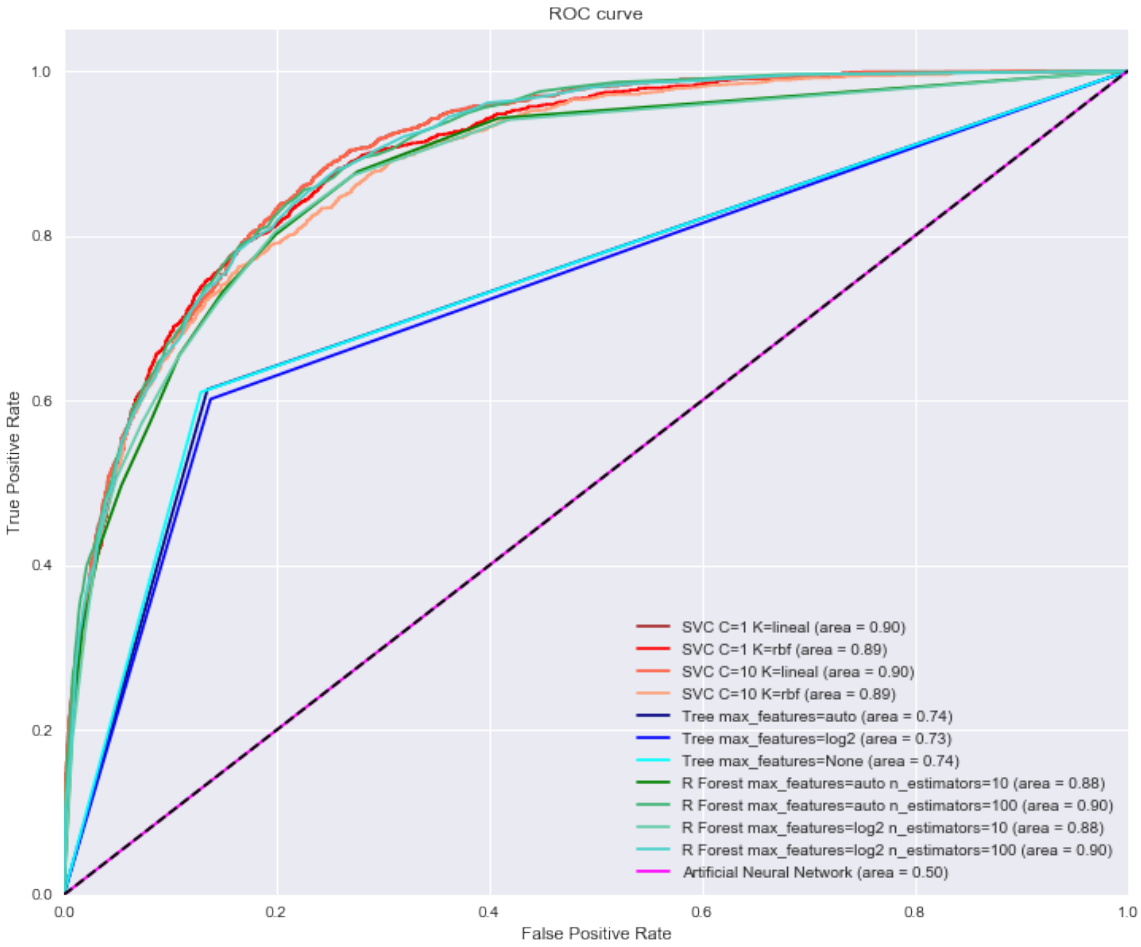


Fig 12. Curvas ROC de los modelos analizados

5 Conclusiones

Podemos observar que la arquitectura que escogimos para la red neuronal fracasó ya que su AUC es de 0.5 y su curva ROC corresponde con la línea de no-discriminación, es decir, tirar una moneda nos daría los mismos resultados. Cabe aclarar que esto no nos dice que una ANN sea una mala opción, solamente la arquitectura seleccionada no fue la adecuada.

Los árboles de decisión presentaron una AUC entre 0.73 y 0.74 siendo la segunda peor opción para este problema en particular.

Finalmente los SVM y los random forest fueron los modelos que mejor se desempeñaron obteniendo una AUC entre 0.87 y 0.90.

Partiendo de la Tabla 3, escogeremos como modelo ganador a la **SVM con C=1 y kernel lineal** ya que presenta la mayor AUC (**0.900885**)

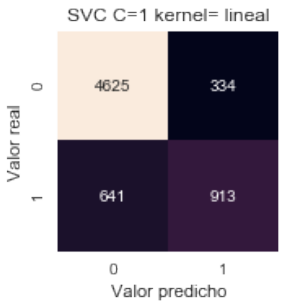


Fig 13. Matriz de confusión del modelo seleccionado

Referencias

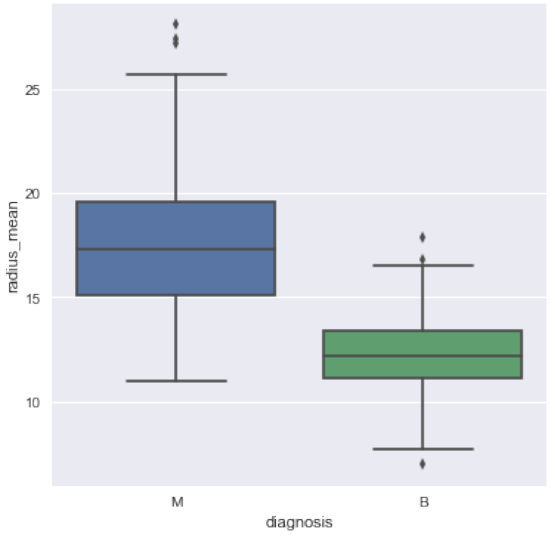
Centro Nacional de Equidad de Género y Salud Reproductiva, *Cancer de Mama*, 2015. http://cnegsr.salud.gob.mx/contenidos/Programas_de_Accion/CancerdelaMujer/cancermama/introduccion_Cama.html

IMSS, *Cancer de Mama,* 2015. http://www.imss.gob.mx/salud-en-linea/cancer-mama

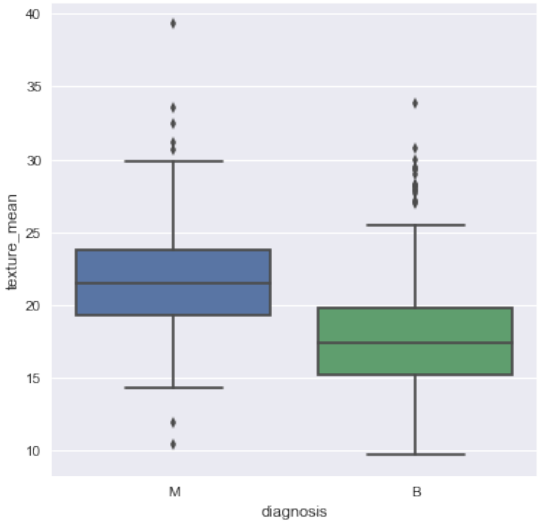
Kaggle, *Breast Cancer Wisconsin (Prognostic) Data Set*, 2017 https://www.kaggle.com/sarahvch/breast-cancer-wisconsin-prognostic-data-set/data

Kaggle, *Adult Income DataSet*, 2017 https://www.kaggle.com/wenruliu/adult-income-dataset/data

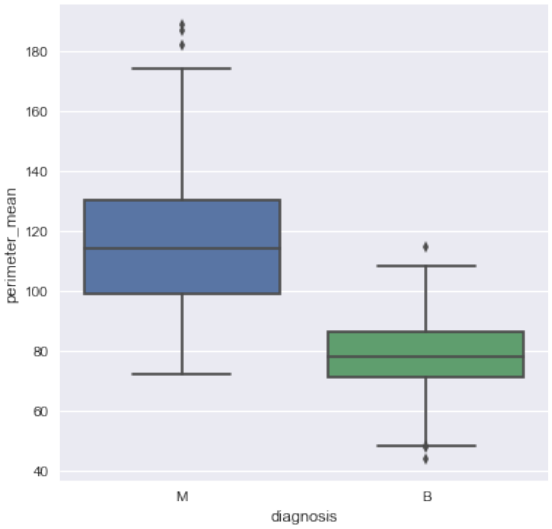
Anexo 1 Gráficas del Análisis Exploratorio de Datos para el Data Set de Cáncer de mama



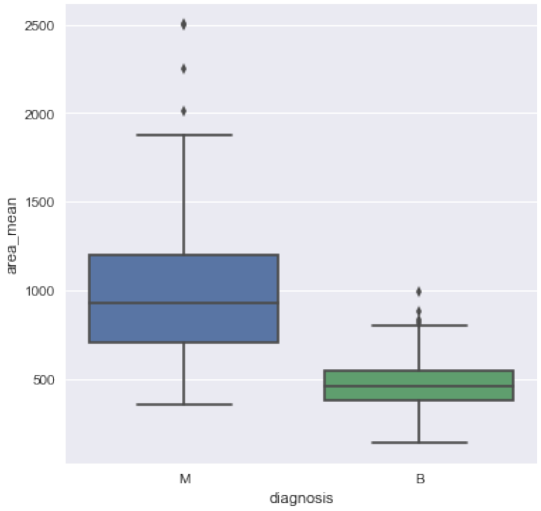
radius_mean vs diagnosis



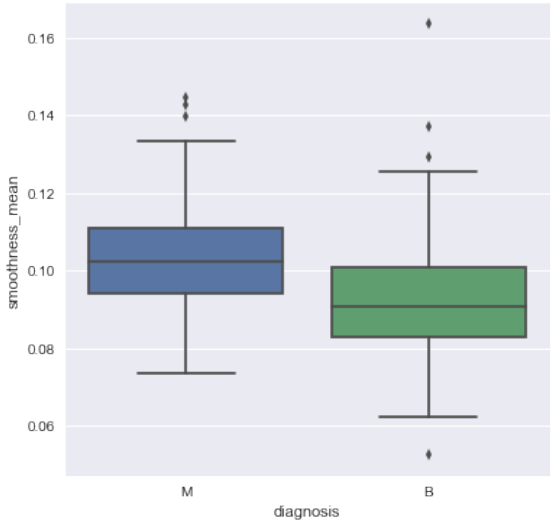
texture_mean vs diagnosis



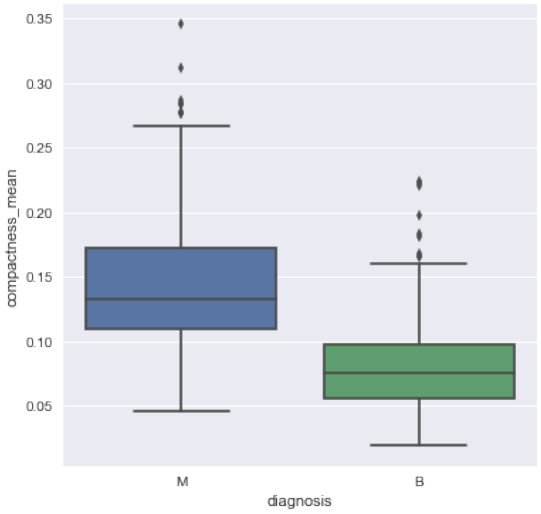
perimeter_mean vs diagnosis



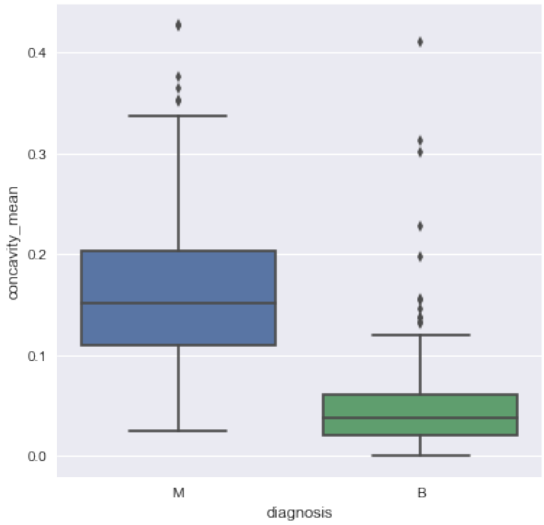
area_mean vs diagnosis



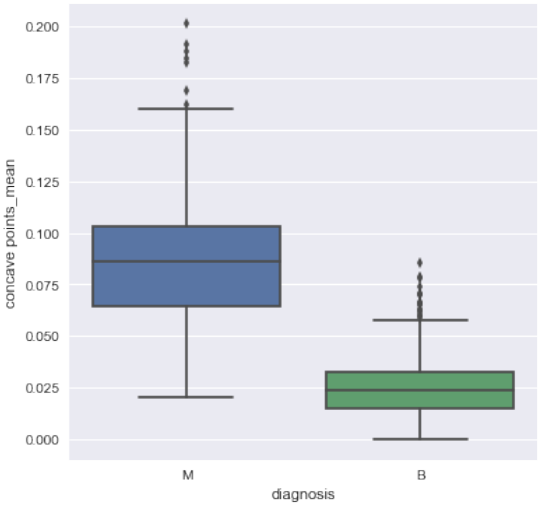
smoothness_mean vs diagnosis



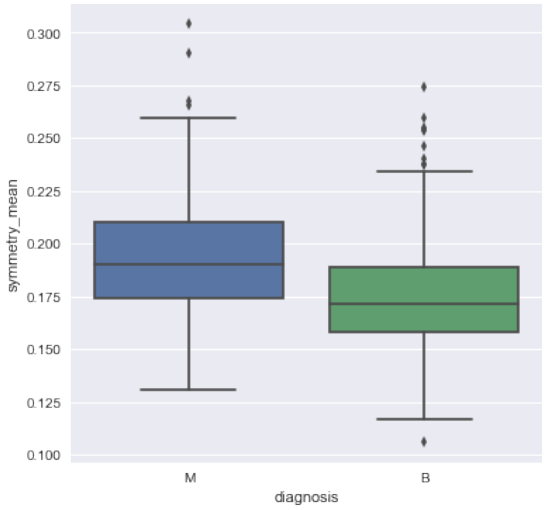
compactness_mean vs diagnosis



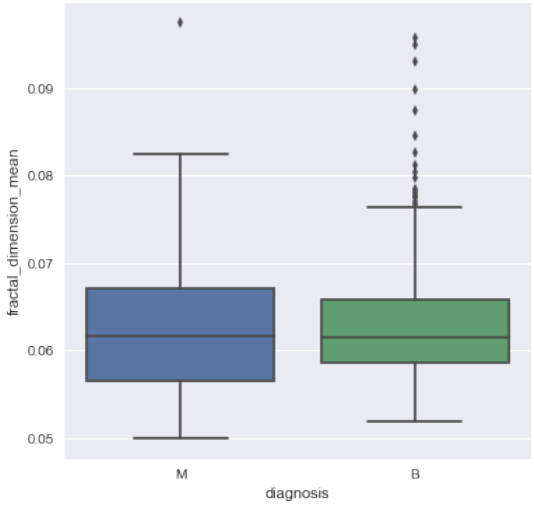
concavity_mean vs diagnosis



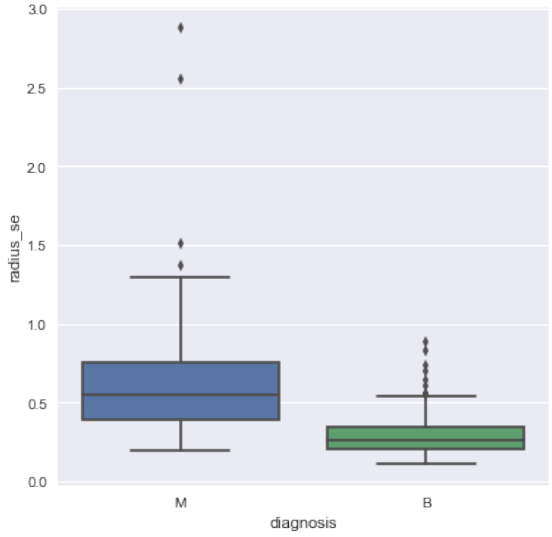
concave points_mean vs diagnosis



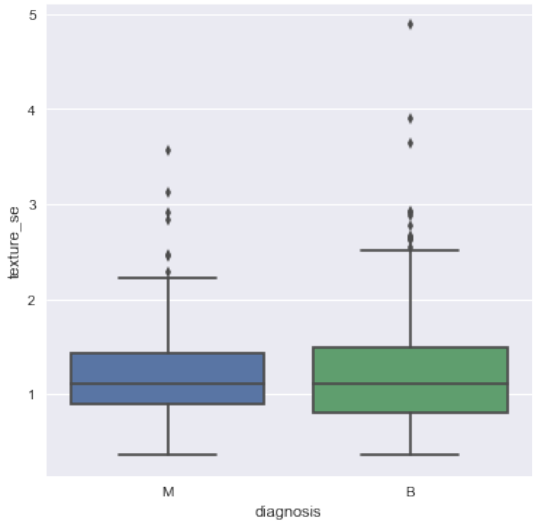
symmetry_mean vs diagnosis



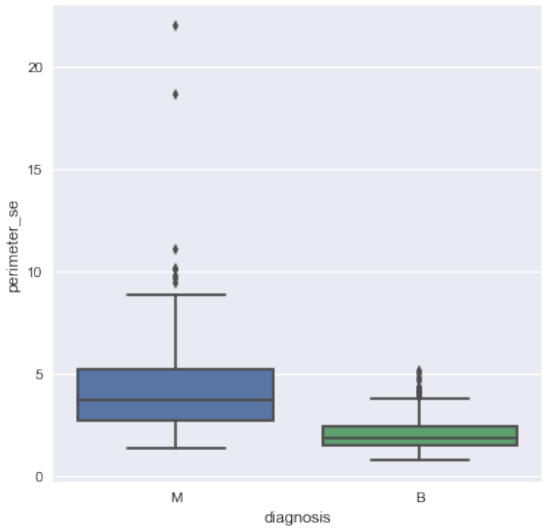
fractal_dimension_mean vs diagnosis



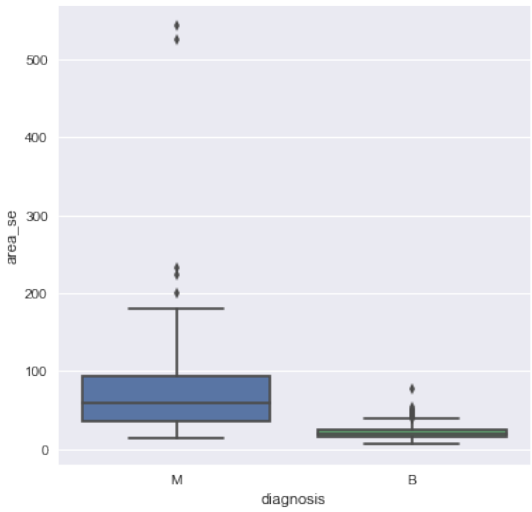
radius_se vs diagnosis



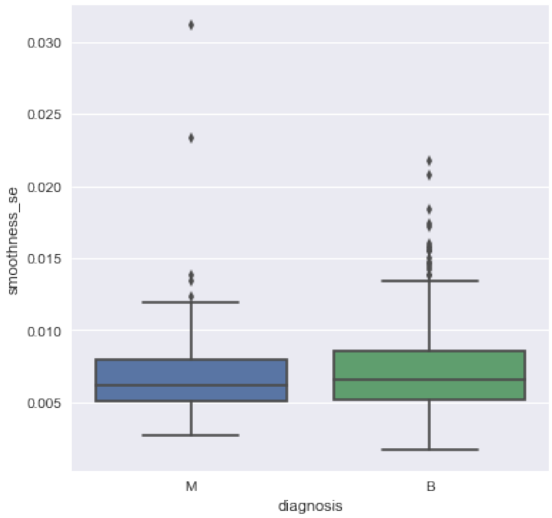
texture_se vs diagnosis



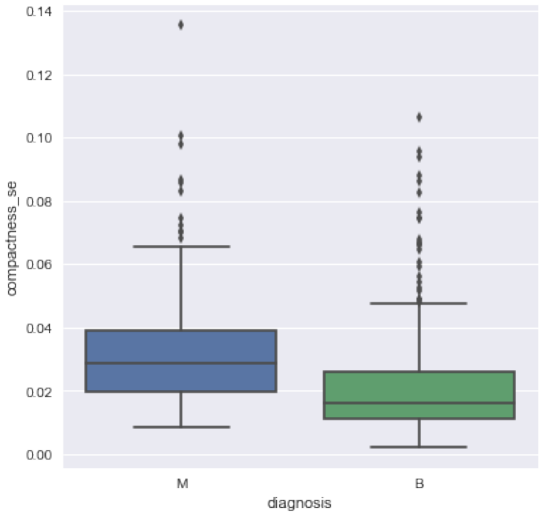
perimeter_se vs diagnosis



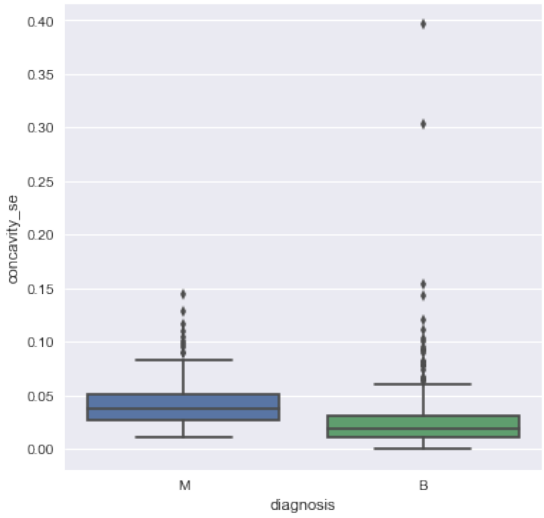
area_se vs diagnosis



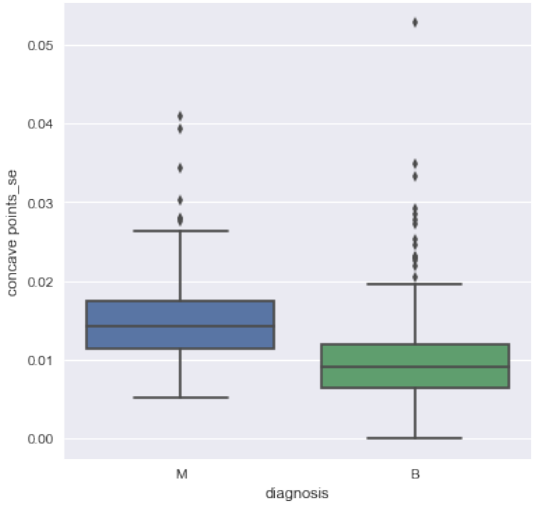
smoothness_se vs diagnosis



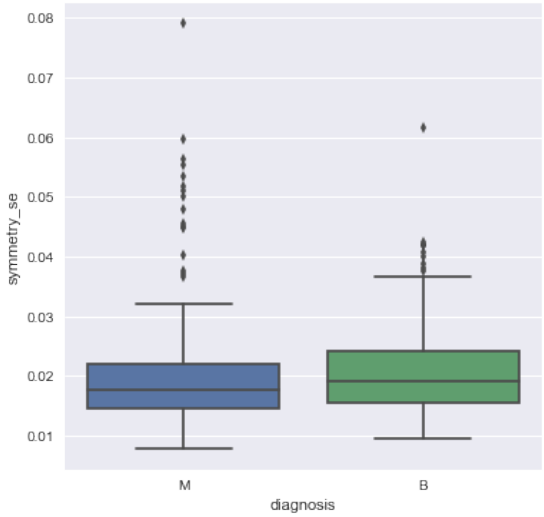
compactness_se vs diagnosis



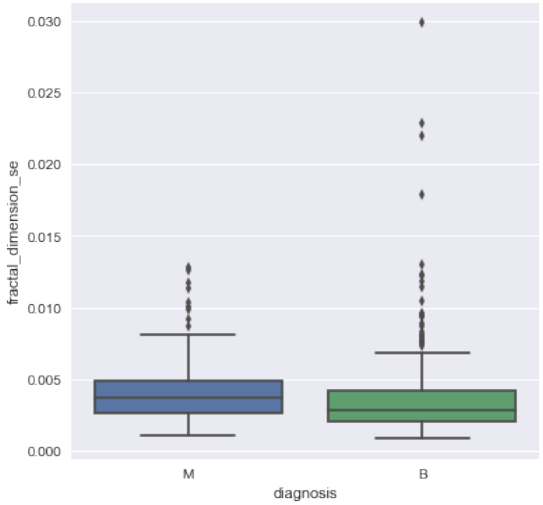
concavity_se vs diagnosis



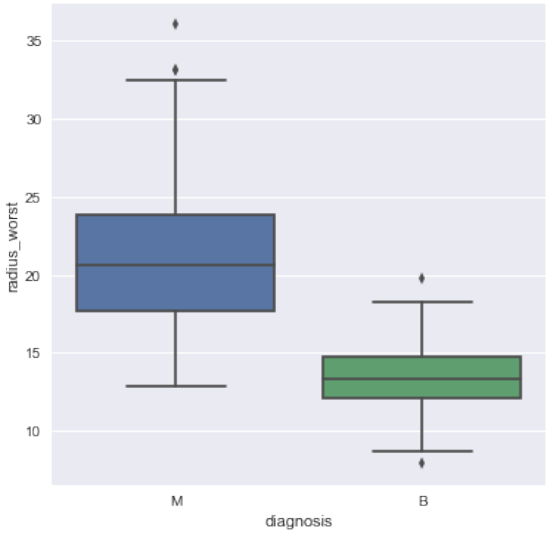
concave points_se vs diagnosis



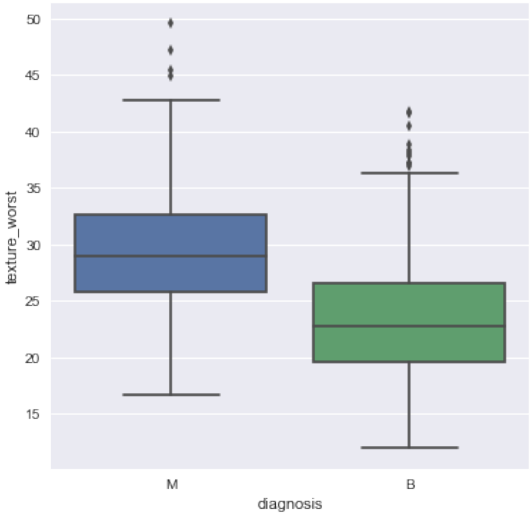
symmetry_se vs diagnosis



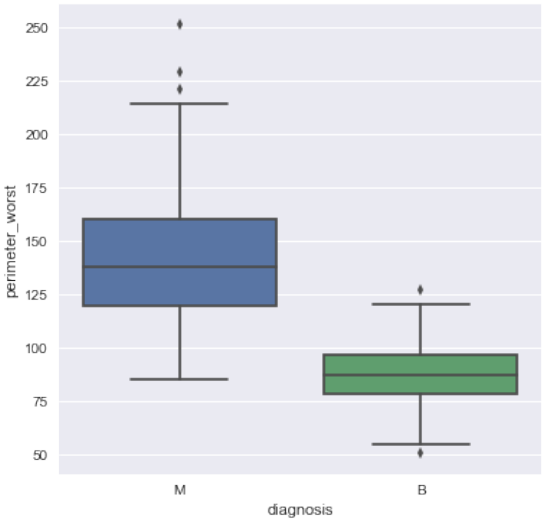
fractal_dimension_se vs diagnosis



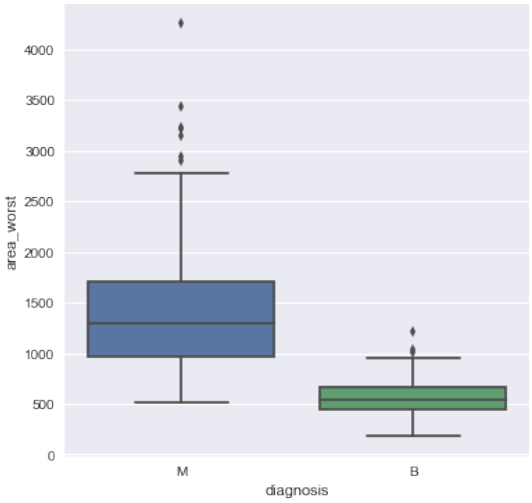
radius_worst vs diagnosis



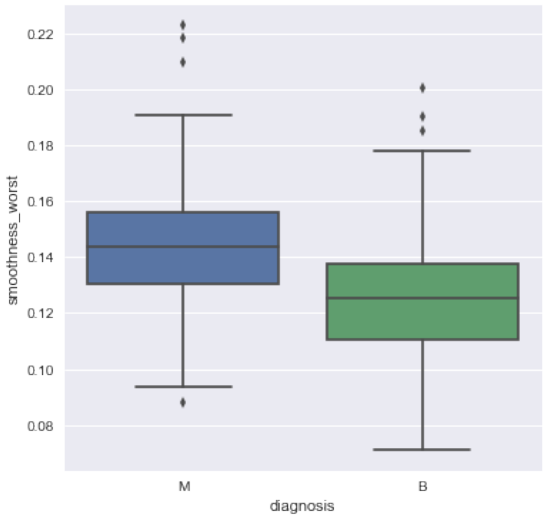
texture_worst vs diagnosis



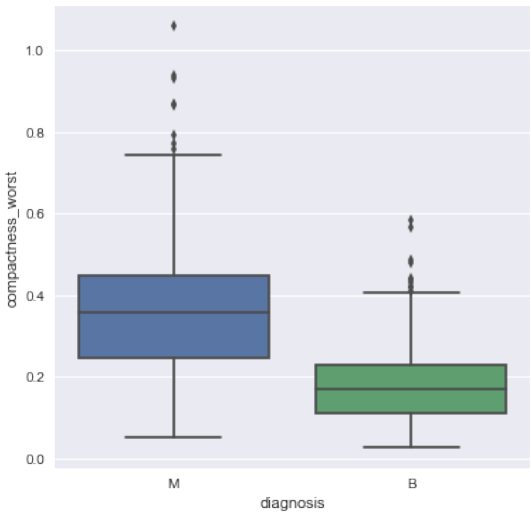
perimeter_worst vs diagnosis



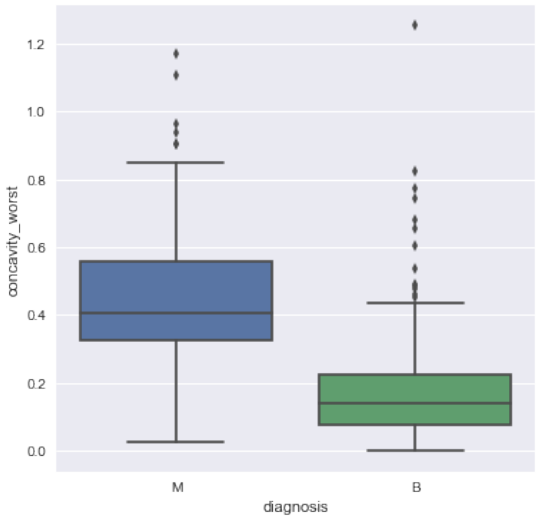
area_worst vs diagnosis



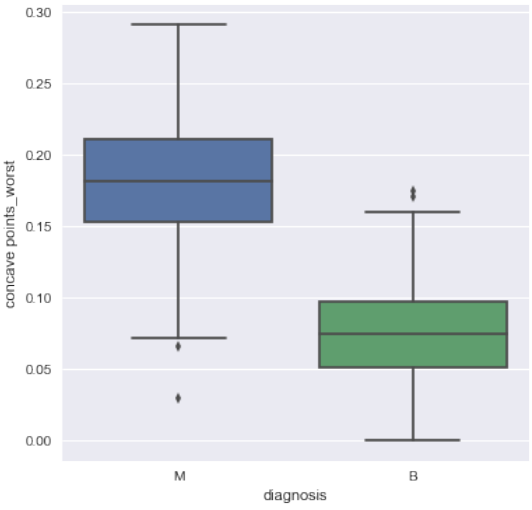
smoothness_worst vs diagnosis



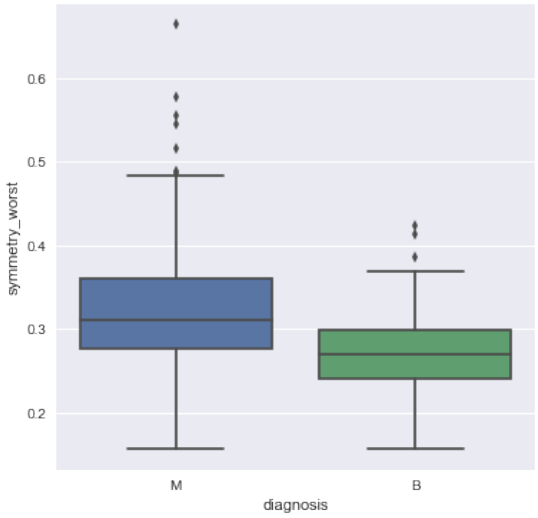
compactness_worst vs diagnosis



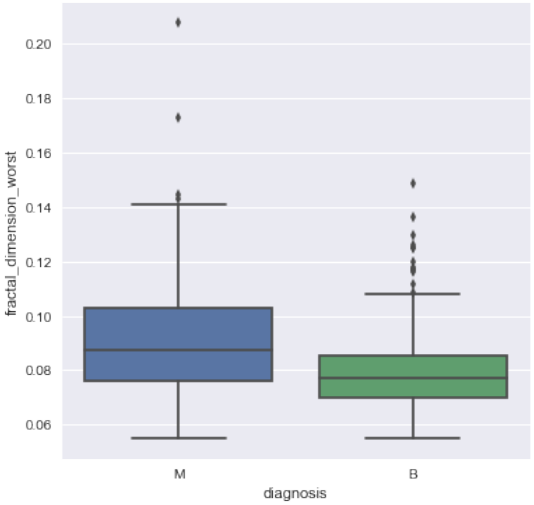
concavity_worst vs diagnosis



concave points_worst vs diagnosis

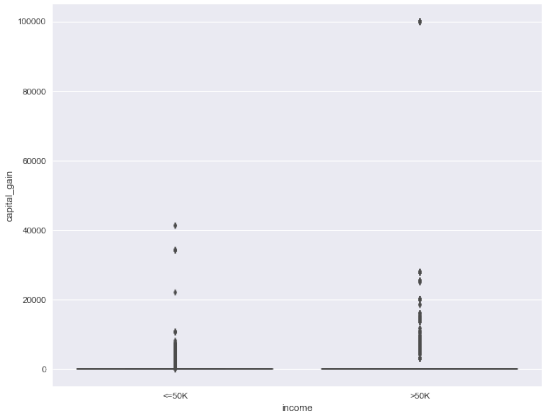


symmetry_worst vs diagnosis

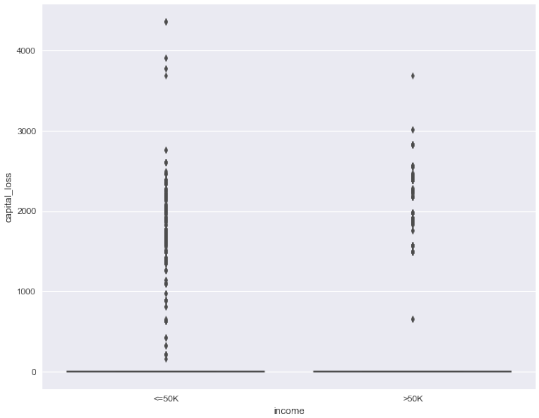


fractal_dimension_worst vs diagnosis

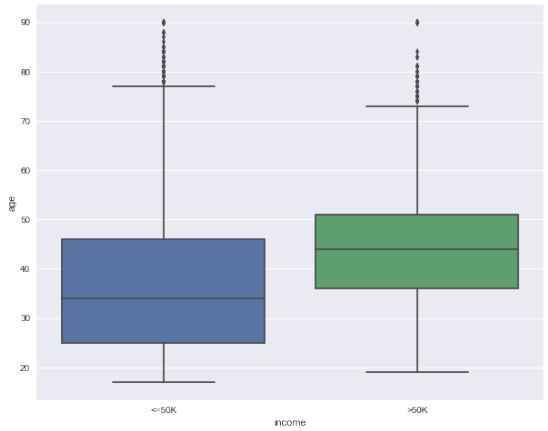
Anexo 2 Gráficas del Análisis Exploratorio de Datos para el Data Set "Adult"



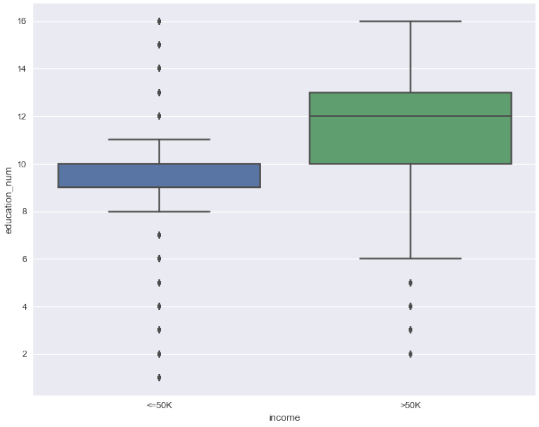
capital_gain vs income



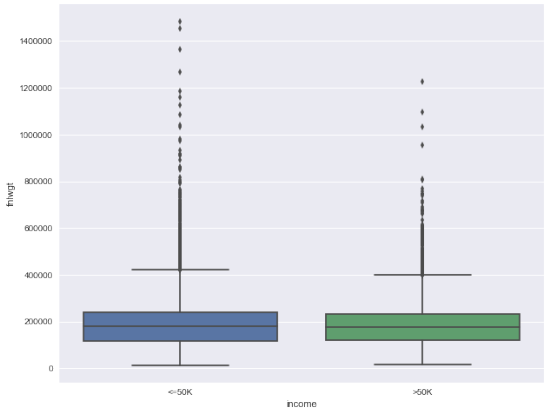
capital_loss vs income



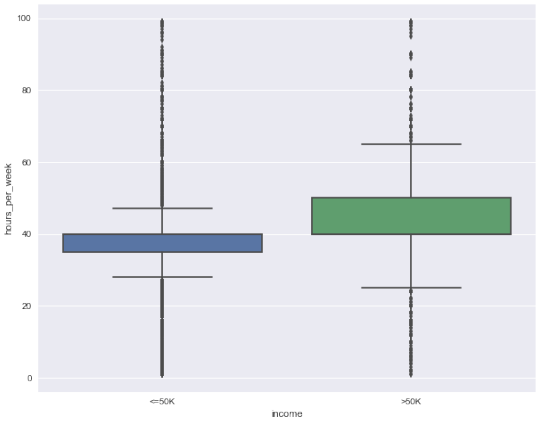
age vs income



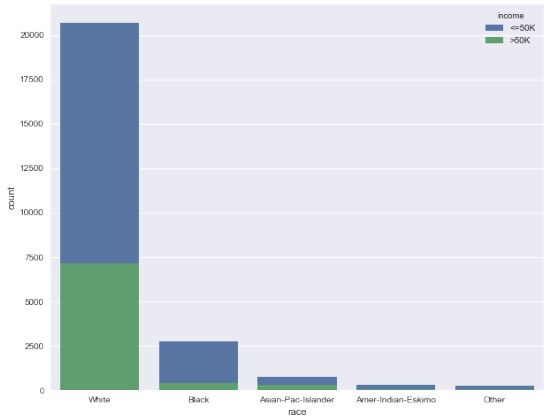
education_num vs income



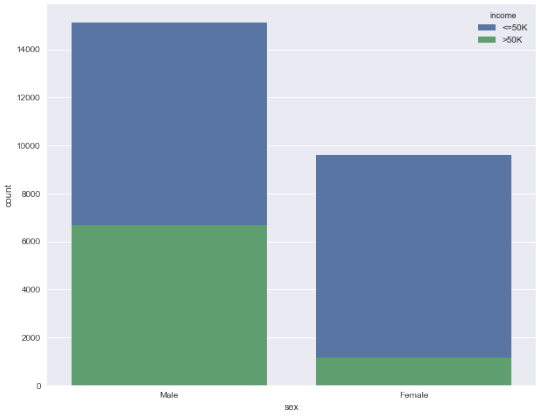
fnlwgt vs income



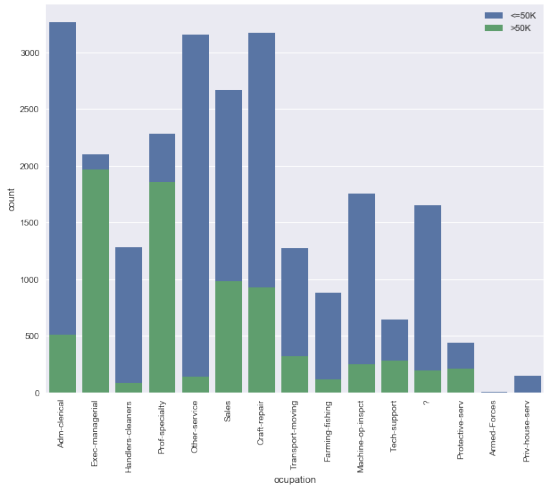
hours_per_week vs income



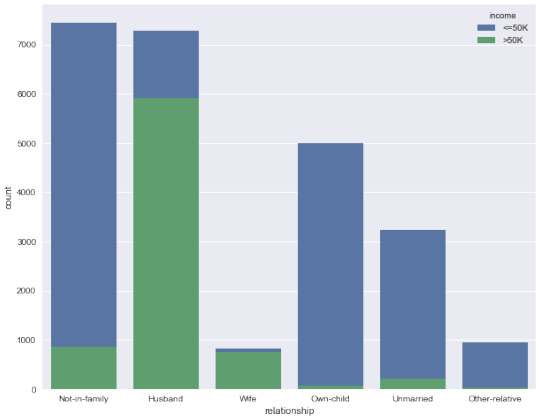
race vs income



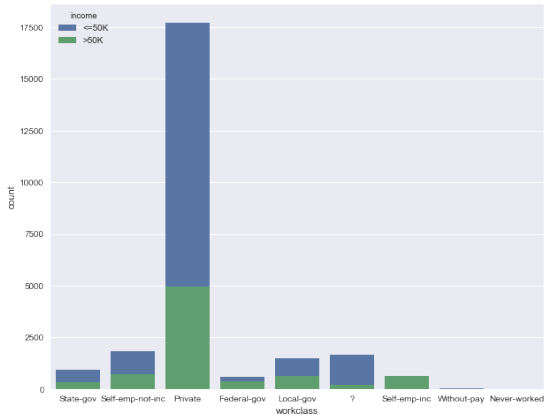
sex vs income



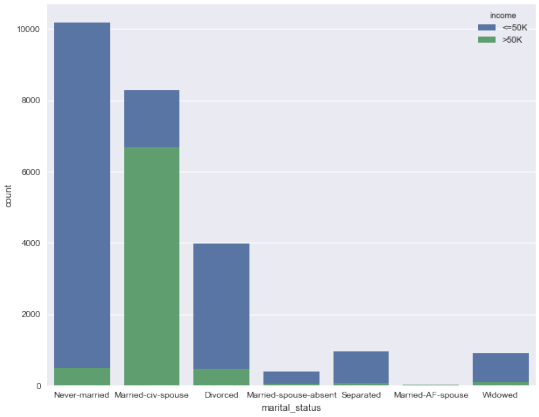
ocupation vs income



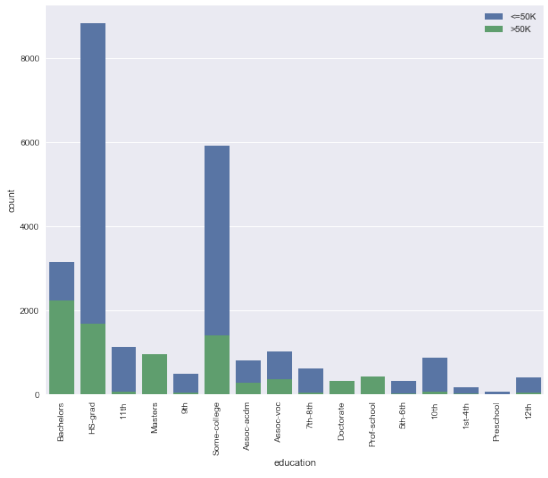
relationship vs income



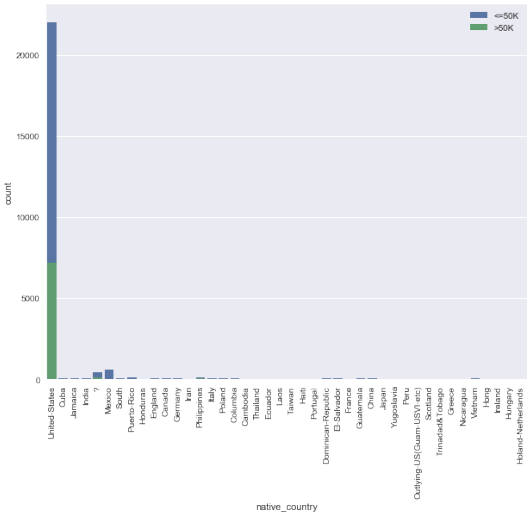
workclass vs income



marital_status vs income



education vs income



native_country vs income