

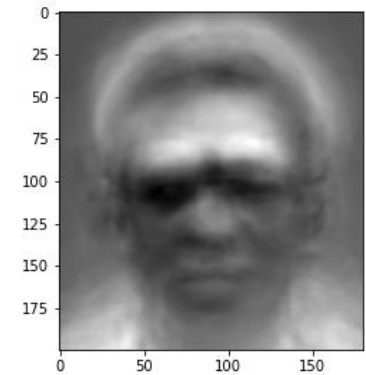
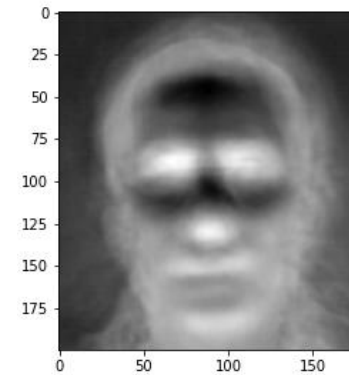
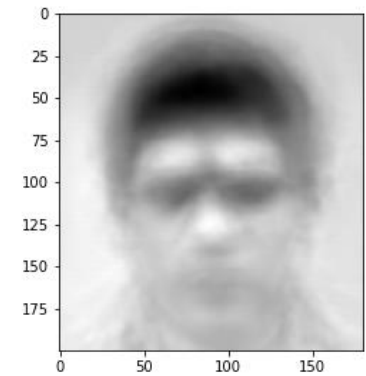
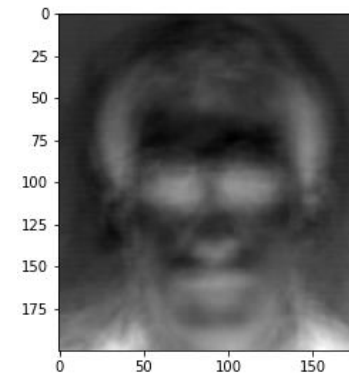
# EIGENFACES NO PARAMÉTRICAS

TRABAJO FINAL CM0896 – ESTADÍSTICA NO PARAMÉTRICA

MIGUEL ANGEL MEJÍA MUÑOZ

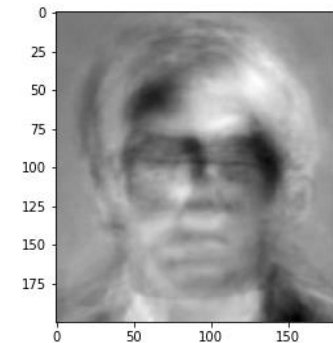
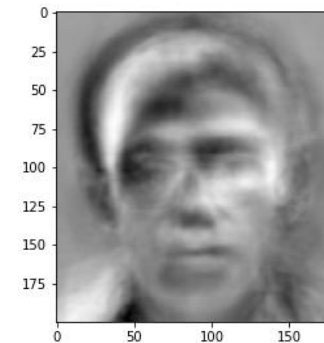
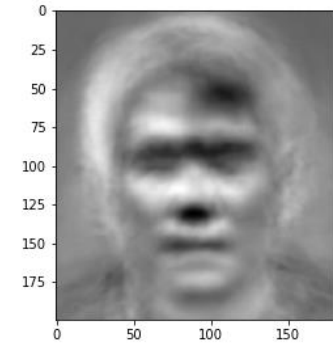
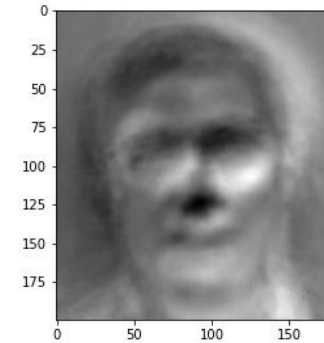
# INTRODUCCIÓN

- Realizar el desarrollo de la técnica de Eigenfaces usando métodos no paramétricos.
- Transformación del conjunto de imágenes al conjunto de eigen-vectores
  - Centrados en Media
  - Centrados en Mediana
  - Covarianza basada en Correlación Spearman y MAD



# EIGENFACES - DEFINICIÓN

- Reducir el espacio de características de un conjunto de imágenes para simplificar las computaciones de este mismo usando los eigen-vectores de la matriz de covarianza resultante de las imágenes.
- Se les llama eigen-faces, dado que estas muestran formas similares a rostros, los cuales corresponden a las características faciales que se abstraen del conjunto.



# PRELIMINARES – DATOS UTILIZADOS



Paisajes



Mujeres



Hombres

*\*Imágenes son transformadas a escala de grises para ser representadas como una matriz bi-dimensional*

*\*\*Imágenes tomadas de las bases de datos de ImageNet para paisajes, y Faces94 para los rostros*

# PRELIMINARES – PRUEBAS DE PROFUNDIDAD DATOS ORIGINALES

## OLS Regression Results

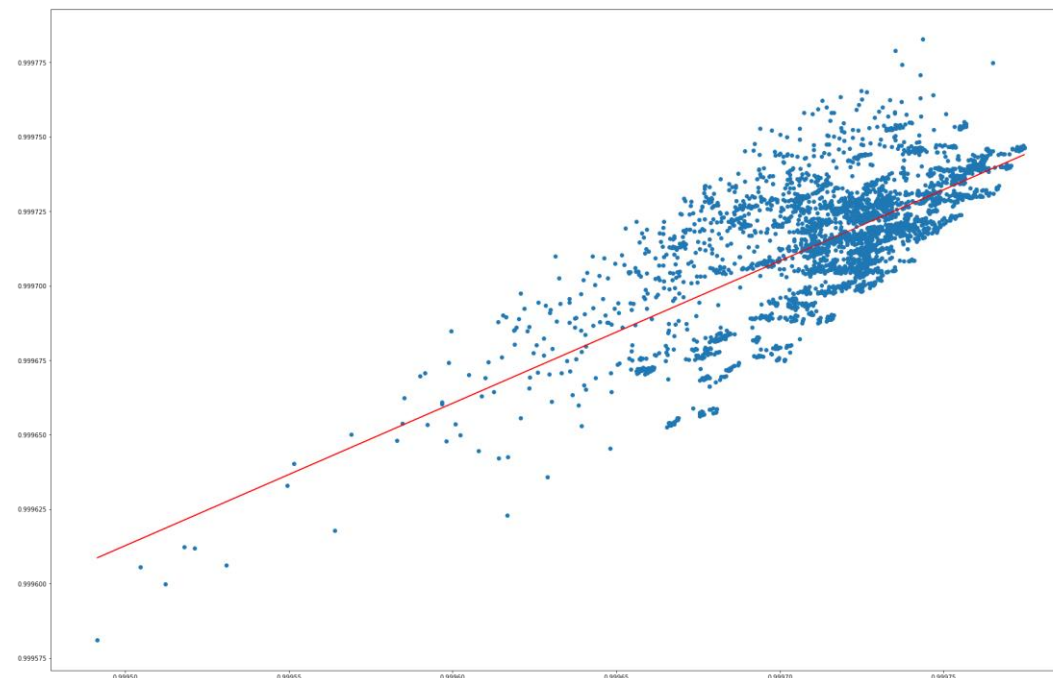
```
=====
Dep. Variable:          y      R-squared:          0.484
Model:                  OLS    Adj. R-squared:      0.484
Method:                 Least Squares    F-statistic:    3299.
Date:                   Tue, 26 May 2020    Prob (F-statistic):    0.00
Time:                   20:29:44    Log-Likelihood:    34037.
No. Observations:      3521    AIC:          -6.807e+04
Df Residuals:          3519    BIC:          -6.806e+04
Df Model:              1
Covariance Type:       nonrobust
=====
```

	coef	std err	t	P> t	[0.025	0.975]
const	0.5216	0.008	62.674	0.000	0.505	0.538
x1	0.4782	0.008	57.439	0.000	0.462	0.495

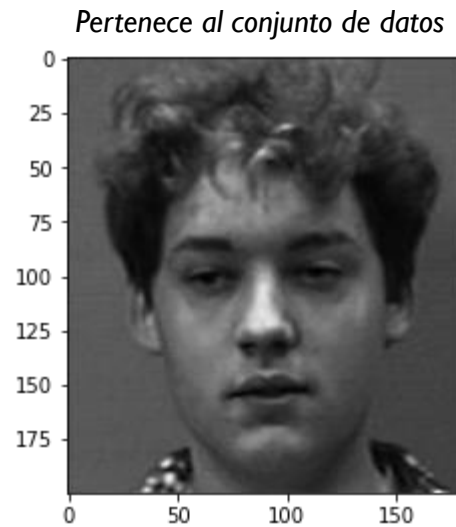
```
=====
Omnibus:                20.891    Durbin-Watson:          0.347
Prob(Omnibus):           0.000    Jarque-Bera (JB):       21.086
Skew:                    0.187    Prob(JB):               2.64e-05
Kurtosis:                3.067    Cond. No.               6.44e+04
=====
```

## Warnings:

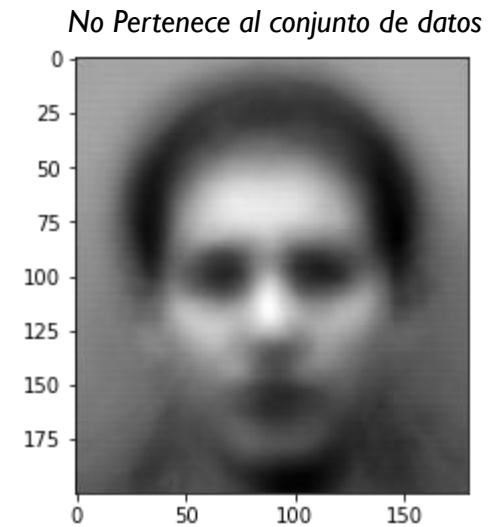
- [1] Standard Errors assume that the covariance matrix of the errors is correctly specified.
- [2] The condition number is large, 6.44e+04. This might indicate that there are strong multicollinearity or other numerical problems.



# PRELIMINARES – IMÁGENES MEDIA Y MEDIANA

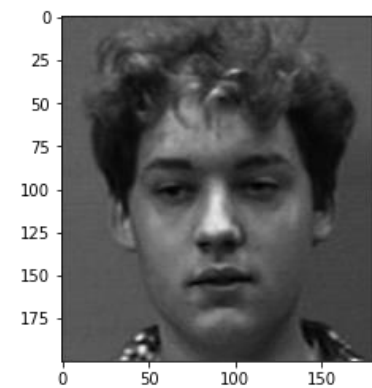


Mediana

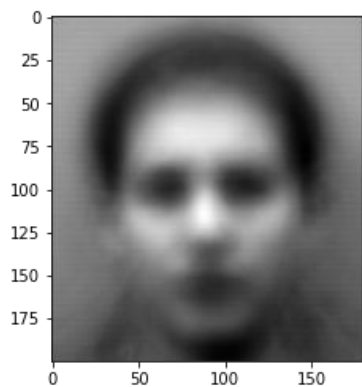
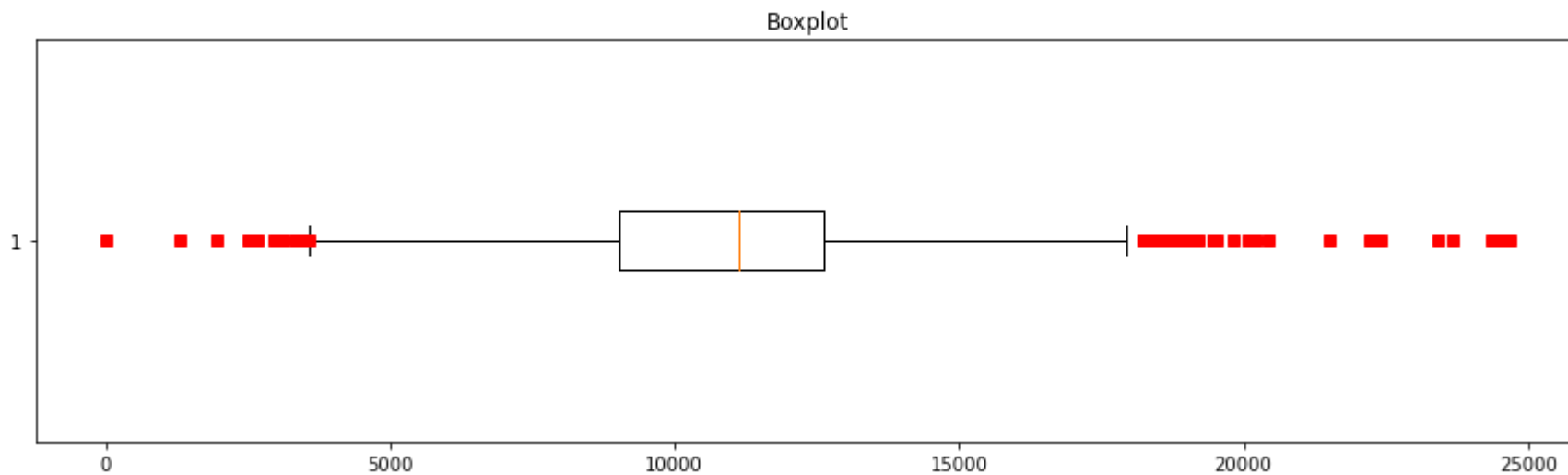


Media

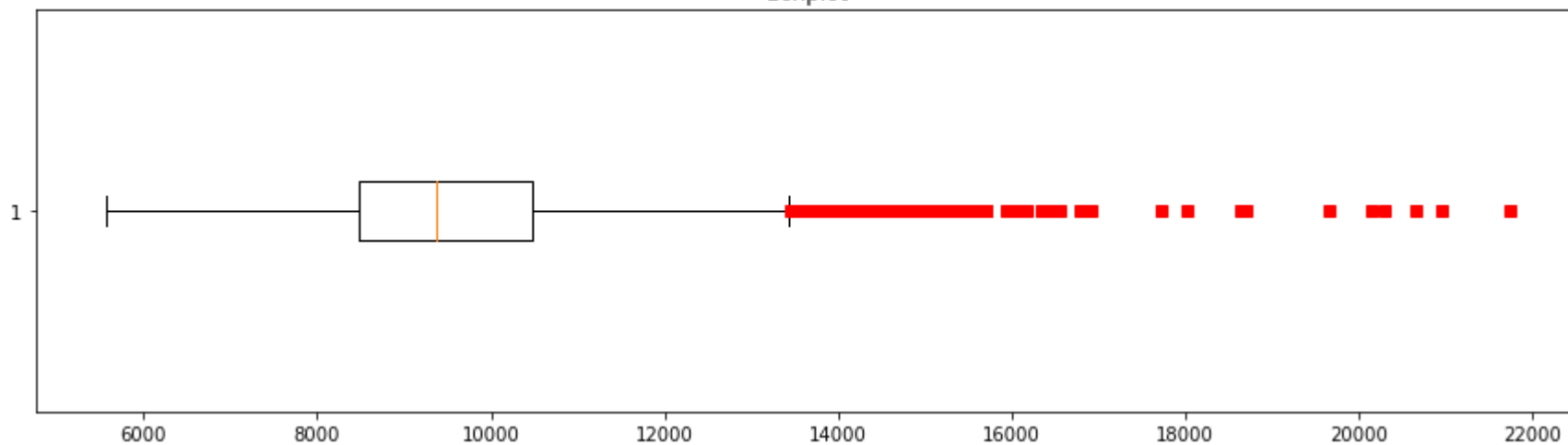
# PRELIMINARES – DISTANCIAS DATOS



Mediana



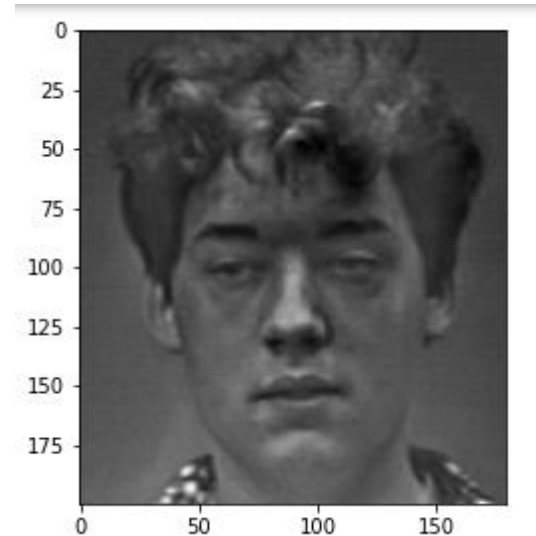
Media



# CÁLCULO MATRICES DE COVARIANZA

Objetivo de ésta investigación

- Datos centrados en la media
- ***Datos centrados en la mediana***
- ***Covarianza robusta basada en Spearman y MAD***





# CÁLCULO DE EIGENFACES



# PRUEBAS DE PROFUNDIDAD DATOS TRANSFORMADOS

## Covarianza tradicional

### OLS Regression Results

```
=====
Dep. Variable:          y      R-squared:          0.396
Model:                  OLS    Adj. R-squared:      0.396
Method:                 Least Squares    F-statistic:    2306.
Date:                   Tue, 26 May 2020    Prob (F-statistic):    0.00
Time:                   03:28:45    Log-Likelihood:    32506.
No. Observations:       3521    AIC:            -6.501e+04
Df Residuals:           3519    BIC:            -6.500e+04
Df Model:                1
Covariance Type:        nonrobust
=====
```

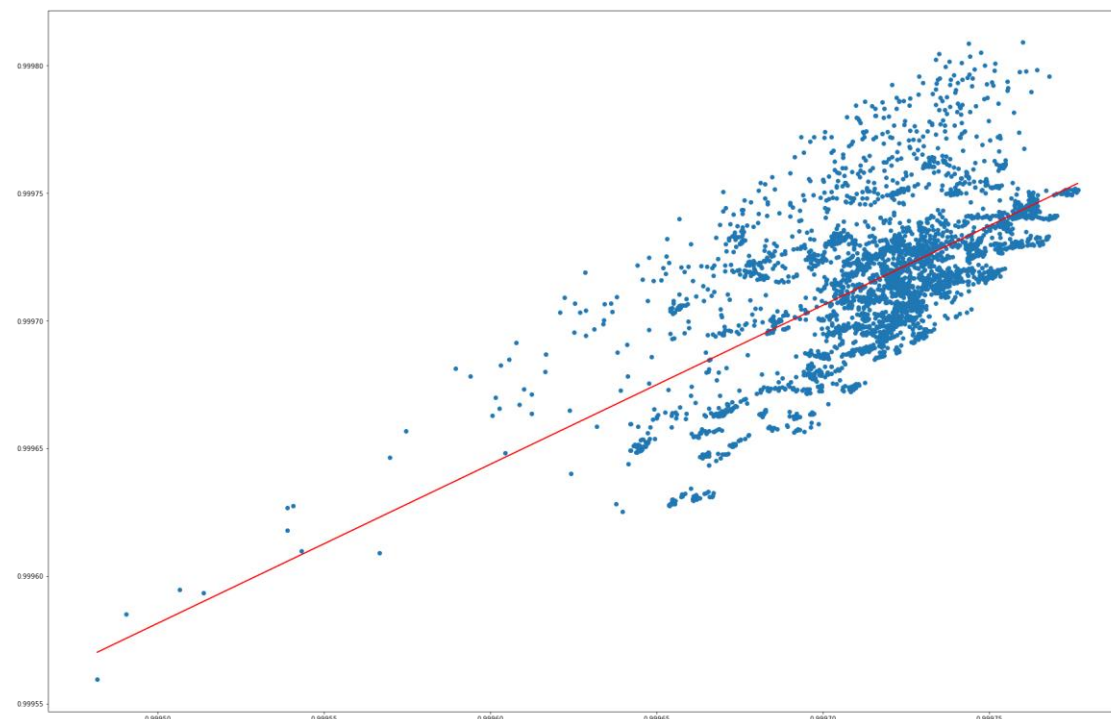
	coef	std err	t	P> t	[0.025	0.975]
const	0.3767	0.013	29.038	0.000	0.351	0.402
x1	0.6232	0.013	48.025	0.000	0.598	0.649

```
=====
```

```
Omnibus:                235.816    Durbin-Watson:          0.260
Prob(Omnibus):           0.000    Jarque-Bera (JB):       283.523
Skew:                    0.679    Prob(JB):               2.71e-62
Kurtosis:                3.293    Cond. No.               6.50e+04
=====
```

#### Warnings:

- [1] Standard Errors assume that the covariance matrix of the errors is correctly specified.
- [2] The condition number is large, 6.5e+04. This might indicate that there are strong multicollinearity or other numerical problems.



# PRUEBAS DE PROFUNDIDAD DATOS TRANSFORMADOS

## Covarianza centrada en mediana

### OLS Regression Results

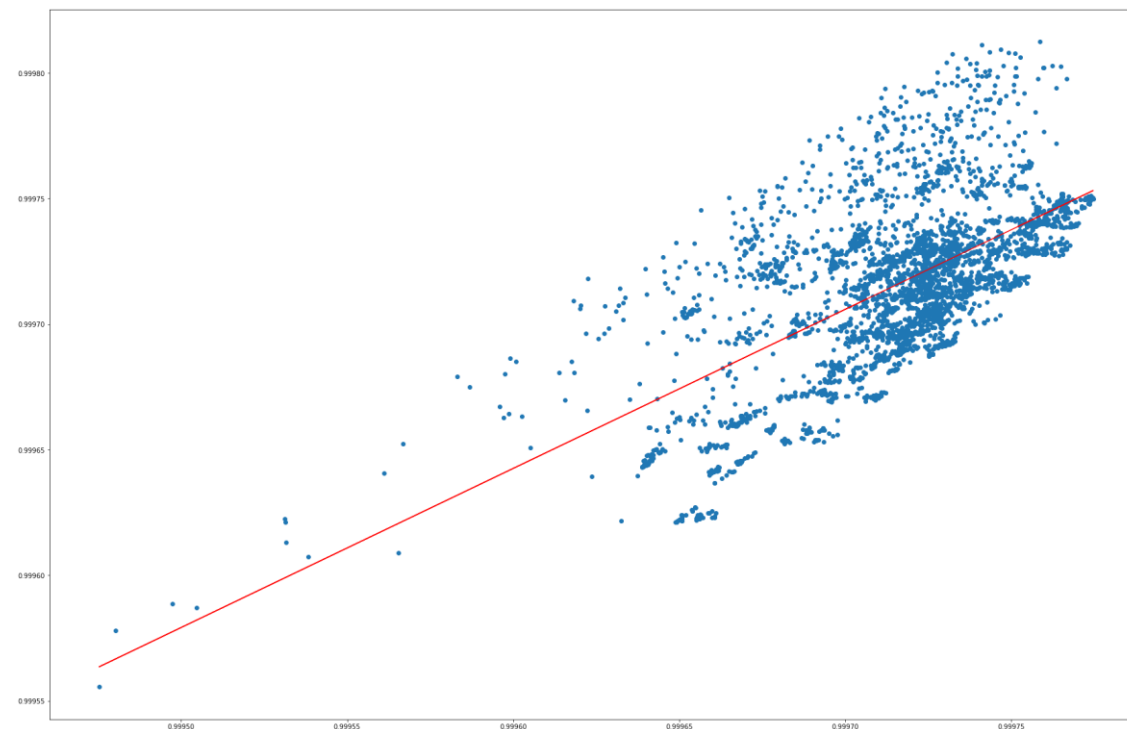
```
=====
Dep. Variable:          y      R-squared:                0.383
Model:                  OLS    Adj. R-squared:           0.383
Method:                 Least Squares    F-statistic:       2183.
Date:                   Tue, 26 May 2020    Prob (F-statistic):   0.00
Time:                   03:28:46    Log-Likelihood:      32225.
No. Observations:       3521    AIC:                 -6.445e+04
Df Residuals:           3519    BIC:                 -6.443e+04
Df Model:                1
Covariance Type:        nonrobust
=====
```

	coef	std err	t	P> t	[0.025	0.975]
const	0.3665	0.014	27.051	0.000	0.340	0.393
x1	0.6334	0.014	46.727	0.000	0.607	0.660

```
=====
Omnibus:                251.186    Durbin-Watson:       0.251
Prob(Omnibus):           0.000    Jarque-Bera (JB):     305.621
Skew:                    0.708    Prob(JB):             4.32e-67
Kurtosis:                3.277    Cond. No.             6.27e+04
=====
```

#### Warnings:

- [1] Standard Errors assume that the covariance matrix of the errors is correctly specified.
- [2] The condition number is large, 6.27e+04. This might indicate that there are strong multicollinearity or other numerical problems.



# PRUEBAS DE PROFUNDIDAD DATOS TRANSFORMADOS

## Covarianza robusta basada en Spearman y MAD

### OLS Regression Results

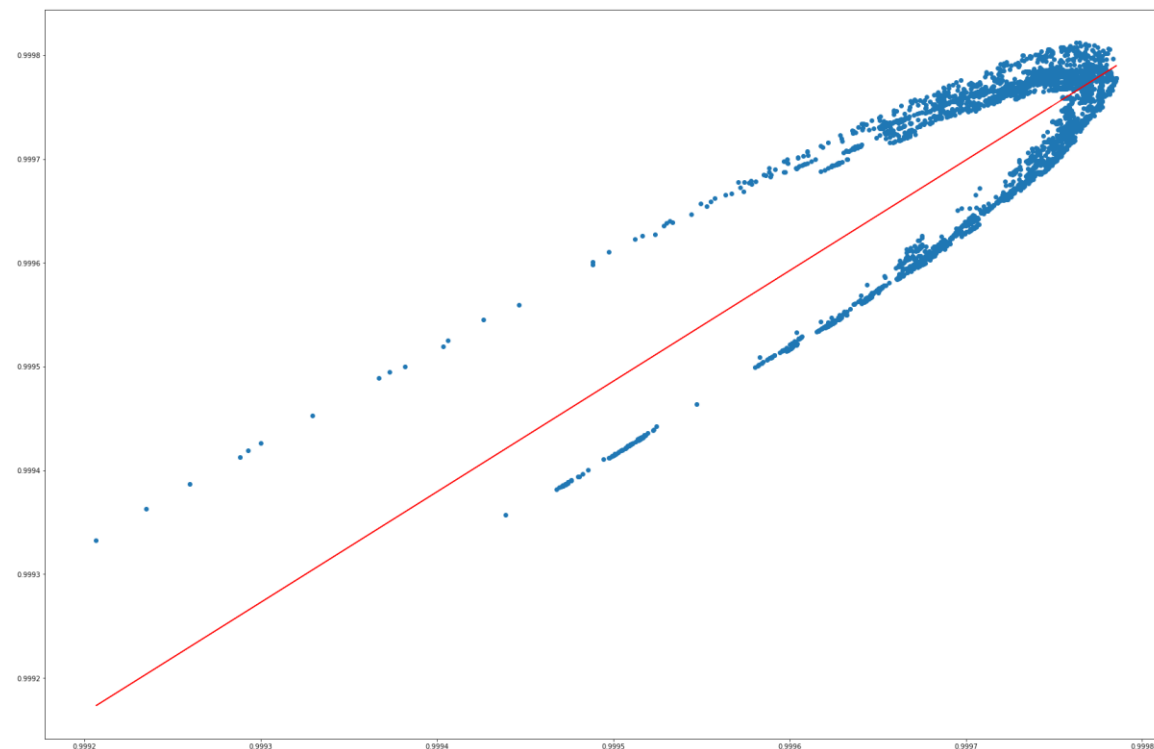
```
=====
Dep. Variable:          y      R-squared:          0.638
Model:                  OLS    Adj. R-squared:      0.638
Method:                 Least Squares    F-statistic:    6197.
Date:                  Tue, 26 May 2020    Prob (F-statistic):    0.00
Time:                  03:28:46    Log-Likelihood:    29801.
No. Observations:      3521    AIC:              -5.960e+04
Df Residuals:          3519    BIC:              -5.958e+04
Df Model:              1
Covariance Type:       nonrobust
=====
```

	coef	std err	t	P> t	[0.025	0.975]
const	-0.0657	0.014	-4.851	0.000	-0.092	-0.039
x1	1.0657	0.014	78.721	0.000	1.039	1.092

```
=====
Omnibus:                684.068    Durbin-Watson:          0.325
Prob(Omnibus):           0.000    Jarque-Bera (JB):        150.146
Skew:                    0.146    Prob(JB):                2.49e-33
Kurtosis:                2.032    Cond. No.                 3.14e+04
=====
```

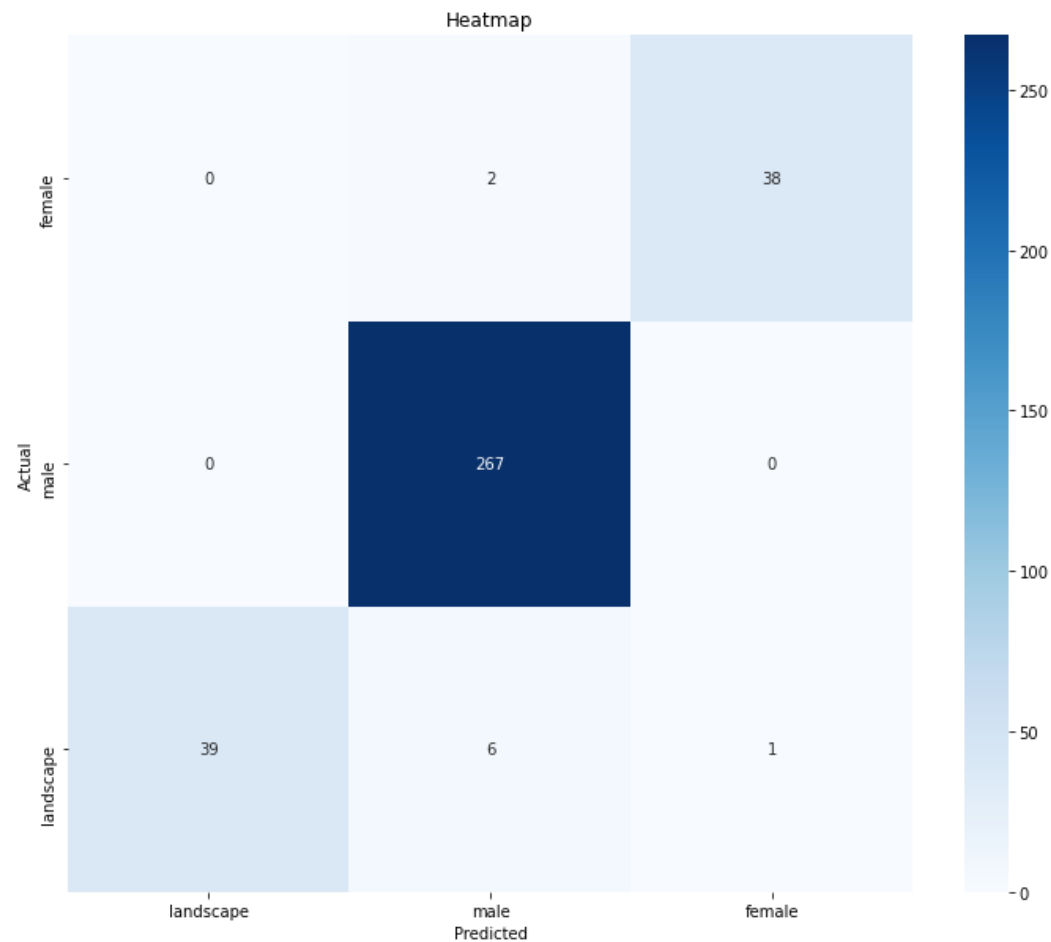
### Warnings:

- [1] Standard Errors assume that the covariance matrix of the errors is correctly specified.
- [2] The condition number is large, 3.14e+04. This might indicate that there are strong multicollinearity or other numerical problems.



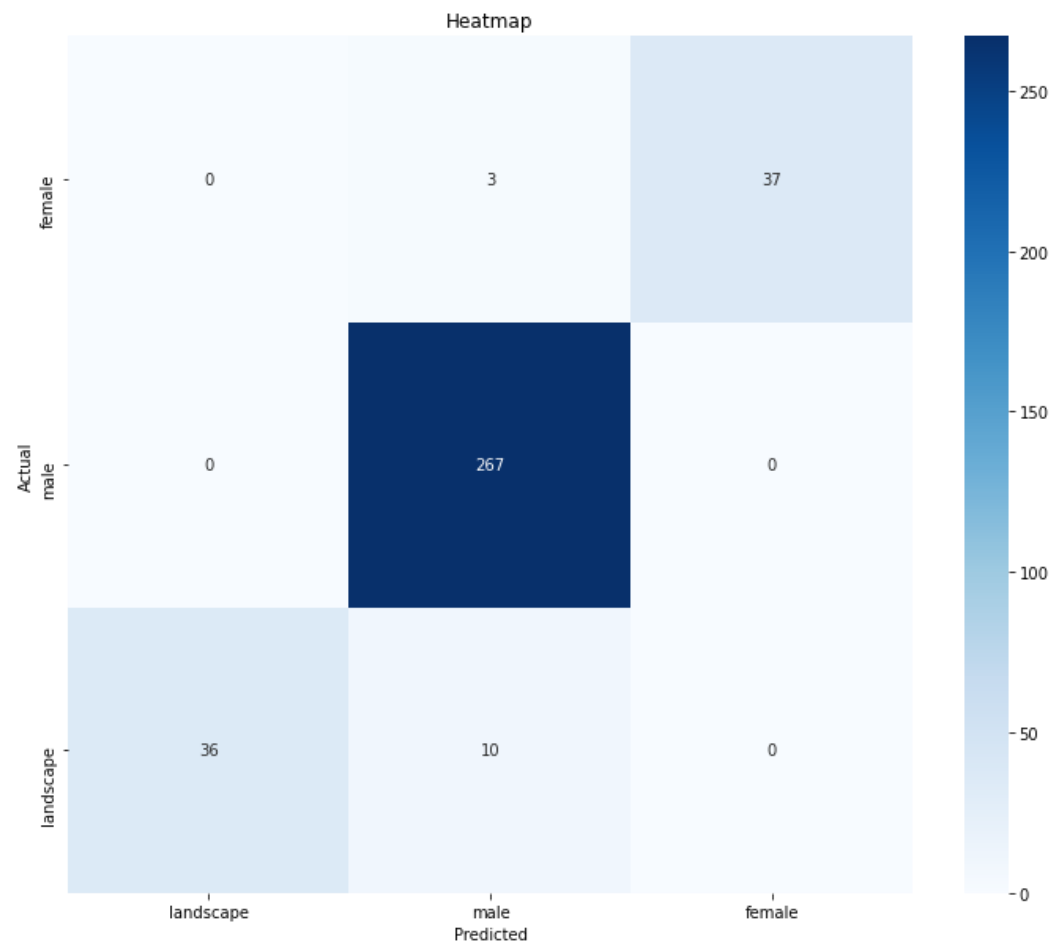
# CLASIFICACIÓN – MAPAS DE CALOR, BOSQUE ALEATORIO

## *Datos Originales*



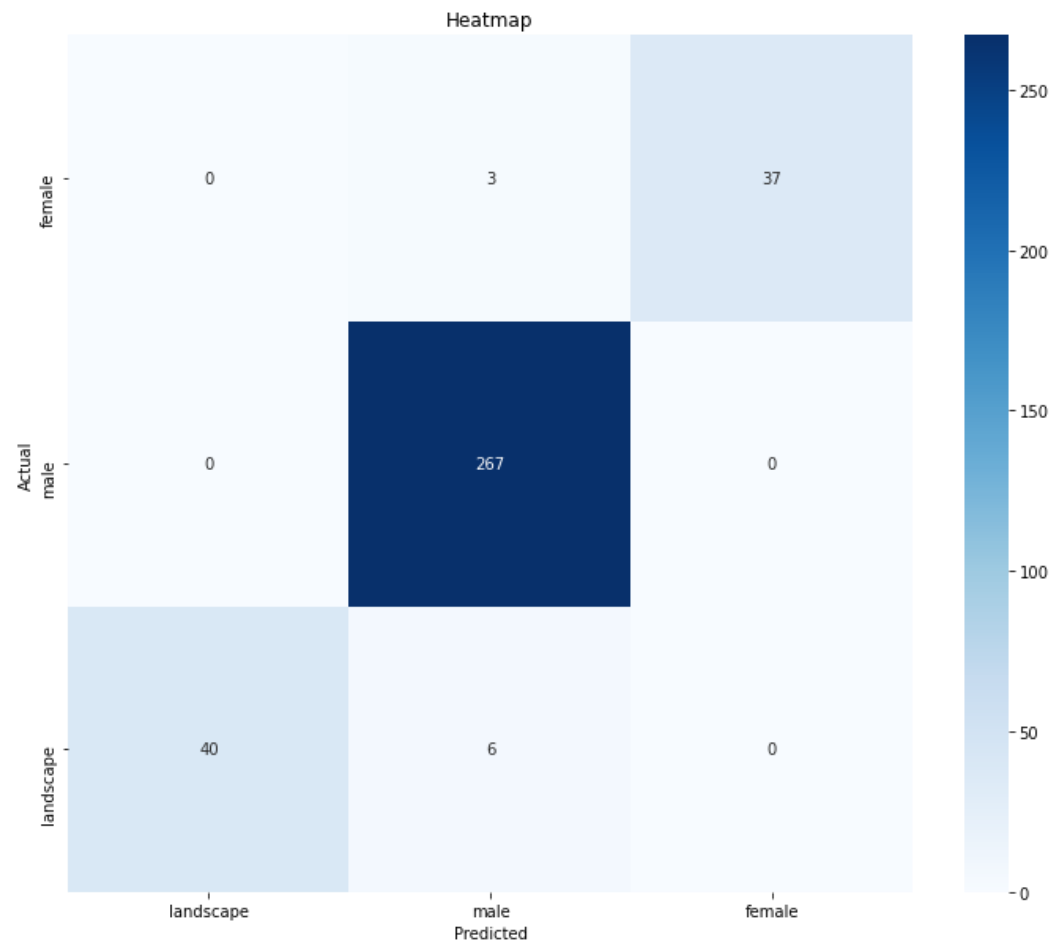
# CLASIFICACIÓN – MAPAS DE CALOR, BOSQUE ALEATORIO

*Eigen-faces centradas en media*



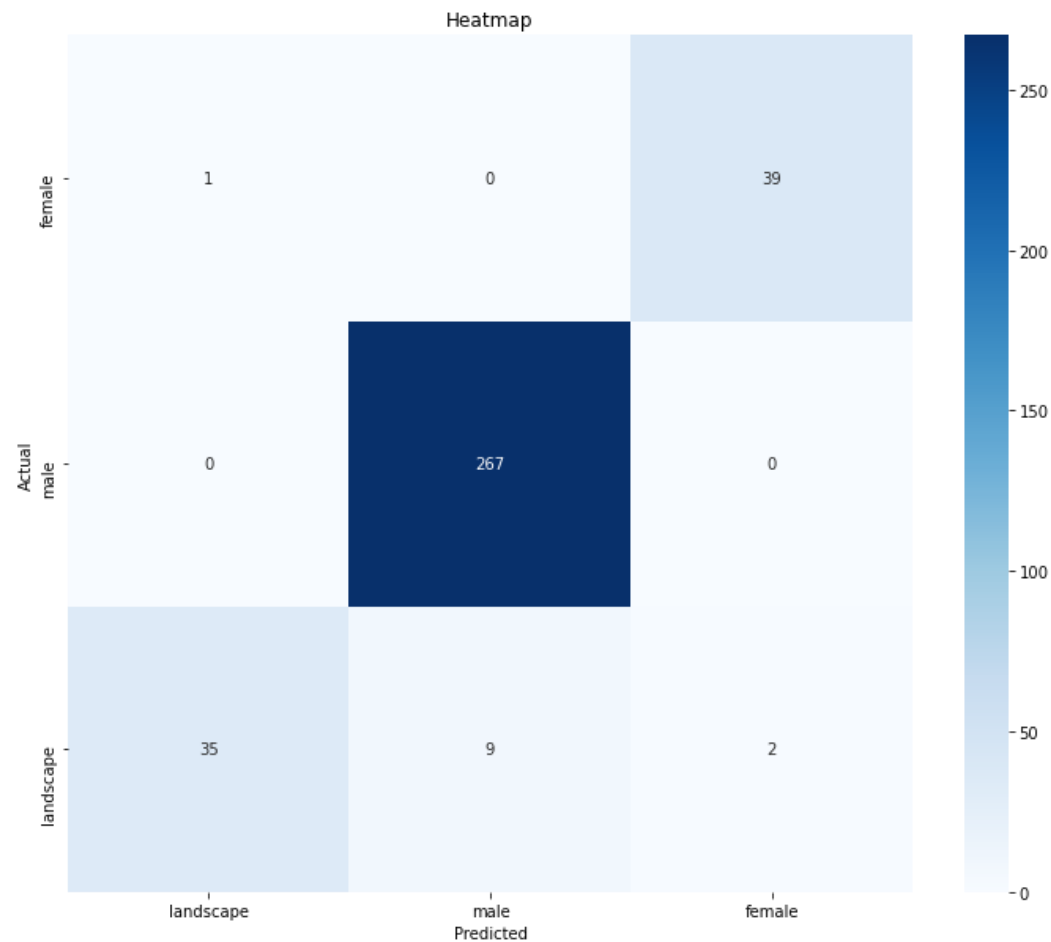
# CLASIFICACIÓN – MAPAS DE CALOR, BOSQUE ALEATORIO

*Eigen-faces centradas en mediana*



# CLASIFICACIÓN – MAPAS DE CALOR, BOSQUE ALEATORIO

*Eigen-faces covarianza robusta*





# CLASIFICACIÓN – RESULTADOS, EXACTITUD

Modelo	Datos Originales		Eigen-faces centradas en media		Eigen-faces centradas en mediana		Eigen-faces covarianza robusta	
	Validación	Prueba	Validación	Prueba	Validación	Prueba	Validación	Prueba
Discriminante Lineal	<b>96.4489%</b>	94.3343%	95.1705%	<b>95.7507%</b>	92.7557%	93.7677%	94.8864%	<b>95.7507%</b>
Regresión Logística	<b>98.2955%</b>	<b>98.0170%</b>	96.4489%	96.6006%	95.8807%	95.1841%	96.8750%	96.6006%
Bosque Aleatorio	97.3011%	<b>97.4504%</b>	98.4375%	96.3173%	<b>98.5795%</b>	<b>97.4504%</b>	97.4432%	96.6006%

*\*Se señalan los valores más altos de la métrica para dicho modelo en validación y pruebas*

# CLASIFICACIÓN – RESULTADOS, PRECISIÓN

	Modelo	Datos Originales		Eigen-faces centradas en media		Eigen-faces centradas en mediana		Eigen-faces covarianza robusta	
		Validación	Prueba	Validación	Prueba	Validación	Prueba	Validación	Prueba
Paisajes	Discriminante Lineal	<b>100%</b>	<b>100%</b>	96.2025%	97.2973%	95.0617%	97.2222%	97.4359%	97.4359%
	Regresión Logística	<b>98.7805%</b>	<b>100%</b>	89.8876%	89.1304%	89.1304%	91.1111%	90.8046%	88.6364%
	Bosque Aleatorio	98.7013%	<b>100%</b>	<b>100%</b>	100%	100%	100%	98.6842%	97.2222%
Hombres	Discriminante Lineal	<b>96.9035%</b>	94.6809%	95.8106%	<b>96.0289%</b>	94.4954%	95.5882%	95.1264%	<b>96.0289%</b>
	Regresión Logística	<b>98.5158%</b>	98.5240%	98.4848%	<b>99.6169%</b>	98.0989%	98.4615%	97.9439%	98.1273%
	Bosque Aleatorio	96.8978%	97.0909%	97.9742%	95.3571%	<b>98.1550%</b>	96.7391%	97.2527%	96.7391%
Mujeres	Discriminante Lineal	<b>90.8046%</b>	88.3721%	89.4737%	<b>92.3077%</b>	78.2051%	80%	90.2778%	91.8919%
	Regresión Logística	<b>96.3855%</b>	93.0233%	90.8046%	86.9565%	89.5349%	81.2500%	96.3415%	<b>95.2381%</b>
	Bosque Aleatorio	98.7342%	97.4359%	<b>100%</b>	<b>100%</b>	<b>100%</b>	<b>100%</b>	97.5610%	95.1220%
Promedio	Discriminante Lineal	<b>95.9027%</b>	94.3510%	93.8289%	<b>95.2113%</b>	89.2541%	90.9368%	94.2800%	95.1189%
	Regresión Logística	<b>97.8939%</b>	<b>97.1824%</b>	93.0590%	91.9013%	92.2547%	90.2742%	95.0300%	94.0006%
	Bosque Aleatorio	98.1111%	98.1756%	99.3247%	98.4524%	<b>99.3850%</b>	<b>98.9130%</b>	97.8326%	96.3611%

*\*Se señalan los valores más altos de la métrica para dicho modelo en validación y pruebas*

# CLASIFICACIÓN – RESULTADOS, EXHAUSTIVIDAD (RECALL)

	Modelo	Datos Originales		Eigen-faces centradas en media		Eigen-faces centradas en mediana		Eigen-faces covarianza robusta	
		Validación	Prueba	Validación	Prueba	Validación	Prueba	Validación	Prueba
Paisajes	Discriminante Lineal	74%	61%	83%	78%	<b>84%</b>	76%	83%	<b>83%</b>
	Regresión Logística	88%	85%	87%	<b>89%</b>	<b>89%</b>	<b>89%</b>	86%	85%
	Bosque Aleatorio	83%	85%	88%	78%	<b>91%</b>	<b>87%</b>	82%	76%
Hombres	Discriminante Lineal	<b>100%</b>	<b>100%</b>	99%	<b>100%</b>	97%	97%	99%	<b>100%</b>
	Regresión Logística	<b>100%</b>	<b>100%</b>	98%	97%	97%	96%	98%	98%
	Bosque Aleatorio	<b>100%</b>	<b>100%</b>	<b>100%</b>	<b>100%</b>	<b>100%</b>	<b>100%</b>	<b>100%</b>	<b>100%</b>
Mujeres	Discriminante Lineal	<b>99%</b>	<b>95%</b>	85%	90%	76%	90%	81%	85%
	Regresión Logística	<b>100%</b>	<b>100%</b>	99%	<b>100%</b>	96%	97%	99%	<b>100%</b>
	Bosque Aleatorio	97%	95%	<b>100%</b>	93%	97%	93%	<b>100%</b>	<b>97%</b>
Promedio	Discriminante Lineal	<b>91%</b>	85%	89%	<b>89%</b>	86%	88%	88%	<b>89%</b>
	Regresión Logística	<b>96%</b>	<b>95%</b>	95%	<b>95%</b>	94%	94%	94%	94%
	Bosque Aleatorio	93%	93%	<b>96%</b>	90%	<b>96%</b>	93%	<b>94%</b>	91%

*\*Se señalan los valores más altos de la métrica para dicho modelo en validación y pruebas*

# CONCLUSIONES

- A pesar de la carga de multidimensionalidad, la información espacial que corresponde a los datos originales se mantiene respecto a los cálculos de distancia pixel a pixel entre cada imagen, por lo que los datos originales, a pesar de representar una serie de computaciones más costosas, conlleva a una de las mejores clasificaciones.
- Las pruebas no paramétricas de pertenencia a la misma población es un excelente determinante para los conjuntos de imágenes, dado que mantiene la información espacial por el cálculo de las distancias pixel a pixel.
- La representación en un espacio reducido, aunque puede llevar a tener unas métricas inferiores que al usar el espacio original, dados los métodos correctos puede llegar a alcanzar niveles de desempeño tolerables, tomando en cuenta la relación entre la reducción del espacio y la pérdida de información.
- Los modelos de clasificación no paramétricos responden de manera más apropiada que los modelos paramétricos, mostrando la robustez de los mismos y cómo estos se presentan balanceados a la hora de clasificar conjuntos de datos con desbalance de clases.
- Para una próxima investigación se sugiere explorar la creación de espacios de eigen-caras con distintas matrices de covarianza no paramétricas: Modificada por Ledoit-Wolff, Co-median, Máxima Kurtosis, etc.

¿Preguntas?

MUCHAS GRACIAS

