Brecha Salarial de Género en México: Análisis y Predicción con Machine Learning de la Tendencia en 2019

WAGEFORECASTERMX

Lara Pacheco,Erick Alberto
Bello Monzoy, José Alfredo
Martín del Campo Gómez, Anahí
Cedillo Padilla, Xochitl Alejandra
Torres Naranjo, Silvia Leticia
Mendoza Carlos

November 23, 2019

Abstract

A pesar de que actualmente muchas de las mujeres se encuentran igualdad de formación y experiencia, diferentes organismos han demostrado que existe una gran brecha salarial de género en el Mundo. Tan solo en América Latina, MÃl'xico es el paś con la peor brecha salarial, según publicó Forbes en un informe de julio de 2019. Dada la actual situacón en México y la agenda que el país tiene para reducir esta brecha surgen preguntas tales como: " ¿ Qué variables en México tienen mayor influencia en el salario? ¿Es diferente según el género?".
El presente trabajo tiene como objetivo generar un modelo, usando m'etodos no lineales, que permita calcular el salario usando variables como el estado, género y rango de edades. La intención es que las predicciones generadas por el modelo sean utilizadas por organizaciones y que ayuden a tomar decisiones informadas en la implementación de políticas públicas estratégicas para disminuir la brecha salarial entre hombres y mujeres.

Key words: Machine Learning, Algoritmos de regresión, Brecha salarial de Género, Wage Forecast

CONTENTS

1 INTRODUCCIÓN

En el análisis de mercado laboral es posible observar la diversificación de ocupaciones basadas en la delimitación geográfica y, en consecuencia, diferentes remuneraciones se generan de acuerdo con el tipo de trabajo. Estas remuneraciones a su vez poseen ciertas diferencias dentro de un mercado competitivo; las mismas se sustentan en la oferta y la demanda de una determinada profesión, e incluso con el nivel de productividad que el capital humano posee. Dada la premisa anterior, cualquier factor que no impacte directamente en el nivel de productividad de un individuo, no debería afectar a su vez en la remuneración que el mismo perciba; dicho de otra manera, la religión, color de piel o GENERO no deberían generar diferencias en las remuneraciones. [5]

A pesar de que en las últimas décadas las mujeres han aumentado su nivel de educación y ocupado posiciones laborales de la misma índole que los hombres, diferentes organismos han demostrado que existe una gran brecha salaria de género en el Mundo. Los números son tan alarmantes que uno de los objetivos del G-20, es reducir la brecha de genero en un 25% para el 2025. [4]

2 Marco Teórico

2.1 Estudios Actuales

Las brechas salariales de género en México se han estudiado asiduamente. Entre los primeros análisis de las brechas salariales se encuentra el de (Alarcón, 1994), quienes utilizaron la muestra urbana de la Encuesta Nacional de Ingreso y Gasto de los Hogares (ENIGH) de 1984, 1989 y 1992. En sus trabajos encontraron que en 1984 las mujeres ganaban 23.3% menos que los hombres; hacia 1989 esta cifra había aumentado a 28.4%, y en 1992 disminuyó a 25.3%. Siguiendo la línea de investigación de Oaxaca (1973) y Blinder (1973), estos autores realizaron una descomposición de la brecha salarial en la media, mediante la estimación de ecuaciones de Mincer (1974) , para analizar tanto la parte de la brecha originada por características observables como la parte provocada por los retornos a tales características. Encontraron también que sólo 27.5% de la brecha se explicaba por diferencias en capital humano en 1984, mientras que en 1989 la proporción fue de 14.4% y en 1992 de 21.2%; es decir que entre 70 y 85% de las brechas se debían a diferencias en los retornos al capital humano, lo cual podría sugerir discriminación en contra de las mujeres o diferencias en productividad que no fueron controladas en la regresión.

Por su parte, Brown, Pagan y Rodríguez-Oreggia (1999) analizaron los cambios en las brechas salariales entre 1987 y 1993 con base en datos de los terceros trimestres de la Encuesta Nacional de Empleo Urbano. Ellos realizaron una descomposición de Wellington (1993) de los cambios de la brecha en el tiempo, y una descomposición de Oaxaca-Blinder para analizar el efecto de la estructura ocupacional en la brecha. Encuentran que la brecha creció en el periodo de un nivel inicial de 20.8%, en 1987, a 22%, en 1993. Este crecimiento en la brecha se debió a cambios en las dotaciones, pues a causa de los cambios en los retornos la brecha se hubiese cerrado. Los autores también encontraron que la mayor parte de la brecha se generó por diferencias en retornos. Sin embargo, lo interesante de sus hallazgos es que la inclusión de controles ocupacionales aumenta la proporción de la brecha explicada por diferencias a los retornos, lo cual, según explican, puede ser resultado de la poca desagregación de las categorías ocupacionales. Es decir, la segregación ocupacional disminuye la brecha salarial en México, lo cual contrasta con los resultados de otros países (Blau, Simpson y Anderson, 1998).

Más recientemente, Pagan y Ullibarri (2000) analizaron la desigualdad salarial entre hombres y mujeres por medio del índice de Jenkins, corrigiendo por selección en la participación laboral de las mujeres. Con base en datos de la ENEU del tercer trimestre de 1995, encontraron que existe mayor desigualdad entre personas con alta y baja escolaridad, así como entre aquellas con mayor experiencia. Por su parte, elaboraron una descomposición del tipo Oaxaca-Blinder mediante la ENIGH 2000, corrigiendo por sesgo de selección con la metodología de Heckman (1974, 1979). Los autores fueron los primeros en incluir en su análisis zonas urbanas y rurales. Hallaron que 85% de la brecha se debe a diferencias en retornos y que ésta es mayor en zonas rurales; de hecho, el efecto de las dotaciones otorga una ventaja a las mujeres. Por último, García y Mendoza (2009) elaboraron una descomposición de Oaxaca-Blinder sin corregir por sesgo de selección y usando datos de la ENOE 2006. Su hallazgo fue una brecha salarial de 12.4% y, al contrario que el resto de la bibliografía, de-

terminaron que 87.6% de la brecha se explica por diferencias en las dotaciones, según la cual el 12.4% restante corresponde a diferencias en los retornos a éstas.

2.2 Análisis desde el Contexto Social

Para el caso de México, la única causa explorada de la brecha no explicada o la brecha de retornos ha sido la liberalización comercial Artecona y Cunningham (2002) encuentran evidencia que sugiere que la liberalización comercial provocó una disminución de la discriminación en las empresas manufactureras que fueron mÃąs afectadas por la liberalización. Por su parte, Aguayo-Téllez, Airola y Juhn (2010) encuentran que la liberalización comercial no afectó los salarios, pero sí tuvo un efecto en el empleo de las mujeres. De esta manera, la evidencia sobre esta posible causal no es muy concluyente. Consideramos que investigaciones futuras deben abordar la cuestión de las causas de los cambios en las brechas salariales de género y de la existencia de "pisos pegajosos" y "techos de cristal". Creemos que el mecanismo expuesto por De la Rica et al. (2008) podría también estar operando en el caso mexicano.

Por otra parte, y siguiendo a Arulampalam et al. (2007) y Christofides et al. (2013) también es necesario explorar el efecto que tuvieron los cambios institucionales de las décadas de 1980 y 1990 (como la caída del salario mínimo real, la negociación colectiva de los salarios y la cobertura sindical) en la brecha salarial de género. Por ejemplo, Arulampalam et al. (2007) sugieren que la dispersión salarial estÃą negativamente relacionada con los "techos de cristal" y positivamente relacionada con los "pisos pegajosos". Si este resultado fuese generalizable a México, deberíamos observar que la disminución de la desigualdad entre 2000 y 2010 se hubiera reflejado en mayores "techos de cristal" y menores "pisos pegajosos", lo cual contrasta con nuestros resultados. Así, es importante analizar cómo la reducción observada en la desigualdad salarial en la década pasada afectó la brecha salarial de género en el contexto mexicano. Otra posible línea de investigación se abre en torno al hallazgo sistemÃątico en la bibliografía sobre México de que la segregación ocupacional de hecho favorece la brecha salarial de género, lo cual es congruente con los resultados de Australia (Barón y Cobb-Clark, 2010), pero no con los de otros países (Blau, Simpson y Anderson, 1998), así como con la creencia generalizada de que la segregación ocupacional es una causal de la existencia de la brecha salarial. Un mayor entendimiento de estas causales nos daría mejores fundamentos para diseñar políticas públicas que promuevan la igualdad de género en el mercado laboral

3 Planteamiento del Problema

3.1 Descriptción del problema

En el análisis de mercado laboral es posible observar la diversificación de ocupaciones basadas en la delimitación geográfica y, en consecuencia, diferentes remuneraciones se generan de acuerdo con el tipo de trabajo. Estas remuneraciones a su vez poseen ciertas diferencias dentro de un mercado competitivo; las mismas se sustentan en la oferta y la demanda de una determinada profesión, e incluso con el nivel de productividad que el capital humano posee. Dada la premisa anterior, cualquier factor que no impacte directamente en el nivel de productividad de un individuo, no debería afectar a su vez en la remuneración que el mismo perciba; dicho de otra manera, la religión, color de piel o GENERO no deberían generar diferencias en las remuneraciones. [5]

A pesar de que en las últimas décadas las mujeres han aumentado su nivel de educación y ocupado posiciones laborales de la misma índole que los hombres, diferentes organismos han demostrado que existe una gran brecha salaria de género en el Mundo. Los números son tan alarmantes que uno de los objetivos del G-20, es reducir la brecha de genero en un 25% para el 2025. [4]

3.2 Justificación

En América Latina, México se encuentra en el último lugar en materia de igualdad de género. de acuerdo con el índice de brechas de género globales, “entre los 56 países estudiados México se encuentra en el lugar número 52, sólo por encima de India, Corea, Jordania, Pakistán, Turquía y Egipto”.[1]

3.3 Objetivos

3.3.1 Objetivo General

- Analizar la tendencia del nivel de sueldo, en base a factores socio-demograficos, entre ellos el género.

3.3.2 Objetivos Particulares

- Generar un dataset de entrenamiento a partir de bases de datos abiertas del gobierno
- Encontrar el modelo con el menor Error Cuadrático Medio
- Contrastar el comportamiento de modelos predictivos no lineales en relación con modelos de prediccón lineal

4 HIPÓTESIS

Las siguientes hipótesis fueron plateadas en base a la disponibilidad de datos y la bibliografía leída:

- El estado, género, y rango de edad son algunas de las variables que tienen mayor influencia en el salario en México.
- Es posible mejorar la capacidad de predición del sueldo usando modelos no lineales.

5 METODOLOGÍA

5.1 DESCRIPCIÓN DE LOS DATOS

5.1.1 LECTURA Y ESTANDARIZACIÓN DE LOS DATOS

El análisis se hizo en base a un dataset de licencia abierta obtenido de los datos abiertos del gobierno. Contiene las siguientes especificaciones:

Campo	Valor
Última actualización	hace 6 horas
Formato	CSV
Licencia	Libre Uso
Estado	Activo
Fecha de última modificación de datos	2019-08-23T00:00:00Z
Periodo de actualización	R/P3M
Periodo cubierto por los datos	De 2005-03-01 a 2019-06-30
Id	c2d04700-b335-4f95-8e23-baf4e04cd6b7
Id del Paquete	96e6a1e1-0de2-4656-9fba-f93b9a176f85
Id de Revisión	21b0794e-99f4-49a6-a6f7-bf14b3a24de0

La serie estadística presenta la población que tiene una actividad economica subordinada y remunerada del pais para cada una de las entidades federativas, desglosada por sexo, grupos de edad y cual es el nivel de ingreso de la población.**??**[3]

	Periodo	Entidad_Federativa	Sexo	Grupo_edad	Nivel_ingreso	Numero_personas
0	20050301	Aguascalientes	Hombre	15 A 24 A¥OS	Menos de 1 s.m.	4284
1	20050301	Aguascalientes	Hombre	15 A 24 A¥OS	1 salario m¡nimo	179
2	20050301	Aguascalientes	Hombre	15 A 24 A¥OS	M s de 2 hasta 3 s.m.	10503
3	20050301	Aguascalientes	Hombre	15 A 24 A¥OS	M s de 3 hasta 5 s.m.	16803
4	20050301	Aguascalientes	Hombre	15 A 24 A¥OS	M s de 5 hasta 10 s.m.	1955

Figure 5.1 Fragmento del dataset utilizado sin procesar.

Con el objetivo de poder utilizar la información fue necesario generar una estandarización para las columnas, estados, espacio temporal. Las columnas se estandarizaron de la siguiente manera:

```
data = pd.read_csv(data_path, encoding='latin1').rename(
    columns = {
        "Periodo": "t",
        "Entidad_Federativa": "state",
        "Sexo": "gender",
```

```
        "Grupo_edad": "age",
        "Nivel_ingreso": "wage_level",
        "Numero_personas": "population"
    }
)
```

Se detectaron las diferencias este los estados y unificó con el siguiente código:

```
# Standardize state variable
standardize_state = {
    'Coahuila': 'Coahuila␣de␣Zaragoza',
    'Ciudad␣de␣M\x82xico': 'Distrito␣Federal',
    'Estado␣de␣M\x82xico': 'M\'exico',
    'Michoac\xa0n': 'Michoac\'an␣de␣Ocampo',
    'Nuevo␣LeÂćn': 'Nuevo␣Le\'on',
    'Queretaro': 'Quer\'etaro',
    'San␣Luis␣PotosÂą': 'San␣Luis␣Potos\'i',
    'Veracruz': 'Veracruz␣de␣Ignacio␣de␣la␣Llave',
    'Yucat\xa0n': 'Yucat\'an'
}

data['state'] = [standardize_state.get(s, s) for s in data.state]
```

Se siguió el mismo proceso con el resto de las variables dentro del dataset:

```
# Standardize year variable
data['year'] = [int(str(y)[:4]) for y in data.t]

# Standardize age variable
standardize_age_dictionary = {age_val: age_val.replace("AÂěOS", "").replace("␣", "") for
data['age'] = [standardize_age_dictionary[age] for age in data.age]
```

De igual forma se limpiaron los datos de informcaión poco relevante para el estudio ('Nacional', No especificado, y NO ESPECIFICADO).

Campo estado (ver figura 5.2)

```
data.Entidad_Federativa.unique()
```

```
array(['Aguascalientes', 'Baja California', 'Baja California Sur',
       'Campeche', 'Chiapas', 'Chihuahua', 'Ciudad de M\x82xico',
       'Coahuila', 'Colima', 'Durango', 'Estado de M\x82xico',
       'Guanajuato', 'Guerrero', 'Hidalgo', 'Jalisco', 'Michoac\xa0n',
       'Morelos', 'Nayarit', 'Nuevo Le¢n', 'Oaxaca', 'Puebla',
       'Queretaro', 'Quintana Roo', 'San Luis Potos¡', 'Sinaloa',
       'Sonora', 'Tabasco', 'Tamaulipas', 'Tlaxcala', 'Veracruz',
       'Yucat\xa0n', 'Zacatecas', 'Nacional'], dtype=object)
```

Figure 5.2 Valores que se limpiaron del dataset a través de la eliminacide sus respectivas las filas

Campo Salario (ver figura 5.3)

```
data.Nivel_ingreso.unique()
```

```
array(['Menos de 1 s.m.', '1 salario m¡nimo', 'M\xa0s de 2 hasta 3 s.m.',
       'M\xa0s de 3 hasta 5 s.m.', 'M\xa0s de 5 hasta 10 s.m.',
       'M\xa0s de 10 s.m.', 'No recibe ingresos', 'No especificado',
       'M\xa0s de 1 hasta 2 s.m.'], dtype=object)
```

Figure 5.3 Valores que se limpiaron del dataset a través de la eliminación de sus respectivas las filas

Campo edad (ver figura 5.4)

```
data.Grupo_edad.unique()
```

```
array(['15 A 24 A¥OS', '25 A 44 A¥OS', '45 A 64 A¥OS', '65 A¥OS Y MAS',
       'NO ESPECIFICADO'], dtype=object)
```

Figure 5.4 Valores que se limpiaron del dataset a través de la eliminación de sus respectivas las filas

5.1.2 Visualización de los datos estandarizados

Despues de estandarizar los datos

	t	state	gender	age	wage
6	2005.00	Aguascalientes	Mujer	45A64	382.773801
262	2005.25	Aguascalientes	Mujer	45A64	381.252663
518	2005.50	Aguascalientes	Mujer	45A64	386.129621
774	2005.75	Aguascalientes	Mujer	45A64	359.659292
1030	2006.00	Aguascalientes	Mujer	45A64	401.856441

Figure 5.5 Datos usados para las graficas de las figuras 5.6, 5.7

5.1.3 Autocorrelación Parcial

Las gráficas de autocorrelación y autocorrelación parcial se usan mucho en el análisis y pronóstico de series de tiempo. Estas son gráficas que resumen gráficamente la fuerza de una relación con una observación en una serie de tiempo con observaciones en pasos de tiempo anteriores. La diferencia entre la autocorrelación y la autocorrelación parcial puede ser difícil y confusa para los principiantes con el pronóstico de series de tiempo.

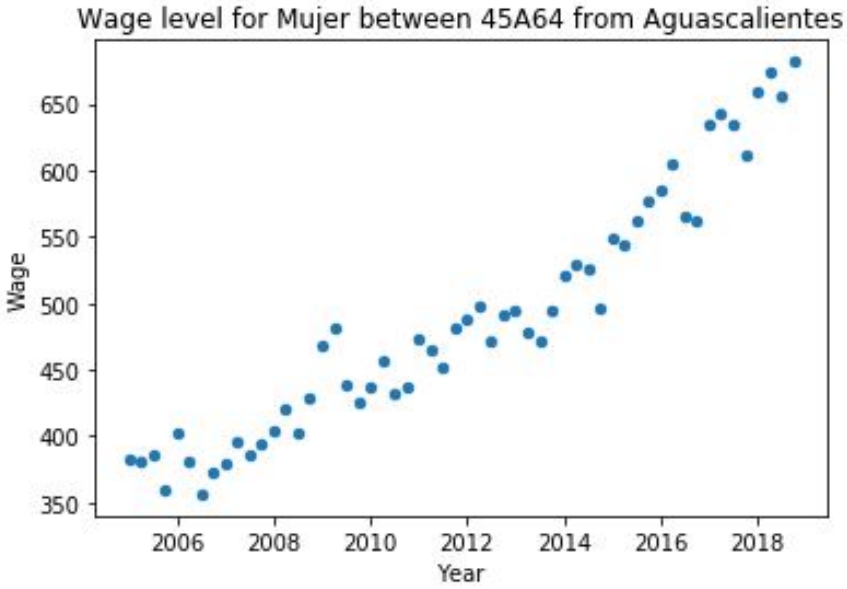

Figure 5.6 Salario en funcion del aÃśo usando un subconjunto de datos aleatorio

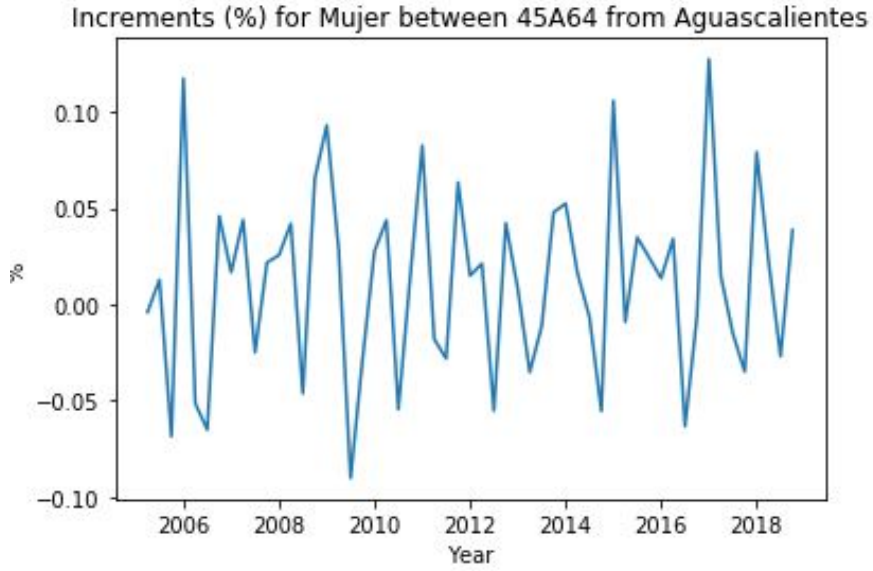

Figure 5.7 Incremento en funcion del año usando un subconjunto de datos aleatorio

Una autocorrelación parcial es un resumen de la relación entre una observación en una serie de tiempo con observaciones en pasos de tiempo anteriores con las relaciones de observaciones intermedias eliminadas.

La autocorrelación parcial en el retraso k es la correlación que resulta después de eliminar

el efecto de cualquier correlación debido a los términos en los retrasos más cortos.

La autocorrelación para una observación y una observación en un paso de tiempo anterior comprende tanto la correlación directa como las correlaciones indirectas. Estas correlaciones indirectas son una función lineal de la correlación de la observación, con observaciones en pasos temporales intermedios.

Son estas correlaciones indirectas las que la función de autocorrelación parcial busca eliminar. Sin entrar en las matemáticas, esta es la intuitivamente la autocorrelación parcial.

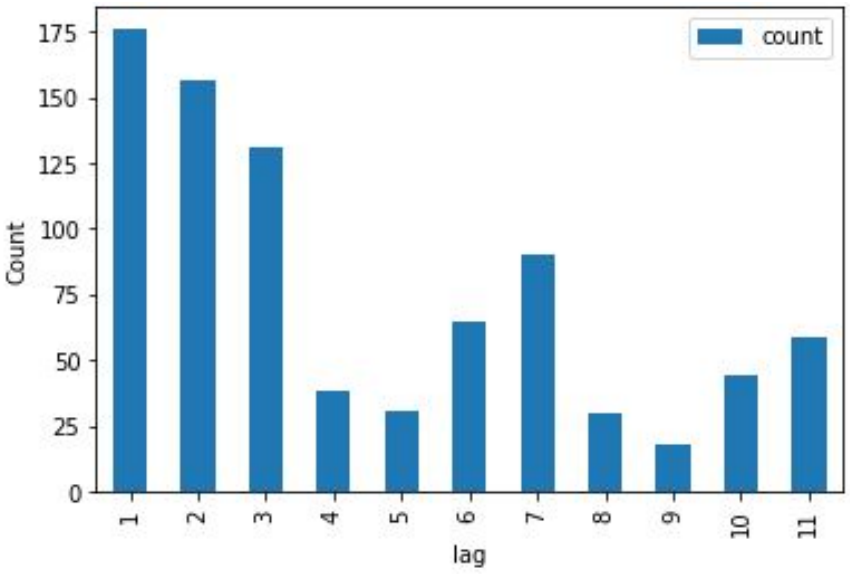

Figure 5.8 Resagos por grupo

5.2 Descripción del Modelo a Utilizar

5.2.1 Regresión Lineal

5.2.2 Decision Tree

Los modelos de árbol de Regresión y Clasificación (CRT, Classification Regression Trees), fueron introducidos en la Estadística por Breiman et al. (1984). Diversos autores utilizan el término "modelos de árbol de regresión" cuando la variable respuesta es cuantitativa y el de "modelos de árbol de clasificació" cuando ésta es cualitativa. Decision tree Regression es un algoritmo de aprendizaje dentro del grupo de aprendizaje supervisado. Este se caracteriza por particionar o dividir los datos en varias clasificaciones de grupos homogéneos respecto a la variable, creando iteraciones de estas clasificaciones del DataFrame hasta conseguir el mejor resultado posible. Este método utiliza los datos de entrenamiento de periodos anteriores y los entrena para conseguir clasificar los nuevos datos. Dentro de las ventajas de este algoritmo se encuentran las siguientes: Facilidad al comprender y explorar datos, debido a

su clasificación y traficación. Dentro de las ventajas de este algoritmo se encuentran: Los sobreajustes son muy comunes en estos modelos y al usar variables continuas, se tiene por sobre entendido que se perderá precisión.

5.2.3 Random Forest Regressor

Ramdom Forest Regressor es una técnica de embolsado y no una técnica de refuerzo. Los árboles en bosques aleatorios se ejecutan en paralelo. No hay interacción entre estos árboles mientras se construyen los árboles. Funciona mediante la construcción de una multitud de árboles de decisión en el momento del entrenamiento y la salida de la clase que es el modo de las clases (clasificación) o predicción media (regresión) de los árboles individuales. Un bosque aleatorio es un metaestimulador (es decir, combina el resultado de múltiples predicciones) que agrega muchos árboles de decisión, con algunas modificaciones útiles:

- El número de características que se pueden dividir en cada nodo está limitado a un porcentaje del total (que se conoce como hiperparámetro). Esto garantiza que el modelo de conjunto no dependa demasiado de ninguna característica individual y hace un uso justo de todas las características potencialmente predictivas.
- Cada árbol extrae una muestra aleatoria del conjunto de datos original al generar sus divisiones, agregando un elemento adicional de aleatoriedad que evita el sobreajuste.

Las modificaciones anteriores ayudan a evitar que los árboles estén demasiado correlacionados. [2]

5.2.4 XGBoost

Es muy parecido al modelo Random Forest, pero esta vez los árboles tienen asociadas una función de pérdidas que tienen que minimizar (o maximizar) para conseguir los mejores resultados (por ello la palabra Boosting). Está basado o es muy parecido al árbol de decisión y es una evolución entre la clasificación y la regresión este se basa en impulsar para maximizar o minimizar la función de pérdidas trabaja sobre bases de datos de gran tamaÃśo así como múltiples variables algo importante es que admite missing values por lo mencionado anteriormente se debe tener en cuenta el equipo para correr bases de datos extensas y usar muchas variables, es recomendable saber las variables que más aportan al algoritmo.

Este clasifica o pronóstica sobre la variable objetivo, utiliza un set de datos utilizando los los árboles de decisiones para potencializar los resultados en base a un procesamiento secuencial y con una función de pérdida que minimiza el error en consecuencia es un pronosticador fuerte. Como en todo algoritmo se deben ajustar los parámetros para obtener un mínimo error de precisión.

5.2.5 NAIVE BAYES

Este algoritmo es del tipo clasificador Se basan en una técnica de clasificación estadística llamada teorema de Bayes, este algoritmo asume que las variables predictorias son independientes entre si, eso quiere decir que los conjuntos de datos pueden no tener relacion entre ellos.

El clasificador Naive Bayes agregar información usando probabilidad condicional con una asunción de independencia entre caracteristicas , este algoritmo esta dentro de la rama de algoritmos supervisados.

Bayes: asume que la presencia de una caracteristica en la clase no esta relacionada a ninguna otra caracteristica. Naive: las propiedades independientes contribuyen al calculo de la propiedad particular.

5.2.6 SUPPORT VECTOR MACHINES

Las máquinas de vectores soporte (SVM, del inglés Support Vector Machines)tienen su origen en los trabajos sobre la teoría del aprendizaje estadístico y fueron introducidas en los aÃśos 90 por Vapnik y sus colaboradores. Aunque originariamente las SVMs fueron pensadas para resolver problemas de clasificación binaria, actualmente se utilizan para resolver diversos tipos de problemas, por ejemplo, la regresión. Desde su introducción, han ido ganando un merecido reconocimiento gracias a sus sólidos fundamentos teóricos.

La idea es seleccionar un hiperplano de separación que equidiste de los ejemplos más cercanos de cada clase para, de esta forma, conseguir lo que se denomina un margen máximo a cada lado del hiperplano. Además, a la hora de definir el hiperplano, sólo se consideran los ejemplos de entrenamiento que distan del hiperplano la distancia margen. Estos ejemplos reciben el nombre de vectores soporte.

5.3 DELIMITACIONES

5.3.1 DATOS

La serie estadística presenta la población que tiene una actividad economica subordinada y remunerada del pais para cada una de las entidades federativas, desglosada por sexo, grupos de edad y cual es el nivel de ingreso de la población.**??**[3]

5.3.2 TEMPORALES

El presente es un estudio longitudinal en donde los datos de 2005 a 2018 se utilizarán como entrada para poder predecir el de salario, del 2019

6 Resultados

6.0.1 Regresión Lineal

Sin la introducción de parámetros se alcanzo el 0.082 de sme:

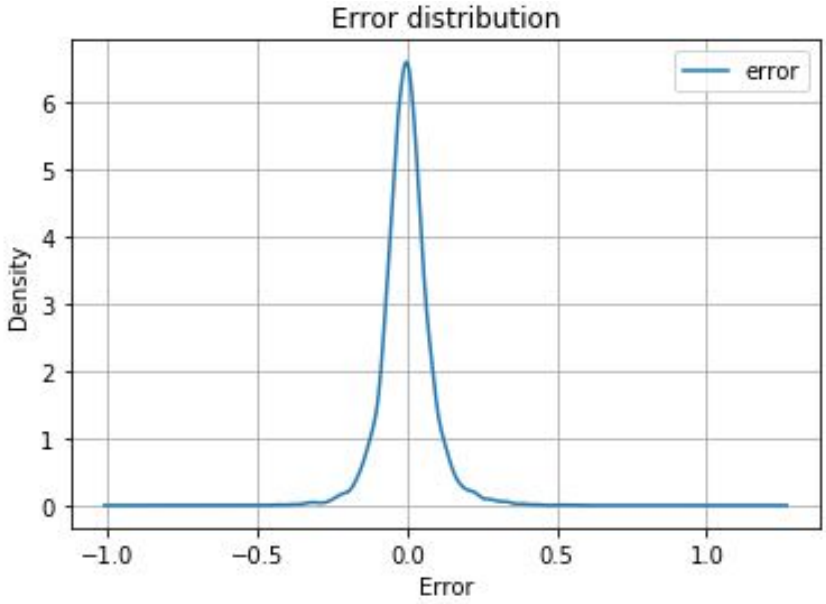

Figure 6.1 Regresión Lineal: distribución de error

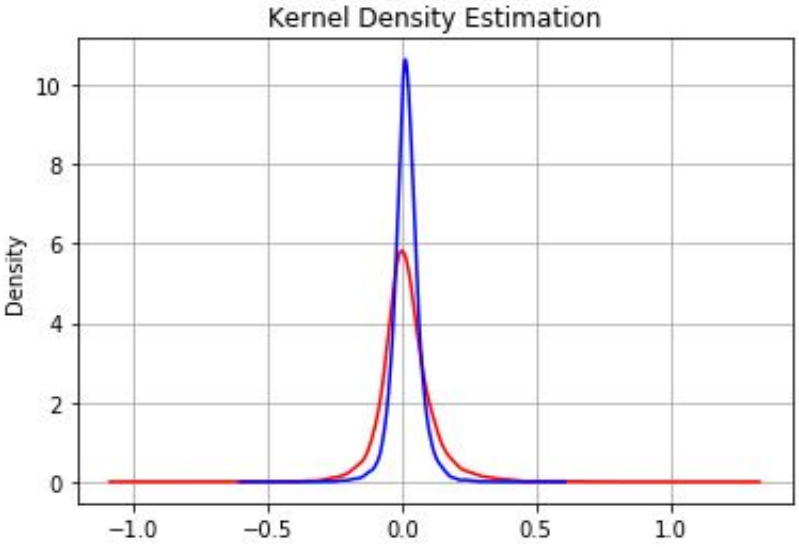

Figure 6.2 Regresión Lineal: Estimación de densidad del núcleo

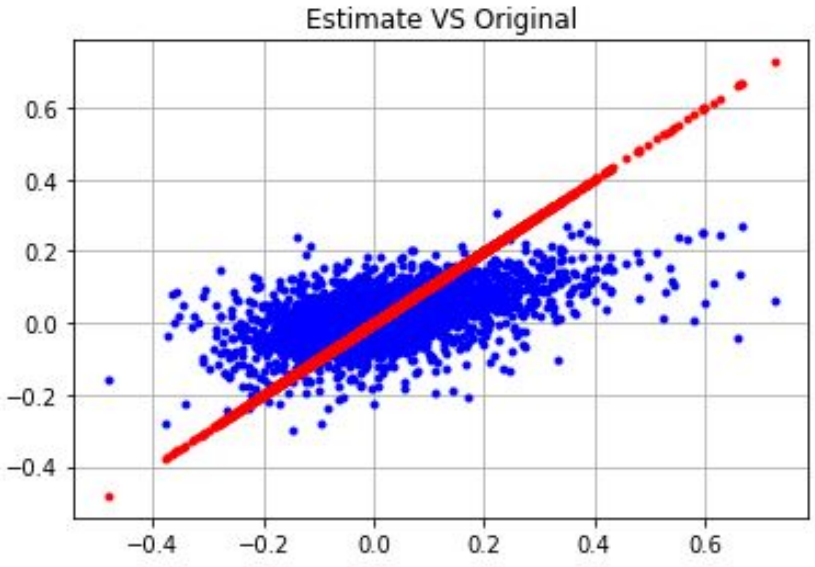

Figure 6.3 Regresión Lineal: Estimate vs Original

6.0.2 Decision Tree

El mejor módelo se obtuvo con los siguientes parámetros con un sme de 0.093583:

```
{       'criterion': mae
        'splitter': best
        'max_depth':2.0
        'max_features':0.24
}
```

Las siguientes gráficas corresponden a los resultados con los respectivos parámetros y usando los set de pruebas:

	y	y_estimate
count	338000.000000	338000.000000
mean	1.834215	1.048366
std	9.903699	1.408911
min	-38.087685	-2.184947
25%	-3.143507	0.501666
50%	1.038663	0.501666
75%	6.151501	1.923532
max	67.095144	7.415433

Figure 6.4 Data description (DT)

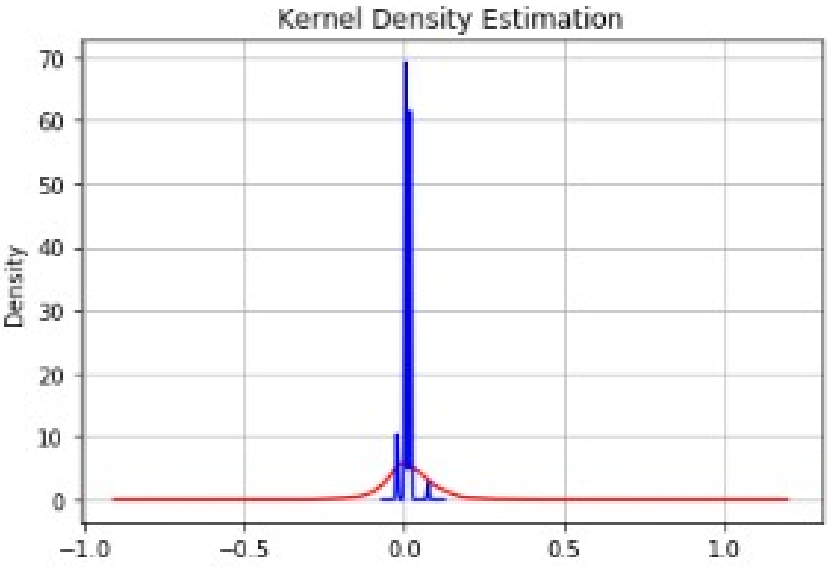

Figure 6.5 Kernel density (DT)

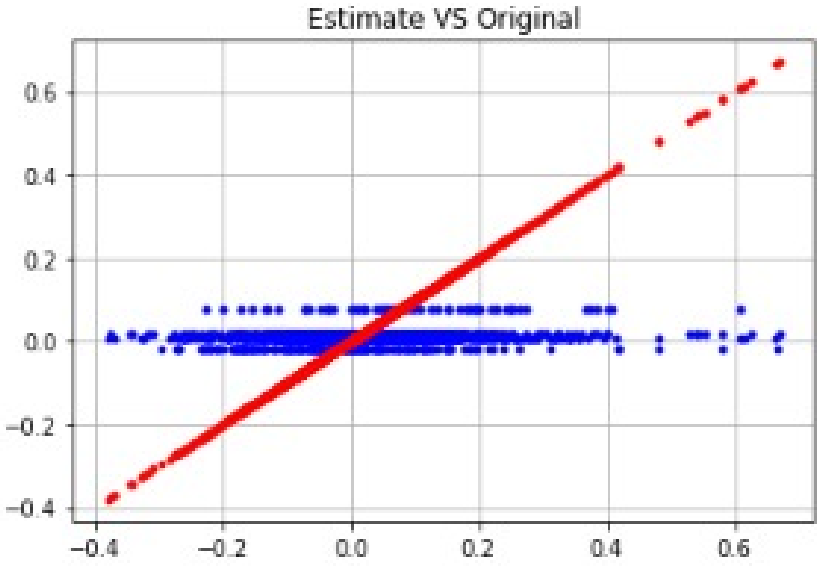

Figure 6.6 Estimate vs Original(DT)

6.0.3 XGBoost

El mejor mşdelo se obtuvo con los siguientes parámetros con un sme de 0.062:

```
xgb_grid.best_params_
{'colsample_bytree': 0.7,
'learning_rate': 0.07,
'max_depth': 5,
'min_child_weight': 4,
'n_estimators': 100,
'nthread': 4,
'objective': 'reg:linear',
'silent': 1,
'subsample': 0.7}
```

Las siguientes gráficas corresponden a los resultados con los respectivos parámetros y usando los set de pruebas:

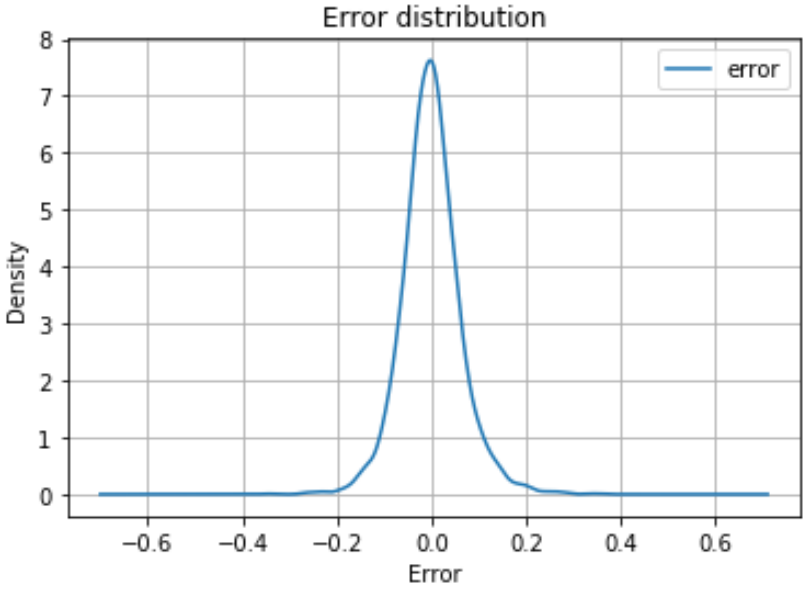

Figure 6.7 Error distribution(XGB)

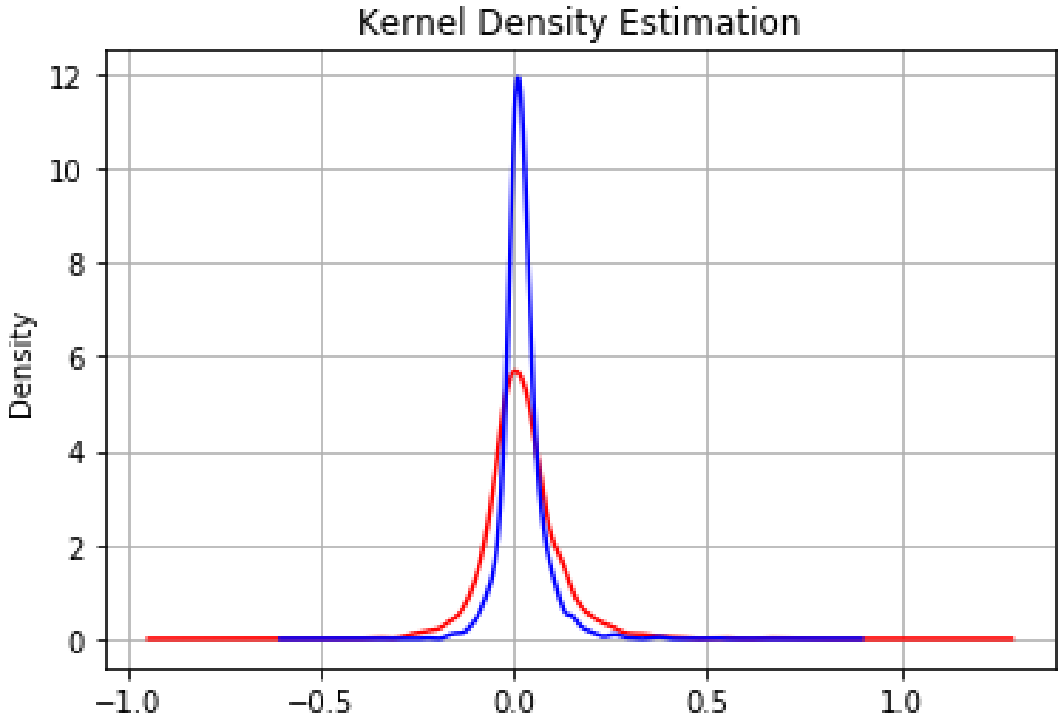

Figure 6.8 Error distribution(XGB)

6.0.4 NAIVE BAYES

Best parameter: smoothing: 307

Las siguientes gráficas corresponden a los resultados con los respectivos parámetros y usando los set de pruebas:

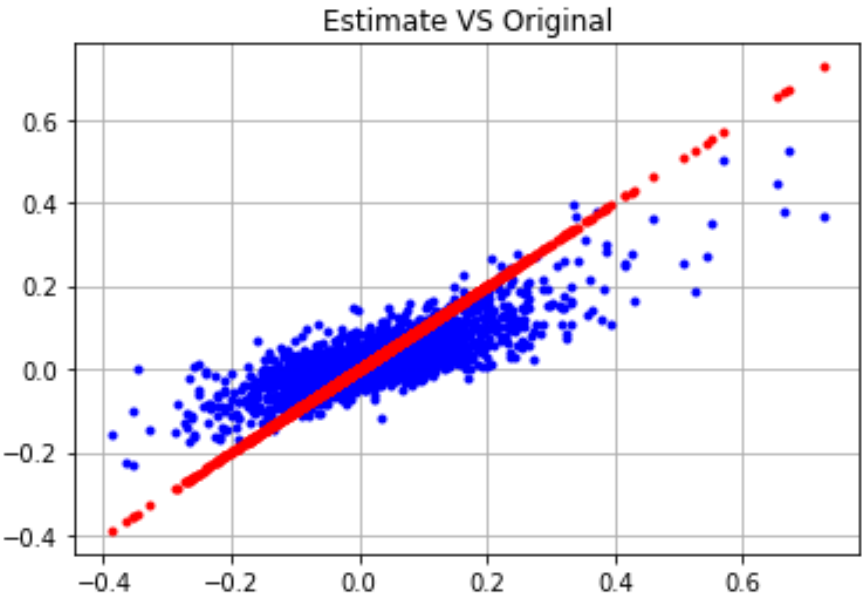

Figure 6.9 Error distribution(XGB)

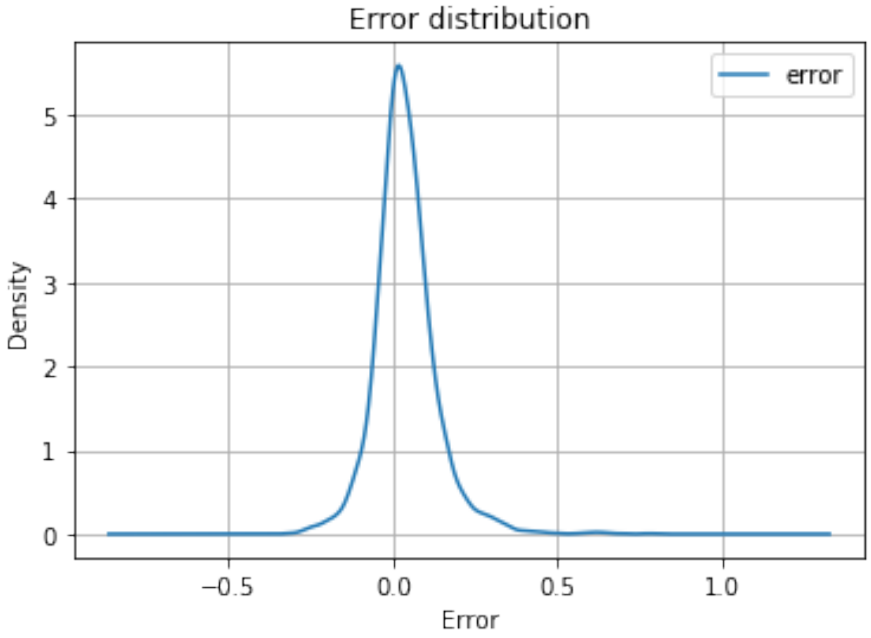

Figure 6.10 Error Distribution (Naive Bayes)

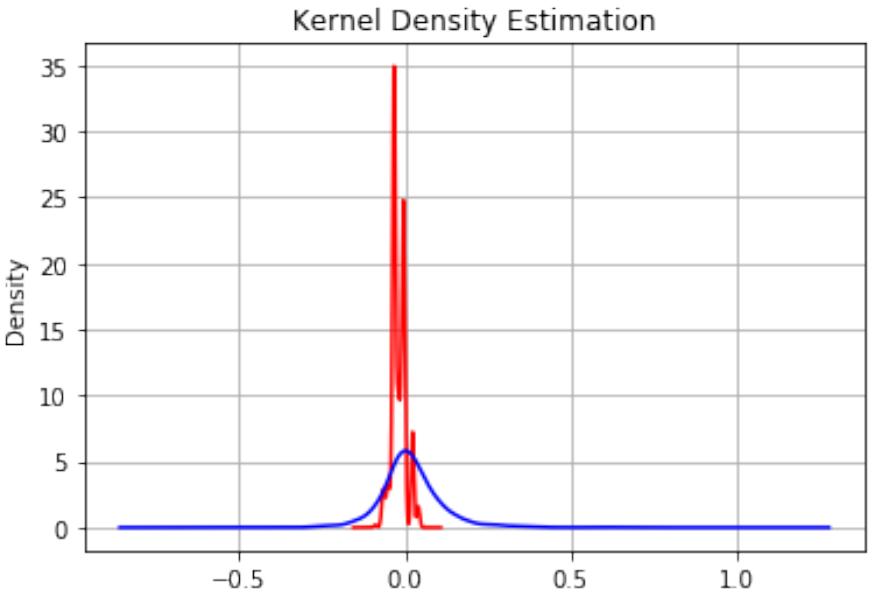

Figure 6.11 Density Estimation (Naive Bayes)

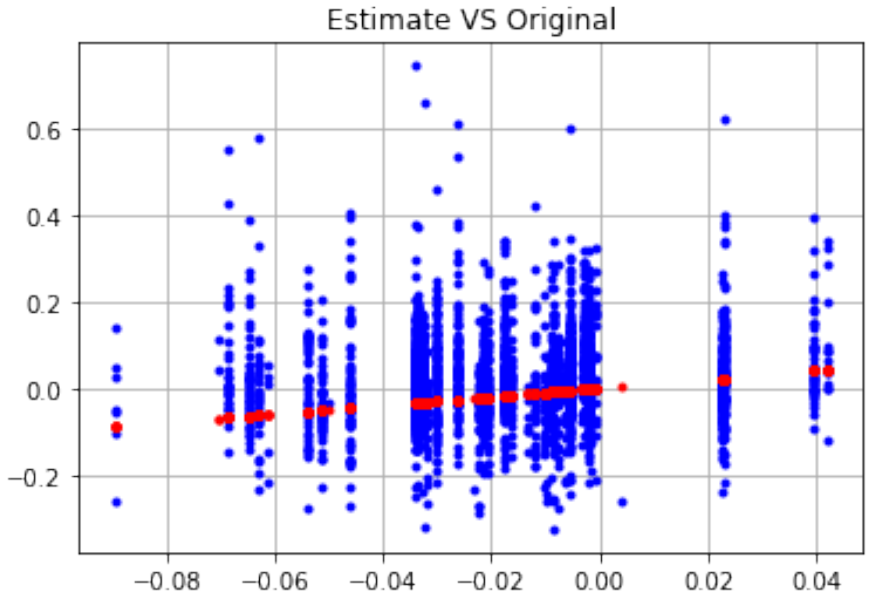

Figure 6.12 Estimation vs Original (Naive Bayes)

6.0.5 Support Vector Machines Regressor

El mejor módelo se obtuvo con los siguientes parámetros con un sme de 0.093:

```
{'kernel':'poly',
 'degree':'2',
 'gamma':'0.55',
 'coef0':'0.1',
 'C':'34',
 'epsilon':'0.2'
}
```

Las siguientes gráficas corresponden a los resultados con los respectivos parámetros y usando los set de pruebas:

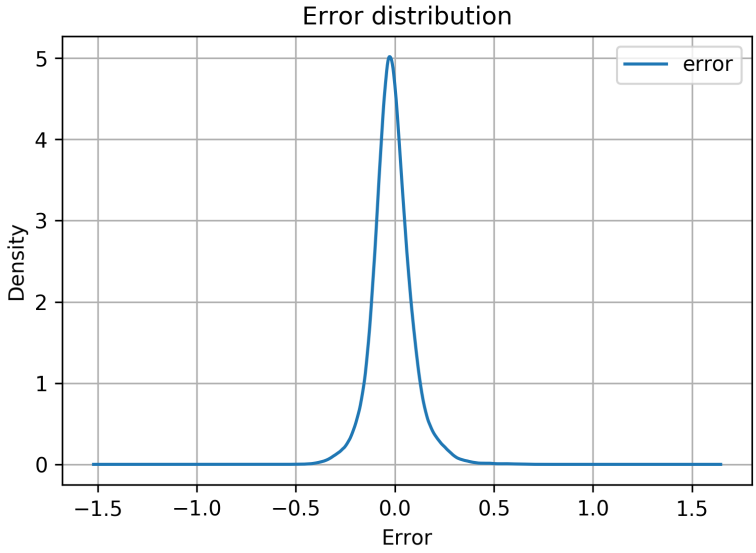

Figure 6.13 Error Distribution (SVMR)

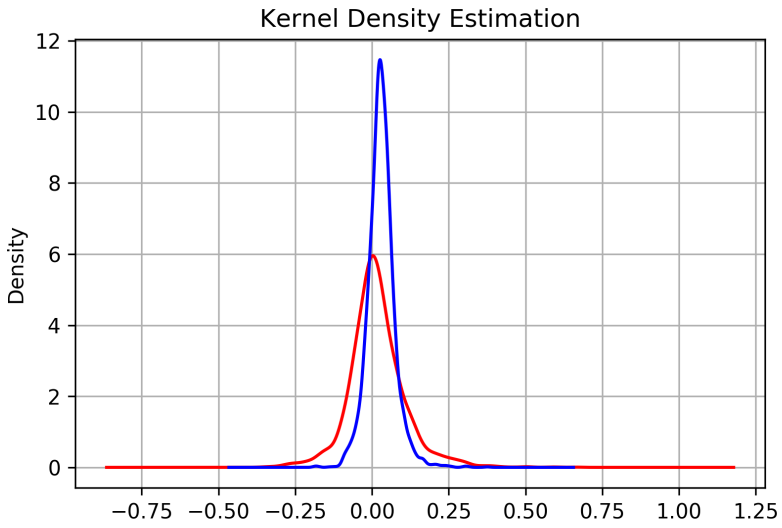

Figure 6.14 Kernel density (SVMR)

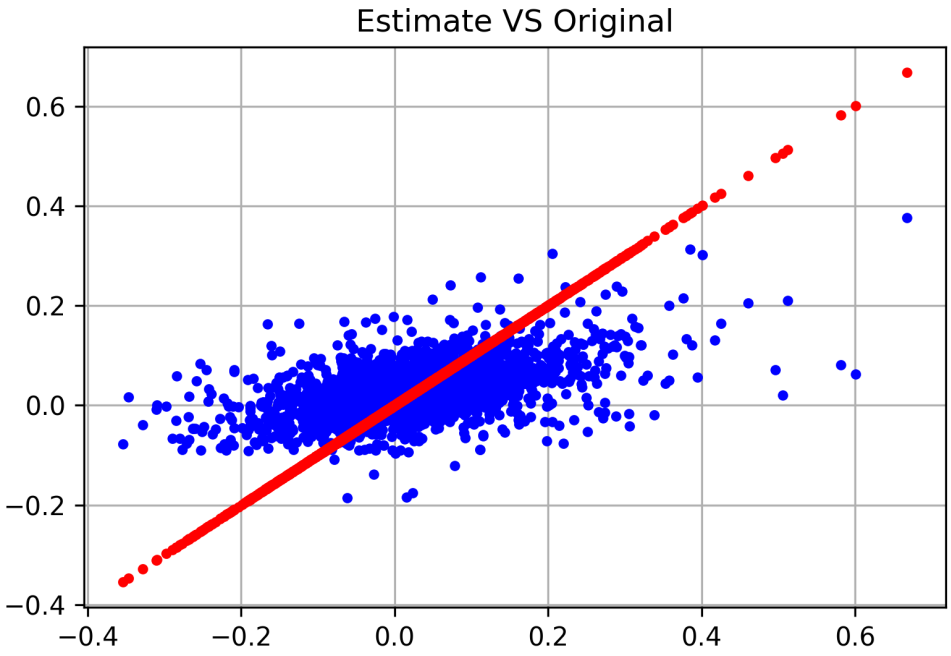

Figure 6.15 Estimate vs Original(SVMR)

7 CONCLUSIONES

A lo largo de este artículo realizamos un proyecto de clasificaciÃşn de aprendizaje de mÃąquina de principio a fin y aprendimos y obtuvimos varias ideas sobre los modelos de clasificaciÃşn y las claves para desarrollar uno con un buen rendimiento. En general el modelo de Regresión lineal tuvo un mejor desempeÃśo a diferencia de los algoritmos con árboles. Sin embargo podemos tambiľn observar que uno de los algoritmos con mejor rendimiento fue el XGBoost con un 0.062 de error por lo que puede ser un buen candidato para el cálculo de salarion dentro de la herramienta.

REFERENCES

[1] E. O. Arceo-GÃand R. M. Campos-VÃ¡zquez. EvoluciÃde la brecha salarial de gÃen MÃ. *El trimestre econÃ*, 81:619 – 653, 09 2014.

[2] L. Breiman. Random forests. *Machine learning*, 45(1):5–32, 2001.

[3] D. A. gob.mx. Indicadores estratÃl'gicos/población ocupada - subordinada y remunerada por nivel de ingresos. urlhttps://datos.gob.mx/busca/dataset/indicadores-estrategicos-poblacion-ocupada–subordinada-y-remunerada-por-nivel-de-ingresos, 2019.

[4] W. of the Ministry of Foreign Affairs of Japan. G20 osaka leaders' declaration. urlhttps://g20.org/en/documents/final$_g$20$_o$saka$_l$eaders$_d$eclaration.html, 2019.

[5] R. E. RodrÃguez PÃand D. Castro-Lugo. DiscriminaciÃsalarial de la mujer en el mercado laboral de MÃy sus regiones. *EconomÃa, sociedad y territorio*, 14:687 – 714, 12 2014.