UANL
UNIVERSIDAD AUTÓNOMA DE NUEVO LEÓN

FCFM
FACULTAD DE CIENCIAS FÍSICO MATEMÁTICAS

Reporte 2

Datos ENIGH

Regression

Materia: MAA(Aprendizaje Automático)
Profesor: José Anastacio Hernández Saldaña
Grupo: 03
Alumno: José de Jesús Almanza Trejo

Contenido

Introducción

A continuación se presenta un análisis descriptivo y aplicación de varios modelos de regresión que pueda calcular el monto de los ingresos trimestrales que perciben las personas encuestadas en el ENIGH a partir de características individuales de cada persona encontradas en algunos de los segmentos de información reportados por esta encuesta.

Se aplicaran los algoritmos de Regresión Lineal, Polinomial, Polinomial con Ridge, Polinomial con Lasso, KNN Regression y Árboles de Decisión sobre una validación cruzada calculando el estadístico de R^2 sobre la base de pruebas para seleccionar el modelo con el mejor desempeño y mostrar sus características.

Definiciones

Construcción de los datos

La (ENIGH, s.f.) cuenta con las siguientes fuentes de información:

Tabla	Descripción	Tipo de datos
(Ingresos, s.f.)	Permite identificar los ingresos y percepciones financieras y de capital de cada uno de los integrantes del hogar.	Ingresos trimestrales, estado donde trabaja
(trabajos, s.f.)	Muestra la condición de actividad de los integrantes del hogar de 12 o más años y algunas características ocupacionales.	Horas trabajadas por semana
(poblacion, s.f.)	Identifica las características sociodemográficas de los integrantes del hogar.	Género, edad, número de hijos, nivel de estudios.

De estas tablas se seleccionaron las siguientes características para realizar el ejercicio:

Variable	Descripción	Tabla	Clasificación	Tipo
folioviv	Folio de la vivienda encuestada.	ingresos	texto	llave
foliohog	Folio del hogar encuestado.	ingresos	texto	llave
numren	Número de referencia de persona.	ingresos	texto	llave
ing_tri	Ingresos trimestrales de la persona.	ingresos	continua	cuantitativa
*entidad	Entidad donde trabaja la persona.	ingresos	discreta	cualitativa
htrab	Horas trabajadas por semana.	trabajos	discreta	cuantitativa
sexo	Género de la persona.	poblacion	Binaria	cualitativa
edad	Edad de la persona.	poblacion	discreta	cuantitativa
*hijos	Construcción de número de hijos.	poblacion	discreta	cuantitativa
*nivelaprob	Nivel de estudios de la persona.	poblacion	discreta	cualitativa

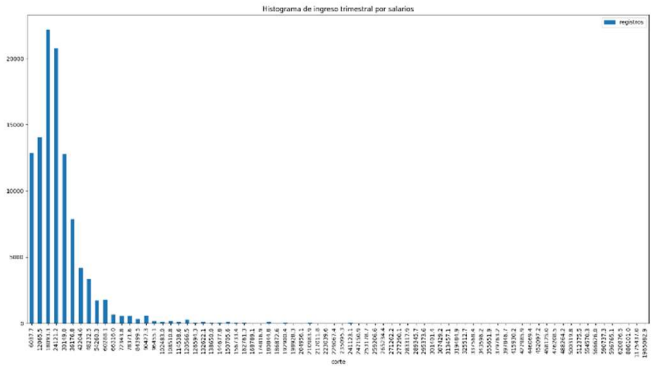
Las variables como entidad y nivel de estudios de la persona se transforman por medio del cálculo del peso de la evidencia (data, s.f.) el cual nos da una medida de la forma en que afecta la variable cualitativa respecto de la variable continua dependiente.

Análisis descriptivo

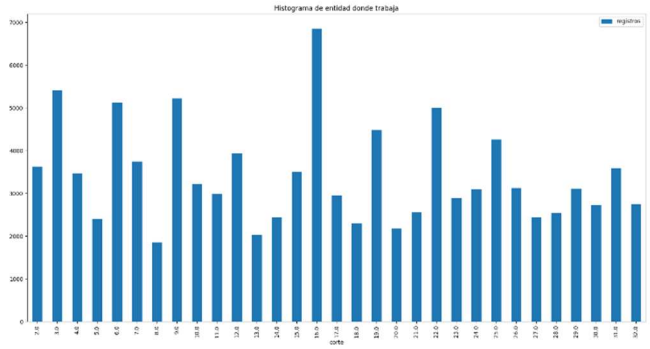
Tenemos las siguientes características de las variables seleccionadas y construidas:

Variable	Conteo	Promedio	Min	Max
ing_tri	105,685	23,627.35	9.83	1,965,083.00
WoE_entidad	105,685	-0.00	-0.39	0.41
htrab	105,685	45.62	1.00	168.00
edad	105,685	37.22	12.00	102.00
sexo	105,685	1.39	1.00	2.00
hijos	105,685	1.03	0.00	10.00
WoE_nivelaprob	105,685	-0.01	-1.48	1.08
WoE_edo_conyug	105,685	-0.00	-0.24	0.23

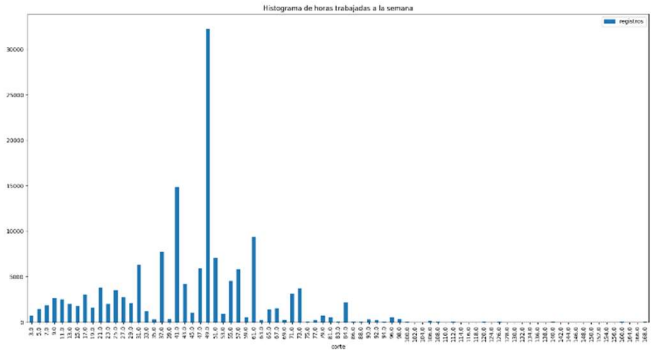
Se muestran algunas de las características de estas variables:



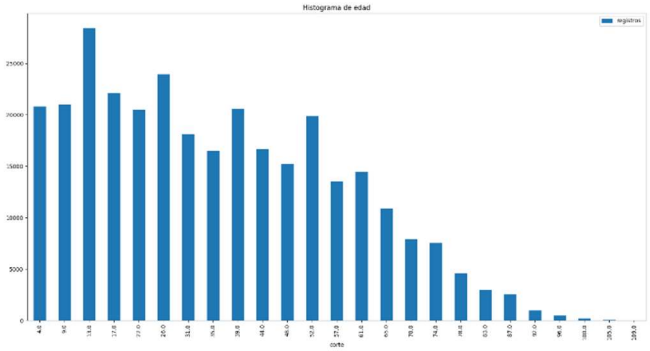
Histograma de los ingresos en el segmento de ingresos.



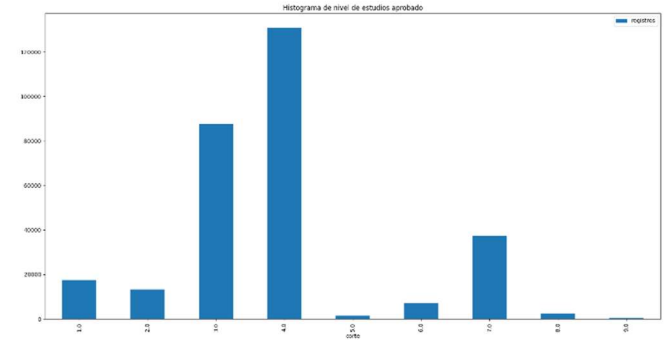
Histograma de la entidad donde trabaja en el segmento de ingresos.



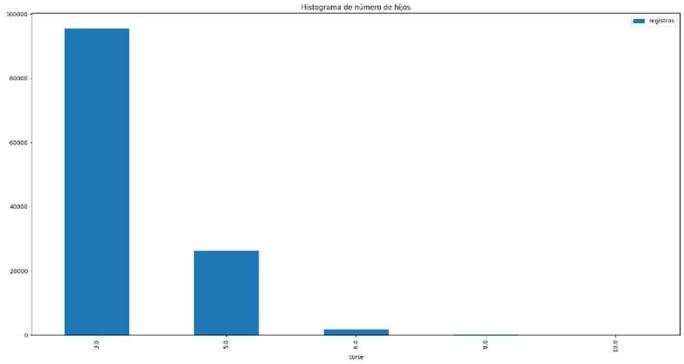
Histograma de las horas trabajadas en la semana del segmento de trabajos



Histograma de la edad en el segmento de población.



Histograma del nivel de estudios aprobado en el segmento de población.



Histograma del número de hijos construida a partir de la tabla de población.

Regresión

Se realizan varios ejercicios de regresión con los diversos algoritmos vistos en la clase, para poder seleccionar el mejor modelo aplicaremos estos algoritmos sobre una base de entrenamiento que consta del 60% de la población y el 40% restante será para medir el desempeño de este en una población no observada anteriormente.

Resultados

Se corren iterativamente algunos de los algoritmos modificando los parámetros que estos reciben para encontrar la mejor solución, encontrando los siguientes mejores modelos por algoritmo con sus características:

Algoritmo	parámetros	R^2 Train	R^2 Test
Regresión Lineal	NA	0.2611	0.2338
Regresión Lineal (datos escalados)	NA	0.2611	0.2338
Polinomial	degree=3	0.3175	0.2782
Polinomial (Ridge)	degree=4, alpha=10000	0.3220	0.2787
Polinomial (Lasso)	degree=4, alpha=100	0.3196	0.2780
K-Nearest Neighbors	neighbors=13	0.2906	0.1597
Decision Tree	depth=16, min_samples_split=50, min_samples_leaf=150	0.3209	0.2751

Del cuadro anterior podemos seleccionar el modelo polinomial con regularización ridge como el mejor modelo generado comparando su R^2 Test.

Bibliografía

data, l. (s.f.). Obtenido de https://www.listendata.com/2019/08/WOE-IV-Continuous-Dependent.html

ENIGH. (s.f.). Obtenido de https://www.inegi.org.mx/rnm/index.php/catalog/901/study-description

Ingresos. (s.f.). Obtenido de https://www.inegi.org.mx/rnm/index.php/catalog/901/data-dictionary/F55?file_name=ingresos

poblacion. (s.f.). Obtenido de https://www.inegi.org.mx/rnm/index.php/catalog/901/data-dictionary/F72?file_name=poblacion

trabajos. (s.f.). Obtenido de https://www.inegi.org.mx/rnm/index.php/catalog/901/data-dictionary/F57?file_name=trabajos