Instituto Politecnico Nacional
Escuela Superior de Computo
Trabajadores en Estados Unidos

Objetivo

El propósito principal de la investigación presente es limpiar el conjunto de datos, este contiene datos censales extraídos de las Encuestas de Población Actual de los años 1994 y 1995 realizadas por la Oficina del Censo de los Estados Unidos, para poder determinar, comprender e identificar el comportamiento de los trabajadores en diversas áreas geográficas, así como los factores que influyen en su vida laboral. Dentro de los campos que localizamos en nuestro conjunto de datos encontramos cuestiones como razones de desempleo, si eran o no inmigrantes, si aun estudiaban o cual había sido su ultimo grado escolar, cuantos de sus familiares aún son menores de 18 años, entre otras cuestiones interesantes que nos pueden mostrar la situación general en la que se encontraban los trabajadores.
Para cumplir nuestro objetivo pondremos en practica los conceptos, las estrategias de limpieza de datos y las herramientas vistas en la unidad de aprendizaje de Data Mining, también la aplicación de técnicas como imputación, normalización y transformación, siempre justificando su decisión, por qué se procedió con cierta técnica o por qué se optó por una herramienta. De la misma maneta usaremos gráficas de los datos más representativos para poder ilustrar los resultados obtenidos y dar una mejor interpretación de estos.

Descripción Conjunto de Datos

El conjunto de datos consiste en 40 atributos, de los cuales 7 son continuos y 33 son nominales. Incluyen variables demográficas y relacionadas con el empleo, sus datos fueron extraídos de las Encuestas de Población Actual de los años 1994 y 1995 realizadas por la Oficina del Censo de los Estados Unidos.
Con el fin de realizar un estudio completo se utilizaran todos los registros del conjunto resultando en un total de 199,523 registros a tratar, pueden ser consultados en
https://archive.ics.uci.edu/ml/datasets/Census-Income+%28KDD%29
En seguida describiremos la información del conjunto de datos:
Propietario Original:
U.S. Census Bureau
http://www.census.gov/
United States Department of Commerce
Donante:
Terran Lane and Ronny Kohavi
Data Mining and Visualization
Silicon Graphics.
terran '@' ecn.purdue.edu, ronnyk '@' sgi.com

Data Mining
3CV11

Integrantes
Aguilar Martínez Oswaldo
Arévalo Andrade Miguel Ángel
Guerrero Espinosa Ximena Mariana

Profesora Ocampo Botello Fabiola

28 de Marzo, 2021

Instituto Politecnico Nacional
Escuela Superior de Computo
Trabajadores en Estados Unidos

Selección de los atributos

Para hacer la seleccion de atributos que usaremos durante la investigacion nos basamos en dos factores en la relevancia que tienen para el estudio.De igual forma buscamos que en los atributos no existiera una cantidad tan grande de datos faltantes.
De esta manera a continuación enlistaremos los 20 atributos seleccionados.

AAGE, ACLSWKR, AHGA, AHRSPAY,AHSCOL, AMARITL, AMJIND, ARACE, ASEX,AUNTYPE, CAPGAIN, CAPLOSS, NOEMP, PARENT, PEFNTVTY, PEMNTVTY, PENATVTY, PRCITSHP, WKSWORK, YEAR.

Con esta eleccion procederemos a desarrollar el diccionario de datos donde definiremos, de cada atributo seleecionado, su nombre, el tipo de dato, su dominio y una breve descripcion del mismo para un mejor entendimiento.

Diccionario de datos

Nombre del Atributo	AAGE
Tipo de dato	Númerico
Dominio	0-90
Descripción	Edad de la persona al momento del registro de la información
Nombre del Atributo	**ACLSWKR**
Tipo de dato	Nominal
Valores	Not in universe, Federal government, Local government, Never worked, Private, Self-employed-incorporated, Self-employed-not incorporated, State government, Without pay
Descripción	Hace referencia a la clase de trabajador que es la persona al momento del registro de la información
Nombre del Atributo	**AHGA**
Tipo de dato	Nominal
Valores	Children, 7th and 8th grade, 9th grade, 10th grade, High school graduate, 11th grade, 12th grade no diploma, 5th or 6th grade, Less than 1st grade, Bachelors degree(BA AB BS), 1st 2nd 3rd or 4th grade, Some college but no degree, Masters degree(MA MS MEng MEd MSW MBA), Associates degree-occup /vocational, Associates degree-academic program, Doctorate degree(PhD EdD), Prof school degree (MD DDS DVM LLB JD)
Descripción	Grado de estudio que tiene el individuo

Data Mining
3CV11

Integrantes
Aguilar Martínez Oswaldo
Arévalo Andrade Miguel Ángel
Guerrero Espinosa Ximena Mariana

Profesora Ocampo Botello Fabiola

28 de Marzo, 2021

Instituto Politecnico Nacional

Escuela Superior de Computo

Trabajadores en Estados Unidos

Nombre del Atributo	AHRSPAY
Tipo de dato	Númerico
Dominio	0-9,999
Descripción	Haace referencia al salario por hora del individuo
Nombre del Atributo	**AHSCOL**
Tipo de dato	Nominal
Valores	Not in universe, High school, College or university
Descripción	Hace referencia a si el inidividuo estaba o no estudiando en al momento del registro de la información
Nombre del Atributo	**AMARITL**
Tipo de dato	Nominal
Valores	Never married, Married-civilian spouse present, Married-spouse absent, Separated, Divorced, Widowed, Married-A F spouse present
Descripción	Hace referencia al estado civil del individuo al momento del registro de la información
Nombre del Atributo	**AMJIND**
Tipo de dato	Nominal
Valores	Not in universe or children, Entertainment, Social services, Agriculture, Education, Public administration, Manufacturing-durable goods, Manufacturing-nondurable goods, Wholesale trade, Retail trade, Finance insurance and real estate, Private household services, Business and repair services, Personal services except private HH, Construction, Medical except hospital, Other professional services, Transportation, Utilities and sanitary services, Mining, Communications, Hospital services, Forestry and fisheries, Armed Forces.
Descripción	Hace referencia al enfoque de la industria en la que se desenvuelve el individuo
Nombre del Atributo	**ARACE**
Tipo de dato	Nominal
Valores	White, Black, Other, Amer Indian Aleut or Eskimo, Asian or Pacific Islander
Descripción	Hace referencia a la raza del individuo
Nombre del Atributo	**ASEX**
Tipo de dato	Nominal
Valores	Female, Male.
Descripción	Hace referencia al genero biologico del individuo
Nombre del Atributo	**AUNTYPE**
Tipo de dato	Nominal
Valores	Not in universe, Re-entrant, Job loser - on layoff, New entrant, Job leaver, Other job loser.
Descripción	Hace referencia a si el individuo tiene o no empleo

Data Mining
3CV11

Integrantes
Aguilar Martínez Oswaldo
Arévalo Andrade Miguel Ángel
Guerrero Espinosa Ximena
Mariana

Profesora Ocampo Botello Fabiola

28 de Marzo, 2021

Instituto Politecnico Nacional

Escuela Superior de Computo

Trabajadores en Estados Unidos

Nombre del Atributo	CAPGAIN
Tipo de dato	Numérico
Dominio	0 - 99,999
Descripción	Hace referencia a las ganancias de capital del individuo
Nombre del Atributo	**CAPLOSS**
Tipo de dato	Numérico
Dominio	0 - 4,608
Descripción	Hace referencia a las perdidas de capital del individuo
Nombre del Atributo	**NOEMP**
Tipo de dato	Numérico
Dominio	0-6
Descripción	Hace referencia a la cantidad de personas que trabajan para el empleador
Nombre del Atributo	**PARENT**
Tipo de dato	Nominal
Valores	Both parents present, Neither parent present, Mother only present, Father only present, Not in universe.
Descripción	Hace referencia a los miembros de la familia que son menores a 18 años
Nombre del Atributo	**PEFNTVTY**
Tipo de dato	Nominal
Valores	Mexico, United-States, Puerto-Rico, Dominican-Republic, Jamaica, Cuba, Portugal, Nicaragua, Peru, Ecuador, Guatemala, Philippines, Canada, Columbia, El-Salvador, Japan, England, Trinadad&Tobago, Honduras, Germany, Taiwan, Outlying-U S (Guam USVI etc), India, Vietnam, China, Hong Kong, Cambodia, France, Laos, Haiti, South Korea, Iran, Greece, Italy, Poland, Thailand, Yugoslavia, Holand-Netherlands, Ireland, Scotland, Hungary, Panama
Descripción	Hace referencia al lugar de nacimiento del papá del individuo
Nombre del Atributo	**PEMNTVTY**
Tipo de dato	Nominal
Valores	India, Mexico, United-States, Puerto-Rico, Dominican-Republic, England, Honduras, Peru, Guatemala, Columbia, El-Salvador, Philippines, France, Ecuador, Nicaragua, Cuba, Outlying-U S (Guam USVI etc), Jamaica, South Korea, China, Germany, Yugoslavia, Canada, Vietnam, Japan, Cambodia, Ireland, Laos, Haiti, Portugal, Taiwan, Holand-Netherlands, Greece, Italy, Poland, Thailand, Trinadad&Tobago, Hungary, Panama, Hong Kong, Scotland, Iran.
Descripción	Hace referencia al lugar de nacimiento de la mamá del individuo

---

Data Mining
3CV11

Integrantes
Aguilar Martínez Oswaldo
Arévalo Andrade Miguel Ángel
Guerrero Espinosa Ximena Mariana

Profesora Ocampo Botello Fabiola

28 de Marzo, 2021

Instituto Politecnico Nacional

Escuela Superior de Computo

Trabajadores en Estados Unidos

Nombre del Atributo	PENATVTY
Tipo de dato	Nominal
Valores	United-States, Mexico, Puerto-Rico, Peru, Canada, South Korea, India, Japan, Haiti, El-Salvador, Dominican-Republic, Portugal, Columbia, England, Thailand, Cuba, Laos, Panama, China, Germany, Vietnam, Italy, Honduras, Outlying-U S (Guam USVI etc), Hungary, Philippines, Poland, Ecuador, Iran, Guatemala, Holand-Netherlands, Taiwan, Nicaragua, France, Jamaica, Scotland, Yugoslavia, Hong Kong, Trinadad&Tobago, Greece, Cambodia, Ireland
Descripción	Hace referencia al lugar de nacimiento del individuo
Nombre del Atributo	**PRCITSHP**
Tipo de dato	Nominal
Valores	Native- Born in the United States, Foreign born- Not a citizen of U S , Native- Born in Puerto Rico or U S Outlying, Native- Born abroad of American Parent(s), Foreign born- U S citizen by naturalization.
Descripción	Hace referencia a la ciudadania del indiciduo
Nombre del Atributo	**WKSWORK**
Tipo de dato	Nominal
Dominio	0-53
Descripción	Hace referencia a la cantidad de semanas que trabaja el indiviuo al año
Nombre del Atributo	**YEAR**
Tipo de dato	Nominal
Valores	94, 95
Descripción	Como fue un estudio de dos años, hace referencia al año en el que se realizo

Data Mining
3CV11

Integrantes
Aguilar Martínez Oswaldo
Arévalo Andrade Miguel Ángel
Guerrero Espinosa Ximena Mariana

Profesora Ocampo Botello Fabiola

28 de Marzo, 2021