Modelo predictivo para estimar al año 2020 la población hispana en la ciudades de Estados Unidos

Leonel Muñoz Cedano
Ingeniero de Sistemas
Universidad Distrital Francisco José de Caldas
Bogotá D.C., Colombia
Email: leoneling@gmail.com

***Resumen*—Los modelos predictivos hoy en día a nivel mundial deben ser parte fundamental en el desarrollo y crecimiento de las organizaciones; sin tener en cuenta el tipo de actividad que realizan, ya que a través de estos modelos se puenden extraer patrones de los datos históricos y transaccionales con el objetivo de identificar riesgos y oportunidades de negocio. En ese sentido, se realizó un análisis minucioso de un Dataset obtenido en Pew Research Center's Hispanic Trends Project, con el fin de plantear un modelo predictivo que permita estimar la población hispana en las ciudades de Estados Unidos.**

***Keywords*—Big Data, Data Mining, Dataset, Modelo predictivo, SMART.**

I. Introducción

El análisis predictivo agrupa una variedad de técnicas estadísticas de modelización, aprendizaje automático y minería de datos que analiza los datos actuales e históricos reales para hacer predicciones acerca del futuro o acontecimientos no conocidos.
Por ello se utilizaron conceptos relacionados con estadistica y "Bigdata" para obtener informacion relevante sobre el conjunto de datos.

II. METODOLOGIA

Para el desarrollo de este documento se planteará las siguientes tareas [1].

II-A. Reconocimiento de la información

- **Identificar el dominio**: Se va a explorar los datos obtenidos de la página Pew Research Center's Hispanic Trends Project [2], en el cual se pueden visualizar la cantidad de ciudadanos hispanos que se han encontrado en varias ciudades de los Estados Unidos y como ha sido el crecimiento de los mismos en diferentes años de la muestra (1990, 2000, 2010, 2011).

- **Variables del DataSet:**
  COUNTY : Ciudad de un estado
  STATE : Estado de USA
  TP : Total de población
  TPNH : Total de población no Hispana
  TPH : Total de población Hispana
  PPH : Porcentaje de población Hispana
  AP : Año de la población

- **Identificar un problema**: El crecimiento poblacional hispano que ha tenido EEUU en los últimos años [2] es muy considerable; y debido al gran impacto socio-económico que esto puede traer en un futuro, se hace necesario poder estimar el crecimiento poblacional hispano en las diferentes ciudades principales de los EEUU. El desarrollo de esta investigación propondrá un modelo predictivo que ayudará a solventar esta problematica.

- **Objetivos SMART**: Los objetivos del proyecto de investigación deben ser orientados con caracteristicas SMART, lo que significa que estos objetivos han de contemplar las siguientes cualidades
  - Specific (Específico)
  - Measurable (Medible)
  - Attainable (Alcanzable)
  - Realistic (Realista)
  - Time-bound (Oportuno)

II-B. Preguntas de investigación

Las preguntas de investigación que se desarrollarán en el proyecto están enmarcadas en los siguientes ámbitos:

- Descriptivas
- Exploratorias
- Inferenciales
- Predictivas

II-C. Analisis exploratorio

- Experimento Aleatorio [3]; Es un proceso de observación mediante el cual se selecciona un elemento de un conjunto de posibles resultados. Un experimento aleatorio es aquel en el que el resultado no se puede predecir con anterioridad a la realización misma del experimento.
- Frecuencia relativa [3]; Sea A un subconjunto del conjunto de posibles resultados de un experimento aleatorio "llamado Ω". Si repetimos N veces el experimento y observamos que en N_A de esas repeticiones se obtuvo un elemento de A, decimos que $f_N(A) = \frac{N_A}{N}$

es la frecuencia relativa del subconjunto A en esas N repeticiones del experimento.

- Medidas de tendencia central[4]
  - Media: la media de las observaciones de un experimento aleatorio $x_1, x_2, ...x_n$ es el promedio aritmétrico de éstas y se denota por;

$$\overline{x} = \sum_{i=1}^{n} \frac{X_i}{n}$$

  - Moda: la moda de un conjunto de observaciones de un experimento aleatorio es elvalor de la observación que ocurre con mayor frecuencia en el conjunto.
  - Mediana: la mediana repreesenta el valor de la variable de posición central en un conjunto de datos ordenados de un experimento aleatorio.
- Varianza [4] : La Varianza de las observaciones $x_1, x_2, ..., x_n$ es en esencia, el promedio del cuadrado de las distancias entre cada observación y la media del conjunto de observaciones. Se denota por:

$$s^2 = \sum_{i=1}^{n} \frac{(x_i - \overline{x})^2}{(n-1)}$$

- Desviación estándar [4]: La desviación estándar es la raiz cuadrada de la varianza y se denota por:

$$s = \sqrt{\sum_{i=1}^{n} \frac{(x_i - \overline{x})^2}{(n-1)}}$$

III. Preguntas de investigación

Las preguntas de investigación juegan un papel importante para el desarrollo de una investigacion de esta indole, ya que a través de ellas se logra una mejor interpretación y definición del problema. Las preguntas de investigación se clasifican en varios tipos de acuerdo al análisis que se desea lograr y en este caso se van a desarrollar las siguientes:

III-A. Preguntas de caracter descriptivo

Cuando se responde las preguntas de carácter descriptivo ya se puede identificar y conocer las características iniciales del conjunto de datos. Las preguntas de caracter descriptivo son:

- ¿Cuál es la Media de ciudadanos en EEUU durante los años 1990, 2000, 2010, 2011?
- ¿Qué ciudad de EEUU tiene la mayor y menor población en el año 1990?
- ¿Qué ciudad de EEUU tiene la mayor y menor población en el año 2000?
- ¿Qué ciudad de EEUU tiene la mayor y menor población en el año 2010?
- ¿Qué ciudad de EEUU tiene la mayor y menor población en el año 2011?
- ¿Cuál es el Promedio de ciudadanos hispanos en ciudades de EEUU en los años 1990, 2000, 2010 y 2011?
- ¿Cuál es la ciudad de EEUU con mayor y menor cantidad de hispanos en el año 1990?
- ¿Qué ciudad de EEUU tiene la mayor y menor cantidad de hispanos en el año 2000?
- ¿Qué ciudad de EEUU tiene la mayor y menor cantidad de hispanos en el año 2010?
- ¿Qué ciudad de EEUU tiene la mayor y menor cantidad de hispanos en el año 2011?

III-B. Preguntas de caracter exploratorio

Las preguntas de caracter exploratorio consisten en la busqueda de patrones o relaciones que soporten una pregunta de investigación.

- ¿Por definir?

III-C. Preguntas de caracter inferencial

Las preguntas de caracter inferencial consisten en el planteamiento de una hipotesis que podria ser resuelta con el analisis respectivo de la informacion

- ¿Por definir?

III-D. Preguntas de caracter predictivo

Las preguntas de caracter predictivo permiten analizar el comportamiento de la informacion a traves del tiempo, con el objetivo de descubrir, proyectar, o realizar hipotesis sobre estados futuros.

- ¿Cuál será el porcentaje de crecimiento poblacional Hispana en las ciudades de EEUU al año 2020?

IV. Análisis exploratorio

El análisis exploratorio es una proceso que se realiza previamente a la aplicacación de cualquier técnica estadística a un conjunto de datos, la cual tiene como el objetivo identificar el comportamiento de los datos a través del análisis de gráficos y estadística básica permitirá explorar las distribución de los datos e identificar características tales como: valores atípicos o outliers, concentraciones de valores, saltos o discontinuidades, forma de la distribución, etc.

IV-A. Analisis inicial

Lo primero que se va analizar es el comportamiento que tienen los datos en los diferentes años en las variables TP, NHP y HP; en se obtiene los siguientes resultados:

Tabla I: Total de la población de EEUU

Statistic	N	Mean	St. Dev.	Min	Max
TotalCiudadanos	12,544	91,700.930	297,470.800	67	9,889,056
TotalNoHispanos	12,544	78,932.380	214,518.700	60	5,511,922
TotalHispanos	12,544	12,768.550	102,278.800	0	4,760,974

Recuerde que el anális principal de esta investigación se enfoca en estimar el crecimiento de la población hispana en algunas ciudades de EEUU, y revisando los resultados anteriores de la variable TPH se evidencia una media muy baja la cual es un valor muy significativo teniendo encuenta los valores de máximo y mínimo de la misma. Por tal razón se ha hace necesario a través de un gráfico poder visaualizar mejor los datos de la variable TPH.

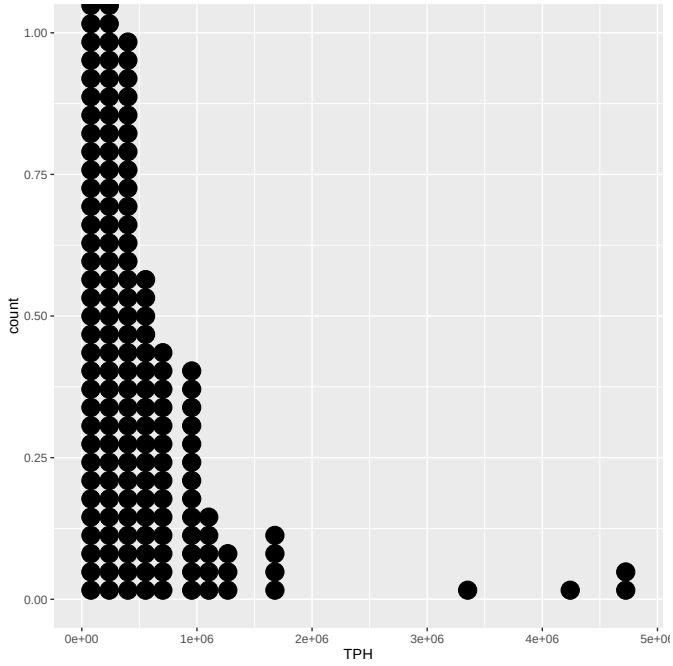


Figura 1: Dotplot de la variable TPH

Posteriormente se procede a realizar un análisis del conjunto de datos más al detalle, de acuerdo con la información que se tiene de los diferentes años de la muestra.

IV-B. Analizando población hispana en el año 1990

Tabla II: Total de la población de EEUU en el año 1990

Statistic	N	Mean	St. Dev.	Min	Max
TotalCiudadanos	3,136	79,300.610	264,006.100	107	8,863,164
TotalNoHispanos	3,136	72,172.490	208,127.900	93	5,511,922
TotalHispanos	3,136	7,128.126	71,748.130	0	3,351,242

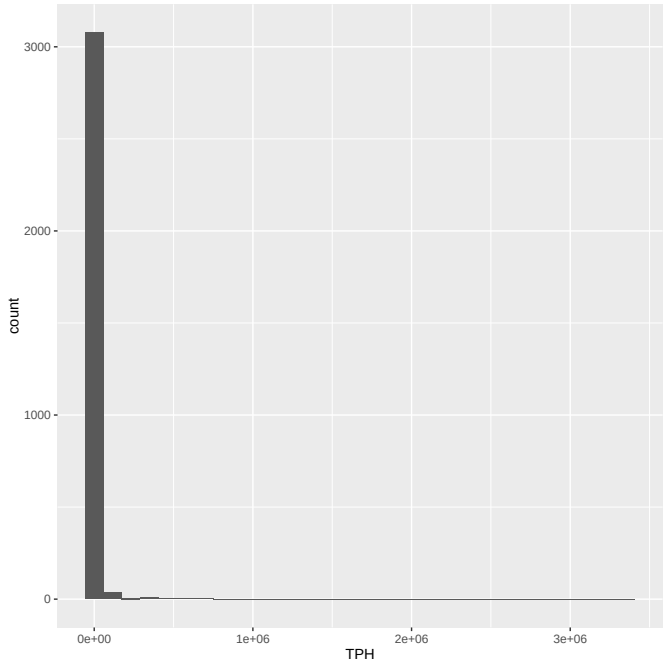


Figura 2: Población Hispana en el año 1990

IV-C. Analizando población hispana en el año 2000

Tabla III: Total de la población de EEUU en el año 2000

Statistic	N	Mean	St. Dev.	Min	Max
TotalCiudadanos	3,136	89,735.040	292,674.700	67	9,519,338
TotalNoHispanos	3,136	78,476.900	214,891.300	60	5,277,125
TotalHispanos	3,136	11,258.140	96,312.440	1	4,242,213

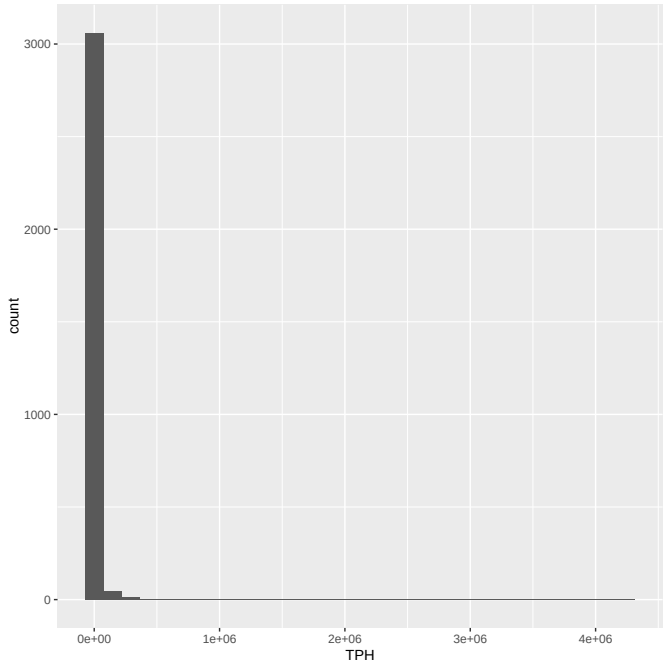


Figura 3: Población Hispana en el año 2000

IV-D. Analizando población hispana en el año 2010

Tabla IV: Total de la población de EEUU en el año 2010

Statistic	N	Mean	St. Dev.	Min	Max
TotalCiudadanos	3,136	98,430.440	313,221.000	82	9,818,605
TotalNoHispanos	3,136	82,336.350	216,856.000	64	5,130,716
TotalHispanos	3,136	16,094.090	115,731.900	0	4,687,889

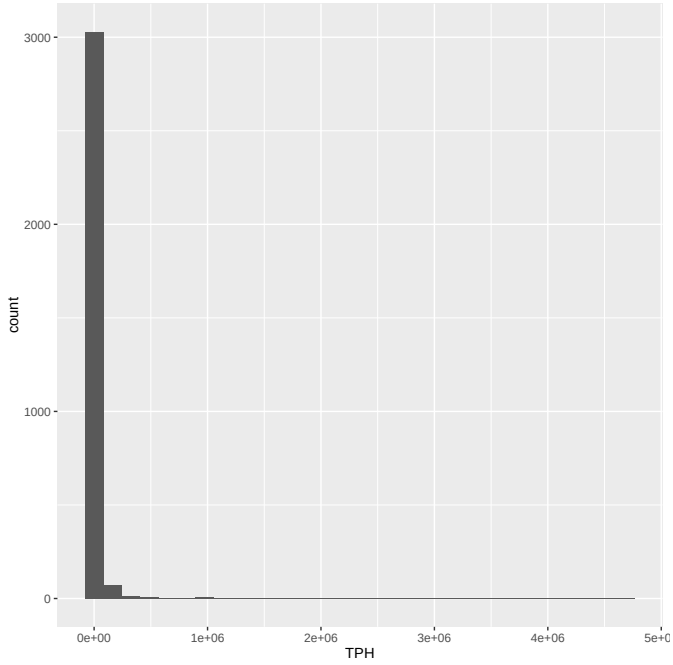


Figura 4: Población Hispana en el año 2010

IV-E. Analizando población hispana en el año 2011

Tabla V: Total de la población de EEUU en el año 2011

Statistic	N	Mean	St. Dev.	Min	Max
TotalCiudadanos	3,136	99,337.630	316,723.400	90	9,889,056
TotalNoHispanos	3,136	82,743.800	217,998.000	76	5,128,082
TotalHispanos	3,136	16,593.830	118,221.400	0	4,760,974

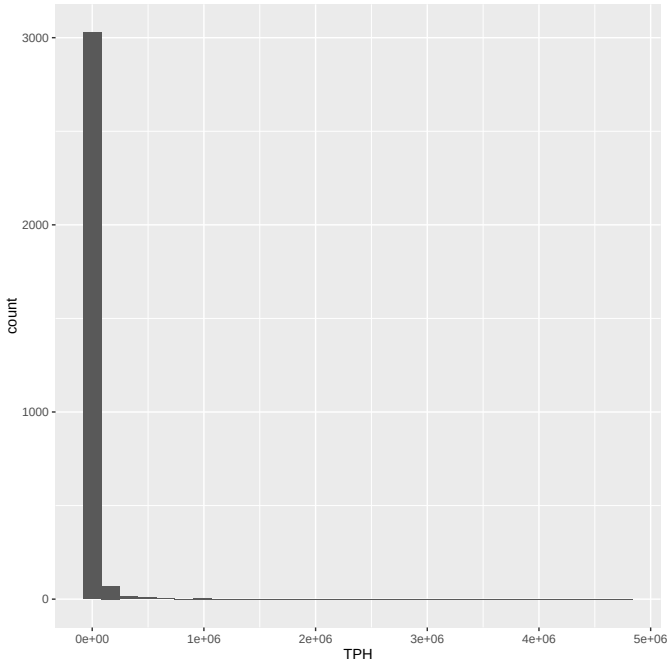


Figura 5: Población Hispana en el año 2011

IV-F. Percentiles del conjunto de datos

El percentil de orden k es el cuantil de orden $k100$. El recorrido intercuantil refleja la variabilidad de las observaciones comprendidas entre los percentiles 25 y 75 en el conjunto de datos. En esta sesión se obtienen los percentiles del 25 %, 50 % y 75 % de las variables Total Población (TP) y Total Población Hispana (TPH) en los diferentes años de la muestra.

	PercentilesTP	PercentilesTPH
25 %	10360.00	67.00
50 %	22224.00	208.00
75 %	54771.50	1159.00

Tabla VI: Percentiles de TP y TPH en el año 1990

	PercentilesTP	PercentilesTPH
25 %	11264.25	155.00
50 %	24658.00	493.00
75 %	61844.25	2411.50

Tabla VII: Percentiles de TP y TPH en el año 2000

	PercentilesTP	PercentilesTPH
25 %	11154.75	262.75
50 %	25901.50	892.00
75 %	67012.50	4226.25

Tabla VIII: Percentiles de TP y TPH en el año 2010

	PercentilesTP	PercentilesTPH
25 %	11145.00	284.00
50 %	25896.00	926.00
75 %	67398.75	4417.75

Tabla IX: Percentiles de TP y TPH en el año 2011

V. Solución de Preguntas

Ahora se procederá a responder todas las preguntas que se plantearón al inicio de la investigación.

V-A. Caracter descriptivo

A continuación se enumeran los promedios de la variable total ploblación en cada uno de los años del DataSet:

- promedio del año 1990: [1] 79300.61
- promedio del año 2000: [1] 89735.04
- promedio del año 2010: [1] 98430.44
- promedio del año 2011: [1] 99337.63

	Ciudad	Estado	Poblacion
1	Los Angeles	California	8863164

Tabla X: Ciudad con más población en el año 1990

	Ciudad	Estado	Poblacion
1	Loving	Texas	107

Tabla XI: Ciudad con menos población en el año 1990

	Ciudad	Estado	Poblacion
1	Los Angeles	California	8863164

Tabla XII: Ciudad con más población en el año 2000

	Ciudad	Estado	Poblacion
1	Loving	Texas	67

Tabla XIII: Ciudad con menos población en el año 2000

A continuación se enumera las medias de la variables Total Población (TP), Total Población No Hispana (TPNH) y Total Ploblación Hispana (TPH) en cada uno de los años del DataSet:

El análisis continua con los siguientes resultados.

V-B. Matriz de correlación

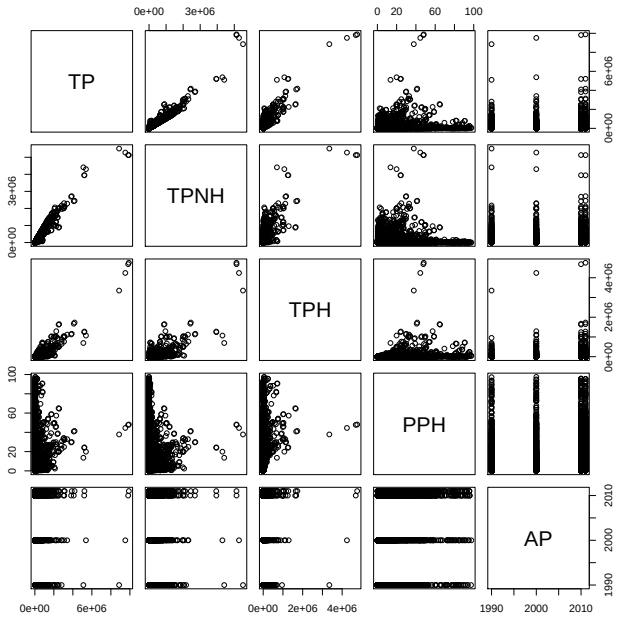
A continuación se visualiza la matriz de correlación que existe entre la variables cuantitativas del conjunto de datos:

	Ciudad	Estado	Poblacion
1	Los Angeles	California	9818605

Tabla XIV: Ciudad con más población en el año 2010

	Ciudad	Estado	Poblacion
1	Loving	Texas	82

Tabla XV: Ciudad con menos pobl. en 2010



	Ciudad	Estado	Poblacion
1	Los Angeles	California	9889056

Tabla XVI: Ciudad con más población en el año 2011

	Ciudad	Estado	Poblacion
1	Kalawao	Hawaii	90

Tabla XVII: Ciudad con menos población en el año 2011

REFERENCIAS

[1] S. Mohanty, M. Jagadeesh and H. Srivatsa, Big Data Imperatives: Enterprise Big Data Warehouse, BI Implementations and Analytics, Published Apress, Isbn: 978-1-4302-4872-9, New York, 2013.

[2] pewhispanic.org, Pew Research Center's Hispanic Trends Project, U.S. Hispanic Population by County, 1980-2011. Disponible en: http://www.pewhispanic.org/2013/08/29/u-s-hispanic-population-by-county-1980-2011/, 2013.

[3] G. C. Canavos, Probabilidad y estadística: Aplicaciones y métodos, Virginia Commonwealth University, Published McGRAW HILL, 1988.

[4] Alzate Marco, 250 Conceptos de Probabilidad, variables aleatorias y procesos estocásticos en redes de comunicaciones, Universidad Distrital Fransisco José de Caldas, pag 15-123, 2005.

	Año 1990	Año 2000	Año 2010	Año 2011
Media TP	79300.61	89735.04	98430.44	99337.63
Media TPNH	72172.49	78476.90	82336.35	82743.80
Media TPH	7128.13	11258.14	16094.09	16593.83

Tabla XVIII: Valores de las medias en TP, TPHN y TPH

	Ciudad	Estado	PoblacionH
1	Los Angeles	California	3351242

Tabla XIX: Ciudad con mayor TPH en el año 1990

	Ciudad	Estado	PoblacionH
1	Garfield	Montana	0
2	Petroleum	Montana	0
3	Arthur	Nebraska	0
4	Blaine	Nebraska	0
5	McPherson	Nebraska	0
6	Wheeler	Nebraska	0
7	Billings	North Dakota	0
8	Campbell	South Dakota	0
9	McPherson	South Dakota	0

Tabla XX: Ciudad con menor TPH en el año 1990

	Ciudad	Estado	PoblacionH
1	Los Angeles	California	4242213

Tabla XXI: Ciudad con mayor TPH en el año 2000

	Ciudad	Estado	PoblacionH
1	Blaine	Nebraska	1
2	Slope	North Dakota	1

Tabla XXII: Ciudad con menor TPH en el año 2000

	Ciudad	Estado	PoblacionH
1	Los Angeles	California	4687889

Tabla XXIII: Ciudad con mayor TPH en el año 2010

	Ciudad	Estado	PoblacionH
1	Blaine	Nebraska	0

Tabla XXIV: Ciudad con menor TPH en el año 2010

	Ciudad	Estado	PoblacionH
1	Los Angeles	California	4760974

Tabla XXV: Ciudad con mayor TPH en el año 2011

	Ciudad	Estado	PoblacionH
1	Blaine	Nebraska	0

Tabla XXVI: Ciudad con menor TPH en el año 2011

	TP	TPNH	TPH	PPH	AP
TP	1.00	0.97	0.87	0.17	0.03
TPNH	0.97	1.00	0.73	0.12	0.02
TPH	0.87	0.73	1.00	0.25	0.04
PPH	0.17	0.12	0.25	1.00	0.13
AP	0.03	0.02	0.04	0.13	1.00

Tabla XXVII: Matrix de correlación