

Modelo predictivo para estimar el crecimiento poblacional hispano al año 2020 en diferentes ciudades de EEUU

Leonel Muñoz Cedano

Ingeniero de Sistemas

Universidad Distrital Francisco José de Caldas

Bogotá D.C., Colombia

Email: leoneling@gmail.com

Resumen—En el desarrollo de esta investigación se utilizan conceptos muy relacionados con la estadística, algoritmos de predicción y maquina de aprendizaje; los cuales están inmersos en la metodología de análisis de BigData[1] que se va disponer, con la que se desarrolla un análisis exhaustivo de los datos y se obtendrán mejores resultados de predicción del conjunto de datos (DataSet). Este DataSet fue extraído desde la página oficial de la compañía Pew Research Center's Hispanic Trends Project[2] quien se encarga de informar de manera independiente al público sobre hechos, actitudes y tendencias de América y el mundo.

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

I. INTRODUCCIÓN

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 pueden extraer patrones de los datos históricos y transaccionales con el objetivo de identificar riesgos y oportunidades de negocio. El análisis predictivo agrupa una variedad de técnicas estadísticas de modelización, aprendizaje automático y minería de datos; las cuales a través de los datos históricos y actuales permiten realizar predicciones acerca del futuro o acontecimientos no conocidos. Teniendo en cuenta lo anterior, y tomando como base para el desarrollo del estudio el DataSet extraído de Pew Research Center y la aplicación de la metodología de análisis de BigData, se analizará el comportamiento poblacional hispano en los diferentes años (1990, 2000, 2010 y 2011), con el objetivo de proponer un modelo predictivo que permita estimar el crecimiento poblacional hispano que tendrán diferentes ciudades de EEUU en el año 2020.

II. METODOLOGIA

Para el desarrollo de este documento; se utilizará algunas tareas de la metodología de análisis de BigData[1].

II-A. Reconocimiento de la información

Identificar el dominio: Se explorará el DataSet obtenido de Pew Research Center's Hispanic Trends Project[2], en el cual se encuentran 12544 observaciones y diferentes variables de información, entre ellas la cantidad de ciudadanos hispanos,

no hispanos y total de población que se ha encontrado en algunas ciudades de los Estados Unidos y como ha sido el comportamiento de los datos en los diferentes años de la muestra (1990, 2000, 2010, 2011).

Variables del DataSet: Las variables que se identificaron en el conjunto de datos son las siguientes:

- **COUNTY:** Ciudad de un estado.
- **STATE:** Estado de EEUU.
- **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 acarrear en un futuro, se hace necesario poder estimar el crecimiento poblacional hispano en las diferentes ciudades principales de los EEUU. Con el desarrollo de esta investigación se propondrá un modelo predictivo que ayudará a solventar esta problemática.

Características de los objetivos SMART: Los objetivos del proyecto de investigación deben ser orientados con las características SMART[3], lo que significa que estos objetivos han de contemplar las siguientes cualidades:

- **Specific (Específico):** Dirigirse a un área específica de mejora.
- **Measurable (Medible):** Cuantificar o al menos sugerir un indicador de progreso.
- **Attainable (Alcanzable):** Identificar qué tipo de habilidades, actitudes u otro tipo de recursos necesitamos para cumplirlas.
- **Realistic (Realista):** Los resultados esperados son acordes con los recursos disponibles.

- Time-bound (Oportuno): Especificar un marco de tiempo para lograr el resultado.

II-B. Tipos de preguntas de una investigación

Las preguntas de investigación[4] que se desarrollarán en el proyecto están enmarcadas en los siguientes ámbitos:

- Descriptivas: Una pregunta descriptiva es la que busca resumir una característica de un conjunto de datos.
- Exploratorias: Las preguntas de carácter exploratorio consisten en la búsqueda de patrones o relaciones que soporten una pregunta de investigación.
- Inferenciales: Una pregunta inferencial consiste en el planteamiento de una hipótesis que podría ser resuelta con el análisis respectivo de la información.
- Predictivas: Las preguntas de carácter predictivo permiten analizar el comportamiento de la información a través del tiempo, con el objetivo de descubrir, proyectar, o realizar hipótesis sobre estados futuros.

II-C. Marco teórico del análisis exploratorio

El análisis exploratorio de los datos son básicamente aquellas funciones estadísticas que permiten visualizar el comportamiento de las observaciones en el DataSet en un proceso de investigación. Las funciones a utilizar son las siguientes:

- Experimento Aleatorio[5]; 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[5]; 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[6];
 - Media: la media de las observaciones de un experimento aleatorio x_1, x_2, \dots, x_n es el promedio aritmético de éstas y se denota por;
$$\bar{x} = \sum_{i=1}^n \frac{X_i}{n}$$
 - Moda: la moda de un conjunto de observaciones de un experimento aleatorio es el valor de la observación que ocurre con mayor frecuencia en el conjunto.
 - Mediana: la mediana representa el valor de la variable de posición central en un conjunto de datos ordenados de un experimento aleatorio.
- Varianza[6]; La Varianza de las observaciones x_1, x_2, \dots, 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:

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

- Desviación estándar[6]; La desviación estándar es la raíz cuadrada de la varianza y se denota por:

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

- Cuartiles[6]; Los cuartiles son dadas una serie de valores x_1, x_2, \dots, x_n ordenados en forma creciente, podemos pensar que su cálculo podría efectuarse:
 - Primer cuartil (Q1) como la mediana de la primera mitad de valores.
 - Segundo cuartil (Q2) como la propia mediana de la serie de valores.
 - Tercer cuartil (Q3) como la mediana de la segunda mitad de valores.

III. PREGUNTAS DE INVESTIGACIÓN

Las preguntas de investigación juegan un papel importante para el desarrollo de una investigación de esta índole, 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 carácter 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 carácter 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 carácter exploratorio

Las preguntas de carácter exploratorio en la investigación son las siguientes:

- ¿El total de población hispana es dependiente del total de población en una ciudad?
- ¿El total de población hispana es dependiente del total de población no hispana en una ciudad?
- ¿El total de población hispana es independiente del total de población en una ciudad?
- ¿El total de población hispana es independiente del total de población no hispana en una ciudad?
- ¿El total de población no hispana es dependiente del total de población en una ciudad?

III-C. Preguntas de carácter inferencial

Las preguntas de carácter inferencial en la investigación son las siguientes:

- ¿El total de población hispana de una ciudad se ve afectado por el total de ciudadanos?
- ¿El total de población hispana de una ciudad se ve afectado por el total de ciudadanos no hispanos?

III-D. Preguntas de carácter predictivo

Las preguntas de carácter predictivo en la investigación son las siguientes:

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

IV. ANÁLISIS EXPLORATORIO DEL DATASET

El análisis exploratorio tiene como objetivo identificar el comportamiento de los datos a través del análisis estadístico. En este análisis se pueden visualizar en gran variedad de tablas y gráficos que permitirán explorar las distribución de los datos e identificar comportamientos y/o características de las observaciones tales como; valores atípicos o outliers, concentraciones de valores, saltos o discontinuidades, forma de la distribución, etc.

IV-A. Análisis inicial

Lo primero que se va analizar es el comportamiento que tienen los datos en los diferentes años en las variables Total Población (TP), Total de Población no Hispana (TPNH) y Total de Población Hispana (TPH); donde se obtiene los siguientes resultados:

Tabla I: Total de la población de EEUU

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

Es importante recordar que el análisis principal de esta investigación se enfoca en estimar el crecimiento de la población

hispana en las ciudades de EEUU, y analizando los resultados obtenidos de la variable TPH, se evidencia una media muy baja con respecto a los valores de máximo y mínimo de la misma. Por tal razón se hace necesario visualizar el comportamiento de los datos de la variable TPH a través del siguiente gráfico.

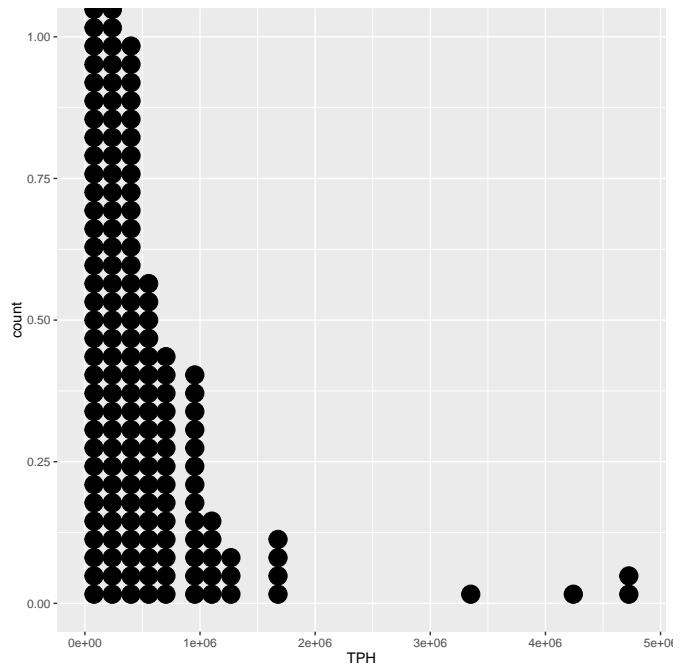


Figura 1: Dotplot de la variable TPH

En la gráfica anterior se observa la existencia de datos atípicos (outliers); los cuales distorsionan el análisis de la información y los resultados, por tal razón estas observaciones atípicas no serán tenidas en cuenta en el desarrollo del modelo predictivo.

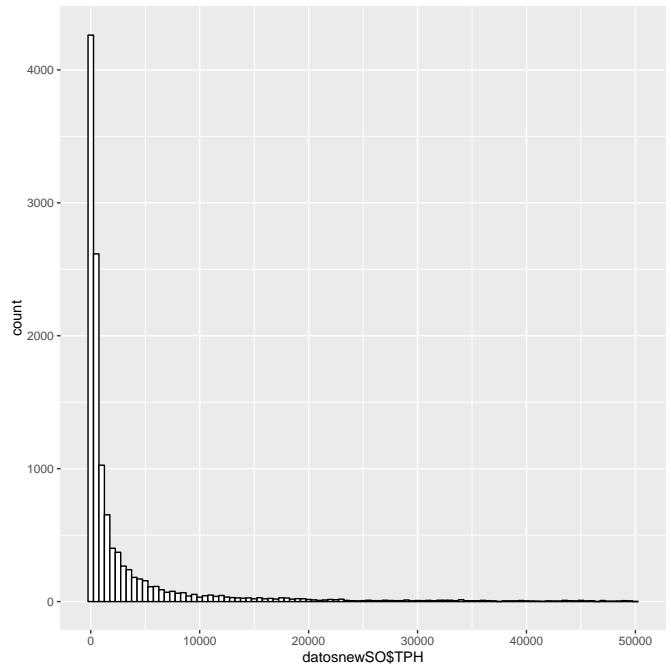


Figura 2: Histograma de la variable TPH sin outliers

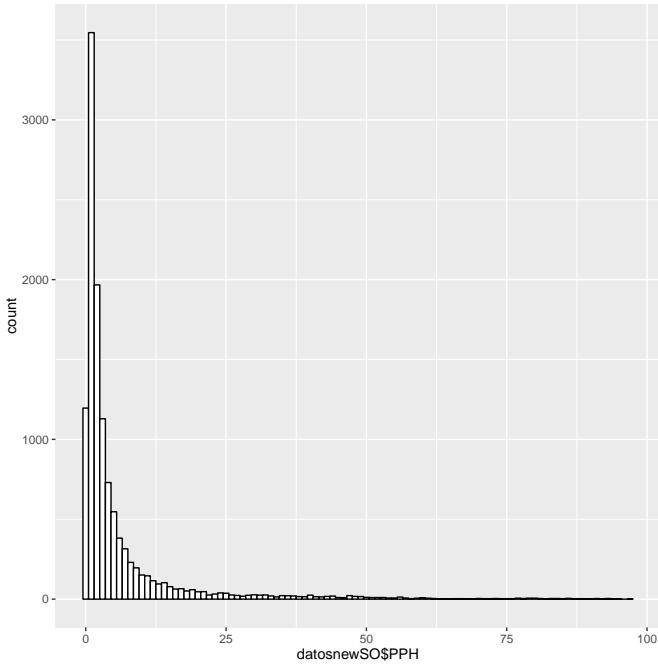


Figura 3: Histograma de la variable PPH sin outliers

Los histogramas obtenidos de las variables TPH y PPH expresan que las observaciones tienen un comportamiento exponencial, pero es muy temprano determinar cual será el tipo de modelo y la cantidad de variables a utilizar en el desarrollo de la predicción. A medida que se avance con el análisis de los datos y los resultados obtenidos en cada proceso se encargarán de determinar cual va ser el mejor modelo a aplicar en la investigación.

IV-B. 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), Total Población Hispana (TPH) y Total Población No Hispana (TPNH).

	Percentiles TP	Percentiles TPH	Percentiles TPNH
25 %	10514.00	146.00	9836.50
50 %	23257.00	514.00	21826.00
75 %	54200.50	2368.00	51145.00

Tabla II: Percentiles de TP, TPH y TPNH

IV-C. Matriz de coeficientes de correlación

La matriz de coeficientes de correlación permite estudiar la relación o comportamiento que existe entre las variables del conjunto de datos. A continuación se muestra el resultado obtenido de la correlación:

	TP	TPNH	TPH	PPH	AP
TP	1.00	1.00	0.65	-0.03	-0.01
TPNH	1.00	1.00	0.61	-0.06	-0.02
TPH	0.65	0.61	1.00	0.39	0.11
PPH	-0.03	-0.06	0.39	1.00	0.12
AP	-0.01	-0.02	0.11	0.12	1.00

Tabla III: Matriz de coeficientes de correlación

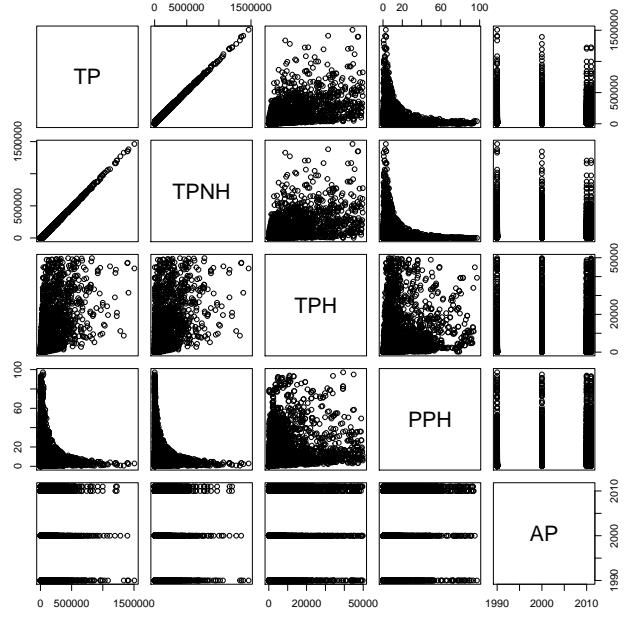


Figura 4: Nube de puntos de correlación

El resultado de la matriz y nube de puntos de correlación, expresan la alta correlación de la variable TPH con respecto a las variables TP y TPNH. Por otro lado, en la nube de puntos de correlación de la variable PPH, se evidencia un comportamiento exponencial con las variables TP y TPNH.

V. SOLUCIÓN DE PREGUNTAS

Han sido varios los resultados obtenidos hasta el momento, pero es necesario comenzar a dar respuesta a las diferentes preguntas que fueron planteadas al inicio de la investigación.

V-A. Preguntas de carácter descriptivo

- ¿Cuál es la Media de ciudadanos en EEUU durante los años 1990, 2000, 2010, 2011?

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

	Año 1990	Año 2000	Año 2010	Año 2011
Media TP	56941.88	58282.22	54864.33	55088.70
Media TPNH	54998.83	55497.86	51221.69	51314.33
Media TPH	1943.06	2784.35	3642.64	3774.37

Tabla IV: Valores de las medias en TP, TPNH y TPH

- ¿Qué ciudad de EEUU tiene la mayor y menor población en el año 1990?

	Ciudad	Estado	Poblacion
1	King	Washington	1507319

Tabla V: Ciudad con más población en el año 1990

Ciudad	Estado	Poblacion
1 Loving	Texas	107

Tabla VI: Ciudad con menos población en el año 1990

- ¿Qué ciudad de EEUU tiene la mayor y menor población en el año 2000?

Ciudad	Estado	Poblacion
1 King	Washington	1507319

Tabla VII: Ciudad con más población en el año 2000

Ciudad	Estado	Poblacion
1 Loving	Texas	67

Tabla VIII: Ciudad con menos población en el año 2000

- ¿Qué ciudad de EEUU tiene la mayor y menor población en el año 2010?

Ciudad de EEUU con la mayor población en el año 2010:

Ciudad	Estado	Poblacion
1 Allegheny	Pennsylvania	1223348

Tabla IX: Ciudad con más población en el año 2010

Ciudad de EEUU con la 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?

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

Ciudad	Estado	Poblacion
1 Loving	Texas	82

Tabla X: Ciudad con menos población en el año 2010

	TP	TPNH	TPH
1	56941.88	54998.83	1943.06
2	58282.22	55497.86	2784.35
3	54864.33	51221.69	3642.64
4	55088.70	51314.33	3774.37

Tabla XI: Valores de las medias en TP, TPNH y TPH

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. T. Doran, There's a S.M.A.R.T. Way to Write Management's Goals and Objectives, Management Review, Vol. 70, Issue 11, pp. 35-36, 1981.
- [4] R. D. Peng and E. Matsui, The Art of Data Science: A guide for anyone who works with data, Published Leanpub, Disponible en:<http://leanpub.com/artofdatascience>, 2015.
- [5] Alzate Marco, 250 Conceptos de Probabilidad, variables aleatorias y procesos estocásticos en redes de comunicaciones, Universidad Distrital Francisco José de Caldas, pag 15-123, 2005.
- [6] G. C. Canavos, Probabilidad y estadística: Aplicaciones y métodos, Virginia Commonwealth University, Published McGRAW HILL, 1988.