Parte 1

Se utiliza el dataset de nombres argentinos históricos del portal nacional de datos. Consta de 3 campos: Nombre, Cantidad y Año.

Toda la implementación fue realizada en Python 3.6.

A) ¿Cuántos nombres distintos hay en todo el dataset, considerando cada nombre por separado? Es decir, un nombre compuesto como María Inés cuenta una vez para María y otra para Inés.

Tomando el contenido del campo *Nombre* del dataset original, en primer lugar, se realizaron algunas transformaciones básicas para el análisis posterior, como ser:

- eliminación de nombres que tuvieran algún número (se consideran errores de digitalización),
- eliminaron de espacios en blanco y
- conversión a minúsculas.

No se realizó ninguna otra modificación en la muestra, por lo que por ejemplo los nombres "José", "Josè" y "Jose", serán considerados distintos para el conteo final.

Luego se generó un nuevo dataset, conteniendo los nombres simples (tomando el espacio en blanco como delimitador) que formaban parte de cada campo nombre original. Nuevamente se realizó la limpieza de algunos caracteres que no se consideraron como parte de los nombres (por ejemplo, comas, puntos, guiones o paréntesis). De éstos nombres simples se eliminaron los que representaban conectores ("del", "de", "los", "lo", "la" y "las") y los que tuvieran un solo carácter.

Descartando los duplicados, se obtienen los nombres simples presentes en el dataset, dando un total de 201.007. A modo de ejemplo se listan algunos:

Nombres
Avaricia
Barbara
Camila
Daniela
Ester

Se puede consultar el listado completo en el archivo nombres_distintos.csv[1], al igual que el código final en nombres_1a.py.

[1] Todos los resultados detallados se encuentran en formato csv en la carpeta *salidas*

B) Construí un clasificador que separe los nombres en masculinos y femeninos ¿Cuán bien funciona? Elegí una métrica apropiada para evaluar el desempeño del clasificador y reportala.

1. Preparación de datos de entrenamiento

Para realizar el entrenamiento del clasificador se utilizó como fuente los nombres permitidos por el registro civil, disponibles en el portal de Datos de la Ciudad de Buenos Aires. Los datos representan 9817 registros, donde se indica nombre y género. Este último puede ser femenino (f), masculino (m) o "ambos" (a), en el caso que se pueda usar indistintamente.

Los nombres permitidos presentan la siguiente distribución de frecuencias:

Género	Cantidad
Masculino	4980
Femenino	4354
Ambos	483

Además, se crearon campos adicionales, utilizando la terminación del nombre. Consisten en 4 variables nuevas, conteniendo de uno a cuatro caracteres finales del nombre. Se detalla, a continuación, un ejemplo de registro con el que se entrena el clasificador:

Variable	Contenido
nombre	Daniela
terminación	a
terminación2	la
terminación3	ela
terminación4	iela
genero	f

2. Elección del clasificador

Previa categorización de todas las variables, se dividió la muestra en 70% para entrenar y 30% para testeo, al azar. Las distribuciones de frecuentas de ambos conjuntos fue:

Entrenamiento	
Género	**Cantidad**
Masculino	3515
Femenino	3068
Ambos	344

Test	
Género	**Cantidad**
Masculino	1465
Femenino	1286
Ambos	139

Con el objetivo de determinar el algoritmo a utilizar, se probaron varios implementados en la librería Sklearn, registrando el accuracy obtenido en cada uno:

Clasificador	Accuracy entrenamiento	Accuracy test
Logistic Regression	0,768	0,776
Nearest Neighbors	0,692	0,537
Linear SVM	1,000	0,509
Gradient Boosting	0,835	**0,818**
Decision Tree	1,000	0,756
Random Forest	0,996	0,809
Neural Net	0,463	0,469
Naive Bayes	0,773	0,782

Se decidió utilizar "Gradient Boosting Classifier", dado que obtuvo en testing un 82% de eficiencia. La siguiente matriz de confusión muestra el resultado de la clasificación en los 2890 registros seleccionados para testear:

	Predicción			
	Ambos	**Femeninos**	**Masculinos**	
Ambos	1	36	102	**139**
Femeninos	5	1060	221	**1286**
Masculinos	3	158	1304	**1465**
	9	**1254**	**1627**	2890

3. Preparación de datos originales

Antes de clasificar se llevaron a cabo varias operaciones para la limpieza y adecuación de los datos originales. Se eliminaron los registros con dígitos dentro del nombre, se reemplazaron algunos caracteres no alfabéticos (como comas, puntos, guiones, etc.), sustituyeron variables acentuadas, y eliminaron palabras consideradas conectores: "del", "de", "lo", "los", "la" y "las" o formadas por un único carácter.

A continuación, se crearon las nuevas variables con las terminaciones detalladas anteriormente y categorizaron los datos igual que en lo datos con género.

4. Clasificador final

Como se busca clasificar los nombres completos, formados en muchos casos por más de un nombre simple, se decidió utilizar el clasificador definido en los tres primeros nombres de cada persona por separado y luego definir el género final según cierta lógica.

Se agrega además una modificación a la clasificación resultante de cada nombre. Definiendo un umbral de 0.75, si la clase resultante no llegara a superarlo y existe un nombre adicional, se clasifica con genero nulo, para evaluarlo posteriormente según la clasificación obtenida con el siguiente nombre.

El género final se define según las siguientes reglas:

Género primer nombre	Género segundo nombre	Género tercer nombre	Género final
Definido	-	-	Predicción primer nombre
Indefinido	Definido	-	Predicción segundo nombre
Ambiguo	Indefinido	Definido o ambiguo	Predicción tercer nombre
Ambiguo	Indefinido	Nulo	Predicción primer nombre
Nulo	Indefinido	Definido o ambiguo	Predicción tercer nombre
Nulo	Ambiguo	Nulo	Predicción segundo nombre

Donde definido es “femenino” o “masculino”, indefinido corresponde a “ambos” o nulo, y ambiguo es cuando la predicción define como categoría el valor “ambos”.

Hay que tener en cuenta que, según esta forma de obtener la clase final, se pueden dar casos en los que la clase final tenga una probabilidad inferior a la de alguno de los nombres parciales. Por ejemplo, cuando el tercer nombre define el género, porque no superaron los dos anteriores el umbral.

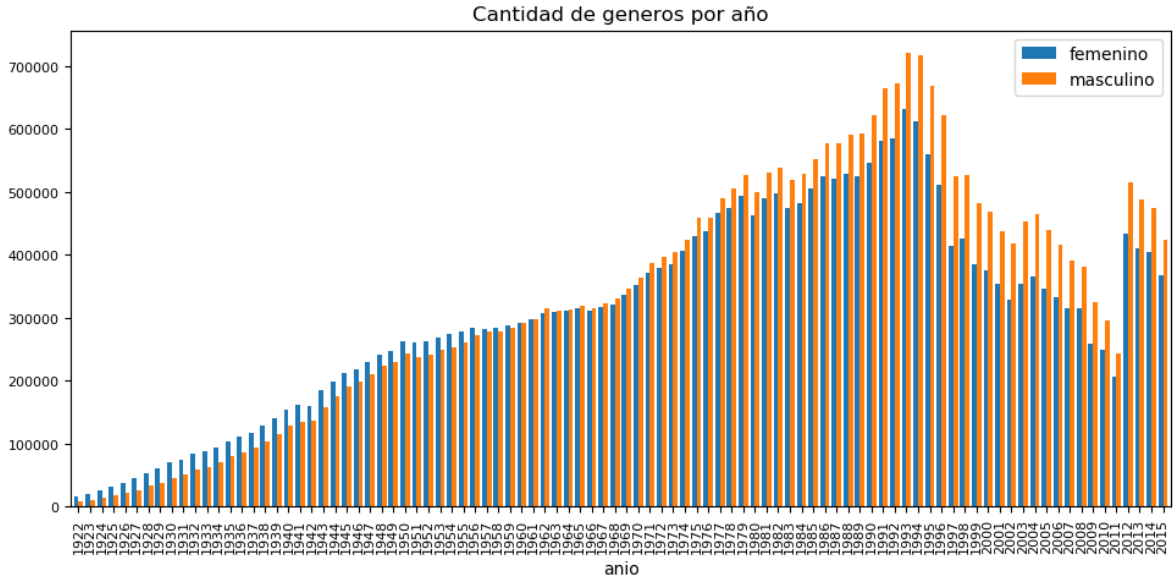
La clasificación final muestra la siguiente distribución de frecuencias:

Género	Cantidad
Masculino	1603960
Femenino	1832525
Ambos	2718

El código completo se encuentra en el archivo nombres_1b.py y la clasificación final completa en original_predict.csv.

C) Usa el clasificador que construiste en el punto anterior y grafica la cantidad estimada de nombres de varón y nombres de mujer para cada año.

Partiendo de la clasificación obtenida en el punto anterior, se generó un dataset sumarizando por año y genero la cantidad de nombres presentes en el archivo histórico.



Al graficarlo se puede observar como durante los primeros años hubo un predominio de nombres femeninos, que luego se revierte en las últimas décadas.

Referencias

Dataset	Nombres argentinos históricos.
Descripción	Contiene nombre y cantidad por año desde 1922 al 2015.
Fuente	Portal nacional de datos
URL	http://datos.gob.ar/dataset/nombres-personas-fisicas
Dataset	Nombres permitidos
Descripción	Nombres permitidos por el registro civil con género: masculino, femenino o ambos.
Fuente	Buenos Aires Data
URL	https://data.buenosaires.gob.ar/dataset/nombres-permitidos

Resultados

Archivo	**Descripción**
Nombres_distintos.csv	Listado de los nombres simples presentes en todo el dataset original (Parte 1a)
test_predict.csv	Predicción del conjunto de test usado para entrenar el dataset (Parte 1b)
original_predict.csv	Predicción del dataset original (Parte 1c)
predict_por_anio_genero.csv	Cantidad de nombres de genero femenino y masculino por año.