Hito 3: Microdatos del Censo Nacional de Población y Vivienda 2017

Kianush Atighi-Moghaddam, Arturo Avendaño, Alonso Rojas, and Diego Vera.

Dpto. Computación e Informática, UFRO.

Temuco, Chile.

***Resumen*—Al trabajar con datos estadísticos de Censos se puede observar fallas humanas realizadas por parte de los censistas, lo que se traduce en datos con valores indeterminados o directamente en blanco. En el siguiente reporte se propondrán soluciones a 2 problemáticas encontradas dentro del Censo Nacional de Chile del año 2017 haciendo uso de estrategias de Análisis de Datos como lo son la EDA, visualización de los Datos, algoritmos de clasificación, etc.**

***Keywords*—EDA, Dataset, Árbol de Decisión, KNN, Visualización de los Datos**

I. Introducción

Los censos son eventos en donde la recolección y compilación de los datos son tareas muy importantes, esto debido a que con estas cifras ocurren los estudios demográficos del país, datos los cuales deben estar presentes y no siempre lo están en estos casos debido a fallas humanas al momento de Censar. Siendo lo anterior el foco principal de estudio.

II. Problemáticas

II-A. Problema 1

El primer problema consiste en predecir la clase *NOM_CAT_ENT* de una manzana, la cual corresponde a su categoría de entidad (ciudad, pueblo, aldea, caserío, etc.). Lo anterior se eligió luego de hacer el análisis exploratorio y observar que muchas filas tienen la columna *NOM_CAT_ENT* con valor indeterminado. Este problema se considera como clasificación multiclase.

II-B. Problema 2

El segundo problema consiste en predecir el atributo *NOM_AREA* que corresponde al tipo de área de una manzana, este atributo indica si la manzana se encuentra en un área rural o urbana. Este problema se identifica como clasificación binaria.

III. Metodología de resolución

Para la resolución de las dos problemáticas planteadas, cada una será abordada de maneras similares pero últimamente diferente.

III-A. Propuesta experimental problema 1:

La primera problemática, como se mencionó dentro de la Introducción de manera superficial, consiste en la predicción mediante modelos de clasificación de la columna *NOM_CAT_ENT* de una manzana dentro del set de datos, tratándose de una clasificación de tipo multiclase.
Por lo tanto, el método de resolución utilizado fue la implementación de un modelo de árbol de decisión, el cual considera los atributos que fueron considerados significativos para la resolución del problema, en este caso, se consideraron atributos tales como el origen del agua que utiliza la manzana, tipos de vivienda y el estado de las construcciones las cuales habitan en esa manzana. En base a lo anterior, se creó un modelo de predicción tipo árbol de decisiones.
Para evitar el sesgo y generar una propuesta de validación cruzada, se utilizaron los mismos atributos para generar un modelo KNN, ya que al indagar el problema de una perspectiva distinta se asegura que las salidas de los modelos sean válidas.

Para medir cuál modelo resulta mejor se utilizará la métrica de precisión, ya que lo que se busca maximizar es el número de aciertos, teniendo en cuenta que los falsos negativos que resulten del modelo no son tan críticos en el contexto del problema. Se considera que de las métricas estudiadas, la precisión es la que otorga un mayor valor a la propuesta, ya que permite clasificar con mayor asertividad la categorización de las manzanas del problema.

III-B. Propuesta experimental problema 2:

A diferencia del problema anterior, en este caso se enfoca en el atributo *NOM_AREA*, además se utiliza el mismo procedimiento (incluyendo los mismos atributos para predicción).

IV. Resultados

IV-A. Problema 1:

En este problema el mejor modelo obtenido fue un árbol de decisión con criterio *gini* y máxima profundidad 7. la precisión promedio fue 0.516, la clase mejor clasificada tiene una precisión de 0.62, la evaluación más pobre la obtuvieron campamento y otros clases que no pudieron ser clasificadas.

IV-B. Problema 2:

Para este problema el mejor modelo obtenido fue un KNN con algoritmo *brute* y numero de vecinos igual a 18. La precisión promedio fue 0.934, además la clase RURAL obtuvo un 0.78 de precisión, finalmente la clase URBANO obtuvo 0.97 de precisión.

V. Análisis de resultados

V-A. Problema 1:

El resultado obtenido no se considera bueno, pues no resuelve el problema de llenar los atributos indeterminados, ya que se consideró que alrededor de un 50 % de datos fallidos en un conjunto de datos oficial del gobierno no es aceptable. Este resultado tiene sentido ya que, como se mencionó anteriormente, la mayoría de las manzanas posee una categoría indeterminada, lo que, junto con la baja correlación entre los atributos utilizados para predecir y la clase, se produjo un alto sesgo en los resultados.

Se observa que el rendimiento del árbol de decisión es superior al modelo KNN en este problema, esto se debe a que se encuentra más optimizado el modelo de árbol para problemas multivariable y debido a la naturaleza misma de los datos del problema, la clusterización de los datos no se encuentra tan definida, además de que KNN por lo general implica un costo computacional más elevado.

En base a las categorías que mejor se predicen (Aldea, Comunidad Indígena y Parcela-Hijuela), se entiende que este comportamiento se debe a la cantidad de datos que contenían esta clasificación, además que comparten el hecho común de contar con agua de origen rural en gran medida, lo cual es un hecho significativo para la predicción. Por otro lado, entre las categorías que peor se predicen (Asentamiento Minero, Asentamiento Pesquero, Campamento y Otros) tienen en común que son muy pocas manzanas las que tenían esas categorías en comparación con el resto, lo que produjo precisiones muy bajas para estas 4 categorías, llegando incluso a tener 0 % de precisión en ciertos casos.

V-B. Problema 2:

La variación entre el peor y mejor modelo es pequeña, pero ya que los modelos obtenían una precisión promedio tan alta se considera que la metodología efectivamente ayudó a elegir un mejor modelo.

El resultado obtenido se considera bueno ya que su precisión es superior a 90 %, esto indica que si se pueden utilizar los atributos del censo para predecir la clase NOM_AREA debido a que, según la métrica elegida, 9 de cada 10 predicciones serán correctas y es posible mitigar el 10 % de pérdida con otras técnicas a futuro.

En este problema se denota que el entrenamiento de KNN es más costoso ya que es un problema multivariable con una salida binaria, es difícil clasificar de forma binaria con KNN ya que el clustering se vuelve muy complejo.

VI. Problemas encontrados

Se puede observar que en los resultados de los modelos las clases con menor cantidad de datos obtuvieron peor precisión que las clases con mayor cantidad de datos, esto ocurre porque no fue revisada la distribución de clases en el análisis exploratorio y por lo tanto no se aplicó alguna técnica correctiva como *undersampling* u *oversampling*.

VII. Discusión a futuro

El proyecto actualmente se encuentra a disposición de aceptar cambios y propuestas de mejora, con tal de mejorar el análisis de la población por medio de técnicas de minería de datos.

En cuanto a aplicaciones a futuro que puedan surgir con los resultados y observaciones obtenidas, se encuentra el planteamiento de utilizar el atributo calculado de Índice de Materialidad como punto entre los manzanares con más urgencia de mejora y renovación de construcción.

También se identifica la posibilidad de explorar este tipo de modelos para otro tipo de problemas asociados al gobierno, por ejemplo, ayudar a la encuesta CADEM a realizar predicciones de votos a lo largo del país.

VIII. Conclusión

Para concluir este experimento, se denota que la intención de este era completar los datos de la plataforma CENSO para mejorar la calidad de los estudios basados en este, por lo tanto se realizó una propuesta experimental que contemplaba el uso de los atributos disponibles para la predicción de la categoría de la manzana, junto con la predicción de su área urbana o rural, siendo estos problemas de clasificación como fue dicho anteriormente. Luego de realizar el experimento, en el primer problema se obtuvo un resultado deplorable de 50 % desde la métrica de precisión, por lo tanto, para este problema se concluye que los atributos no presentaban una relación fuerte con la clase, algo que se podía percibir más no asegurar en la matriz de correlación. Sin embargo, en el segundo problema se obtuvo un resultado satisfactorio, con un 90 % de precisión, lo que permite utilizar modelo de tal problema para predecir esa categoría en base a variables como el origen del agua, la calidad de materialidad y el tipo de vivienda.

En base a lo anterior, se concluye que el estudio fue significativo para modelar el problema y plantearlo para futuros estudios con otras variables que permitan refinar el resultado. Por otra parte, se obtuvo un significativo resultado en el problema de clasificación de categoría urbana o rural, ya que este contaba con datos consistentes y en la totalidad del dataframe, siendo exitoso en resolver el problema y útil para su utilización en futuros estudios.

Referencias

[1] J. ARIAS DE BLOIS. “Censo de Población.” Biblioteca Virtual en Población. https://ccp.ucr.ac.cr/bvp/texto/13/censos.htm (accessed: Oct. 31, 2022).

[2] Hito 3 Censo2017 . GitLab. Available at: https://gitlab.com/k.atighimoghadda01/hito_2_censo2017 (Accessed: November 6, 2022).