



UNIVERSIDAD
DE LA FRONTERA

Análisis de caso Manzanar - Censo 2017 - Revisión 2

Arturo Avendaño - Alonso Rojas - Kianush Atighi-Moghaddam - Diego Vera





Mejoras Hito 1

- Recorte del dataset para optimizar la carga y la comprensión de los datos.
- Se genera un nuevo dato que corresponde al índice de materialidad de la manzana, se utiliza una formulación donde se asignan pesos a los valores de materialidad.
- Mejora de la visualización de los datos.
- Nueva aplicación y preguntas de investigación más cohesionadas.
- Normalización de atributos para mejorar la precisión del modelo.

Recorte del dataset

Se recorta el dataset a los atributos que serán utilizados en la investigación.

De esta manera se evitan cargas innecesarias y es menos propenso al error humano.

| | PERSONAS | VIVIENDA_PARTICULAR | VIVIENDA_COLECTIVA | VIVIENDA_PARTICULAR_OCUPADA | TOTAL_VIV | TIPO_VIV_CASA | TIPO_VIV_DPTO | TIPO_VIV_TRADICIONAL |
|----|----------|---------------------|--------------------|-----------------------------|-----------|---------------|---------------|----------------------|
| 1 | 173 | 0.9841270 | 0.015873016 | 0.8571429 | 63 | 0.58730159 | 0.06349206 | 0.000000000 |
| 2 | 70 | 1.0000000 | 0.000000000 | 0.9565217 | 23 | 1.00000000 | 0.00000000 | 0.000000000 |
| 3 | 128 | 1.0000000 | 0.000000000 | 1.0000000 | 33 | 1.00000000 | 0.00000000 | 0.000000000 |
| 4 | 229 | 1.0000000 | 0.000000000 | 0.8923077 | 65 | 1.00000000 | 0.00000000 | 0.000000000 |
| 6 | 138 | 1.0000000 | 0.000000000 | 0.8918919 | 37 | 1.00000000 | 0.00000000 | 0.000000000 |
| 7 | 132 | 1.0000000 | 0.000000000 | 0.7500000 | 52 | 0.67307692 | 0.00000000 | 0.000000000 |
| 9 | 106 | 1.0000000 | 0.000000000 | 0.9642857 | 28 | 1.00000000 | 0.00000000 | 0.000000000 |
| 10 | 237 | 1.0000000 | 0.000000000 | 0.9866667 | 75 | 0.97333333 | 0.00000000 | 0.000000000 |
| 13 | 166 | 1.0000000 | 0.000000000 | 0.9074074 | 54 | 1.00000000 | 0.00000000 | 0.000000000 |
| 14 | 100 | 1.0000000 | 0.000000000 | 0.9642857 | 28 | 1.00000000 | 0.00000000 | 0.000000000 |
| 15 | 105 | 1.0000000 | 0.000000000 | 0.8928571 | 28 | 0.85714286 | 0.00000000 | 0.000000000 |
| 16 | 655 | 1.0000000 | 0.000000000 | 0.6039886 | 351 | 0.13390313 | 0.85470085 | 0.000000000 |
| 18 | 118 | 1.0000000 | 0.000000000 | 0.9032258 | 31 | 0.74193548 | 0.00000000 | 0.000000000 |
| 19 | 127 | 1.0000000 | 0.000000000 | 0.9736842 | 38 | 0.78947368 | 0.00000000 | 0.000000000 |
| 20 | 624 | 0.9980315 | 0.001968504 | 0.3877953 | 508 | 0.25000000 | 0.74606299 | 0.001968504 |
| 21 | 0 | 1.0000000 | 0.000000000 | 0.0000000 | 1 | 1.00000000 | 0.00000000 | 0.000000000 |
| 22 | 178 | 1.0000000 | 0.000000000 | 0.9767442 | 43 | 0.95348837 | 0.00000000 | 0.000000000 |
| 23 | 125 | 1.0000000 | 0.000000000 | 0.9714286 | 35 | 1.00000000 | 0.00000000 | 0.000000000 |
| 25 | 108 | 1.0000000 | 0.000000000 | 0.8297872 | 47 | 0.00000000 | 1.00000000 | 0.000000000 |
| 26 | 99 | 1.0000000 | 0.000000000 | 0.9166667 | 36 | 0.77777778 | 0.00000000 | 0.000000000 |
| 29 | 0 | 1.0000000 | 0.000000000 | 0.0000000 | 2 | 1.00000000 | 0.00000000 | 0.000000000 |
| 30 | 82 | 1.0000000 | 0.000000000 | 0.8846154 | 26 | 0.96153846 | 0.00000000 | 0.000000000 |
| 33 | 182 | 1.0000000 | 0.000000000 | 0.7758621 | 58 | 0.77586207 | 0.13793103 | 0.000000000 |
| 34 | 68 | 1.0000000 | 0.000000000 | 0.8076923 | 26 | 1.00000000 | 0.00000000 | 0.000000000 |
| 35 | 77 | 1.0000000 | 0.000000000 | 0.9090909 | 22 | 1.00000000 | 0.00000000 | 0.000000000 |

Índice de materialidad

Se utiliza una nueva métrica generada en base al tipo de materialidad de cada manzana, que describe de manera porcentual el nivel de solidez de las estructuras de la manzana.

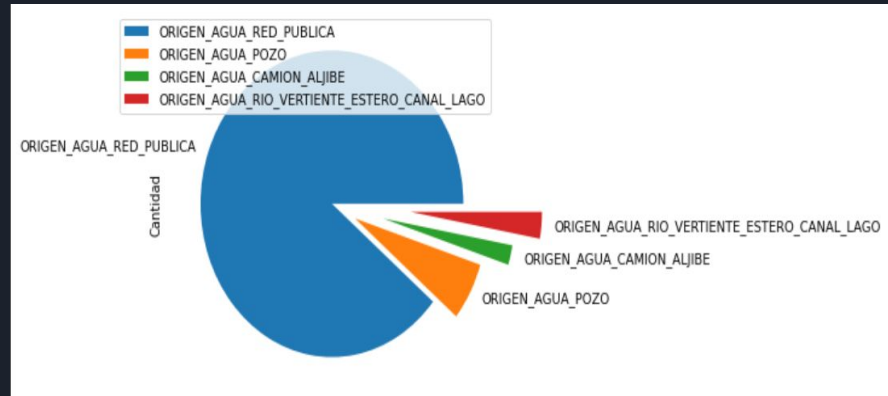
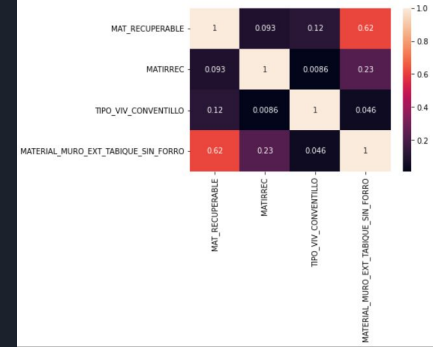
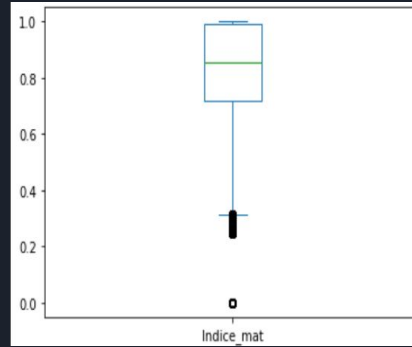
Será utilizado como referencia para posteriores implementaciones.

$$\text{Indice mat.} = \frac{\text{MAT.ACCEPT.} + \text{MAT.RECUP.} + \text{MAT.IRREC.}}{\text{MAT.ACCEPT.} + 2 \cdot \text{MAT.RECUP.} + 4 \cdot \text{MAT.IRREC.}}$$

| MAT_ACCEPTABLE | MAT_RECUPERABLE | MATIRREC | indice_mat |
|----------------|-----------------|----------------|---------------|
| Min. :0.0000 | Min. :0.00000 | Min. :0.0000 | Min. :0.250 |
| 1st Qu.:0.5383 | 1st Qu.:0.00000 | 1st Qu.:0.0000 | 1st Qu.:0.750 |
| Median :0.7353 | Median :0.09434 | Median :0.0000 | Median :0.865 |
| Mean :0.6520 | Mean :0.14009 | Mean :0.0131 | Mean :0.843 |
| 3rd Qu.:0.8625 | 3rd Qu.:0.22222 | 3rd Qu.:0.0000 | 3rd Qu.:0.964 |
| Max. :1.0000 | Max. :1.00000 | Max. :1.0000 | Max. :1.000 |
| | | | NA's :5371 |

Mejora de la visualización de los datos

- Gráfico de torta del origen del agua para identificar la segmentación del atributo.
- Boxplot del índice de materialidad para realizar una comprobación lógica de la utilización de la nueva variable.
- Matriz de correlación para comprobar datos de interés.
- Variados histogramas para revisar frecuencia y peso de los atributos más significativos para el estudio.
- Descripción del Dataset filtrado.





Contexto de investigación

- Una de las preguntas que buscamos responder es la posibilidad de predecir la clase de la manzana la cual corresponde al atributo “NOM_CAT_ENT”, esto nos permite mejorar la calidad del dataset ya que completa datos que actualmente no están disponibles. Se abordará a través del uso de algoritmos de clasificación ya que se detecta como un problema de clasificación de clase.
- La otra pregunta que se presenta en este estudio es la utilización de un índice de calidad de la manzana ya que permite resumir varios atributos en un solo valor significativo.

Normalización de atributos

| | PERSONAS | VIVIENDA_PARTICULAR | VIVIENDA_COLECTIVA | VIVIENDA_PARTICULAR_OCUPADA | TOTAL_VIV | TIPO_VIV_CASA | TIPO_VIV_DPTO | TIPO_VIV_TRADICIONAL |
|----|----------|---------------------|--------------------|-----------------------------|-----------|---------------|---------------|----------------------|
| 1 | 173 | 62 | 1 | 54 | 63 | 37 | 4 | 0 |
| 2 | 70 | 23 | 0 | 22 | 23 | 23 | 0 | 0 |
| 3 | 128 | 33 | 0 | 33 | 33 | 33 | 0 | 0 |
| 4 | 229 | 65 | 0 | 58 | 65 | 65 | 0 | 0 |
| 6 | 138 | 37 | 0 | 33 | 37 | 37 | 0 | 0 |
| 7 | 132 | 52 | 0 | 39 | 52 | 35 | 0 | 0 |
| 9 | 106 | 28 | 0 | 27 | 28 | 28 | 0 | 0 |
| 10 | 237 | 75 | 0 | 74 | 75 | 73 | 0 | 0 |
| 13 | 166 | 54 | 0 | 49 | 54 | 54 | 0 | 0 |
| 14 | 100 | 28 | 0 | 27 | 28 | 28 | 0 | 0 |
| 15 | 105 | 28 | 0 | 25 | 28 | 24 | 0 | 0 |
| 16 | 655 | 351 | 0 | 212 | 351 | 47 | 300 | 0 |
| 18 | 118 | 31 | 0 | 28 | 31 | 23 | 0 | 0 |
| 19 | 127 | 38 | 0 | 37 | 38 | 30 | 0 | 0 |
| 20 | 624 | 507 | 1 | 197 | 508 | 127 | 379 | 1 |
| 21 | 0 | 1 | 0 | 0 | 1 | 1 | 0 | 0 |
| 22 | 178 | 43 | 0 | 42 | 43 | 41 | 0 | 0 |
| 23 | 125 | 35 | 0 | 34 | 35 | 35 | 0 | 0 |
| 25 | 108 | 47 | 0 | 35 | 47 | 0 | 47 | 0 |
| 26 | 99 | 36 | 0 | 33 | 36 | 0 | 0 | 0 |
| 29 | 0 | 0 | 0 | 0 | 2 | 2 | 0 | 0 |
| 30 | 82 | 26 | 0 | 23 | 26 | 25 | 0 | 0 |
| 33 | 182 | 58 | 0 | 45 | 58 | 45 | 8 | 0 |
| 34 | 68 | 26 | 0 | 21 | 26 | 26 | 0 | 0 |
| 35 | 77 | 22 | 0 | 20 | 22 | 22 | 0 | 0 |
| 36 | 117 | 32 | 0 | 30 | 32 | 32 | 0 | 0 |
| 42 | 58 | 17 | 0 | 16 | 17 | 17 | 0 | 0 |
| 43 | 132 | 39 | 0 | 35 | 39 | 38 | 0 | 0 |
| 44 | 174 | 54 | 0 | 52 | 54 | 54 | 0 | 0 |



| | PERSONAS | VIVIENDA_PARTICULAR | VIVIENDA_COLECTIVA | VIVIENDA_PARTICULAR_OCUPADA | TOTAL_VIV | TIPO_VIV_CASA | TIPO_VIV_DPTO | TIPO_VIV_TRADICIONAL |
|----|----------|---------------------|--------------------|-----------------------------|-----------|---------------|---------------|----------------------|
| 1 | 173 | 0.9841270 | 0.015873016 | 0.8571429 | 63 | 0.58730159 | 0.06349206 | 0.000000000 |
| 2 | 70 | 1.0000000 | 0.000000000 | 0.9565217 | 23 | 1.00000000 | 0.000000000 | 0.000000000 |
| 3 | 128 | 1.0000000 | 0.000000000 | 1.0000000 | 33 | 1.00000000 | 0.000000000 | 0.000000000 |
| 4 | 229 | 1.0000000 | 0.000000000 | 0.8923077 | 65 | 1.00000000 | 0.000000000 | 0.000000000 |
| 6 | 138 | 1.0000000 | 0.000000000 | 0.8918919 | 37 | 1.00000000 | 0.000000000 | 0.000000000 |
| 7 | 132 | 1.0000000 | 0.000000000 | 0.7500000 | 52 | 0.67367692 | 0.000000000 | 0.000000000 |
| 9 | 106 | 1.0000000 | 0.000000000 | 0.9642857 | 28 | 1.00000000 | 0.000000000 | 0.000000000 |
| 10 | 237 | 1.0000000 | 0.000000000 | 0.9666667 | 75 | 0.97333333 | 0.000000000 | 0.000000000 |
| 13 | 166 | 1.0000000 | 0.000000000 | 0.9074074 | 54 | 1.00000000 | 0.000000000 | 0.000000000 |
| 14 | 100 | 1.0000000 | 0.000000000 | 0.9642857 | 28 | 1.00000000 | 0.000000000 | 0.000000000 |
| 15 | 105 | 1.0000000 | 0.000000000 | 0.8928571 | 28 | 0.85714286 | 0.000000000 | 0.000000000 |
| 16 | 655 | 1.0000000 | 0.000000000 | 0.6039886 | 351 | 0.13390313 | 0.85470085 | 0.000000000 |
| 18 | 118 | 1.0000000 | 0.000000000 | 0.9032258 | 31 | 0.74193548 | 0.000000000 | 0.000000000 |
| 19 | 127 | 1.0000000 | 0.000000000 | 0.9736842 | 38 | 0.78947368 | 0.000000000 | 0.000000000 |
| 20 | 624 | 0.9980315 | 0.001968504 | 0.3877953 | 508 | 0.25000000 | 0.74606299 | 0.001968504 |
| 21 | 0 | 1.0000000 | 0.000000000 | 0.0000000 | 1 | 1.00000000 | 0.000000000 | 0.000000000 |
| 22 | 178 | 1.0000000 | 0.000000000 | 0.9767442 | 43 | 0.95348837 | 0.000000000 | 0.000000000 |
| 23 | 125 | 1.0000000 | 0.000000000 | 0.9714286 | 35 | 1.00000000 | 0.000000000 | 0.000000000 |
| 25 | 108 | 1.0000000 | 0.000000000 | 0.8297872 | 47 | 0.00000000 | 1.000000000 | 0.000000000 |
| 26 | 99 | 1.0000000 | 0.000000000 | 0.9166667 | 36 | 0.77777778 | 0.000000000 | 0.000000000 |
| 29 | 0 | 1.0000000 | 0.000000000 | 0.0000000 | 2 | 1.00000000 | 0.000000000 | 0.000000000 |
| 30 | 82 | 1.0000000 | 0.000000000 | 0.8846134 | 26 | 0.96153846 | 0.000000000 | 0.000000000 |
| 33 | 182 | 1.0000000 | 0.000000000 | 0.7758621 | 58 | 0.77586207 | 0.13793103 | 0.000000000 |
| 34 | 68 | 1.0000000 | 0.000000000 | 0.8076523 | 26 | 1.00000000 | 0.000000000 | 0.000000000 |
| 35 | 77 | 1.0000000 | 0.000000000 | 0.9090909 | 22 | 1.00000000 | 0.000000000 | 0.000000000 |
| 36 | 117 | 1.0000000 | 0.000000000 | 0.9375000 | 32 | 1.00000000 | 0.000000000 | 0.000000000 |
| 42 | 58 | 1.0000000 | 0.000000000 | 0.9411765 | 17 | 1.00000000 | 0.000000000 | 0.000000000 |
| 43 | 132 | 1.0000000 | 0.000000000 | 0.8974359 | 39 | 0.97435897 | 0.000000000 | 0.000000000 |
| 44 | 174 | 1.0000000 | 0.000000000 | 0.9142857 | 54 | 1.00000000 | 0.000000000 | 0.000000000 |

Se normalizan los datos basados en la cantidad de viviendas, para evitar que una manzana con mayor cantidad de viviendas modifique el modelo.




Propuesta experimental

Se implementará la primera pregunta planteada que busca predecir la categoría de la manzana a partir de la variable “NOM_CAT_ENT”. Para ello lo que se hará es entrenar el subconjunto de datos que posea una categoría definida, para luego predecir la categoría de las filas que poseen valor indeterminado de aquel atributo. Para ello se utilizarán los algoritmos de Árbol de Decisión y KNN.

Es importante tener en cuenta que la proporción de manzanas con categoría indeterminada es mucho mayor a las que poseen categoría definida, lo que producirá un cierto sesgo en sus resultados.

Para comparar el resultado de predicción de los modelos se utilizará la métrica “precision” con tal de buscar la mayor asertividad dentro de la predicción.

“Indeterminado”  “Aldea/Caserío/Parcela/etc.”

Resultado preliminar

Resultados preliminares del modelo definido en la propuesta experimental.

En el caso del “decision tree” se obtuvo un buen rendimiento para una primera aproximación de alrededor de un 60%, lo cual nos permite avanzar y utilizar esta metodología para continuar investigando y asignando valores a la categoría de las muestras no identificadas.

Se observa que el rendimiento de KNN es inferior en la predicción de la variable de clase (“NOM_CAT_ENT”) por lo tanto se seleccionan los resultados del primer modelo.

| | precision | recall | f1-score | support |
|-------------------------|-----------|--------|----------|---------|
| Aldea | 0.65 | 0.61 | 0.63 | 218 |
| Asentamiento Minero | 0.33 | 0.31 | 0.32 | 75 |
| Asentamiento Pesquero | 0.38 | 0.41 | 0.40 | 75 |
| Campamento | 0.00 | 0.00 | 0.00 | 16 |
| Caserío | 0.38 | 0.39 | 0.39 | 1101 |
| Comunidad Indígena | 0.37 | 0.38 | 0.38 | 647 |
| Fundo-Estancia-Hacienda | 0.40 | 0.56 | 0.46 | 1029 |
| Otros | 0.12 | 0.03 | 0.05 | 91 |
| Parcela de Agrado | 0.53 | 0.55 | 0.54 | 490 |
| Parcela-Hijuela | 0.60 | 0.54 | 0.57 | 3307 |
| Veranada-Majada-Aguada | 0.39 | 0.30 | 0.34 | 148 |
| accuracy | | | 0.49 | 7197 |
| macro avg | 0.38 | 0.37 | 0.37 | 7197 |
| weighted avg | 0.50 | 0.49 | 0.49 | 7197 |

Decision tree - metrics

| | precision | recall | f1-score | support |
|-------------------------|-----------|--------|----------|---------|
| Aldea | 0.57 | 0.76 | 0.65 | 218 |
| Asentamiento Minero | 0.24 | 0.16 | 0.19 | 75 |
| Asentamiento Pesquero | 0.52 | 0.41 | 0.46 | 75 |
| Campamento | 0.00 | 0.00 | 0.00 | 16 |
| Caserío | 0.31 | 0.32 | 0.32 | 1101 |
| Comunidad Indígena | 0.21 | 0.13 | 0.16 | 647 |
| Fundo-Estancia-Hacienda | 0.43 | 0.55 | 0.48 | 1029 |
| Otros | 0.25 | 0.04 | 0.07 | 91 |
| Parcela de Agrado | 0.36 | 0.16 | 0.22 | 490 |
| Parcela-Hijuela | 0.56 | 0.61 | 0.59 | 3307 |
| Veranada-Majada-Aguada | 0.41 | 0.30 | 0.35 | 148 |
| accuracy | | | 0.47 | 7197 |
| macro avg | 0.35 | 0.31 | 0.32 | 7197 |
| weighted avg | 0.45 | 0.47 | 0.45 | 7197 |

KNN - metrics



Conclusión

En base al experimento realizado se observa que la capacidad de predicción del modelo en base a los atributos elegidos alcanza alrededor del 60%, lo cual nos indica que el modelo es favorable y cumple con acertar más del 50% de las predicciones. Al revisar la predicción de datos no controlados, se observa una tendencia a clasificar sectores rurales, lo cual tiene relevancia ya que es lógicamente probable que esas zonas de difícil acceso no hayan sido catalogadas, no así la ciudad la cual tiene menos incidencia según el modelo. Se considera que este experimento cumple las expectativas de un modelo que vale la pena seguir optimizando para posteriormente ser utilizado en el estudio del CENSO.