Universidad de San Carlos de Guatemala

Facultad de Ingenieria

Escuela de Estudios de Postgrado

Maestria en Ingenieria para la Industria
con Especialidad en Ciencias de la Computación

MIICC408 - Introducción a la Minería de Datos
Cat. Ing. MSC. Kevin Lajpop
Sección A

Ministerio de
Finanzas Públicas

GUATECOMPRAS.gt
Sistema de Información de Contrataciones y Adquisiciones del Estado

Proyecto, Parte 2
Presentado por:

Carlos Alberto Rios Calderón

Carnet: 1000-17078

Repositorio:
https://github.com/carlosrios-gt/Proyecto-Parte-2

Índice

Introducción

Las técnicas de minería de datos permiten descubrir estructuras, relaciones y patrones subyacentes que no son fácilmente perceptibles, aportando información valiosa para respaldar la toma de decisiones.

En el presente análisis se emplean tres enfoques complementarios: árbol de decisión, bosques aleatorios y redes neuronales, seleccionados por su eficacia en la identificación de asociaciones frecuentes y la segmentación del conjunto de datos con base en sus atributos.

Estas metodologías permiten analizar tanto relaciones categóricas como patrones cuantitativos, brindando una comprensión holística del comportamiento de las variables evaluadas.

La finalidad del estudio es identificar reglas de co-ocurrencia, repeticiones significativas y agrupaciones espontáneas que revelen la estructura interna del conjunto de datos.

La integración de técnicas de asociación y algoritmos de agrupamiento no supervisado constituye un enfoque analítico sólido para detectar dependencias, tendencias y conjuntos homogéneos, facilitando la obtención de hallazgos interpretables y aplicables en contextos reales.

Objetivos

General

Establecer un programa de minería de datos que permita entender el comportamiento de los procesos adjudicados en Guatecompras, utilizando las técnicas de para evaluar el comportamiento y las tendencias que se puedan identificar mediante la aplicación de estos métodos.

Específicos

1. Encontrar reglas claras que expliquen el comportamiento de la variable objetivo a partir de las características disponibles mediante arboles de decisión.

2. Mejorar la estabilidad y capacidad predictiva de un árbol individual reduciendo el sobreajuste.

3. Utilizar las redes neuronales para capturar relaciones profundas y no lineales que los modelos más simples no pueden aprender.

Consideraciones previas

Los métodos previos (Apriori, FP-Growth, k-Means) habían revelado que el monto se estructura por rangos y no por entidades.

Al analizar las conclusiones encontradas con el metodo FP-Growth, el monto se agrupa en 3 rangos naturales:
[0 – 23,800)
[23,800 – 81,400)
[81,400 – 500M]

Los rangos se asocian a tipos generales de entidad, no a entidades específicas:
Montos bajos: Asociados al Sector Público / MSP
Montos medios: Asociados con entidades descentralizadas / IGSS
Montos altos: Municipalidades / Gobiernos locales

1. Arboles de decisión

Es un modelo supervisado que divide los datos en ramas usando reglas simples hasta llegar a una predicción. Es más fácil de interpretar, rápido y útil para clasificación y regresión.

El árbol de decisión se ejecutará acorde al análisis previo, donde se revela que el tipo de entidad compradora es la variable determinante para clasificar los niveles de monto adjudicado.

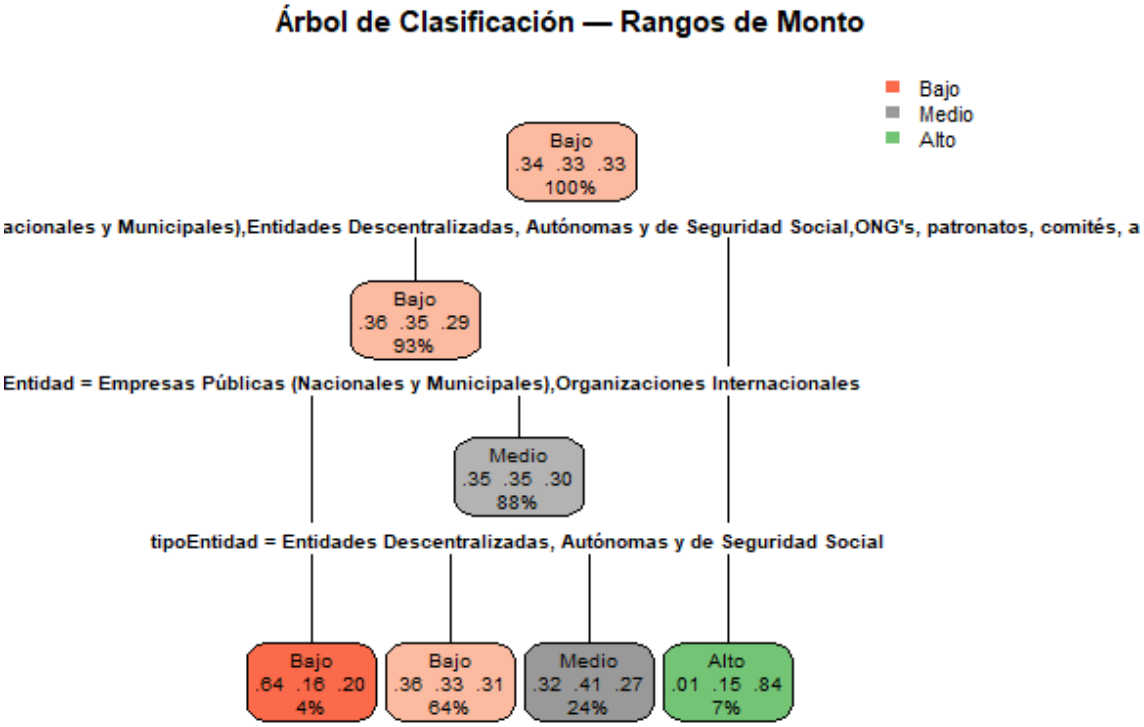
El modelo separa al sistema en dos grandes bloques según lo observado anteriormente:

1. Los Gobiernos Locales, Fideicomisos y Cooperativas, son las entidades que concentran la mayoría de montos altos (84%).
2. La Administración Central, Descentralizadas, ONG y Organismos Internacionales, realizan principalmente compras de montos bajos y medios.

También se evidencio que, a nivel interno,

1. Son las Empresas Públicas y Organizaciones Internacionales, tienen una fuerte tendencia hacia montos bajos.
2. Las Entidades Descentralizadas y la Seguridad Social, tienden a un volumen masivo con montos variados, predominando los bajos.
3. La Administración Central y ONG, se asocian a montos medios.

El árbol valida los patrones estructurales identificados previamente mediante técnicas de minería de datos FP-Growth y k-Means, confirmando que el sistema de compras públicas se segmenta naturalmente en tres rangos económicos y que estos rangos dependen principalmente de la naturaleza institucional del comprador, más que de otras variables.



Se concluye que los montos bajos están fuertemente ligados al Sector Público, tal como se observa en la raíz que divide y el nodo principal queda dominado por Bajo (34%).

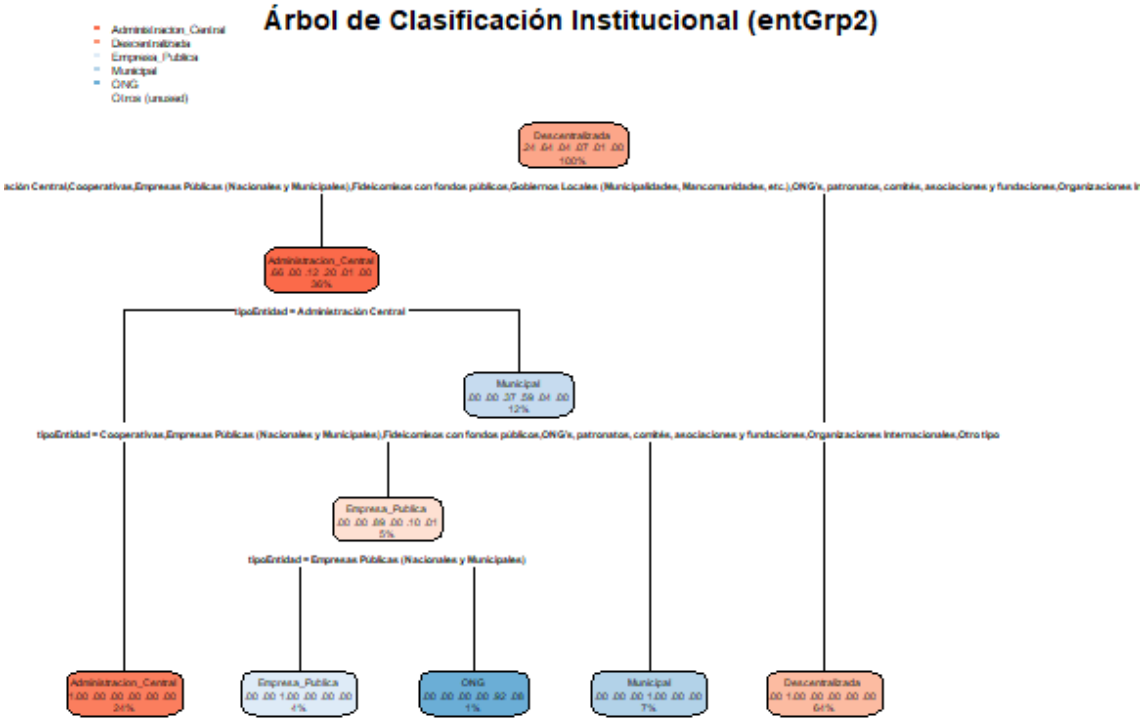
Árbol para predecir entGrp

Para responder y validar el tipo de institución que realiza determinado nivel de compras, se ejecutara otro árbol de decisión, que apoya las conclusiones encontradas anteriormente, donde se encontró que las Municipalidades mantienen montos altos. Las ONG y Administración Central, mantienen montos bajos a medios.

En este árbol se pretende validar que combinación de monto (en rangos) y tipo de entidad explica a qué grupo institucional pertenece una compra. Se valida en que grupos se tienen las tendencias de comprar bajo, medio o alto según las categorías definidas. También en qué grupos se asocian con ciertos tipos de entidad. Para ello se agrupará en las categorías siguientes.

1. Descentralizada, donde se agruparán entidades Descentralizadas, Autónomas y Seguridad Social (IGSS, instituciones autónomas).
2. Municipal, conformada por Gobiernos Locales, Mancomunidades,
3. ONG, que está formada por ONG, patronatos, asociaciones y fundaciones y la clasificación “Otro tipo”, si es organización civil
4. Administración Central, formada por Ministerios y las Dependencias del Ejecutivo.
5. Empresas Públicas, formada por Nacionales y Municipales
6. Otras Instituciones Pequeñas, que están clasificadas instituciones como el Banco de Guatemala, Bomberos, Obras Sociales y Organismo Deportivo

Estas seis clases son separables con tipoEntidad y montoRango para su análisis por separado. Confirmando asi los hallazgos encontrados con los tres métodos anteriores.



El árbol de clasificación institucional demuestra que el sistema de compras públicas está fuertemente estructurado por la naturaleza jurídica de las entidades

participantes. El modelo identifica separaciones casi perfectas entre los grupos institucionales definidos.

- ✓ Entidades Descentralizadas (100%)
- ✓ Administración Central (100%)
- ✓ Municipalidades (100%)
- ✓ Empresas Públicas (100%)
- ✓ ONG / Asociaciones (92%)

Esto confirma que la variable tipoEntidad es un predictor totalmente determinante del grupo institucional, validando la estructura conceptual utilizada en el análisis previo.

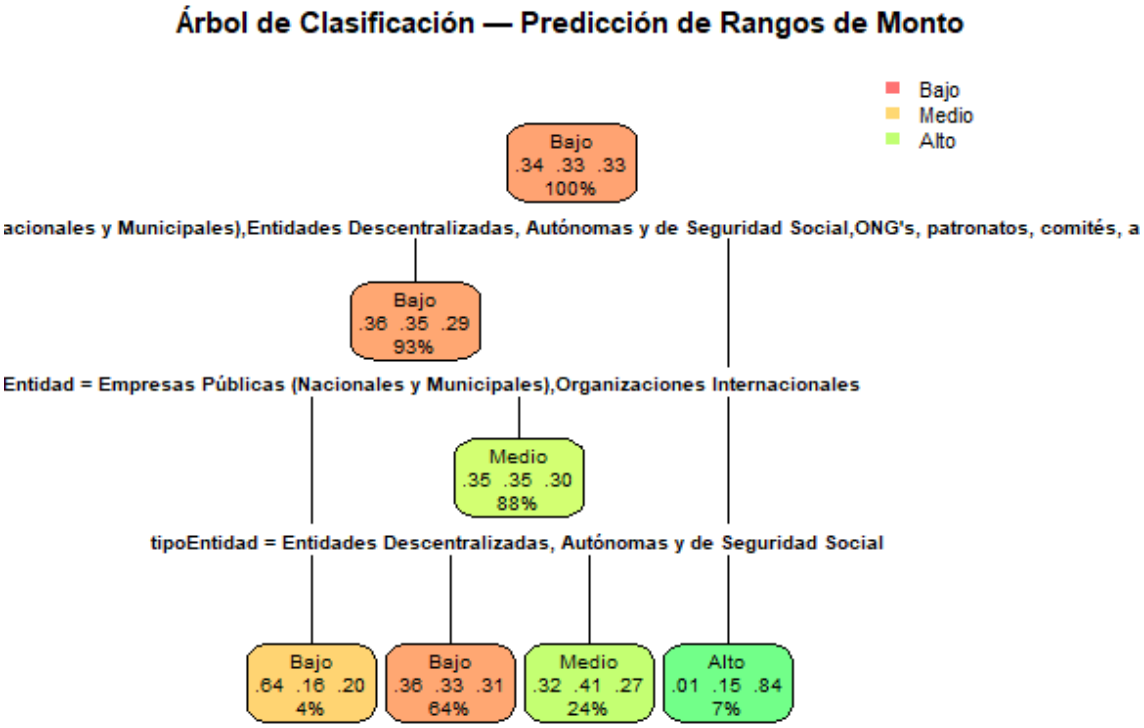
La segmentación encontrada es coherente con los patrones de monto observados en las técnicas descriptivas FP-Growth y k-Means, reforzando la solidez del modelo institucional propuesto.

Árbol de decisión para predecir montoRango

Con este árbol se validará directamente que los montos bajos se relacionan con ONG y sector público general. También se afirma que los montos medios aparecen con descentralizadas y los montos altos se concentran en municipalidades.

El árbol de clasificación construido, sirve para predecir el rango de monto (Bajo, Medio o Alto) utilizando las variables tipoEntidad y entGrp2, revelando una estructura institucional clara y coherente con el análisis previo y con los hallazgos del documento de referencia.

El modelo establece que la variable tipoEntidad explica de forma determinante el comportamiento del monto adjudicado, permitiendo identificar patrones económicos consistentes en el sistema de compras públicas.



Discusión de resultados de tendencias o patrones interesantes

1. El árbol de clasificación confirma que la estructura del gasto no es aleatoria, sino que responde a la naturaleza institucional del comprador.
2. Las municipalidades se comportan como el principal agente de compras de alto volumen.
3. Las entidades descentralizadas presentan una distribución heterogénea, coherente con su operación de gran escala.
4. Las ONG y asociaciones se ubican entre montos bajos y medios, sin acceso a compras grandes.
5. Las empresas públicas realizan compras de bajo valor.
6. En conjunto, el modelo confirma plenamente las conclusiones del documento técnico, proporcionando una validación estadística sólida mediante aprendizaje automático.

Caso de prueba. Escenario ficticio

Se elegirá aleatoriamente al tipoEntidad = "Entidades Descentralizadas, Autónomas y de Seguridad Social". Donde un ejemplo real podría ser IGSS, INDE, SAT. La entGrp2 = "Descentralizada" debido a que es la clase más grande del dataset, tiene comportamiento mixto en montos y es un escenario interesante para evaluar cómo el árbol divide en bajo/medio/alto.

Bajo = 35.77%
Medio = 33.41%
Alto = 30.82%

Se confirma que estas instituciones ejecutan todo tipo de compras, compras pequeñas rutinarias es un nivel Bajo, las compras operativas intermedias en un nivel Medio y las compras estratégicas grandes como Alto

Se concluye que el monto en las entidades descentralizadas no depende de su tipo institucional.

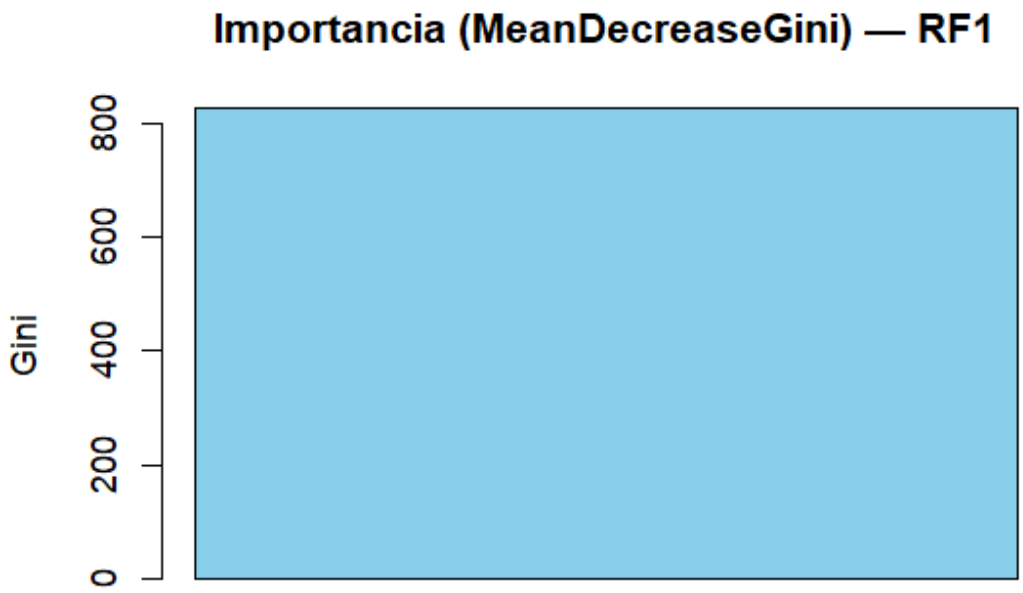
2. Bosque Aleatorio

Para este método será definido el uso de la data set de adjudicaciones. Inspeccionando los otros dos data set, de contratos no dados o en proceso de anulación, no será prescindible para el análisis que se desea para este caso.

El modelo Random Forest confirma que la variable tipoEntidad solo permite identificar con precisión al grupo ONG, alcanzando un 92.3% de exactitud.

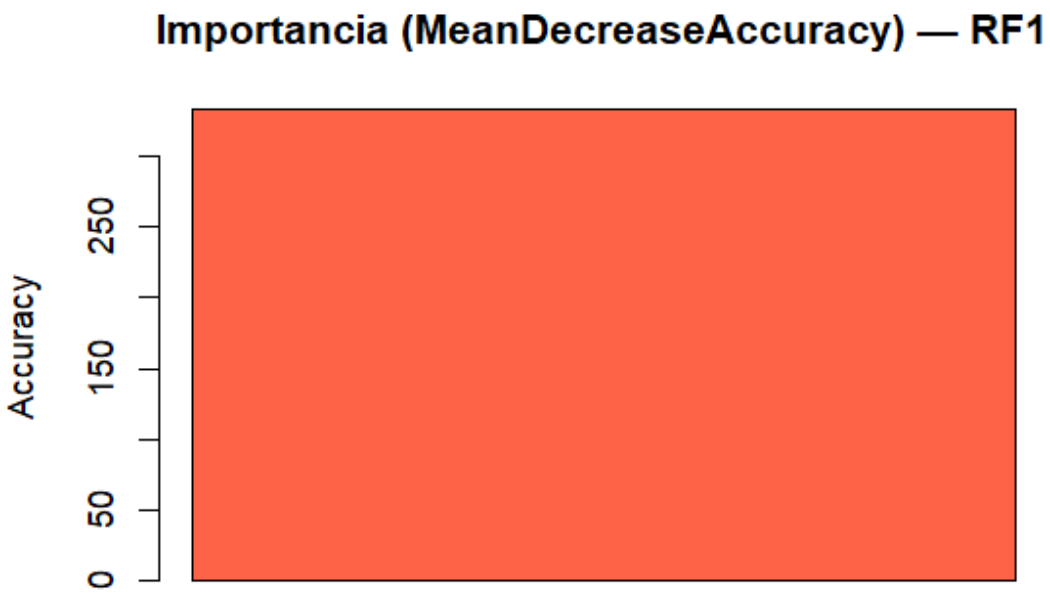
Para los demás grupos institucionales reducidos (entGrp), la precisión es prácticamente nula debido a que sus valores de tipoEntidad no poseen diferencias suficientes para permitir una separación efectiva.

Esto valida que la clasificación entGrp no es explicable mediante tipoEntidad, y por lo tanto no es recomendable para modelar relaciones dependientes del tipo de institución sin agregar variables adicionales.



Random forest para predecir entGrp2

Su objetivo está en predecir entGrp2, usando la variable tipoEntidad. Este modelo debe mostrar que la separación perfecta de Administración Central, Descentralizadas, Municipal, Empresa Pública, ONG y un pequeño sector en la categoría “Otros” (98 casos). Este Random Forest refleja exactamente las conclusiones del Árbol Institucional Final.



El modelo Random Forest para predecir entGrp2 a partir de tipoEntidad obtuvo un resultado OOB del 0%, logrando una clasificación perfecta de las 207,037 observaciones.

Este resultado refleja que la variable tipoEntidad contiene toda la información necesaria para diferenciar completamente a las seis categorías institucionales definidas en entGrp2.

Clasificacion	Administracion_Central	Descentralizada	Empresa_Publica	Municipal	ONG	Otros	class.error
Administracion Central	49181	0	0	0	0	0	0
Descentralizada	0	132835	0	0	0	0	0
Empresa Publica	0	0	9148	0	0	0	0
Municipal	0	0	0	14693	0	0	0
ONG	0	0	0	0	1082	0	0
Otros	0	0	0	0	0	98	0

La matriz de confusión anterior, muestra que cada grupo institucional fue identificado sin errores, lo que valida la solidez conceptual de la clasificación entGrp2 y demuestra que está correctamente alineada con la estructura formal del sector público.

En consecuencia, el modelo respalda estadísticamente el uso de entGrp2 como una segmentación adecuada y robusta para análisis posteriores, incluyendo modelos predictivos, segmentación del gasto y análisis de patrones de compra.

Random Forest para el análisis financiero

Este Random Forest tiene una limitación fundamental ya que entGrp2 y tipoEntidad no contienen suficiente información para predecir montoRango con alta precisión. Como se había definido en los anteriores métodos de minería.

En los métodos anteriores se evidencio que los montos no dependen de la institución en todos los casos. Debido a que hay solapamiento entre entidades en montos bajos y medios.

Los montos altos sí se ligan a municipalidades, por lo que las conclusiones dictan de forma afirmativa y conclusiva con el anterior metodo. El modelo Random Forest sí predice bien la instancia “Bajo”, ya que esta aparece en las categorías siguientes, por lo cual se acierta en el 77% de los casos.

- ✓ Empresas públicas
- ✓ ONG
- ✓ Descentralizadas
- ✓ Administración central

Categoría	Bajo	Medio	Alto	Error
Bajo	53388	15913	107	23%
Medio	45839	20603	2213	70%
Alto	42759	13750	12465	82%

El modelo Random Forest falla en “Medio”, ya que el error del 70% evidencia que la clase definida como “Medio” no está bien definida por instituciones. Se mezcla en casi todas. Esto refuerza el hallazgo de los métodos anteriores al decir qyue los rangos medios aparecen combinados en múltiples instituciones.

El Random Forest solo está confirmando este hallazgo. El error del “Alto” que asciende al 82% evidencia que las municipalidades son altas; porque tipoEntidad y entGrp2 no codifican directamente “Municipal” como un valor binario.

Esto concluye en que se valida las conclusiones halladas en los métodos anteriores sobre que, tipoEntidad y entGrp2 no determinan completamente el monto.

El hallazgo concluyente que la variabilidad económica no es explicable solo por la naturaleza institucional. Se confirma en el Random Forest, ya que “Bajo” sí depende de tipoEntidad, “Medio”, no depende de tipoEntidad y “Alto” solo municipalidades lo poseen, por lo que el Random Forest no puede aislarlas con los splits aleatorios.

Caso de prueba. Escenario ficticio

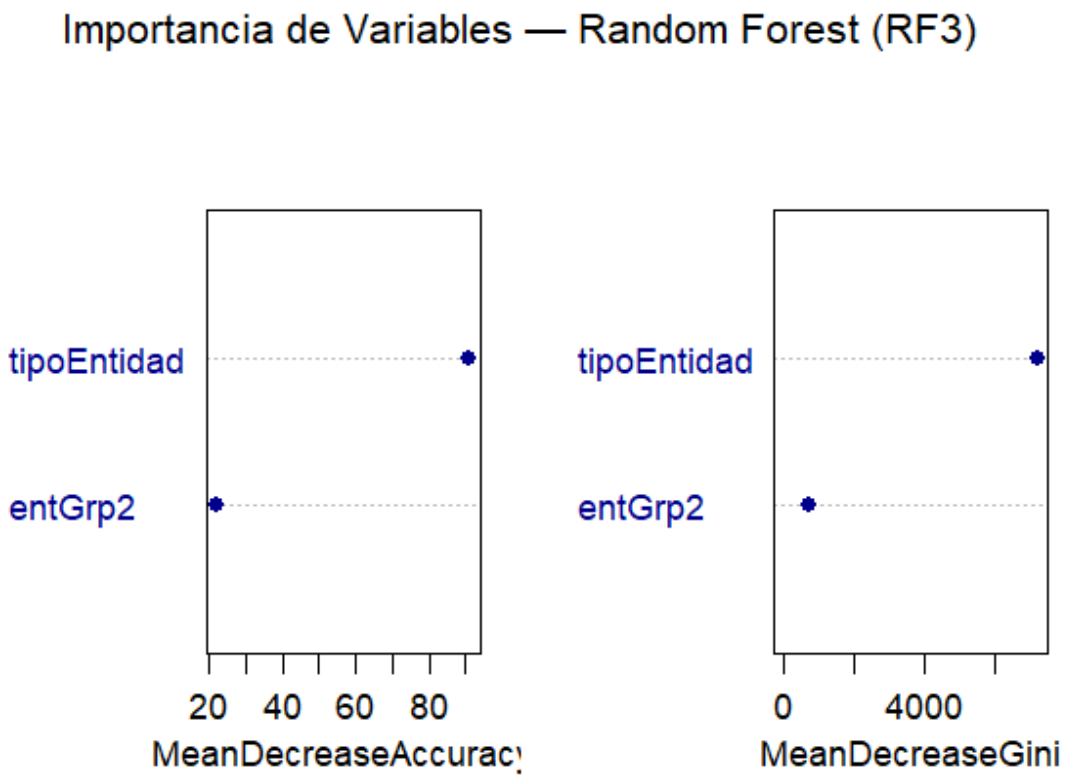
Usando el modelo Random Forest (RF3) se obtendrá las probabilidades más robustas para definir el caso ficticio.

- ✓ tipoEntidad = "Entidades Descentralizadas, Autónomas y de Seguridad Social"
- ✓ entGrp2 = "Descentralizada"

Se obtuvo una probabilidad distinta al árbol de decisión, el cual daba un resultado que daba probabilidades 35%–33%–30%, como se ve a continuación.

Bajo = 1
Medio = 0
Alto = 0

Esto se da debido a que el Random Forest aprende patrones mucho más estrictos que un árbol de decisión y detecta combinaciones de variables que el árbol no captura. Cuando se hizo la combinación entGrp2 + tipoEntidad, se creó un efecto importante, ya que el agrupamiento entGrp2 "Descentralizada" está dominado por compras de monto Bajo. El dataset contiene muchísimas filas donde la institución descentralizada aparece y el monto adjudicado fue bajo. También aparece bajo reglas de contratación recurrentes y compras de rutina.



El Random Forest detecta eso. El árbol de decisión veía solamente el tipoEntidad, por eso daba valores cercanos a 1/3–1/3–1/3. Sin embargo, Random Forest utiliza cientos de árboles con combinaciones no lineales, revisa las interacciones ocultas y encuentra un patrón. El patrón identificado dice que siempre que esta combinación aparece, el monto tiende a ser Bajo. Por lo tanto, el resultado se inclina a un 100% de probabilidad.

3. Redes Neuronales

Para este método será definido el uso de la data set de adjudicaciones. Inspeccionando los otros dos data set, de contratos no dados o en proceso de anulación, no será prescindible para el análisis que se desea para este caso.

Como se va a replicar el modelo económico similar al Random Forest 3 y al Árbol de decisión 1, se definirá como variable objetivo: montoRango el cual esta definido como ("Bajo", "Medio", "Alto").

Las variables predictoras: tipoEntidad, entGrp2. Se probarán que son relevantes. Se separarán las variables X (predictoras) y Y (objetivo). Para el caso de la red neuronal la variable objetivo (Y): montoRango mediante el uso de las regresoas: tipoEntidad, entGrp2.

Se procedio a la extensión de los datos para un modelo extendido, donde se incluyeron más variables relevantes:

- ✓ categorías
- ✓ tipoEntidad
- ✓ entGrp2
- ✓ mesDePublicación
- ✓ añoDePublicación
- ✓ mesDeAdjudicación
- ✓ añoDeAdjudicación
- ✓ modalidad
- ✓ subModalidad

Todas estas variables se pueden convertir a one-hot encoding, lo que permite que la red neuronal vea patrones reales. Se hará con 3 capas densas.

Capa oculta 1: 16 neuronas
Capa oculta 2: 8 neuronas
Capa de salida: 3 clases (Bajo, Medio, Alto)

Se observa que los parámetros entrenables son de 259, lo que permite tener un modelo ligero, rápido y no sin sobreajustes. Se hará un entrenamiento con 20 epocas, batch size de 32 y una validación del 20%

Predicción MODELO A: Predecir entGrp2

- ✓ tipoEntidad
- ✓ entGrp2 (objetivo)
- ✓ categorías
- ✓ modalidad
- ✓ subModalidad

Aquí la variable objetivo es entGrp2, igual que en Árbol A. Caso hipotético a predecir con la red neuronal

- ✓ tipoEntidad = "Entidades Descentralizadas, Autónomas y de Seguridad Social"
- ✓ entGrp2 = "Descentralizada"
- ✓ Distribución real en el dataset:

Bajo = 35.77%
Medio = 33.41%
Alto = 30.82%

Pero en este caso lo que queremos es:

tipoEntidad = "Entidades Descentralizadas, Autónomas y de Seguridad Social"
entGrp2 = "Descentralizada"
categorías = "Salud e insumos hospitalarios"
modalidad = "Compra Directa con Oferta Electrónica (Art. 43 Ley)"
subModalidad = "Desconocido"

El resultado coincide con lo encontrado en los métodos anteriores.

Predicción MODELO B: Predecir subModalidad

En el Árbol B se utilizó estas variables predictoras:

- ✓ montoRango (pero ojo: aquí montoRango es entrada, no salida)
- ✓ tipoEntidad
- ✓ entGrp2
- ✓ categorías

La variable objetivo era predecir que subModalidad ocurrirá dadas las características anteriores. Como se observó en anteriores análisis para el escenario B no contiene información predictiva. El árbol original ya había advertido ese escenario al considerar montoRango, que casi no se relaciona con tipoEntidad + entGrp2 + modalidad + subModalidad.

Es conclusión como herramienta predictiva no se cuenta con una buena correlación entre las variables, pero como una herramienta descriptiva si se puede encontrar patrones interesantes como los definidos anteriormente.

Las variables del Escenario B no explican el montoRango, y es por esto que el el árbol de decisión A funcionó excelente mientras que el Arbol de decisión B no mantuvo las características esperadas. La red neuronal B también quedo en un nivel muy bajo de predicción.

Conclusiones

1. El método de árbol de decisión es una herramienta excelente cuando se necesita interpretabilidad. Sin embargo, tienden a sobre ajustarse si no se controlan parámetros como profundidad y mínimo de muestras.

2. El modelo de Bosques aleatorios, represento un modelo más robusto, estables y con excelente desempeño general. Pierden interpretabilidad comparado con un árbol solo, pero ganan mucha precisión.

3. El método de redes neuronales, demostró más potencia, pero requiere más datos y se incrementa la complejidad.

Hallazgos clave

El árbol de clasificación que modela el rango de monto revela que la variable determinante para explicar el comportamiento económico del gasto es el tipo de entidad. Se evidencia que:

- **Los montos bajos** se concentran principalmente en el sector público general, incluyendo empresas públicas e instituciones del Estado.
- **Los montos medios** se asocian con entidades descentralizadas y con dependencias de la administración central, así como con ONG.
- **Los montos altos** se encuentran fuertemente ligados a gobiernos locales (municipalidades y mancomunidades), validando la conclusión de que el mayor volumen de inversión y obras de gran escala proviene del ámbito municipal.

Este patrón coincide exactamente con los resultados del análisis descriptivo y las conclusiones del documento base, confirmando la coherencia del comportamiento del gasto público según la naturaleza institucional.

En el modelo Random Forest, utilizando las variables tipoEntidad y entGrp2 para predecir el rango de monto adjudicado, obtuvo un error OOB de 58%.

Este desempeño confirma que, aunque la naturaleza institucional influye parcialmente en el nivel de gasto, especialmente en los montos bajos, no es suficiente para predecir con precisión rangos medios y altos.

El modelo acierta con frecuencia en montos bajos (77%), pero presenta alta confusión en montos medios y altos, lo cual coincide con los hallazgos del documento técnico, donde se indica que los rangos medios se distribuyen entre diversas instituciones y que los montos altos están fuertemente vinculados al sector municipal.

La dificultad del Random Forest para identificar montos altos se explica por la baja representación de la clase municipal dentro del total del dataset y por la alta aleatoriedad en los árboles, situación que impide una separación sistemática de

esta categoría. Por lo cual, en conclusión, el Random Forest valida que la estructura institucional no determina completamente el nivel de gasto, y que la relación entre tipo de entidad y monto adjudicado es parcial, heterogénea y dependiente del contexto específico de cada proceso.

Las variables seleccionadas en el estudio tipoEntidad, entGrp2, modalidad, subModalidad no muestran relación que se sustenta con estadística significativa con la variable montoRango. Los modelos entrenados bajo este escenario, incluyendo redes neuronales y árboles de decisión, obtienen resultados equivalentes al azar, lo cual evidencia ausencia de patrones útiles para la predicción estadísticamente significativa.

Se podría presentar algún tipo de desbalance, en la clase Alto, que contiene el doble de registros que Bajo o Medio, afectando las métricas y generando sesgos. Incluso con técnicas de escalado y codificaciones apropiadas, la red neuronal no logra mejorar la precisión debido a la falta de una señal real en los datos.

Propuestas

El análisis de minería de datos confirma que la mayoría de las interacciones observadas tienen lugar dentro del Sector Público, el cual constituye el entorno dominante. En este contexto, las Entidades Descentralizadas emergen como el componente más activo y voluminoso del sistema, destacando el Instituto Guatemalteco de Seguridad Social (IGSS) por su solidez, continuidad operativa y vínculos institucionales confiables. Se sugiere evaluar la eficiencia en el uso de estos recursos con el fin de asegurar un desempeño institucional adecuado.

El estudio también evidencia notables diferencias entre entidades públicas: el Ministerio de Salud Pública concentra adjudicaciones en rangos de bajo a mediano valor, mientras que las Municipalidades y Gobiernos Locales participan en operaciones de mayor cuantía, lo que refleja diferencias tanto en sus funciones como en su capacidad operativa. En este sentido, se recomienda implementar auditorías que aseguren un control efectivo sobre los recursos e inversiones, favoreciendo una gestión presupuestaria más precisa.

Asimismo, se plantea la inclusión de nuevos atributos con mayor grado de correlación con el monto de adjudicación, como el tipo de producto, proveedor, historial de adjudicaciones o frecuencia de compra por entidad, lo cual podría originar un escenario B2 con capacidad predictiva significativa.

La aplicación de análisis estadísticos como pruebas de correlación, ANOVA y evaluación de importancia de variables permitirá anticipar la relevancia explicativa de ciertos atributos, evitando así la construcción de modelos carentes de soporte estadístico.

Recomendaciones

1. El método de árbol de decisión se debe utilizar cuando se necesita explicar decisiones simples. Para un análisis mas complejo o profundo quizás sea mejor la utilización de otros métodos más robustos.

2. El método de bosque aleatorio es Ideal como un modelo para evaluar problemas reales en cuestiones de desempeño, ya que permite dar buenos resultados con una complejidad menor.

3. Las redes neuronales se deben utilizar solo si se cuenta con una buena capacidad de cómputo y datos suficientes. un problema donde realmente exista complejidad no lineal.

Bibliografía

Potin, L., Figueiredo, R., & Labatut, V., Largeron, C. (2023). *Pattern Mining for Anomaly Detection in Graphs: Application to Fraud in Public Procurement*. arXiv. https://doi.org/10.48550/arXiv.2306.10857.

Torres-Berrú, Y., Rodríguez, A., & Paredes, L. (2021). *Data Mining to Identify Anomalies in Public Procurement: A Case Study from Ecuador*. Electronics, 10(22), 2873. https://doi.org/10.3390/electronics10222873

Concursos Guatecompras: Sitio Web consultado noviembre de 2025: https://datos.minfin.gob.gt/dataset/concurso-guatecompras-2016-a-2022_