# Problema a resolver

Rest-on es una guía turística que está armando en su web una sección de restoranes, que sirva para asesorar a sus clientes en las características y calidad de los restoranes. A estos efectos, los contratan a ustedes para apoyarlos en los algoritmos de soporte de su web, con dos objetivos:

1. Poder recomendarles a sus clientes restoranes similares a los que han consultado
2. Rankear a los restoranes, en términos relativos a los restoranes comparables.

A estos efectos, cuenta con una base de datos de restoranes, a la cual se puede acceder en el siguiente link: <https://archive.ics.uci.edu/ml/datasets/Restaurant+%26+consumer+data>

# Entendimiento del caso de negocio y planificación del trabajo

## Discuta el desafío planteado en términos de la importancia que tiene para una empresa cualquiera y para Rest-on en particular.

Ofrecer recomendaciones que generan valor para los clientes es un objetivo para Rest-ON. No es solo recomendar una opción al cliente, sino que esta desarrolle una conexión con la marca y los clientes vean en la empresa un asesoramiento fundamentado. Por tanto, es objetivo realizar un este análisis permitiendo que los clientes no solo encuentren lo que están buscando, sino que se identifiquen con el servicio.

Este desafió está presente en todas aquellas empresas en su valor de negocio esté vinculado con la fidelidad de los clientes que trabajan con la marca.

## Identifique conceptual y técnicamente qué tipo de modelos serían necesarios para atender este problema

Para ofrecer una información que genere valor para el cliente se deben clasificar correctamente los restaurantes a partir de las características que estos poseen y los servicios que ofrece, también se puede tener en cuenta las características propias de los usuarios que visitan los restaurantes que permita realizar una segmentación en función de las características de los usuarios.

Todo dependerá del análisis del negocio detallando donde se identifique las variables que pueden formar parte de la segmentación de los restaurantes y poder ofrecer una calidad del servicio superior a los turistas que vienen a recibir la experiencia de Rest-ON

## Plantee hipótesis respecto a qué dimensiones de análisis y variables pueden ser relevantes para lograr el objetivo

Analizando el modelo de negocio de la empresa planteamos como hipótesis que:

Para ofrecer mejores recomendaciones es necesario conocer las características de los restaurantes, entender que tipos de cliente asisten y establecer las peculiaridades de la comida que ofrecen

**Dimensiones propuestas:**

* Características de los restaurantes
* Características de la comida
* Datos de los clientes que concurren

**Como variables dentro de cada dimensión identificamos las siguientes:**

* Características de la comida: Esta dimensión abarca las particularidades de la comida respecto al tipo de menú, la forma en que se sirve, restricciones alimenticias, entre otros.
* Características de los restaurantes – aquí se recoge la tipología del restaurante, en relación a aspectos operativos y físicos, tales como ubicación, disponibilidad de estacionamiento, medios de pagos aceptados, y otras cualidades que lo hagan atractivo.
* Datos del cliente: Deberíamos conocer el perfil y las costumbres de los clientes que concurren al restaurante, que nos permita clasificarlos teniendo en cuenta características de los clientes que concurren y en qué modo lo hacen.

*Ejemplo: Si es con familia, o amigos, si prefieren un ambiente relajado, etc*

## Identifique, en base a lo anterior, qué tipo de datos precisaría obtener para llevar a cabo este trabajo

|  |  |  |
| --- | --- | --- |
| Dimensión | Atributos | Observaciones |
| Características de la comida | Servicios ofrecidos. Almuerzo, Cena, Merienda, Menú Ejecutivo | Se consultará por separado por cada uno de los servicios que ofrece. |
| Consumo de bebida alcohólica (Ejemplo Vino, cerveza, tragos) | Cuál es la bebida alcohólica que más se consume. |
| Estilo de la comida: italiana, Mediterránea |  |
| Tipo de servicio: A la mesa, Bufet, fast-food |  |
| Restricciones alimenticias | Se consultará por separado disponibilidad de menú para celiacos, vegetarianos, veganos, diabéticos |
| Características de los restaurantes | Medios de pagos | Listado de opciones de medios de pago con los que cuenta el restaurante |
| Ubicación Geográfica | Coordenadas geográficas para medir cercanías entre restaurantes |
| Ambiente / Estilo | Se refiere al diseño o ambiente. Ej. Con espacios abiertos, vintage, naturalista, entre otros. |
| Música ambiente o No |  |
| Rango de precios | Se clasificarán los precios en categorías. |
| Área de juegos para niños |  |
| Área para fumadores |  |
| Calificación otorgada al servicio (rating 5 estrellas) |  |
| Disponibilidad de parking |  |
| Datos del cliente | Distribución de edades | Por rangos |
| Horarios de mayor concurrencia |  |
| Estado civil |  |
| Con quien concurren: Con familia, con hijos, con pareja, con amigos |  |
| Sexo de mayor concurrencia |  |
| Frecuencias de visitas (promedio de repetición mensual) |  |

## Arme un plan de trabajo, identificando las diferentes etapas y actividades para lograr el objetivo

Se propone para el desarrollo de esta investigación utilizar la metodología CRISP–DM que abarca desde el entendimiento del negocio hasta el despliegue de la solución sin dejar de lado las tareas de gestión de proyectos que permiten el monitoreo y control del tiempo, costos y riesgos asociados.

De acuerdo con la metodología CRISP-DM se proponen las siguientes fases:

|  |
| --- |
| 1. Comprensión del negocio |
| 1. Comprensión de los datos |
| 1. Preparación de los datos |
| 1. Modelado |
| 1. Evaluación |
| 1. Implementación |

**Comprensión del negocio**: El plan para esta fase consta de comprender el negocio y los objetivos comerciales de la empresa. Adicionalmente abarcaremos tareas tales como describir el problema que se desea resolver, la naturaleza del negocio y el aporte que se desea obtener con el trabajo.

**Comprensión de los datos**: Se realizará un análisis de todos los dataset, revisando las variables y cantidad de registros de cada uno de ellos, presencia de N/A, estructura de los datos, errores en los mismos y utilidad que aporta cada una de las variables para el objetivo del trabajo.

**Preparación de los datos:** En función de los resultados obtenidos en la fase anterior trabajaremos en la transformación de los datos para obtener las variables en el formato adecuado, crear nuevas a partir de la transformación de las ya existentes, y corrección mediante alguna metodología de los registros con NA.

**Modelado:** En esta etapa se evaluarán los modelos a utilizar para realizar el análisis de cluster de los restaurantes. Por la naturaleza del problema se tratará de un modelo no supervisado. Evaluaremos modelos aglomerativos y divisivos.

**Evaluación:** Se utilizarán técnicas como silueta y elbow para ajustar los modelos y verificar los resultados.

**Implementación:** Por último se pondrá el modelo a disposición para ser utilizado evaluando la forma más eficiente de hacerlo.

Si bien todas las etapas comentadas anteriormente se hacen en forma cronológica, el modelo prevé la posibilidad de regresar a una fase anterior para revisarla y retomar nuevamente el proceso.

Gestión de proyecto:  
Para la implementación del proyecto se realizara entre el equipo un control de toda la versión del proyecto y se genera un repositorio para el seguimiento y control de código. Utilizando buenas prácticas reduce la posible generación de errores en los controles de cambio.

|  |  |  |  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
| Fase | sem - 1 | | sem - 2 | | sem - 3 | | sem - 4 | | sem - 5 | | sem - 6 | |
| Comprensión del negocio |  |  |  |  |  |  |  |  |  |  |  |  |
| Comprensión de los datos |  |  |  |  |  |  |  |  |  |  |  |  |
| Preparación de los datos |  |  |  |  |  |  |  |  |  |  |  |  |
| Modelado |  |  |  |  |  |  |  |  |  |  |  |  |
| Evaluación |  |  |  |  |  |  |  |  |  |  |  |  |
| Implementación |  |  |  |  |  |  |  |  |  |  |  |  |

## ¿Cómo será utilizado el resultado del trabajo por el cliente?

El resultado o solución va a ser utilizado por la agencia turística para incluirlo en su motor de búsqueda y ofrecer un servicio que beneficie a sus usuarios, generando más volumen de visitas a su sitio y fidelizándolos. También será usado por los asesores de la empresa utilizarán el sitio para ofrecer a los vacacionistas mejores recomendaciones según sus necesidades y las características de los restaurantes.

Los restaurantes en la medida que este motor de búsqueda y recomendaciones tenga mayor difusión se beneficiaran por tener mayor publicidad y el perfil de clientes que se sienta más cómodo respecto la propuesta que ofrece.

El usuario en la medida que el motor de búsqueda sea reconocido y fiable en cuanto a las recomendaciones podrá disfrutar de opciones más adecuadas a sus preferencias.

## En base a lo anterior, ¿Qué áreas se deberían involucrar en el proyecto y qué rol cumplirían?

Para la ejecución exitosa del proyecto tendría que participar todas las áreas que estarán en el ciclo completo del proyecto, desde la definición hasta su puesta en producción y mantenimiento del proyecto. En tres las áreas identificadas están comercial, marketing y el área de soporte técnico que son quienes más conocen a los usuarios de la guía. Por otro lado el área de operaciones y IT al conocer como manejar la gestión de la empresa realizarán la captura de datos y recopilación de datos de los diferentes restaurantes. Finalmente, el área financiera deberá verificar los impactos económicos de las iniciativas y el equipo de Analistas de Datos para interactuar con los restantes interlocutores.

# Extracción, Transformación y Carga de datos

## Diseñe la estructura de tabla datos analítica: ¿qué va a ser cada fila? cuáles van a ser las columnas.

## ¿Cómo se construye cada una de las variables en la tabla de datos? ¿De qué tablas surge y qué transformaciones de datos son necesarias?

## Construya la tabla de datos analíticas en base a la especificación anterior

La tabla analítica resultante de esta etapa contará de ID de restaurantes en las filas y variables en las columnas.

En el archivo adjunto “Datos complementarios.xlsx” se especifican las características de las variables seleccionadas, cuál será la fuente de información y que transformación se le hará a cada una.

Finalmente se ensamblan todas las variable en la tabla analítica a partir de los ID de restaurantes de la tabla de geoplaces. Esto es debido a que a pesar de que tiene notoriamente menos registros que los restantes dataset, contiene información muy relevante al momento de realizar el cluster, por lo que decidimos reducir la cantidad de restaurantes en pos de enriquecer el análisis.

Se trabaja en la corrección de los registros con NA en las variables donde están presentes definiéndolos con un valor diferente a los existentes de forma de tenerlos identificados.

A modo aclaratorio comentamos que los dataset de los usuarios no fueron utilizados ya que no están identificados como usuarios de los restaurantes por lo que no aportan información relevante para el perfil de clientes y adicionalmente no agregan información que ya no este contenida en los datasets de los restaurantes.

# Exploración y análisis descriptivo

## Realice un análisis descriptivo de la tabla de datos.

Se obtiene la tabla con las variables ya definidas y los datos corregidos habiendo chequeado la existencia de ceros y NA.

## ¿Descarta a priori alguna variable a partir de este análisis?

Respecto al medio de pago decidimos reducir la cantidad de variables y quedarnos solo con la variable de cantidad. Anteriormente también habíamos descartado la variable relativa a la disponibilidad de internet ya que si bien es relevante había muy pocos restaurantes que contaban con este registro de información.

## ¿Puede sospechar qué variables serán útiles para el problema identificado?

Entendemos que variables relativas a la ubicación, tipos de comidas y precio serán importantes para el análisis.

## ¿Es necesario tomar alguna decisión respecto al universo o a la exclusión de valores atípicos?

Para las variables categóricas no existen datos atípicos ya que la única que presentaba observaciones de este tipo era la de tipos de comida y la reclasificamos reduciendo el número de niveles.

Para las variables numéricas analizamos los outliers a partir de la distancia de Mahalanobis, y los monitorearemos una vez que tengamos los resultados de la clusterización.

# Modelado y Evaluación

## ¿Qué variables pre-seleccionaría en base al análisis funcional y estadístico?

Una vez realizado el análisis anterior en relación a la transformación, eliminación y creación de nuevas variables decidimos quedarnos con todas las que se encuentran en la nueva tabla de datos.

## Estime al menos dos modelos (usando diferentes algoritmos y/o variables), compárelos (en base a criterios estadísticos y funcionales) y seleccione el que entienda más adecuado.

Utilizaremos los algoritmos de hclust como clustering aglomerativo y diana como clustering divisivo.

Para calcular las distancias de Gower utilizaremos Daisy ya que tenemos variables mixtas dentro de la tabla de datos. Adicionalmente utilizaremos PCAmixed para calcular las mismas utilizando como dimensión la cantidad de componentes principales que acumulan el 90% de la varianza.

## Caracterice y describa los resultados a los cuales se arribaron en el caso anterior, desde el punto de vista del negocio.

jhhhhhhhhhhhhhhhhhhhhhhhhhhhhhhhhhhhhhhhhhhhhhhhhhhhhhhjjj

# Implementación

## ¿Cómo entiende que el modelo podría ser utilizado por la web para cumplir el objetivo funcional?

En función de lo transmitido en el requerimiento el usuario debería ingresar el nombre de un restaurante y el sitio web le devolvería los restaurantes similares en función del análisis de clusters realizado

## En base a lo anterior, ¿cómo debería ser la implementación de este modelo (estimación resultados en tiempo real, procesamiento batch, etc)?

Dentro de los tipos de implementación vistos en clase este modelo se podría clasificar como un análisis puntual que no precisa ser corrido con frecuencia. De todas formas sería importante cada cierto tiempo ir enriqueciendo las bases con información de los restaurantes y más aún cuando debimos trabajar solo con 130 registros por no contar con información relevante para los restantes.

## ¿Cómo se podrían armar, usando los resultados de este modelo, los rankings de restoranes “en términos relativos” o “específicos a un grupo”? Si bien no es necesario realizar los desarrollos, identifique el proceso, actividades, variables a utilizar, etc y describa con el mayor detalle posible.

Los resultados del modelo se deberían mostrar en la web al ingresar el nombre de un restaurante por parte del usuario devolviendo todos los registros correspondientes a ese cluster. Entendemos que el cliente nos debería especificar el criterio para mostrar los restaurantes correspondientes a ese cluster pero sugerimos hacerlo por distancia de gower comenzado por la que tenga una menor, considerando que una menor distancia implica mayor similitud entre los mismos.

Otro criterio de ordenamiento podría ser la ubicación geográfica, el tipo de comida recomendando en primer lugar los que ofrecen la misma comida que ingresó el usuario o por rating del restaurante.

## ¿Qué ventaja tendría esto frente a usar un ranking general?

En el caso de la distancia de gower por ejemplo, la ventaja es que si no conocemos específicamente el atributo que según el usuario hace que un restaurante sea similar a otro, este parámetro contempla todas las variables mejorando la probabilidad de devolver uno similar considerando todas las variables contenidas en el dataset.

## ¿Cómo se podría evaluar el resultado (éxito) del trabajo?

Más allá de los métodos estadísticos como la silueta que se muestra en el desarrollo del trabajo, a nivel de negocio sería importante conocer la opinión del usuario luego de haber concurrido al restaurante recomendado de forma de conocer el grado de satisfacción de la recomendación.