Estimación Transit Times

1. Definiciones y especificaciones de requerimientos.

a. Definición general del proyecto:

Bayer es una empresa multinacional con competencias en salud y agricultura. Dispone de una planta modelo de procesamiento de maíz que está ubicada en Ruta 31 Km 170, Rojas, Provincia de Buenos Aires.
Durante las campañas de cosecha se transporta el maíz desde los productores, en distintos establecimientos o campos del país, hacia su planta de procesamiento. El objetivo del proyecto es poder estimar mediante datos históricos la duración de esos viajes según el siguiente detalle:

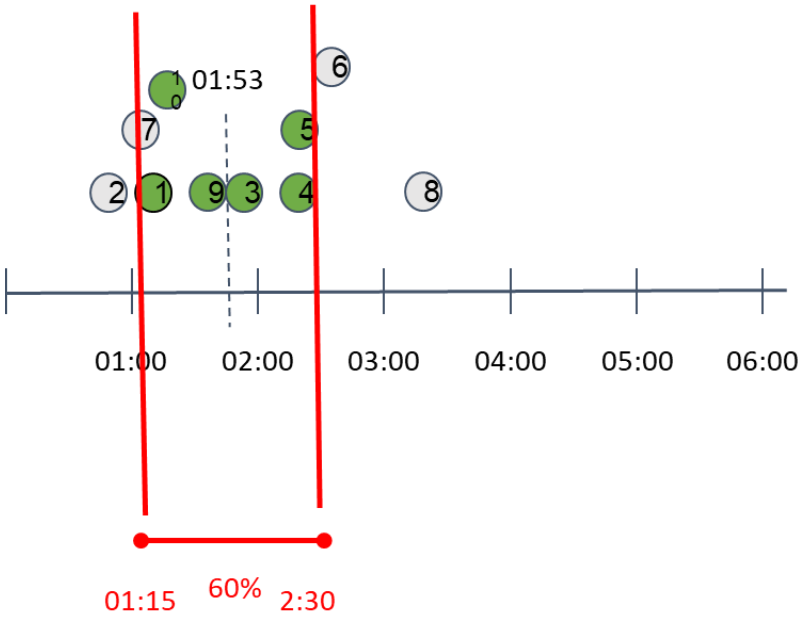
- Establecer la duración mínima y máxima de transit times para cada zona.
- Duración única de un viaje desde el establecimiento o campo, a la planta de procesado.

b. Alcance:

El proyecto inició el 01 julio de 2022 y para su desarrollo se establecen un total de 4 spring de 20 días cada uno.
Para una primera etapa se establece el desarrollo e implementación de los modelos que den solución a los objetivos planteados. Seguido de una etapa de automatización, testing que permitirá realimentación continua y flexibilización para realizar cambios que los usuarios crean pertinentes y por último el despliegue.
Se prevé su finalización y puesta en marcha antes del inicio de la campaña de este año.

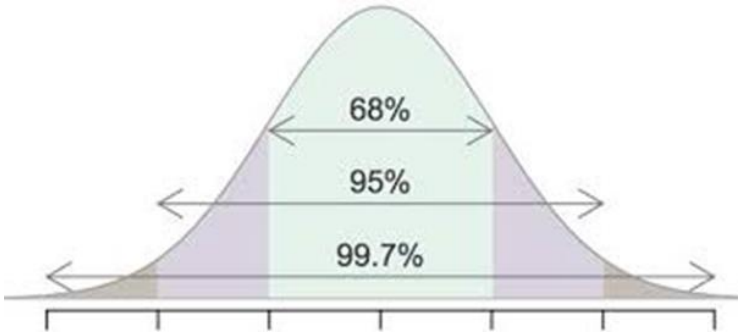
2. Modelo Estimación Transit Times Mínimos y Máximos

La estimación se establece a partir de 2 métodos para calcular los transit times mínimos y máximos de duración de viajes para las Zonas y Establecimientos:



En la gráfica anterior podemos observar una representación de los tiempos límites. Cada pelota es un viaje tomado al azar de datos históricos. Los viajes que tienen una duración entre estos valores límites son los viajes que llegan “a tiempo” mientras que los que llegan antes del mínimo establecido llegan “temprano” y contrariamente los que llegan después del tiempo máximo son viajes que llegan “tarde”. Los métodos para su calculo son:

- **Método Estadístico:** en base a datos históricos se establecen los valores de transit time mínimo y máximo para garantizar un porcentaje de datos entre dos intervalos de confianza alrededor del desempeño medio .



La información histórica juega un rol clave para este modelo y conforma la base con la que se obtiene la distribución de probabilidad de cada subzona y se calcula los intervalos de confianza para un nivel de significancia establecido. Por defecto este nivel de significancia sugerido es de 70% siendo el nivel óptimo. Los márgenes restantes son un 10% los transit times que llegan antes, 10% llegadas tardes. Por último, un 10% de los datos restantes se corresponden con valores atípicos (accidentes, incidentes, roturas, demoras exageradas, etc).

Para poder realizar esta inferencia mientras más grande sea la muestra de datos históricos de cada establecimiento, más se asemejará al desempeño general, por este motivo se establece un tamaño mínimo de muestra de 30 datos.

Este modelo se caracteriza por:

- Su comportamiento basado en datos (hechos).
- Se ajusta al desempeño general alrededor del promedio de duración de los viajes.
- Necesita de al menos 30 datos históricos, para tener confianza en los resultados.
- Mejor adaptación para zonas con gran variabilidad de distancias.
- Flexible, adaptable al desempeño general de los viajes.
- Parámetro porcentaje de % confianza.

- **Método algorítmico:** en este caso para calcular los valores mínimos y máximos de transit time se utilizan 5 parámetros.

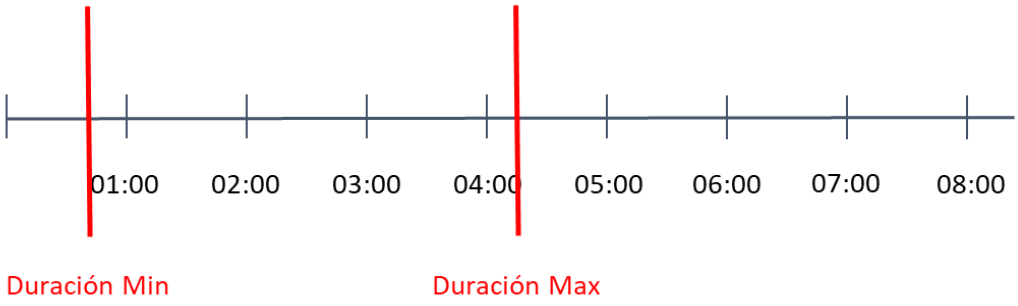
  Los primeros parámetros surgen de los procedimientos de descansos para cada viaje. Se establece entonces la lógica según lo indicado por el procedimiento de negocio, donde para cada 3 hs de viaje corresponde 1 hs de descanso. Corresponde entonces el parámetro

  - **hs_manejo**: cantidad de horas continúas permitidas sin descanso (3 por defecto).
  - **hs_descanso**: cantidad de horas de descanso luego de las horas permitidas de manejo.
  - **vel_media**: es la velocidad que se utiliza para calcular el descanso para cada establecimiento. Se calcula la duración de viaje que surge de dividir la distancia del establecimiento por esta velocidad y se aplica la lógica de negocio.

  El cuarto y quinto parámetro sirven para establecer la velocidad constante con las que se calcularán las duraciones mínimas y máximas para cada zona y establecimiento.

  Para los casos de un establecimiento en particular las ventanas de duración se calculan:

Min = (Distancia/ Vel Max) + Descanso Max=(Distancia/ Vel Min) + Descanso



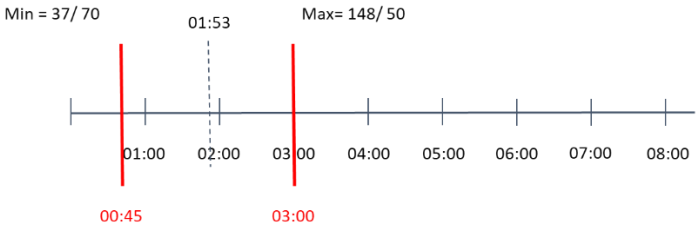
  Entonces la duración mínima del viaje esta dada por la distancia del establecimiento a planta de Bayer dividido su velocidad máxima más el descanso que le corresponde. Similar es el calculo de la duración máxima, pero sustituyendo la velocidad máxima por la velocidad mínima.

  El otro caso es para obtener las ventas de duración para una zona en donde seguimos el mismo esquema, pero considerando la menor distancia para duración mínima y la mayor distancia para la duración máxima de los establecimientos que conforman la zona, por ejemplo una zona con 2 establecimientos:

	Establecimiento	Distancia
Min	El Largo	37
Max	Santa Escolastica	148

F2

Vel Min = 50 km/h
Vel Max = 70 km/h
Vel Promedio = 60 km/h
Descanso = 1hs cada 3 de viaje

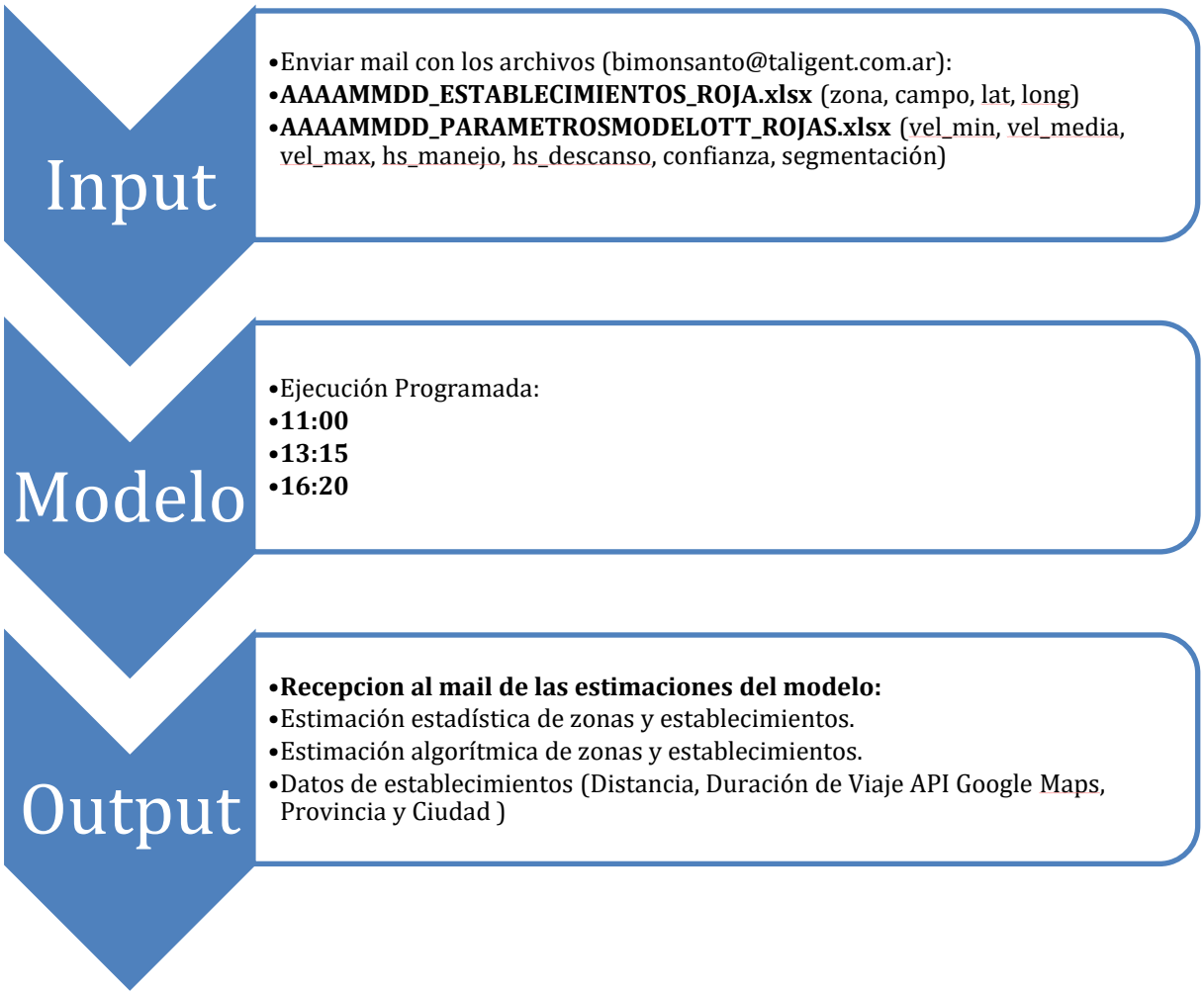


Características del método algorítmico:

- Comportamiento basado en distancia y velocidad.
- Indistinto al desempeño promedio.
- No necesita datos históricos.
- Ventanas mayores para zonas con gran variabilidad de distancias.
- Fijo, constante, no cambia con el desempeño general de los viajes.
- Parámetro Vel Min y Vel Max.

Ejecución del Modelo Estimación Transit Times Mínimos y Maximos

La ejecución del modelo de transit times está automatizado y programado para ejecutarse 3 veces por día en el horario de **11:00, 13:15, 16:20**, solo se ejecutara si encuentra archivos nuevos para ser procesados. A continuación, detallaremos los pasos a seguir:



a. Un usuario deberá enviar un mail a la casilla de correo **bimonsanto@taligent.com.ar**, adjuntando los inputs del modelo:

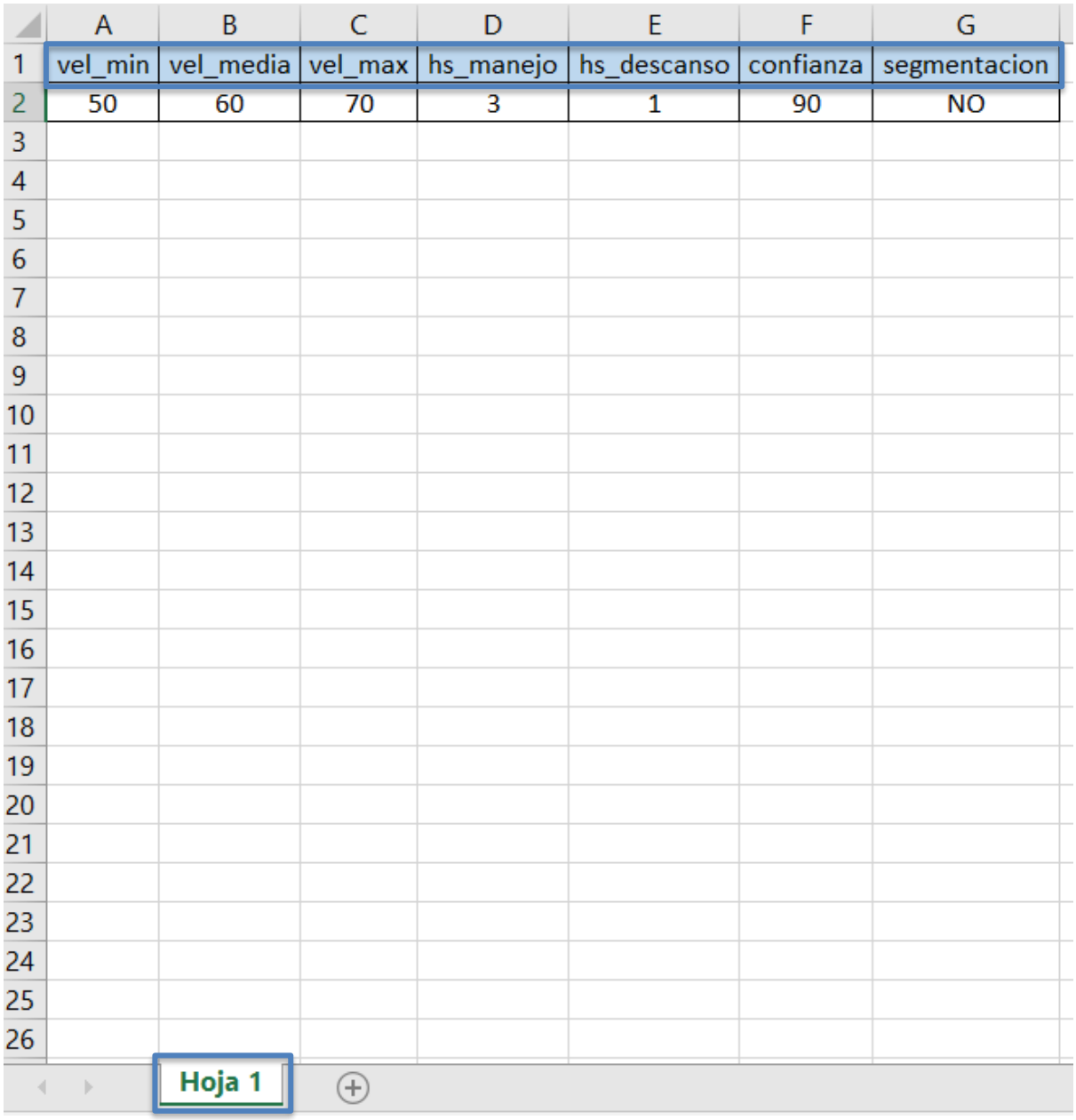
- Un archivo de nombre **"AAAAMMDD_ESTABLECIMIENTOS_ROJAS.xlsx"** (donde AAAA es el año en formato completo - MM número del mes con dos dígitos - DD numero de dia con dos dígitos de la fecha de envío del archivo). Nótese que este es un archivo de Excel. Contiene en la primer hoja los datos de **"Zona"**(Letra que caracteriza la zona logística a la cual pertenece el establecimiento), **"Campo"** (Nombre del establecimiento) este es el valor clave para su unión con los datos históricos y por último una columna con nombre **"Lat"** y **"Long"** que contiene la latitud y longitud indispensable para la geolocalización. Veamos un ejemplo:

	A	B	C	D
1	Zona	Campo	Lat	Long
2	C	Don Miguel	-27,76027778	-65,47388889
3	C	Luces del Noa	-27,932264	-65,463123
4	C	Maranzana	-27,86668725	-65,295075
5	C	Mistol Ancho	-27,970565	-65,39233575
6	C	San Antonio	-27,89611111	-65,31972222
7	E	Mana	-32,83728912	-59,75664138
8	E	Maria Lola	-32,8435926	-59,79777521
9	F5	Chiapero	-32,634937	-65,190027
10	F2	Don Pedro	-33,66049722	-60,42078889
11	F2	Don Reinaldo	-34,634764	-60,057856
12	F2	Don Tuco	-34,03723	-59,997533
13	F2	Doña Clementina	-33,96950833	-59,95343611
14	F6	El Buho	-30,684997	-63,559947
15	F2	El Clavo	-34,583686	-59,945333
16	F1	El Largo	-34,034475	-60,96463333
17	F3	El Milagro	-35,845565	-60,492036
18	F4	El Mirador Este	-33,2852	-63,45895
19	F4	El Piquete	-33,326293	-63,838094
20	F5	Emetres	-32,501117	-65,177378
21	F2	La Amalia II	-34,641692	-60,110969
22	F3	La Aurora	-32,828662	-59,704698
23	F3	La Cabaña 2	-32,76126667	-59,83988056
24	F5	La Candelaria	-32,44135	-65,077981
25	F2	La Capitana	-33,708911	-60,093586
26	F2	La Federala	-34,35939	-59,732653
27	F2	La Genoveva	-33,634366	-60,067745
28	F5	La Karina	-32,625049	-65,148419
29	F3	La Malena	-32,74269444	-59,83466111

Hoja 1

A partir de estos datos, se realiza una búsqueda en la **API de Google Maps**, geolocalizando los campos por su Lat y Long. La finalidad es contar con una fuente de información fiable y unificada. Luego de este proceso se obtiene la distancia a la planta, en km de Bayer, la duración promedio del viaje en condiciones óptimas, la ciudad de origen y la provincia de origen.

- Opcionalmente se podrá enviar un segundo archivo con los parámetros configurables del modelo. El nombre del archivo debe ser "**AAAAMMDD_PARAMETROSMODELOTT_ROJAS.xlxs".** Veamos un ejemplo:



	A	B	C	D	E	F	G
1	vel_min	vel_media	vel_max	hs_manejo	hs_descanso	confianza	segmentacion
2	50	60	70	3	1	90	NO

Hoja 1

Estos parámetros permiten adaptar el modelo a las necesidades del negocio, estos son, **“vel_min”** establece la velocidad mínima para el cálculo de la duración máxima de llegada del método algorítmico, **“vel_max”** similar al anterior pero para la duración mínima de llegada a planta, **“vel_media”** se utiliza para calcular los tiempos de descansos, **“hs_de_manejo”** (debe ser entero) hace referencia a la cantidad de horas de manejo permitidas sin descanso, **“hs_de_descaso”** (debe ser entero) establece las horas de descanso permitidas luego de cumplir con las horas de manejo en el ejemplo anterior vemos que cada 3 horas de manejo corresponde 1 de descanso, **“confianza”** es el intervalo de confianza para calcular las ventanas a partir del método estadístico y por último “**segmentación”** (recibe sólo dos parámetros “SI” o “NO”) si se establece en “SI” el método aplicara la segmentación o agrupación de las subzonas mediante un algoritmo de machine learning.

b. Luego del envío de los archivos el modelo se ejecutará en los horarios programados y los resultados obtenidos serán enviados al mail de los usuarios, previamente establecidos. El formato de salida es un archivo Excel que contiene 5 libros:

- Libro **"T.T Subzonas Est"**, contiene la estimación estadística del transit times mínimos y máximos para las distintas zonas.

	A	B	C	D
1	**T.T Subzonas Estadistico**			
2	**SubZone**	**T.T Min**	**T.T Max**	**Confianza**
3	C	16:30	23:00	Si
4	E	05:00	07:00	Si
5	F1	00:30	01:30	Si
6	F2	01:30	03:00	Si
7	F3	05:00	06:30	Si

- Libro **"T.T Subzonas Const"**, contiene la estimación algorítmica del transit times mínimos y máximos para las distintas zonas.

	A	B	C
1	**T.T Subzonas Cosntante**		
2	**Zona**	**T.T Min**	**T.T Max**
3	C	18:00	26:00
4	E	05:30	08:00
5	F1	00:30	02:00
6	F2	02:00	04:30

- Libro **"T.T Establecimientos Est"**, contiene la estimación estadística de las ventanas mínimas y máximas para cada una de los campos del input.

	A	B	C	D	E
1	T.T Establecimientos Estadistico				
2	Establecimient	T.T Min	T.T Max	Confianza	Descanso
3	Don Miguel	16:16	22:38	Si	05:30
4	Luces del Noa	17:02	22:21	Si	05:30
5	Maranzana	18:19	25:39	Nuevo	05:30
6	Mistol Ancho	18:11	25:28	Nuevo	05:30
7	San Antonio	16:31	22:59	Si	05:30

- Libro **“T.T Establecimientos Est”**, contiene la estimación algorítmica de las ventanas mínimas y máximas para cada uno de los campos del input.

	A	B	C	D
1	T.T Establecimientos Constante			
2	Establecimient	T.T Min	T.T Max	Descanso
3	Don Miguel	18:37	26:04	05:30
4	Luces del Noa	18:28	25:52	05:30
5	Maranzana	18:19	25:39	05:30
6	Mistol Ancho	18:11	25:28	05:30

- Libro **“Establecimientos”**, contiene información de utilidad de los campos, extraída de la geolocalización de la API de Google Maps a partir de la Lat y Long. Si en el Input inicial hay una dirección de ubicación (Lat, Long) que no es correcta estos campos serán ceros. En tal caso se deberá revalidar la latitud y longitud e iniciar de nuevo con el paso 1, para obtener una salida completa.

	A	B	C	D	E	F	G	H	I
1	Datos Establecimientos								
2	Establecimiento	SubZona	Provincia	Ciudad	Latitud	Longitud	Distancia a Planta	uracion Goog	Descanso
3	Don Miguel	C	Tucumán	La Cocha	-27,76027778	-65,47388889	978	11 h 18 min	05:30
4	Luces del Noa	C	Catamarca	Santa Rosa	-27,932264	-65,463123	970	11 h 11 min	05:30
5	Maranzana	C	Tucumán	Via sin nombre	-27,86668725	-65,295075	962	11 h 3 min	05:30

- Libro **“Parámetros Modelo”**, mostrara configuración de los parámetros con el que ha sido ejecutado el modelo.

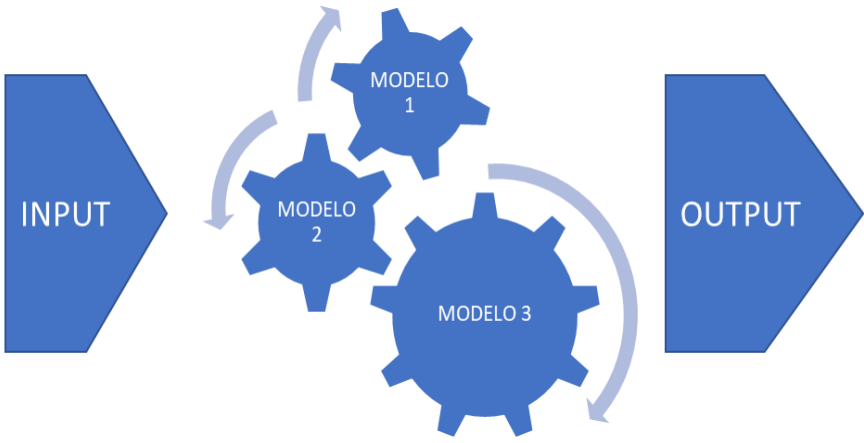
	A	B	C	D	E	F	G
1	Parametros de configuración del modelo						
2	vel_min	vel_media	vel_max	hs_manejo	hs_descanso	confianza	segmentacion
3	50	60	70	3	1	80	SI

3. Modelo Estimación Transit Time

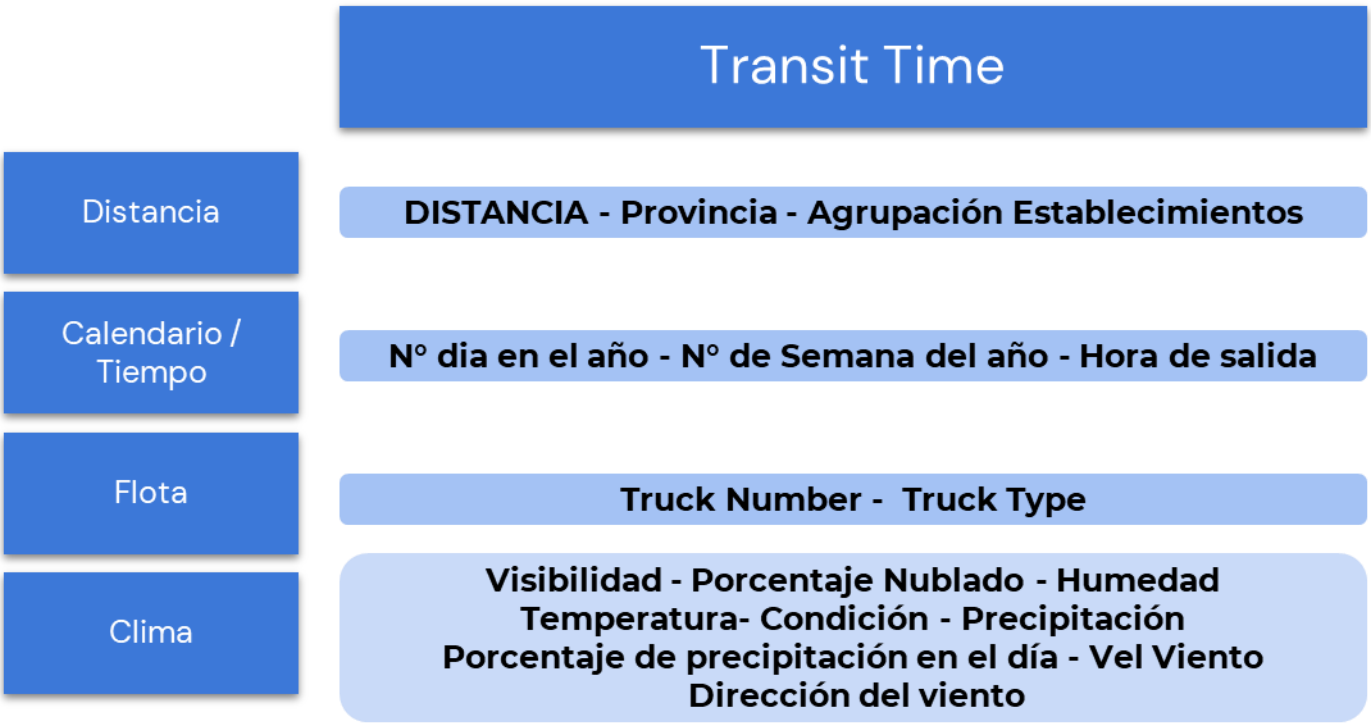
El objetivo de este modelo es estimar a través de información histórica, los tiempos de viajes de los camiones que transportan maíz desde los Establecimientos a la planta de procesamiento de Bayer “María Eugenia”, en resumen, conocer cuanto demorara un futuro viaje según ciertas variables y condiciones. Para lograrlo se realizo un proceso de DataScience para desarrollar un modelo de Machine Learning que realiza la regresión (estimación). Se ejecutaron las siguientes fases de manera iterativa incremental durante el desarrollo:

- Carga de Datos
- Análisis Exploratorio de Datos
- Limpieza de Datos
- Ingeniería de Variables
- Selección de variables
- Implementación de modelos
- Evaluación y Optimización de modelos
- Interpretación de Resultados

Se probaron 10 algoritmos de Machine Learnig, se evaluaron mediante su error cuadrático (MSE), que mide el promedio de los errores elevados al cuadrado (error = tiempo real – tiempo estimado por modelo). Se seleccionaron los 3 modelos con mejor métricas en su evaluación y se elaboro un tercer modelo de ensamble **“stacking”.** Cada modelo produce una predicción diferente. Las predicciones de los distintos modelos se combinan para obtener una única predicción. La ventaja que obtenemos al combinar modelos diferentes es que como cada modelo funciona de forma diferente, sus errores tienden a compensarse. Esto resulta en un menor error de generalización.



Como se observa en la imagen para poder realizar una predicción debemos partir de datos de entrada o variables. Estas variables fueron definidas por la importancia que tienen para la duración de un viaje. El input del modelo es el siguiente:



Detallaremos esas variables para comprender mejor su importancia en la estimación:

- Distancia: es la distancia en km de un campo especifico a la planta de Bayer.
- Provincia: corresponde a la provincia argentina donde esta ubicado el campo.
- Agrupación de Establecimientos: es un grupo generado a partir de una segmentación (agrupación) de los campos según su distancia a planta y la duración promedio de viaje extraído de la API de google maps. Esto permite tener información esencial para generalizar correctamente las predicciones de nuevos campos y de los cuales no hay datos históricos.
- Numero de día del año: es un valor entre 0-365 donde 0 es el primer día del año y 365 el ultimo día del año.
- Numero de semana del año: es un numero 1 – 52 que informa el numero de la semana de salida del viaje.
- Hora de salida: tiempo en minutos del horario de salida del camión.
- Truck number: es el identificador único del camión de Bayer.
- Truck Type: es el tipo de camión.
- Visibilidad: visibilidad de km en el día.
- Porcentaje nublado: porcentaje que se mantiene nublado en el día.
- Humedad: porcentaje de la humedad relativa en el día.
- Temperatura: temperatura mínima del día.
- Precipitación: cantidad estimada de precipitación a caer en el día.
- Porcentaje de precipitación: Porcentaje del día que estuvo lloviendo.
- Vel Viento: velocidad promedio del viento del día.
- Dirección del viento: dirección del viento en grados.

Funcionamiento Modelo Estimación Transit Time

El modelo recibe como entrada las siguientes variables:

Nro	SubZone	Establecimiento	Start Dt.	Truck N°	Truck Type
1	X	San Manuel	29/8/2022 08:00	100356	CH
2	U1	Anita	30/8/2022 12:00	100017	CH
3	X	San Bartolome	29/8/2022 17:00	130956	CH
4	E	Mana	30/8/2022 21:00	100372	CH

A partir de estos datos se completa el resto de información necesaria:

- Distancia, Provincia se busca del maestro de establecimientos generado desde la API de Google Maps.
- Numero de día, numero de semana y hora de salida se genera a partir del Start Dt de salida del viaje.
- Las variables de clima se procesan de la repuesta de la API de Clima https://www.visualcrossing.com, completando lo necesario para el modelo.

Obtener todos los datos necesarios es trasparente al usuario, solo debe conocer los datos iniciales y el resto lo hará el proceso de ejecución del modelo. Por ultimo se obtendrán las estimaciones de duración de viajes de los establecimientos especificados:

Estimacion TT								
Nro	SubZone	Establecimiento	Distancia	Start Dt.	End Dt.	Truck N°	Truck Type	nsit Time Estim
1	X	san manuel	558	2022-08-29 08:00:00	2022-08-29 17:12:12	100356	CH	09:12
2	U1	anita	593	2022-08-30 12:00:00	2022-08-30 21:44:15	100017	CH	09:44
3	X	san bartolome	99	2022-08-29 17:00:00	2022-08-29 18:38:25	130956	CH	01:38
4	E	mana	280	2022-08-30 21:00:00	2022-08-31 02:39:38	100372	CH	05:39

Después de la ejecución del modelo, podemos visualizar la estimación del viaje en la columna “Transit Time Estimado” y saber el día y horario de llegada aproximado a planta en la columna “End Dt”.