Trabajo Fin de Máster

Máster en Ciencia de Datos e Ingeniería de Datos en la Nube

Predicción de duración de dosificación para corrección de colas en procesos automáticos de fabricación de pienso.

Autor: José Luis Casado Valero

Tutor: Luis de la Ossa Jiménez

Co-Tutor:

Octubre, 2021

Dedicado a mi familia, mi apoyo,
mi guía.

Declaración de Autoría

Yo, José Luis Casado Valero con DNI 74512698-N, declaro que soy el único autor del trabajo fin de grado titulado “Predicción de duración de dosificación para corrección de colas en procesos automáticos de fabricación de pienso” y que el citado trabajo no infringe las leyes en vigor sobre propiedad intelectual y que todo el material no original contenido en dicho trabajo está apropiadamente atribuido a sus legítimos autores.

Albacete, a 15 de octubre de 2021

Fdo: José Luis Casado Valero

Resumen

Este trabajo fin de máster está fundamentado en los datos de una aplicación real desarrollada en mi trabajo actual. Se trata de un software de automatización de procesos para una fábrica agroalimentaria de producción de pienso animal.

En la industria agroalimentaria, y en concreto en fábricas de producción de pienso animal, existen gran cantidad de procesos susceptibles de ser automatizados, y ya que la maquinaria utilizada no es de muy alta precisión en cuanto a las mediciones, los errores en cuanto a stock, dosificaciones, estimaciones, etc. son muy habituales.

Una fábrica de pienso en líneas generales es perecida a un "Telepizza": hay una recepción de materias primas (cereales, trigo, soja, grasas, líquidos, …) que son almacenadas en silos (depósitos); mediante unas fórmulas o recetas se marcan los ingredientes que forman cada pienso compuesto; y son los procesos automáticos los que dosifican, mezclan y tratan los ingredientes para formar piensos compuestos, que son expedidos a las granjas para el consumo animal.

En gran parte de los procesos que forman este ciclo de vida, se utilizan diferentes sondas y sensores. Sin embargo, en algunos de ellos no se disponen de esos elementos, por lo que aumenta el nivel de incertidumbre y con esto el número de errores en muchos de las decisiones en tiempo real que hay que tomar en la automatización.

Uno de los procesos más importantes, y que se produce en diferentes fases de la fabricación, es la dosificación de materias. Los silos comentados anteriormente tienen en

su parte inferior una rasera / compuerta o una rosca sinfín que al ser activadas dejan pasar la materia, esta cae por gravedad en una báscula de pesaje que nos indica la cantidad dosificada. Entre el elemento mecánico (rasera o rosca) activado y la báscula hay una distancia, por lo que una vez que se para o cierra el elemento sigue cayendo materia a la báscula durante un tiempo, lo cual nos provoca desviaciones en las mediciones.

Para corregir este fenómeno se utilizan lo que denominamos “colas de caída”, que es la cantidad de materia que cae desde que desactivamos el elemento hasta que se estabiliza la báscula. Durante el proceso automático, se desactiva el elemento mecánico un poco antes de que la báscula marque la cantidad deseada, en concreto el valor de esa “cola de caída”, para intentar que cuando la báscula se estabilice marque la cantidad deseada.

Las colas de caída varían constantemente en el tiempo, dependen mucho de la báscula, del elemento dosificador, de la materia, lo que se ha dosificado antes, etc. e incluso de la temperatura y humedad. El cálculo actual de colas se basa en los datos de algunas de las variables anteriormente mencionadas, que han sido registrados en las últimas dosificaciones realizadas.

El sistema del que se dispone actualmente es correcto y funciona aceptablemente, pero en ciertas circunstancias el método de cálculo es insuficiente, y se producen ciertas desviaciones que llevan a errores en las dosificaciones, debido a ciertas variables que influyen en la dosificación y de las que no se tiene información, como la temperatura, humedad, etc.

La idea de este trabajo fin de máster es, teniendo información de las dosificaciones realizadas en los últimos años en una fábrica, en torno a los 0,4 millones, realizar un trabajo de limpieza, estudio, creación de características y obtención de un modelo que sea capaz de predecir la duración temporal de la próxima dosificación y con ello corregir la cola de caída utilizada y evitar las desviaciones esporádicas.

Agradecimientos

A todas las personas que han participado en esta experiencia, alumnos, profesores y a todo aquel que la haya hecho posible, a E.R. Ingeniería, por su confianza y respaldo, a Luis de la Ossa, por embarcarme en esta trepidante aventura que me ha hecho rememorar mi época en la universidad y reciclar mi conocimiento para afrontar el futuro, y especialmente a mi familia por su comprensión y ánimo.

Índice general

Índice de figuras

Índice de tablas

Capítulo 1

Introducción

1.1 Introducción

Como ya sabemos y se puede observar en la Figura 1.1, el proceso de dosificación de materia utilizado en una fábrica de piensos se realiza mediante la activación de elementos mecánicos de diferentes tipos, provocando que la materia almacenada en un silo/depósito caiga por gravedad en una báscula de pesaje, que nos indica el peso de la materia dosificada.

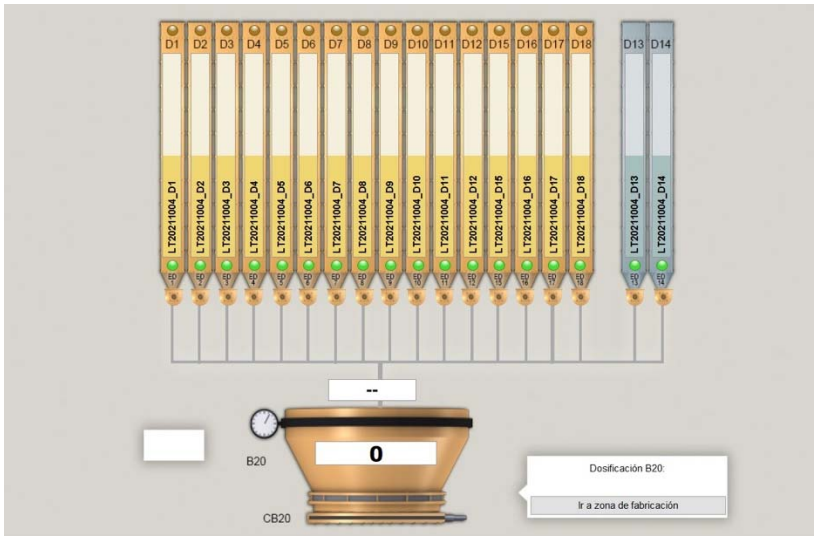

Figura 1.1 Pantalla de dosificación de materia

El problema en este tipo de procesos es que entre el elemento dosificador y la báscula hay cierta distancia, por lo que desde que el elemento se desactiva hasta que báscula deja de variar su peso pasa cierto tiempo. La cantidad de materia que cae en ese periodo de tiempo es lo que denominamos “cola de caída”.

Si no contemplamos este fenómeno, existirán errores en todas las pesadas realizadas. Para corregirlo, lo que se hace es desactivar el elemento dosificador cuando la báscula marca la cantidad de materia objetivo, menos el valor de la “cola de caída”, de modo que la báscula marcará el peso deseado cuando se estabilice.

El sistema utilizado en la actualidad tiene en cuenta una serie de variables comunes a todas las dosificaciones, como son: materia a dosificar, silo del que se dosifica, materia que se quiere obtener, etc., y se basa en las últimas pesadas, o mejor dicho en la cola que hubiesen necesitado las últimas pesadas para esas variables.

Estas colas, al tener en cuenta siempre las últimas pesadas, se van actualizando después de cada una de ellas, por lo que van variando con el tiempo. Para el caso de las primeras pesadas de las que no tengamos datos anteriores, se utilizan unas colas por defecto que nos sirven para inicializar los datos almacenados para ese conjunto de variables.

Lo bueno de este sistema es que aprende muy rápido y los resultados obtenidos son bastante aceptables, pero en ciertas circunstancias hace que se produzcan desviaciones muy elevadas en las pesadas ya que no tienen en cuenta ciertos factores por los que se pueden ver afectadas.

1.2 Objetivos

El objetivo de este trabajo no es en sí modificar el método de cálculo de colas, ya que como he comentado, su funcionamiento es aceptable, sino obtener un modelo de predicción que lo complemente.

La idea es, basándonos en los datos recopilados de las dosificaciones realizadas hasta la fecha, poder predecir la duración de la próxima pesada, es decir, el tiempo que debería estar activado el elemento dosificador, y con este valor, corregir esa cola calculada para mejorar la dosificación.

La duración de cada una de las pesadas es un indicativo de la cantidad de materia que cae a la báscula, siempre teniendo en cuenta la situación de partida e incluso las dosificaciones anteriores. Si somos capaces de predecir con exactitud la duración de un proceso de dosificación, tendríamos otra herramienta para realizar el proceso automático, aunque también estaría afectado por otro tipo de desviaciones. El objetivo que nos planteamos es utilizar dos métodos paralelos para que se complemente y así que cada uno de ellos corrija las desviaciones del otro.

Hay muchos factores que pueden influir y que nos pueden ayudar a refinar el sistema y predecir cuando una dosificación puede sufrir una desviación, como por ejemplo la fecha en la que se produce, la hora, el día de la semana, la pesada o pesadas que se han realizado anteriormente, temperatura, humedad, etc.

El modelo resultante de este trabajo será utilizado para predecir con anterioridad a cada una de las pesadas la duración que va a tener la dosificación según las circunstancias y el momento en el que se va a realizar, dicha duración complementará / corregirá la cola calculada con el método actual, y conseguiremos reducir las desviaciones que se producen en la actualidad.

1.3 Estructura del proyecto

Este trabajo será dividido en diferentes etapas, que básicamente son las etapas de un proyecto de machine learning, aunque anteriormente nos centraremos en explicar el significado de los datos de partida con el fin de comprender mejor el problema y el objetivo que pretendemos conseguir.

A continuación, haremos un análisis exploratorio de los datos, identificando los diferentes tipos de datos, las variables numéricas, las variables categóricas, etc. Realizaremos una

limpieza de estos, buscando valores perdidos, outliers e intentaremos crear ciertas características que pueden ayudar a caracterizar el problema.

Es siguiente paso será realizar un preprocesamiento de los datos, imputando valores perdidos, estandarizando y escalando variables numéricas, etc.

Entrenaremos diferentes modelos y realizaremos su validación correspondiente para quedarnos con el que mejores resultados obtengamos y que pueda ser pasado a producción dentro de nuestros procesos.

Capítulo 2

Presentación de los datos

En este apartado describiremos la información de partida para la obtención del modelo final con el fin de entender la naturaleza del problema y poder desarrollar diferentes ideas o estrategias en las fases posteriores de análisis y preprocesamiento.

2.1 Fuente de datos

Los datos de los que disponemos son extraídos de la propia aplicación que gestiona los procesos de automatización de la fábrica de piensos. Dentro de la amplia legislación que rige la industria agroalimentaria, existe una estricta normativa relativa a la trazabilidad. Este tipo de procesos deben tener la capacidad seguir el rastro a toda materia prima que forme parte de este, desde que es adquirida hasta que es consumida.

Como hemos comentado, en una fábrica de pienso existen gran cantidad de procesos de transferencia y transformación de materia, y todos ellos deben quedar registrados para una posible consulta de trazabilidad. En el caso concreto que nos ocupa, la dosificación de materias primas, se produce una transferencia de materia y a su vez un mezclado de esta, formándose una materia compuesta. Cada una de las dosificaciones es registrada y la información que obtenida se describe en la Tabla 2.1.

Tabla 2.1. Muestra de datos de partida basados en dosificaciones anteriores

Nombre columna	Ejemplo	Descripción
Fecha_inicio	2013-08-14 16:51:48	Fecha de inicio de la dosificación
Fecha_fin	2013-08-14 16:52:39	Fecha de fin de la dosificación
Cantidad_solicitada	235.026	Cantidad objetivo a dosificar
Cantidad_dosificada	236.5	Cantidad real dosificada al finalizar
Mezcla	1	Nº de mezcla dentro de la fabricación
Pesada	0	Nº de pesada dentro de la mezcla
Peso_inicial	0.0	Peso que marca la báscula antes de iniciar la dosificación
Manual	0	Booleano que indica si la dosificación ha sido manual o automática
Id_lote_destino	232527	Lote de fabricación.
Materia_origen	13	Código de la materia origen.
Materia_destino	66	Código de la materia destino
Id_silo	90	Silo origen del que se dosifica
Tipo_materia	Prima	Tipo de materia origen
Tipo_destino	Premezcla	Tipo de materia destino
Densidad	1.0	Densidad de la materia origen
Tam_mezcla	500	Suma teórica total de los ingredientes al finalizar todas las dosificaciones
Desviacion	1.47	Error real en la dosificación

Como se puede observa, la información almacenada refleja el punto de partida de la dosificación, los elementos que intervienen, el tipo de materia a obtener y el origen de esta, y finalmente los resultados obtenidos de esa dosificación.

Un dato que puede que sea bastante relevante, pero del que no disponemos, es la cola de caída utilizada, que viene fijada por las dosificaciones anteriores en unas circunstancias similares.

En los datos que se muestran en el ejemplo se describe una dosificación realizada en agosto de 2013, en la que se partía con la báscula vacía, se dosifica una materia prima con código 13, para obtener después de mezclar todas las materias dosificadas una materia con código 66. Teóricamente queremos obtener una cantidad de 500 Kg cuando tengamos todas las materias en la báscula y de esta concretamente queremos dosificar 235,026 Kg. Después de unos 51 segundos activando el elemento dosificador, se registra

que la báscula ha variado de peso 236.5 Kg, por lo que se ha producido una desviación de 1,47 Kg sobre el objetivo.

En este caso, la desviación está dentro de los límites de tolerancia y ha sido bastante aceptable, pero en ocasiones esa desviación puede superar esos límites provocando un deterioro de la calidad de la materia a producir. Estas desviaciones ocasionales puede que se produzcan debido a que, en el cálculo de las colas de caída, no se estén teniendo en cuenta todas las variables necesarias, esto es así, para acelerar ese proceso de aprendizaje y se puedan obtener buenos resultados desde el principio.

En las siguientes etapas de este trabajo intentaremos caracterizar esas desviaciones ocasionales con toda la información de la que se dispone de las dosificaciones y alguna más que podamos aportar de nuestra experiencia y así poder entrenar un modelo que sea capaz de predecir la duración real de la dosificación.

Capítulo 3

Análisis exploratorio

En este capítulo realizaremos un análisis exploratorio de los datos: tipos de datos, cantidad de datos, correlación entre variables numéricas, variables cualitativas, etc. para poder diseñar un correcto preprocesado de datos en fases posteriores. Para ello nos serviremos de un cuaderno de JupiterLab y mediante las librerías de pandas, matplotlib y seaborn iremos realizando el análisis.

3.1 Análisis exploratorio

La etapa de exploración de los datos es una de las más importantes dentro del proceso de machine learning, ya que nos sirve para tener un conocimiento más profundo de los datos y de la relación entre ellos. Esto nos permitirá orientar mejor el procesamiento y la elección de modelos a entrenar.

En primer lugar, cargaremos nuestros datos en un dataframe de pandas para su manipulación y análisis. Contamos con 395.477 registros comprobando que efectivamente todas las columnas se han cargado correctamente y que los tipos de datos son los correctos. Verificaremos que no existen valores perdidos, eliminando los registros en los que la columna manual tiene valor 1 (indica que el registro pertenece a una

dosificación realizada manualmente), los cuales puede generar ruido en nuestros entrenamientos de modelos.

3.2 Filtrado de datos y detección de outliers.

En el siguiente paso estudiaremos la distribución de valores y los estadísticos de las variables numéricas que más caracteriza nuestro problema, como son la cantidad solicitada y la dosificada, el peso inicial de la báscula, el tamaño de la mezcla, la duración de la dosificación y la desviación final de la pesada, los datos obtenidos se muestran en la Tabla 3.1.

Tabla 3.1. Distribución y estadísticos de las principales variables numéricas

	C. Solicitada	C. Dosificada	P. Inicial	T. mezcla	Duración	Desviación
Count	395477	395477	395477	395477	395477	395477
Mean	801.251	800.430	1342.961	4000.817	37.613	-0.8203
Std	744.934	745.126	1249.587	211.905	37.362	30.4486
Min	0.000	0.000	-64.000	400.000	0.000	-2623.09
25%	167.984	168.000	11.000	4000.000	16.000	-3.000
50%	668.040	664.000	1574.000	4000.000	25.000	0.0040
75%	1163.096	1168.000	2600.000	4040.000	49.000	3.0198
Max	3121.984	3127.000	3995.000	4500.000	7786.000	1179.891

Para visualizar mejor los datos obtenidos se ha obtenido un gráfico de cajas de las variables anteriores como el de la Figura 3.1.

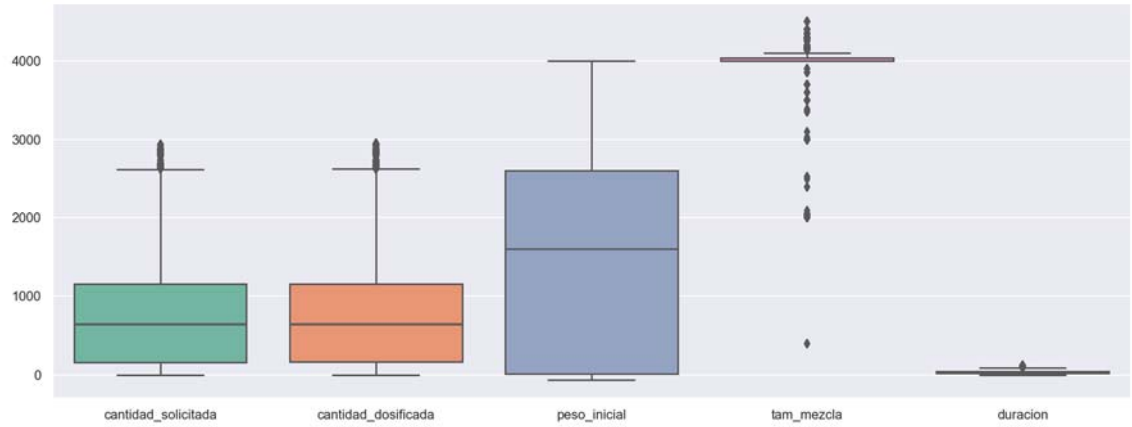

Figura 3.1. Distribución y estadísticos de las principales variables numéricas

A la vista de los datos podemos observar que la cantidad solicitada normalmente no son superiores a unos 1200Kg, esto sucede porque normalmente las fórmulas llevan uno o dos productos base, de los que se dosifica una gran cantidad, entre 1.000 o 1.500 Kg y posteriormente se van añadiendo cantidades más pequeñas de otros ingredientes. Cantidades solicitadas superiores a 1300 Kg las consideraremos outliers, ya que no son habituales y normalmente puede que sean pruebas o errores.

Como era de esperar, los datos de la cantidad dosificada se asemejan bastante a los de la cantidad solicitada, y es por ello, que las desviaciones son muy cercanas a 0. Ya que las desviaciones esporádicas pueden rondar los 100 Kg desecharemos los registros con desviaciones superiores a esta cantidad. También se desecharán dosificaciones mayores de 1300 Kg, ya que también se pueden deber a pruebas o a errores mecánicos.

Los datos relativos al peso inicial denotan que normalmente la mayoría de las dosificaciones se producen cuando la báscula esta vacía o tiene poco peso, un poco acorde con las cantidades dosificadas, no son normales dosificaciones cuando la báscula tiene ya un peso superior a 2.700 Kg, por lo que desecharemos esos registros al considerarlos outliers.

En cuanto al tamaño de mezcla observamos que los datos son más o menos normales, la mayoría de las mezclas se fabrican a 4000 Kg, esto es lo que sumarían teóricamente todos los ingredientes, y las premezclas a 500 Kg. Existen valores intermedios dentro de eses rango, pero son normales y correctos.

En relación a la duración de la dosificación, en segundos, que es la variable que queremos predecir, vemos que de media duran en torno a los 37 seg, y la desviación estándar es también de unos 37 seg. El tiempo estimado para dosificar todos los productos es como mucho de 4 minutos para que la fábrica tenga un rendimiento óptimo, y normalmente hay que dosificar varios ingredientes, por lo que duraciones superiores a esos 120 seg. las consideraremos outliers y las filtraremos.

Con la información obtenida y las decisiones tomas pasamos se filtran los datos que consideramos ruido, quedándonos para nuestro aprendizaje lo que cumplen los siguientes criterios:

- Cantidad solicitada mayor de 0.
- Cantidad solicitada menor de 1300.
- Cantidad dosificada mayor de 1.
- Cantidad dosificada menor de 1400.
- Desviación > -100.
- Desviación < 100.
- Peso inicial de la báscula menor de 2700.
- Duración menor de 120 segundos.

Como resultado nos queda una base de datos con 235.045 registros.

3.3 Análisis de las columnas de información y creación de características.

Una vez filtrados los outliers analizaremos las columnas de las que disponemos y estudiaremos la posibilidad de crear nuevas características que nos puedan caracterizar el problema.

En el apartado anterior, con el fin de estudiar sus estadísticos, ya se había incorporado una columna denominada `'duración'`, la cual es calculada y equivale a los segundos que transcurren entre la fecha de inicio y la fecha fin de la dosificación, es decir, el tiempo que está activo el elemento dosificador. Esta columna es la que queremos predecir con nuestro modelo.

Por otro lado, es fácil pensar que el orden de cada una de las dosificaciones dentro de todos los ingredientes puede ser un dato que caracterice el problema y nos sirva para mejorar nuestros modelos, por ello, generaremos una nueva característica con esta información de nuestros datos, la cual llamaremos `'orden'`.

Filtrando la información de nuestro dataset, en primer lugar, debemos desechar las columnas `'fecha_fin'` y `'cantidad_dosificada',` ya que si las utilizásemos en nuestro entrenamiento, cometeríamos una fuga de datos, ya que es información que se genera una vez finalizada la dosificación.

Del mismo modo, la columna `'manual'` no nos aporta ninguna información, ya que siempre tiene el mismo valor, nos sirvió para eliminar los registros de las dosificaciones manuales, pero llegados a este punto, no es de utilidad.

En cuanto a las variables `'densidad'`, `'tipo_materia'` y `'tipo_destino'`, están estrechamente relacionadas con `'materia_origen'` (ya tiene intrínseca la información de tipo_materia y densidad, las cuales no varían) y `'materia_destino'` que siempre será del mismo tipo (`'tipo_destino'`). Por tanto, prescindiremos de estas columnas, ya que sería información redundante.

La columna `'desviación'` nos indica el error entre la cantidad solicitada y dosificada, realmente no sabemos la desviación que se va a producir y no es nuestro objetivo, ya que basaremos nuestras predicciones en el tiempo de las dosificaciones, por lo que esta variable también la eliminaremos.

Por último, prescindiremos también de las columnas `'pesada'` y `'mezcla',` ya que cada vez que se cambia de pesada o mezcla las condiciones de las dosificaciones vuelven a ser las mismas, partiendo de la báscula vacía, por lo que no aportan gran información.

A continuación, se listan las columnas a eliminar:

- Fecha_fin.
- Cantidad_dosificada.
- Mezcla.
- Pesada.
- Manual.
- Densidad.
- Tipo_materia.

- Tipo_destino.
- Desviacion.

En principio no tenemos ningún indicio de por qué circunstancias pueden producirse las desviaciones, pero es fácil pensar que puedan venir marcadas por algún patrón temporal, por lo que crearemos una serie de características en este sentido, intentando aportar información al modelo. Para ello, descompondremos la fecha de la dosificación en diferentes columnas:

- Hora: hora del día a la que se produce la dosificación.
- Dow: día de la semana a la que se produce la dosificación.

Por último, en cuanto a la creación de características se refiere, también es fácil relacionar los errores en las dosificaciones con cambios en las propiedades de las materias y que estos cambios se pueden producir por cambios en las condiciones de temperatura y humedad dentro de la fábrica.

En la actualidad no se dispone de sondas de medición de temperatura y humedad en la planta, y para poder incorporar esta información a nuestro dataset, hemos obtenido la información de la Agencia Estatal de Meteorología (AEMET).

AEMET ofrece un servicio API REST denominado AEMET OpenData, a través del cual podemos acceder a datos históricos meteorológicos de multitud de estaciones meteorologías.

En este caso hemos descargado los datos históricos de dos estaciones meteorológicas más o menos equidistantes a la fábrica, hemos calculado la media de los valores de las dos y hemos agregado a nuestros datos la variable de temperatura media del día en el que se produjo cada una de las dosificaciones. Nos hubiese gustado adjuntar también datos de humedad, ya que puede alterar también las propiedades de densidad y peso de las materias, pero no se ofrece en este servicio concreto. A la nueva característica generada la hemos denominado `'tmed'`.

3.4 Estudio de la correlación.

Una vez seleccionadas las características y filtrados los datos, estudiaremos grado de correlación entre variables, en la Tabla 3.2 y en la Figura 3.2 podemos observar los resultados.

Tabla 3.2. Correlación entre variables

	C.S	P.I.	M.O.	M.D.	SILO	T.M.	DUR.	ORD.	HOR	DOW	TMED
C.S.	1.000	-0.668	-0.288	0.107	0.093	0.031	0.656	-0.675	0.049	0.003	-0.042
P.I.	-0.668	1.000	0.228	-0.010	-0.242	0.068	-0.277	0.841	-0.031	0.0006	0.002
M.O.	-0.288	0.228	1.000	0.166	0.051	-0.005	-0.167	0.312	0.011	-0.028	-0.012
M.D.	0.107	-0.010	0.166	1.000	-0.016	0.011	0.124	0.046	0.074	-0.047	-0.051
SILO.	0.093	-0.242	0.051	-0.016	1.000	-0.001	-0.115	-0.223	0.009	-0.020	0.004
T.M.	0.031	0.068	-0.005	0.011	-0.001	1.000	-0.027	-0.009	-0.028	0.035	0.125
DUR.	0.656	-0.277	-0.167	0.124	-0.115	-0.027	1.000	-0.335	0.050	-0.004	-0.053
ORD.	-0.675	0.841	0.312	0.046	-0.223	-0.009	-0.335	1.000	-0.024	0.004	-0.011
HOR.	0.049	-0.031	0.0119	0.074	0.009	-0.028	0.050	-0.024	1.000	0.047	0.049
DOW.	0.003	0.000	-0.028	-0.047	-0.020	0.035	-0.004	0.004	0.047	1.000	0.027
TMED.	-0.042	0.002	-0.012	-0.051	0.004	0.125	-0.053	-0.011	0.049	0.027	1.000

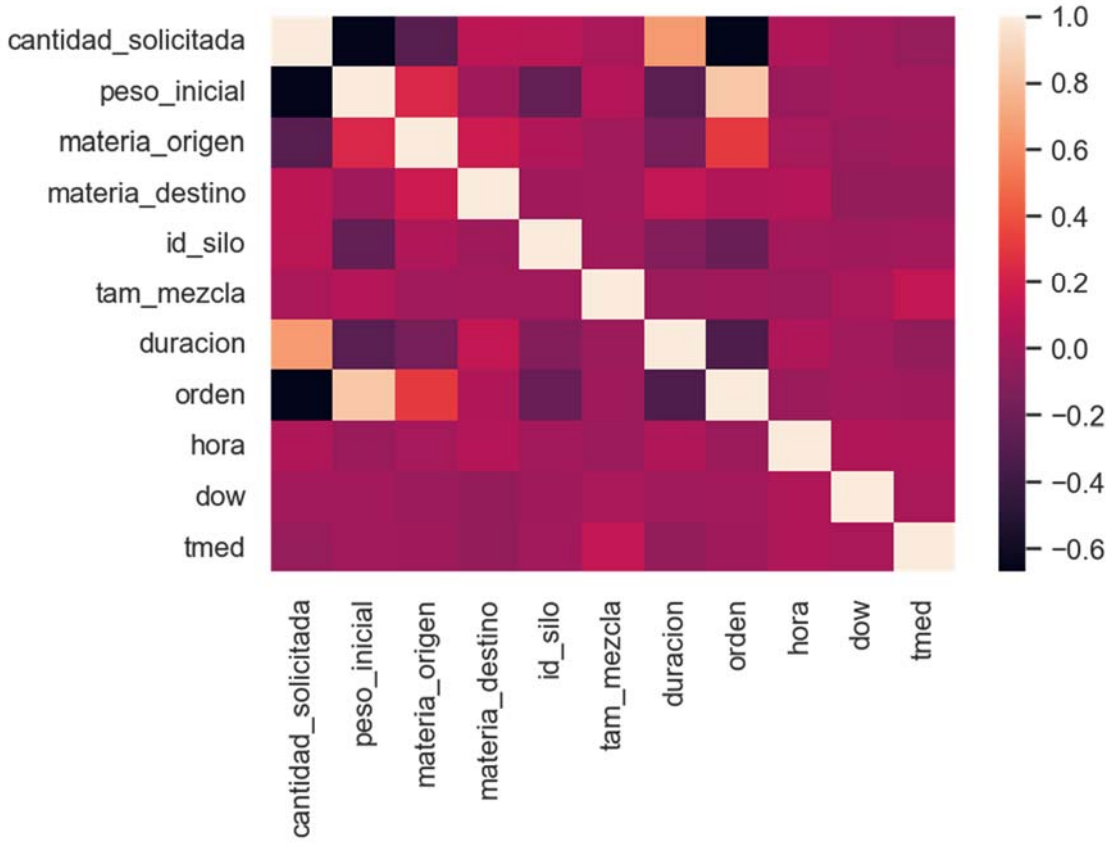

Figura 3.2 Correlación entre variables

Podemos observar que la mayor correlación se produce entre el peso inicial y la materia origen, eso es normal, ya que como he comentado antes, hay ciertos productos que sirven como base y se suelen dosificar al principio, el resto irán posteriormente en orden ascendente de cantidad solicitada, por lo que normalmente, estos productos base, se dosifican con pesos iniciales similares en la báscula. Por la misma razón el peso inicial está correlacionado con la cantidad solicitada, ya que las dosificaciones se producen por cantidad solicitada ascendentemente.

En cuanto a la variable a predecir que es la duración, existe una correlación clara con la cantidad solicitada, ya que cuanto mayor sea esa cantidad, más durará la dosificación y con peso inicial y orden, ya que primero se dosifican las materias de las que más cantidad se requiere, por lo que, a menor orden y peso en la báscula, mayor duración.

En la Figura 3.3, se puede ver la relación mencionada anteriormente, aunque no es totalmente clara y hay gran cantidad de ruido que la altera. Puede que esta circunstancia se deba a que hay gran variedad de materias y que estas pueden almacenarse y dosificarse desde cualquier silo.

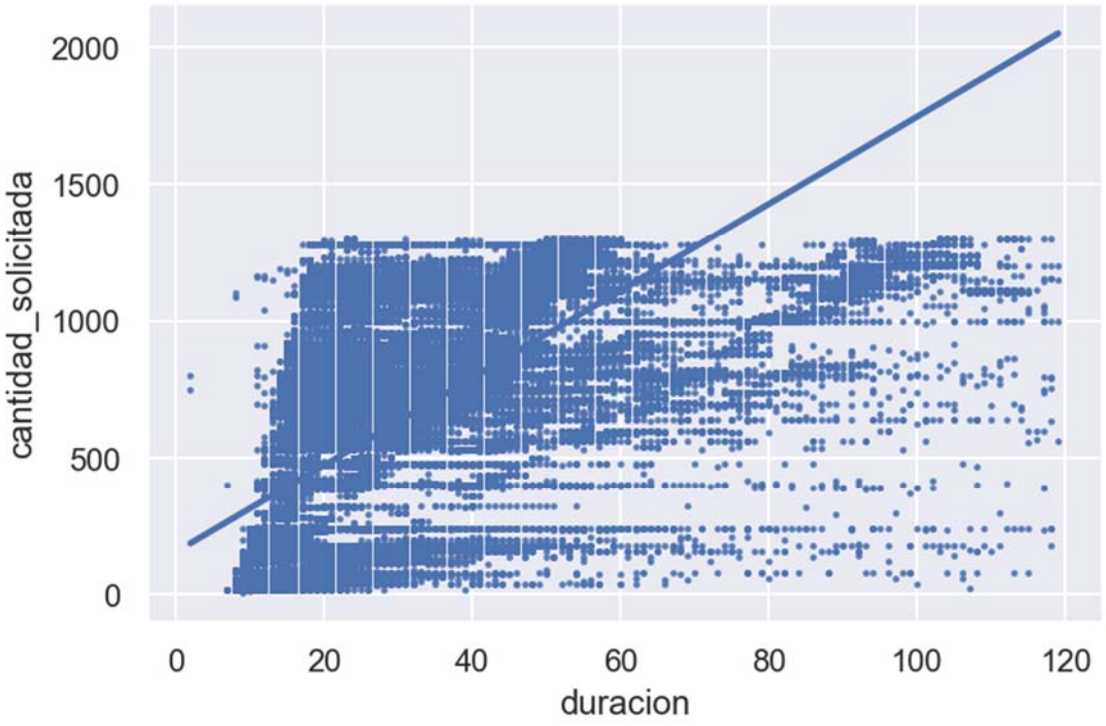

Figura 3.3 Relación entre cantidad solicitada y duración

Es interesante visualizar como se comporta esa relación duración / cantidad solicitada para cada uno de los silos, y como se puede ver en la **Figura 3.4** si parece que hay comportamientos diferentes para cada uno de los silos con diferentes materias, aunque hay bastante ruido y no parece muy definido.

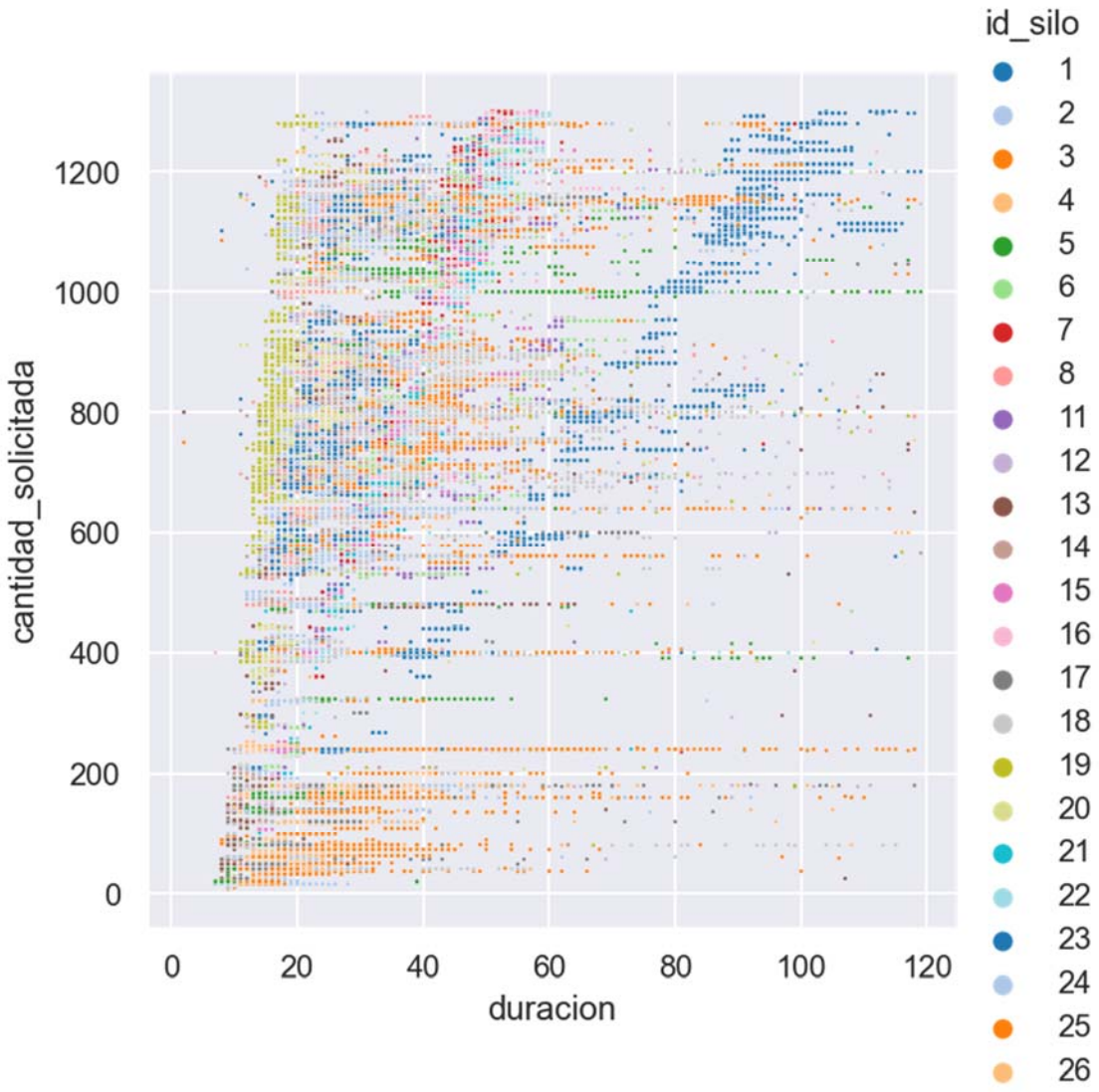

Figura 3.4 Relación duración / cantidad dosificada por silo

Para profundizar en el estudio de la relación entre la cantidad solicitada y la duración de la dosificación estudiaremos que sucede con una materia concreta en un solo silo. Como se puede apreciar en la **Figura 3.5**, parece que sí que se marca más esa relación o

tendencia. Lo que es curioso es que para una sola materia en un silo se visualicen dos comportamientos diferentes, esto puede tener dos explicaciones: en muchas ocasiones, a materias con las mismas cualidades nutritivas las llaman del mismo modo aunque sus características físicas de peso específico, volumen, humedad, etc. sean diferentes o, por otro lado, que en un silo se cambie o se modifique el comportamiento del elemento mecánico que dosifica de él, lo que haría que la materia saliese del silo a otra velocidad.

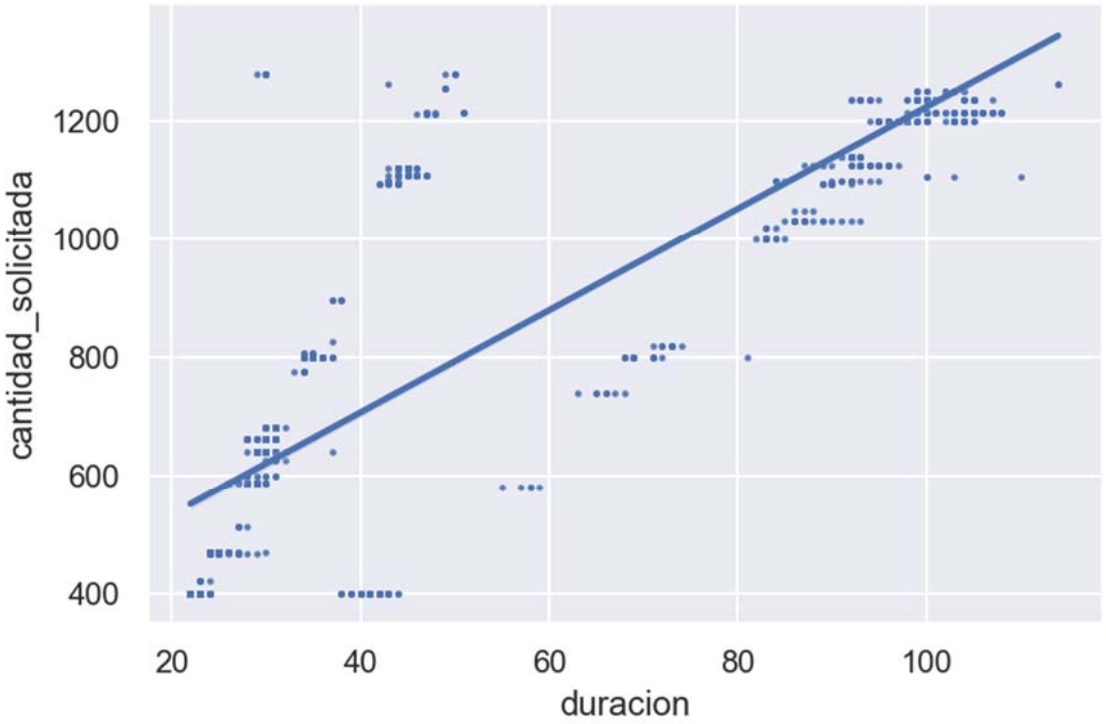

Figura 3.5 Relación entre cantidad solicitada y duración para un silo y materia concreto

Capítulo 4

Preprocesamiento

En este capítulo abordaremos el preprocesamiento de los datos, mediante el cual realizaremos diferentes transformaciones de los datos para facilitar el aprendizaje de los algoritmos de machine learning. Esto es importante para evitar el sobreajuste de modelos debido a ciertas características de los datos.

4.1 Preprocesamiento

En primer lugar, dividiremos las variables en numéricas y categóricas, ya que recibirán un tratamiento diferente en este preprocesamiento.

- Variables Categóricas:
  - Materia_origen: Diferentes materias primas que pueden ser dosificadas, puede tomar un número finito de valores.
  - Materia_destino: Similar a la materia origen, son las diferentes fórmulas o recetas que queremos fábricas, también un número acotado de valores.
  - Id_silo: Silo del que se realiza la dosificación.
  - Orden: Orden de la dosificación dentro de una mezcla.
  - Hora y día de la semana: También pueden tomar un número limitado de valores por lo que las consideraremos variables categóricas.

- Variables Numérica:
  - Cantidad_solicitada: Cantidad de materia objetivo a dosificar.
  - Peso_inicial: Peso actual de la báscula.
  - Tam_mezcla: Cantidad total teórica a dosificar sumando todos los productos.
  - Tmed: Temperatura ambiente media del día.

Una vez identificados los tipos de las diferentes variables diseñaremos un Pipeline como el de la Figura 4.1 para el preprocesamiento de estos. Para las variables numéricas programaremos dos tareas, una imputación de valores nulos, y un escalado standard de los valores. En cuanto a las variables categóricas también imputaremos los valores perdidos y realizaremos un onehotencoder para binarizar todas las variables categóricas.

Figura 4.1. Pipeline de preprocesamiento de variables

4.2 División train y test

Una vez diseñado el preprocesamiento haremos una división de nuestros datos en train y test, el conjunto de datos de train lo utilizaremos para entrenar nuestros modelos y el de test para validarlos. La partición de los datos la haremos de un modo aleatorio, quedándonos con un 66% de los datos para entrenamiento y un 33% para el test.

Capítulo 5

Entrenamiento de Modelos

Una vez analizados los datos, generadas nuevas características, preprocesados y divididos en train y test, nuestra siguiente tarea será entrenar diferentes modelos de regresión para poder compararlos y quedarnos con el que mejores resultados obtengamos para pasarlo a producción. En todos los modelos entrenados utilizaremos el error absoluto medio como medida de evaluación, ya que nos será de utilidad tanto para cuantificar y comparar los modelos, como para disponer de una medida que nos permita valorar el error desde el punto de vista operativo. Para entrenar los diferentes modelos nos serviremos de los modelos que nos proporciona la librería sklearn para Python.

5.1 Regresión Lineal

La regresión lineal es el modelo de regresión más básico mediante el que se intenta representar la relación lineal entre cada una de las variables y la variable a predecir, para la validación del modelo utilizaremos el error absoluto medio que nos expresa claramente la desviación que se produce en las predicciones.

Como era de esperar, es difícil que un modelo de regresión lineal simple se ajuste al un problema tan complejo y variable como hemos visto en la exploración de los datos.

El error absoluto medio obtenido para los diferentes conjuntos de datos son los siguientes:

- Entrenamiento: 4.18 Seg.
- Test: 4.15 Seg.

En términos generales vemos que el modelo no consigue la expresividad suficiente y queda alejado de unas buenas predicciones.

5.2 Ridge

La regresión lineal con regularización intenta evitar el sobreajuste, en nuestro caso hemos visto que la en regresión lineal simple no se ha producido este fenómeno, pero los resultados no han sido óptimos, quizá esa regularización nos ayude a filtrar o dar menos peso a las variables que menos caracterizan nuestro problema y así conseguir mejores resultados.

En este caso realizaremos una búsqueda de hiperparámetros para el parámetro Alpha, que determina el peso de la penalización, y para la imputación de valores perdidos de los valores numéricos, para ello realizaremos una validación cruzada con una parte de los datos de test y posteriormente entrenaremos el modelo con los mejores parámetros obtenidos. Como método de validación utilizaremos nuevamente el error absoluto medio. La búsqueda la realizaremos con un subconjunto de los datos de train con el fin de agilizar la búsqueda.

El conjunto de hiperparámetros utilizados para búsqueda de los óptimos abarcará la estrategia de imputación de valores perdidos en variables numéricas y el parámetro alpha de la regularización con los siguientes valores:

- Estrategias de imputación:
  - Mean , Median
- Alpha:
  - 1 , 0.1 , 0.01 , 0.001 , 0.0001 , 0

La validación cruzada nos indica que el mejor valor de alpha es 1 y que la estrategia de imputación 'mean', por lo que procedemos a entrenar el modelo Ridge con esa configuración y la totalidad de datos del conjunto de train.

El error absoluto medio obtenido para los diferentes conjuntos de datos son los siguientes:

- Entrenamiento: 4.18 Seg.
- Test: 4.15 Seg.

Observamos que los resultados obtenidos son similares a la regresión lineal simple, seguramente debido a que al no haberse producido un sobreajuste en ese modelo no hemos conseguido mejorarlo.

Esperábamos mejorar los resultados del modelo anterior, pero puede que la complejidad y el ruido sean difíciles de caracterizar con este tipo de modelos.

5.3 Random Forest

Intentaremos mejorar los resultados obtenidos anteriormente entrenando un modelo mediante Random forest, explotando así la aleatoriedad de los datos y esperando la mejora comentada.

Random Forest es un ensemble, un conjunto de modelos que combinan sus predicciones para obtener una única predicción. Es uno de los algoritmos de aprendizaje que suele obtener mejores resultados en problemas de predicción con bases de datos de gran tamaño. Es capaz de manejar cientos de variables, estimando cuales de ellas son importantes y cuales no.

En este caso realizaremos una búsqueda de hiperparámetros con una validación cruzada sobre un subconjunto de 2000 registros del conjunto de train, intentando fijar el número de estimadores, máximo número de atributos y máximo número de niveles de profundidad, con la siguiente configuración:

- Número de estimadores:
  - 50 , 100 , 200 , 1000
- Número de atributos:
  - ‘auto’, 3 , 5 , 7 , 10
- Profundidad:
  - None , 3 , 5 , 10 , 20

Los mejores valores para los parámetros de entrenamiento del modelo son 200 para el número de estimadores, ‘auto’ para el número de atributos máximo y 20 para la máxima profundidad del árbol. Con estos valores, procedemos a entrenar el modelo Random Forest con el conjunto de datos de train.

El error absoluto medio obtenido para los diferentes conjuntos de datos son los siguientes:

- Entrenamiento: 1.31 Seg.
- Test: 1.55 Seg.

Observamos que los resultados obtenidos mejoran bastante los modelos anteriores, seguramente debido a que Random Forest se comporta mejor con grandes cantidades

de datos, gran número de variables y es capaz de discriminar el ruido existente en los datos.

También se puede deducir que este modelo sobre ajusta un poco más que los anteriores a los datos de train, aunque no excesivamente.

Este modelo ya puede considerarse valido para nuestro objetivo, ya que 1,5 segundos de desviación no son excesivos en las predicciones y como ya veremos en el capitulo de implantación en producción nos servirá para ajustar el momento de parada de la dosificación.

5.4 Random Forest para un silo concreto

Con el modelo anterior ya tenemos una buena aproximación a nuestro problema, pero como vimos en el análisis exploratorio, parece que las dosificaciones se comportan de diferente mantera en función del silo en el que se producen, por lo que vamos a entrenar un modelo similar al anterior, pero con los datos referentes a un solo silo.

Parámetros de entrenamiento del modelo serán los mismos que para el caso anterior, 200 para el número de estimadores, 'auto' para el número de atributos máximo y 20 para la máxima profundidad del árbol. Con estos valores, procedemos a entrenar el modelo Random Forest con el conjunto de datos de train para el silo 1 en este caso.

El error absoluto medio obtenido para los diferentes conjuntos de datos son los siguientes:

- Entrenamiento: 0.39 Seg.
- Test: 0.87 Seg.

Observamos que los resultados obtenidos mejoran bastante el modelo anterior, aunque también aumenta el sobreajuste, esto ratifica lo que esgrimía la exploración de los datos, el comportamiento de cada uno de los silos es particular y hace que al entrenar los modelos con datos de todos los silos sea más difícil caracterizar el problema.

5.5 Random Forest para un silo y materia concretos

Hemos visto que particularizando el problema para un silo los resultados mejoraban notablemente. Esto nos lleva a dar un paso más y es lógico pensar que el comportamiento de diferentes materias dentro de un mismo silo debe ser diferente.

Por ello, vamos a repetir la operación anterior utilizando únicamente los datos de una materia concreta en un silo concreto para intentar mejorar los resultados del modelo anterior.

Parámetros de entrenamiento del modelo serán los mismos que para el caso anterior, 200 para el número de estimadores, 'auto' para el número de atributos máximo y 20 para la máxima profundidad del árbol. Con estos valores, procedemos a entrenar el modelo Random Forest con el conjunto de datos de train para el silo 1 y la materia 10 en este caso.

El error absoluto medio obtenido para los diferentes conjuntos de datos son los siguientes:

- Entrenamiento: 0.35 Seg.
- Test: 0.88 Seg.

Observamos que los resultados obtenidos mejoran en entrenamiento, pero empeoran en test, aunque la diferencia es inapreciable, esto nos hace pensar que realmente la materia no influye mucho en la dosificación, es más determinante el silo o mejor dicho el elemento dosificador, su naturaleza física y su configuración de funcionamiento.

Capítulo 6

Comparación de modelos

En este capítulo compararemos los resultados obtenidos para los diferentes modelos y decidiremos cual de ellos pasamos a producción para predecir las desviaciones en las dosificaciones.

6.1 Tabla comparativa de modelos

La siguiente Tabla 6.1, nos ayuda a visualizar los resultados de validación de los diferentes modelos tanto para los datos de entrenamiento como para test, en todos los casos hemos utilizado el error absoluto medio.

Tabla 6.1. Comparación de datos de validación de modelos

Modelo	Error Entrenamiento	Error Test
R. Lineal	4.18 Seg.	4.15 Seg.
Ridge	4.18 Seg.	4.15 Seg.
Random Forest	1.31 Seg.	1.55 Seg.
Random Forest 1 silo	0.39 Seg.	0.87 Seg.
Random Forest 1 silo 1 materia	0.35 Seg.	0.88 Seg.

Apreciamos claramente que el modelo con el que mejores resultados hemos obtenido es Random Forest, y en concreto los modelos que utilizaban datos más específicos con

información del silo y la materia. En este caso nos quedaremos con el modelo entrenado con los datos de un solo silo, sin diferenciar entre materias, ya que separando los datos también por materias no se mejora el anterior y nos llevaría a necesitar gran cantidad de modelos.

Capítulo 7

Adaptación del modelo a producción

En este capítulo describiremos como adaptaremos el modelo seleccionado y lo utilizaremos para conseguir nuestro objetivo de mejorar la precisión de las dosificaciones corrigiendo las colas de caída.

7.1 Método de corrección

Como ya sabemos, en la actualidad, nuestro método actual de dosificación de materia se realiza controlando el peso de la báscula en tiempo real, activando un elemento dosificador al inicio del proceso y desactivándolo un poco antes de llegar al peso requerido por el efecto de la cola de caída (cantidad de materia que cae a la báscula desde que desactivas el elemento dosificador hasta que esta se estabiliza).

Para el cálculo de esa cola de caída se utilizan las últimas pesadas realizadas en la báscula para el silo en concreto, pero como hemos expuesto hay otros factores que pueden influir en las dosificaciones y que no se están teniendo en cuenta, como la cantidad de materia que ya tiene la báscula, el orden de la dosificación, la temperatura, el día, etc.

Hemos conseguido obtener un modelo que nos predice con bastante exactitud el tiempo que va a durar la dosificación añadiendo a nuestros cálculos las variables que no

utilizamos al calcular la cola de caída y ese tiempo lo utilizaremos para corregir la cola en tiempo real y afinar y mejorar las dosificaciones.

Para complementar las dos aproximaciones lo que haremos será controlar el proceso con las dos medidas, comenzaremos la pesada activando el elemento dosificador y tendremos dos objetivos para pararlo, llegar a un peso calculado restando a la cantidad objetivo el tamaño de la cola de caída o que se cumpla el tiempo de activación calculado por nuestro modelo.

Vigilaremos tanto el peso en tiempo real de la báscula como el tiempo transcurrido desde que activamos la dosificación y el primer objetivo que se consiga (peso báscula objetivo o tiempo de dosificación) hará comenzar nuestro proceso de corrección.

El objetivo de peso de la báscula siempre tendrá prioridad y será el que corregiremos en según el nivel de complimiento del tiempo de dosificación. En la Figura 7.1 podemos observar las diferentes situaciones con las que nos podemos encontrar, ante las cuales deberemos ir tomando diferentes correcciones:

A. El tiempo de dosificación calculado por el modelo se cumple antes que lleguemos a parar por peso objetivo. En este punto calcularemos la cantidad que nos queda para llegar a nuestro peso objetivo y lo reduciremos un 50% de esa cantidad para que la dosificación pare antes de lo calculado originalmente.
B. El tiempo de dosificación acaba cuando nos estamos aproximando a nuestro peso objetivo. No realizaremos ninguna acción y dejaremos acabar la dosificación como estaba configurada.
C. El tiempo de dosificación no ha expirado cuando llegamos a nuestro objetivo de peso. Haremos la operación inversa a la primera, es decir, calculamos la cantidad que se dosificaría en el tiempo que falta por dosificar y aumentamos en 50% de esa cantidad el peso objetivo para que la dosificación acabe con posterioridad al cálculo original.

Figura 7.1 Casuística de corrección de cola

Pongamos un ejemplo para facilitar la compresión de este proceso de corrección: Partamos de una situación en la que la báscula tiene ya 1210 Kg de pesadas anteriores, queremos dosificar 500 Kg y según nuestro cálculo de cola de caída deberíamos aplicarle una corrección de 20 Kg. Por otro lado, nuestro modelo de predicción nos dice que la dosificación debe tardar 35 segundos.

Con estos datos, nuestro peso final de la báscula debería ser 1710 Kg (1210 + 500), con lo que deberíamos parar el elemento de dosificación cuando la báscula marque 1690 Kg. (1710 – 20 de cola). Activamos el elemento dosificador controlando el peso de la báscula y el tiempo transcurrido y se nos pueden dar las siguientes situaciones:

CASO A: Se cumple 35 segundos de dosificación y todavía no estamos cerca de los 1690 Kg objetivo. Cambiaremos nuestro objetivo de peso, restándole el 50% de lo que nos reste por dosificar, si por ejemplo la báscula marca en este momento 1600 Kg, nuestro objetivo final pasará a ser de 1645Kg. (1600 + ((1690 - 1600) / 2))

CASO B: Se cumple 35 segundos de dosificación y la báscula está próxima a marcar los 1690 Kg de objetivo. No corregimos nada y nuestro objetivo final seguirán siendo los 1690 Kg.

CASO C: La báscula marca los 1690 Kg objetivo y no se han cumplido los 35 segundos de dosificación estimados por nuestro modelo. Cambiaremos nuestro objetivo de peso, aumentándolo un 50% de lo que supuestamente se dosificaría si agotásemos el tiempo calculado, si por ejemplo nos restan 2 segundos y calculamos que caen 5 Kg/seg (sería en total 10 Kg), nuestro nuevo objetivo final sería 1695 Kg. (1690 + ((2 * 5) / 2))

Capítulo 8

Conclusiones y Trabajo Futuro

8.1 Conclusiones

Como ya adelantábamos en la introducción del trabajo, en la industria agroalimentaria, la mecánica utilizada no es de alta precisión, por lo que se hace difícil realizar operaciones de automatización con gran exactitud, sobre todo en el proceso de dosificación de materia que es el que nos ocupa.

El propio diseño de los elementos y su disposición física dentro de la fábrica forman un papel importante e influyen notablemente en el resultado final de todos los procesos. En el caso de la dosificación desde los silos a la báscula, según hemos esgrimido del trabajo, parece que es más notorio. Un silo tiene un elemento de dosificación concreto que funciona de una manera diferente al resto, aunque sea de una naturaleza similar. Esto puede deberse a la propia disposición dentro de la zona de la báscula. Estamos hablando de básculas que pueden medir en torno a 10-12 metros de longitud y 4-5 de anchura, por lo que parece que los resultados de la dosificación dependen en gran medida de la zona en la que esté situado el silo y de las condiciones iniciales del proceso.

Durante este trabajo se ha podido deducir claramente que hay varios factores que influyen en el proceso de dosificación y que pueden alterar el resultado final de la misma:

El peso inicial de la báscula, el orden de la dosificación, la temperatura, el día de la semana, la hora, la materia a dosificar, etc.

Gracias a incluir todos estos parámetros a modelo hemos conseguido caracterizar el problema y obtener unos resultados válidos para mejorar nuestros procesos, que como hemos visto, no sustituye al anterior método, pero si lo complementa y lo mejora.

Cabe resaltar que en un proceso de machine learning, la fase de análisis es crucial, sobre todo para entender el problema y para orientar en mayor medida los modelos entrenar y realizar un preprocesamiento de los datos más eficiente. También resaltar la importancia de la generación de características que hace que nuestros datos sean más expresivos y fáciles de caracterizar.

Este es un caso de éxito en el que la ciencia de datos y el machine learning son capaces de mejorar procesos automáticos dentro de la industria agroalimentaria y gracias a ello la calidad de los productos fabricados con la consiguiente mejora de negocio que aporta a la empresa.

8.2 Trabajo futuro

Técnicas de machine learning podría ser aplicadas en multitud de situaciones en los procesos automáticos dentro de la industria agroalimentaria, ya que como hemos reiterado en varias ocasiones los elementos mecánicos generan bastante incertidumbre y los elementos de medición son escasos, además hay multitud de análisis de información que nos podría ayudar a la toma de decisiones.

Centrándonos en nuestro problema de dosificación, podríamos buscar la manera de mejorarlo analizando y generando un modelo que sea capaz de sugerirnos qué materias debemos almacenar en cada silo. Según hemos visto en este trabajo, el silo es un factor importante en el resultado del proceso, pero sí que es cierto que hay materias que marcan diferentes patrones de funcionamiento en diferentes silos.

Almacenar las materias en los silos en los que menos desviaciones se produzcan mejorará la calidad de estas y en general nuestro sistema automático.

Bibliografía

Web oficial librería scikit-learn en Python.

https://scikit-learn.org

Web oficial de librería matplotlib.

https://matplotlib.org

Web oficial de librería seabor.

https://seaborn.pydata.org

Web oficial API REST OpenData de AEMET.

http://www.aemet.es/es/datos_abiertos/AEMET_OpenData