Aprendizaje Automático (2014-2015)
3º Grado en Ingeniería Informática
Universidad de Granada

Proyecto Final: Airfoil Self-Noise Data Set

Pedro Antonio Ruiz Cuesta
Ignacio Martín Requena

23 de junio de 2015

Índice

Índice de figuras

1. Descripción del problema

Nuestro objetivo durante el desarrollo de todo el proyecto será estudiar la base de datos *Airfoil Self-Noise Data Set* con el fin de poder analizar, predecir y obtener aquellos atributos relevantes para reducir el nivel de presión mediante técnicas de regresion el nivel de presión de sonido escalado (Scaled sound pressure level) medido en decibelios.

Para ello utilizaremos diferentes técnicas de regresión con el fin de contrastar y analizar qué modelo se ajusta mejor para resolver el problema.

La documentación de todo el proceso de experimentos y pruebas a partir del cual se obtuvo la base de datos que usaremos la podemos encontrar en un documento oficial publicado por la NASA en 1989 con el titulo de *Airfoild Self-Noise and Prediction*.

2. Descripción de la base de datos usada

Nuestra base de datos se compone de:

- 1503 observaciones
- 6 variables:
    1. Frecuencia, en Hercios. (**Frequency**)
    2. Ángulo de ataque, en grados. (**AOA**)
    3. Longitud de la cuerda, en metros. (**CL**)
    4. Velocidad de la corriente de aire, en metros por segundo. (**FSV**)
    5. Desplazamiento lateral, en metros. (**SSDT**)
    6. Nivel de presión escalado del sonido, en decibelios. (**SSPL**)

Nosotros utilizaremos una sola respuesta, el nivel de presión escalado del sonido, representado frente a uno o mas del resto de predictores de nuestra base de datos.

Todas las variables de la base de datos contienen datos de tipo real menos la frecuencia, que es de tipo entero.

3. Resumen del análisis de datos realizado para determinar las técnicas seleccionadas

En el análisis de nuestra base de datos hemos obtenido los siguientes valores para nuestra **variable respuesta (SSPL)**:

3.1. Valores estadísticos descriptivos

Frequency	AOA	CL	FSV	SSDT	SSPL
Min. : 200	Min. : 0.000	Min. :0.0254	Min. :31.70	Min. :0.0004007	Min. :103.4
1st Qu.: 800	1st Qu.: 2.000	1st Qu.:0.0508	1st Qu.:39.60	1st Qu.:0.0025351	1st Qu.:120.2
Median : 1600	Median : 5.400	Median :0.1016	Median :39.60	Median :0.0049574	Median :125.7
Mean : 2886	Mean : 6.782	Mean :0.1365	Mean :50.86	Mean :0.0111399	Mean :124.8
3rd Qu.: 4000	3rd Qu.: 9.900	3rd Qu.:0.2286	3rd Qu.:71.30	3rd Qu.:0.0155759	3rd Qu.:130.0
Max. :20000	Max. :22.200	Max. :0.3048	Max. :71.30	Max. :0.0584113	Max. :141.0

Figura 3.1: Valores estadísticos descriptivos

Como podemos ver en la **Figura 3.1** el rango de valores de SSPL oscila en esta base de datos entre 103.4 y 141.0 lo que nos lleva a pensar que los valores van a estar muy concentrados (en un rango de apenas 40 decibelios). La mayor cantidad de valores va a estar entorno a 125.7, el valor de la mediana para esta variable. Otro dato a destacar es que la media posee un valor de 124.8, lo cual nos lleva a intuir que, al estar aproximadamente en la mitad entre el valor mínimo y el máximo, es posible que nuestros datos se encuentren dispersos entre estos dos valores, o lo que es lo mismo, no se encuentran muy concentrados cerca del valor mínimo o del máximo.

Otro valor estadistico que nos ayuda a obtener información sobre los posibles modelos a ajustar es la **matriz de correlación**:

	Frequency	AOA	CL	FSV	SSDT	SSPL
Frequency	1.000000000	-0.27276454	-0.003660639	0.133663831	-0.230107353	-0.3907114
AOA	-0.272764536	1.00000000	-0.504868150	0.058759565	0.753393785	-0.1561075
CL	-0.003660639	-0.50486815	1.000000000	0.003786629	-0.220842431	-0.2361615
FSV	0.133663831	0.05875957	0.003786629	1.000000000	-0.003974013	0.1251028
SSDT	-0.230107353	0.75339378	-0.220842431	-0.003974013	1.000000000	-0.3126695
SSPL	-0.390711412	-0.15610753	-0.236161512	0.125102801	-0.312669506	1.0000000

Figura 3.2: Matriz de correlación

En la **Figura 3.2** podemos ver como a priori ningún predictor simple esta muy relacionado con la respuesta , ya que su coeficiente de correlación están lejanos a 1 o -1. Aun así se puede observar que con las variables Frecuency y SSDT quizá se podría obtener resultados relativamente válidos. Como el coeficiente de correlacion nos especifica el grado de relacion entre variables para un modelo lineal simple podemos intuir que el modelo real al que se ajustan los datos probablemente no sea este. Aun así, como tenemos rela-

tivamente pocos predictores es viable utilizarlos todos o algunas combinaciones de estos para realizar el ajuste.

3.2. Visualizaciónes gráficas

En primer lugar representaremos gráficamente todas las variables con todas para dar un primer vistazo a la relación que existe entre ellas.

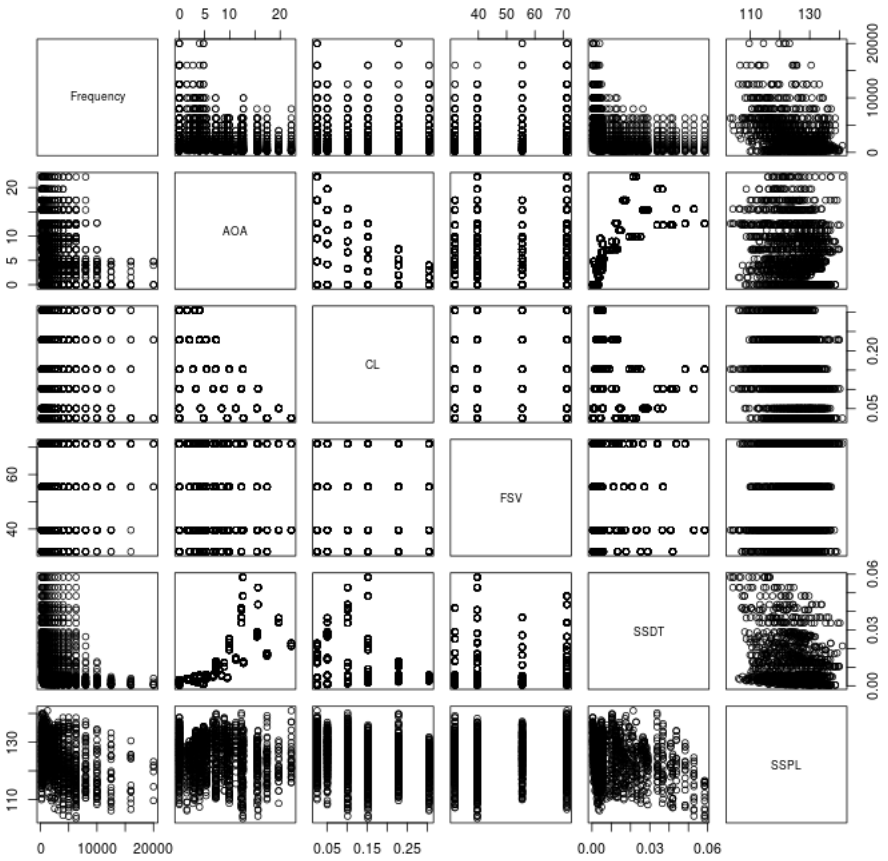

Figura 3.3: Relacion gráfica entre pares de variables

A simple vista la relación entre nuestra respuesta y los predictores no se observa que no se observa que exista a simple vista, tal y como hemos visto en la **Figura 3.1**. Fijandonos en el resto de relaciones vemos como entre ADA y SSDT si existe una relación logarítmica, lo que nos puede ayudar a la hora de buscar alguna combinación entre variables

Todo lo comentado anteriormente sobre la **Figura 3.3** se puede ver graficamente en el siguiente **histograma**:

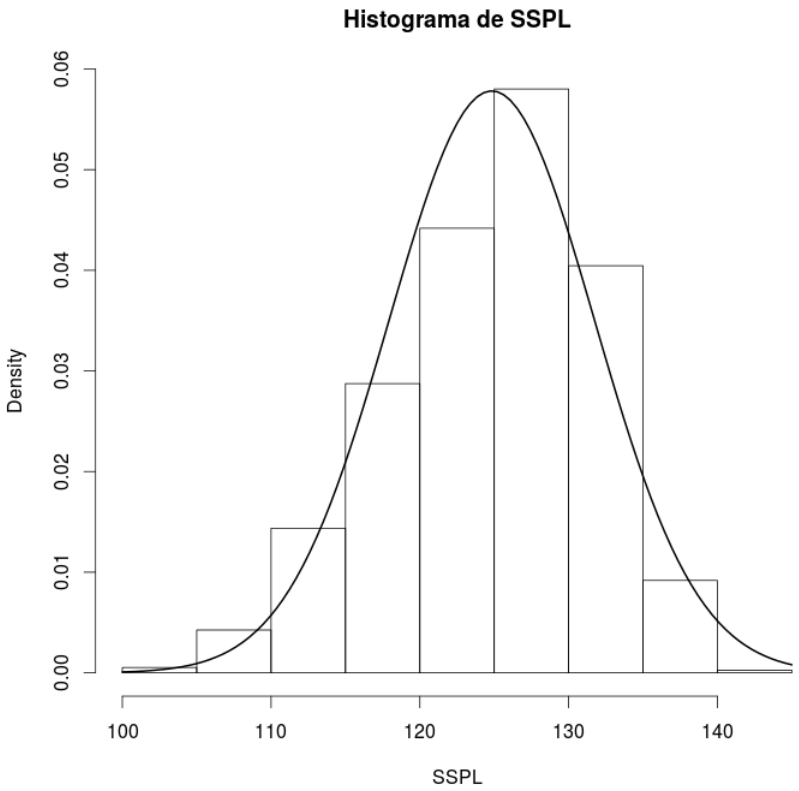

Figura 3.4: Histograma SSPL

Como se esperaba, el mayor número de observaciones con un valor SSPL cercano a la mediana no lo encontramos justo en el centro si no que se encuentra un poco desplazado, tal y como nos indicaba la mediana teniendo un valor mayor que la media.

Otro gráfico que nos puede ayudar a la hora de analizar las variables es el **diagrama de cajas**:

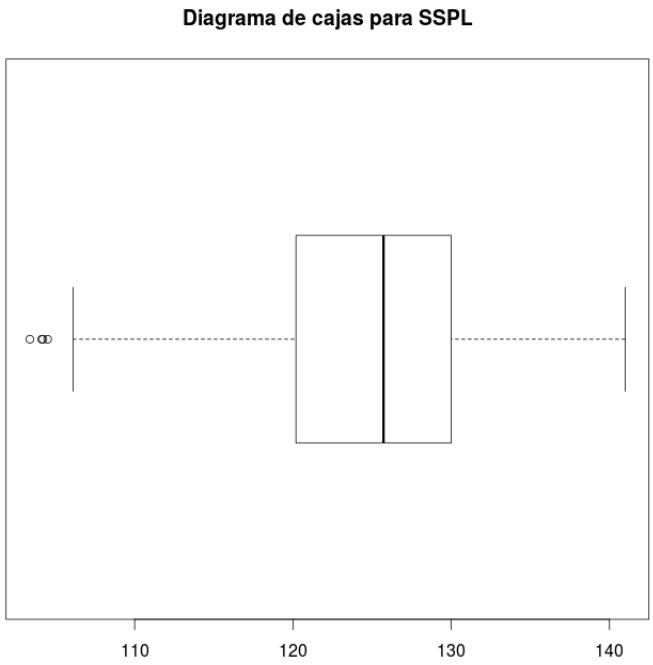

Figura 3.5: Diagrama de cajas SSPL

En este diagrama observamos que la varianza de los valores menores que la mediana va a ser bastante mayor que los que superan a esta. En concreto, si calculamos la varianza de cada uno de los dos subconjuntos obtenemos que para los valores por debajo de la mediana su varianza es 2,27 veces mayor que los que superan la mediana (22,64 frente a 9,97).

4. Resumen de la metodología llevada a cabo en el ajuste de los modelos empleados

4.1. Elección de datos de test y train

Para la selección de los conjuntos de train y test hemos dividido la base de datos aleatoriamente en dos subconjuntos con el 80 % de las observaciones originales y un 20 % respectivamente.

4.2. Análisis de de la dependencia entre variables

Ahora vamos a analizar las variables que nuestra base de datos posee para elegir aquellas que de verdad nos aporten información a la hora de ajustar nuestros modelos, es decir, aquellas variables o combinaciones de variables que hacen nuestro modelo lo mas simple

y eficaz posible.

En primer lugar representaremos como el coeficiente RSQ va creciendo en función del número de variables seleccionadas para un modelo de regresión, incluyendo las combinaciones de las variables originales:

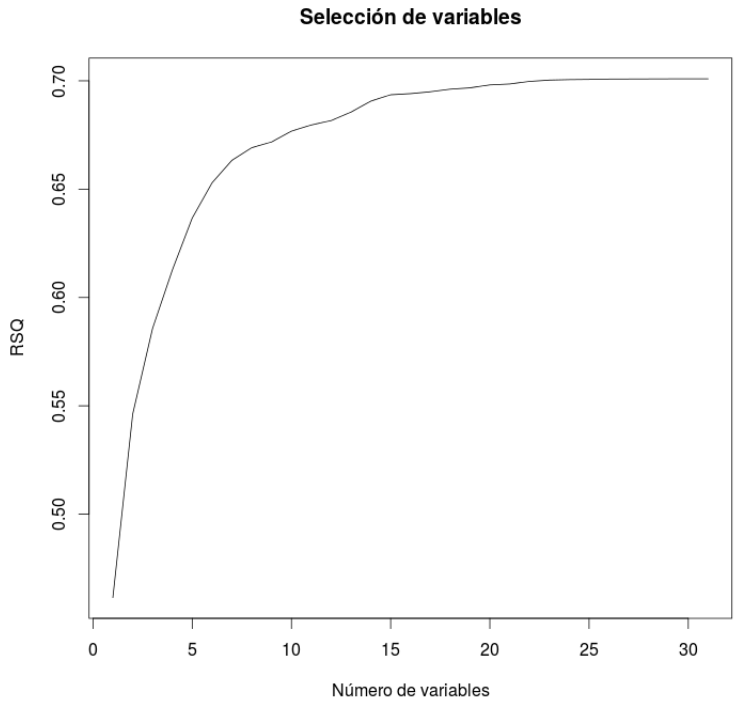

Figura 4.1: Numero de variables usadas vs RSQ

Como podemos ver en la **Figura 4.1** el coeficiente RSQ para modelos lineales va a ser cercano a 0.7, ya que este valor es el resultado de usar todas las posibles variables incluyendo las combinaciones entre ellas que sería el máximo ajustable que podriamos conseguir, por lo que con nuestros datos a lo sumo conseguiremos que la capacidad del modelo lineal para explicar la variación en la variable respuesta no será superior al 70 %. Para modelos no lineales no podemos predecir a priori cual va a ser el limite máximo de RSQ.

Con estos datos deducimos que el incremento de RSQ a partir de usar mas de 20 variables es poco significativo e incluye complejidad a nuestro modelo, por tanto nuestro numero de variables a elegir será este límite.

Aplicando el método de selección de busqueda exhaustiva que trata de ajustar un numero definido de modelos con todas las combinaciones de variables predictoras posibles

concretaremos las 20 variables a usar.

Frequency	AOA	FSV
Frequency:AOA	Frequency:CL	AOA:CL
Frequency:FSV	AOA:FSV	CL:FSV
Frequency:SSDT	CL:SSDT	FSV:SSDT
Frequency:AOA:CL	Frequency:CL:FSV	Frequency:AOA:SSDT
Frequency:CL:SSDT	AOA:FSV:SSDT	Frequency:AOA:CL:FSV
Frequency:AOA:CL:SSDT		Frequency:CL:FSV:SSDT

Tabla 4.1: Variables simples y combinadas usadas en el modelo

4.3. Elección de técnicas a utilizar

A raíz de lo expuesto en el apartado anterior y a la vista de los resultados del analisis de datos no podemos intuir la tendencia que van a seguir los datos, ya que en su representación, la variable SSPL no muestra una forma definida. Por tanto se ha elegido una técnica de regresión lineal múltiple debido a su simplicidad y a que es un buen metodo para preveer la dificultad de ajuste de las diferentes técnicas a los datos.

Por otro lado se ha optado por aplicar una técnica de Support Vector Machine, ya que al ser versátil aporta una mayor potencia y capacidad de decisión (al poseer diferentes nucleos para cambiar el tipo de ajuste) y por tanto puede que de mejores resultados.

5. Detalles de la metodología de entrenamiento para cada uno de los modelos empleados

Para realizar el entrenamiento de las técnicas empleadas se han construido dos subconjuntos de muestras de la base de datos. El primero de ellos, el conjunto de train lo usamos para realizar el ajuste del modelo (conjunto de train) y, una vez ajustado el modelo, con el subconjunto restrante. Como ya se concretó en el apartado 4.2, la distribución de los conjuntos de train y test está determinada en una relación 80-20 % respectivamente.

Una vez separada la base de datos en los dos subconjuntos comentados anteriormente, nos centramos en el conjunto de train, con el que obtendremos las relaciones que guardan los predictores con la respuesta. El siguiente paso será elegir, de entre todos los posibles predictores a usar, los que más nos convengan, buscando un compromiso complejidad y ajuste sobre los datos supervisados. Para hacer esta elección de predictores se ha decidido aplicar una técnica de búsqueda exhaustiva ya comentada anteriormente.

Elegido el modelo a ajustar, se procede a realizar dicho ajuste sobre el conjunto de train, este ajuste será el que haga las predicciones a partir de la función ajustada.

Después de esto, se realizan las predicciones sobre el conjunto de test. Dichas predicciones, serán las que nos den la bondad del ajuste realizado por la técnica seleccionada.

6. Comparación entre sí de las distintas técnicas usadas con valoración de la idoneidad de cada una para el problema en estudio.

Para el estudio de la comparación entre los modelos utilizados nos centraremos únicamente en los valores obtenidos para el conjunto de test, que son los realmente relevantes.

6.1. Comparación gráfica de las predicciones

Puede ser interesante representar cada una de las predicciones frente a los datos reales para hacernos una idea de que modelo tendrá mayor bondad de ajuste.

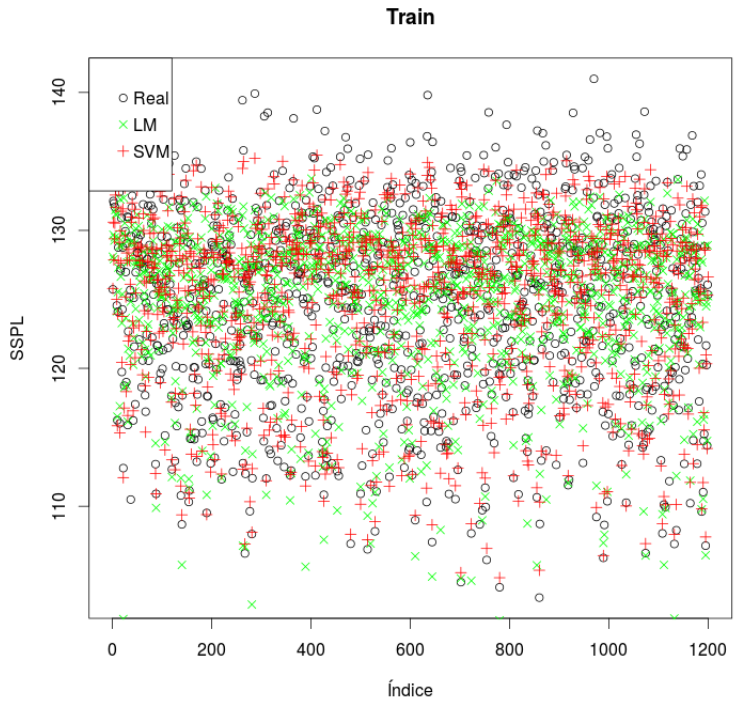

Figura 6.1: Valores reales y valores predichos para train

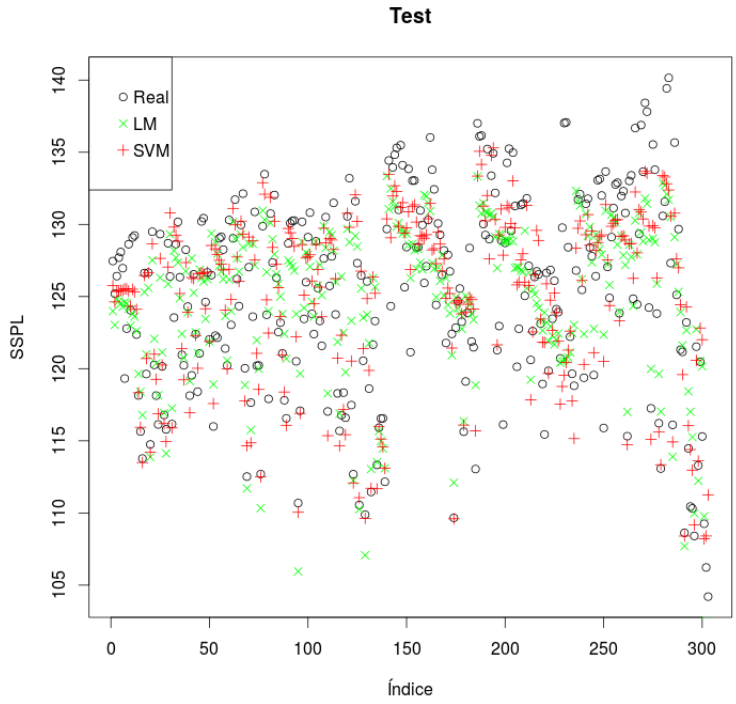

Figura 6.2: Valores reales y valores predichos para test

A priori vemos que LM tiene algunas predicciones que no se acercan a los datos reales, por ejemplo en zonas en la que los datos están más dispersos. En contraste, SVM no se aleja demasiado de las nubes de puntos reales y, en zonas con alta dispersión su capacidad de predicción es aceptable.

Para la comparacion entre las técnicas de ajuste se han obtenido los siguientes valores estadísticos:

	RSQ	**MSE**	**SD**
LM	0.7	14.35	5.76
SVM	0.81	7.69	6.21

Tabla 6.1: Comparación de bondad de ajuste para train

	RSQ	MSE	SD
LM	0.65	14.68	5.57
SVM	0.75	9.66	5.99

Tabla 6.2: Comparación de bondad de ajuste para test

6.2. R^2

Este valor estadístico representa el tanto por ciento de la varianza de los datos que el modelo es capaz de explicar.

Al principio comentamos que un modelo lineal no podia tener un valor de R^2 superior a 0.7, en cambio vemos como SVM supera tal valor. Esto es debido a que se ha utilizado un nucleo radial para el entrenamiento del ajuste. Se hicieron algunas pruebas con otros núcleos como el polinomial, sigmoidal y lineal, pero los resultados obtenidos no fueron superiores al los del uso de un núcleo radial.

Basandonos en los valores obtenidos para cada modelo podemos observar como SVM es capaz de explicar los datos introducidos un 15,38 % más que LM.

6.3. Error medio cuadrático

Es un valor estadístico que mide el error en la estimación de un modelo determinado, es decir, es un forma de evaluar la diferencia entre un estimador y el valor real de la cantidad que se quiere calcular. El MSE mide el promedio del cuadrado del error, siendo el error el valor en la que el estimador difiere de la cantidad a ser estimada.

Como se puede observar en la **Tabla 6.2**, el valor de este estadístico es menor en SVM que en LM, y por tanto el error que se comete en las predicciones de SVM será menor que en las de LM.

6.4. Desviación típica

La desviación típica es la "distancia" entre la media de un conjunto de datos y el valor de cada uno de los datos. Este valor puede sernos útil si comparamos las desviaciones típicas entre las predicciones de cada ajuste y la desviación típica de que realmente tienen los datos (nuestro conjunto de test).

Como la desviación típica real del conjunto de test es de 6.91, cuanto mas cercano esté el valor de la desviación estándar de una desviación más cerca estaremos de realizar unas predicciones acordes con la realidad. En nuestro caso, como podemos ver en la **Tabla 6.1**, SVM se acerca más al valor real de la desviación típica que LM y, por tanto, tendrá una mayor capacidad de predicción.

7. Valoración final de los resultados obtenidos.

Cuando se comenzó a hacer el análisis exploratorio en los primeros pasos se pensaba que no se podría sacar una predicción decente, ya que los datos de la variable respuesta estaban muy repartidos y la dependencia entre variables no era significativa a simple vista (**Figura 3.3**).

Realizando un análisis de dependencia entre variables y combinandolas entre ellas se consiguió obtener un modelo que aportaba más información y, por tanto, mejores resultados.

Con las técnicas seleccionadas la bondad de ajuste conseguida se puede considerar aceptable en relacion a lo esperado. Los resultados obtenidos en las predicciones de las dos técnicas de ajuste aplicadas (LM y SVM) así como los valores estadísticos obtenidos (R^2, MSE y SD) muestran como la diferencia entre las predicciones y los valores reales es asumible.

8. Conclusiones.

Al ser un problema con un alto grado de incertidumbre y de interacción entre los atributos, el error asociado a la medición de los valores de nuestro modelo será alto y, por tanto,

Como el objetivo final de este estudio es intentar predecir el nivel de presion de sonido con el fin de, a partir de los modelos de predicción obtenidos, poder reducirlo, se llega a la conclusión de que si se mejoran los valores de frecuencia (predictor más influyente), ángulo de ataque y la velocidad de la corriente de aire (variables más frecuentes en nuestro modelo, especificadas en la **Tabla 4.2**) el nivel de presión escalado del sonido mejorará también.