Logotipo, nombre de la empresa

Descripción generada automáticamente

Máster en Data Science

2020-2021

Trabajo Fin de Máster

*“Data Science en AIM Ibérica”*

Jesús Paredes Vallés

ÍNDICE

[INTRODUCCIÓN 3](#_Toc77621112)

[DATOS 4](#_Toc77621113)

[METODOLOGÍA 4](#_Toc77621114)

[1. Modelo predictivo 5](#_Toc77621115)

[2. “Enzaprost” 7](#_Toc77621116)

[3. Analizadores 10](#_Toc77621117)

[4. Diluyentes 12](#_Toc77621118)

[5. Fecha de nacimiento vs Calidad Seminal y Supervivencia 15](#_Toc77621119)

[RESUMEN DE RESULTADOS Y CONCLUSIONES 16](#_Toc77621120)

[1. Modelo Predictivo 16](#_Toc77621121)

[2. “Enzaprost” 17](#_Toc77621122)

[3. Analizadores 18](#_Toc77621123)

[4. Diluyentes 20](#_Toc77621124)

[5. Fecha de nacimiento vs Calidad Seminal y Supervivencia 21](#_Toc77621125)

[MANUAL DE USUARIO 21](#_Toc77621126)

# INTRODUCCIÓN

* Situación general de la industria porcina

##### En los últimos años, el porcino español se ha convertido en un sector clave para la economía de nuestro país. Con una producción más o menos estable por encima de los 4 Mt, se ha logrado consolidar la privilegiada posición estratégica del sector en el contexto del sistema agroalimentario español como uno de los soportes básicos de toda la economía nacional.

##### En total, el sector porcino, con una facturación superior a los 15.000 M€, representa el 36,4% de la producción final ganadera española y el 17,4% de la producción final agraria, solo superado por frutas y hortalizas. Asimismo, supone el 1,4%del PIB nacional y el 14% del Producto Interior Bruto Industrial.

Por su parte, en el ámbito internacional, factura más de 5.000 M€, que nos sitúan como el segundo exportador europeo y el tercero a nivel mundial y como uno de los grandes sectores exportadores de la economía española. En este sentido, si echamos la vista atrás vemos el destacado crecimiento del sector en el exterior en los últimos 10 años, superior al 66% en volumen y al 108% en valor.

Pero si hay un dato significativo es el que refleja la balanza comercial agroalimentaria, positiva en más de 4.500 M€ al importar carne de cerdo por unos 500 M€ al año, la décima parte de lo que exportamos. Esos son fondos que vienen a España y revierten en nuestras empresas, nuestros ganaderos y nuestros trabajadores. Y lo hacen especialmente en las zonas rurales, que acogen en gran parte la actividad ganadera y empresarial del sector porcino español, convirtiéndose de esta forma en un motor de la economía y el empleo de miles de pueblos de toda la geografía española, en los que se concentran buena parte de los más de 300.000 empleos directos que genera el sector porcino español y los más de 1 millón de puestos de trabajo indirectos que penden de su actividad.

Las más de 80.000 granjas de porcino de nuestro país suponen el 10% de todas las explotaciones agrarias de España y aportan el 10,3% de todo el empleo agrario español, dando futuro a nuestras zonas rurales, que son en definitiva las zonas a las que estamos ligados por la propia naturaleza del sector.

* Situación general de la industria de inseminación porcina

Las mejoras en los sistemas productivos, la nutrición, las instalaciones, la mejora genética, los avances el control de las enfermedades y el desarrollo en el campo de la reproducción porcina han permitido alcanzar niveles de producción de carne en cantidad y calidad difícilmente imaginables años atrás. Desde el inicio en el uso de la inseminación artificial porcina en las primeras décadas del siglo XX hasta nuestros días, se ha producido una gran evolución en la aplicación y desarrollo de la técnica, asociados a las evoluciones que ha sufrido la industria porcina.

La inseminación artificial porcina permite la amplia difusión de los genes de animales genéticamente valiosos, controlar la calidad seminal de los reproductores, evitar y controlar enfermedades y reducir los costes de mantenimiento de los verracos. La utilización de esta técnica en el ganado porcino tiene un alto grado de implantación en los países desarrollados. Se estima que en España su utilización está por encima del 95%, de manera que unos 7500 verracos alojados en los centros de inseminación, produciendo unos 12,5 millones de dosis año serían los responsables de esas cubriciones. Aunque según los últimos datos disponibles, en España hay 32.400 verracos y un total de los 2,47 millones de cerdas reproductoras del censo nacional.

# DATOS

Los datos han sido proporcionados por la empresa en formato .xlsx (Excel) sumando un total de 11 archivos diferentes, los cuales recogen información desde finales del año 2012 hasta comienzos del año 2021, sumando un total de 153699 registros la base de datos más extensa. Esta base de datos contiene registros por cada extracción realizada a cada verraco en determinados días, almacenando información como puede ser la calidad de la muestra, la motilidad, el número de dosis potenciales y reales extraíbles de la misma, así como información relativa al verraco (raza, origen, procedencia, centro en el que se encuentra, su número de identificación, etc).

# METODOLOGÍA

El desarrollo de este proyecto de Data Science se ha basado en la resolución de cinco cuestiones propuestas por la empresa con el objetivo de extraer de sus bases de datos información clave para guiar nuevos comportamientos que permitan obtener una ventaja competitiva frente a sus rivales dentro del sector.

Así mismo, dichas cuestiones se pueden resumir en:

1. Modelo de Machine Learning con el objetivo de predecir la productividad de los verracos en función a sus diferentes características.
2. Influencia del medicamento “Enzaprost” sobre la calidad seminal y la supervivencia del verraco en el centro.
3. Diferencias entre los analizadores.
4. Influencia de los diluyentes en las muestras.
5. Relación entre la fecha de nacimiento de los verracos y su calidad seminal, así como con su supervivencia en el centro.

## Modelo predictivo

El objetivo de este apartado es realizar un modelo predictivo en base a determinadas características que posteriormente se mencionarán. Debido a la alta competencia en el sector porcino, conocer de antemano si un verraco será productivo o no en base a determinadas características puede suponer una ventaja competitivo muy importante frente a los rivales dentro de la industria. Así mismo, este modelo intentará predecir con la mayor precisión posible dicho resultado en base a las características de los verracos que se adjuntan a continuación:

* Raza del verraco.
* Genética del verraco.
* Procedencia del verraco.
* Dosis obtenidas en cada extracción.
* Centro en el que los verracos se encuentran.

Entrando más en materia, el primer paso es realizar una exploración inicial de los datos. De esta forma, primero se extrae el año y el mes de la fecha de cada una de las extracciones registradas en la base de datos “EYACULADOS”. A continuación, se procede a analizar la cantidad de muestras diferentes que hay en las variables anteriormente citadas, con el objetivo de ir conociendo los datos, así como apreciar posibles futuros puntos de error.

El siguiente paso consiste en definir cuando un verraco es productivo o no. Para ello (y contando con la información aportada por la empresa), se define como productivo (se asigna un valor de 1 en la columna “PRODUCTIVO”) a aquellos verracos que hayan producido más de 80 dosis en el año de la correspondiente extracción.

La siguiente etapa del modelo consiste en realizar una limpieza de los datos, evitando así tratar con NaN, NaT, espacios vacíos, registros carentes de sentido o duplicados, entre otros errores posibles.

Tabla

Descripción generada automáticamente

Al ser variables categóricas se ha decidido eliminar aquellos registros que contengan NaN, puesto que éstos no pueden ser sustituidos por valores numéricos. Además, el impacto sobre el tamaño global de los datos del modelo no es muy significativo.

No se encuentra nada extraño en las variables de Centro, Dosis en las extracciones, Productivo y Genética, encontrándose errores en las variables Procedencia y Genética.

Tabla

Descripción generada automáticamente

En cuanto a la procedencia, tres mismos registros estaban almacenados con algunos caracteres distintos. Y, por último, existían dos razas denominadas “-“ y “0”, las cuales han sido eliminadas.

Tabla

Descripción generada automáticamente

El último paso antes de entrenar al modelo es realizar un correcto Features Engineering. Esta etapa consiste en codificar los valores de las variables categóricas en variables numéricas que puedan ser introducidas al modelo. Para ello, el procedimiento será siempre el mismo para todas ellas, haciéndolo siempre sobre el conjunto de entrenamiento. Inicialmente se separan todos los datos en dos grupos; uno de entrenamiento denominado “train” y otro de comprobación de los resultados, denominado “test”. A continuación, se agrupan todos los verracos por sus razas y por su productividad (1 productivo, 0 no productivo) y se realiza la operación de la media sobre ellas, mostrando así un número en tanto por uno que representa la probabilidad de ser productivo siento de determinada raza, genética, centro, etc.

Interfaz de usuario gráfica, Texto, Aplicación

Descripción generada automáticamente

Una vez hecho este proceso para todas las variables categóricas que intervengan en el modelo, el siguiente paso consistirá en introducirlas al conjunto de “train” realizando una función denominada “merge”.

Texto

Descripción generada automáticamente

Se ha escogido el método “inner” para hacer la función “merge” porque así se pueden obtener todas la información relativa a las dos palabras llave en las dos bases de datos.

El último paso que se ha realizado antes de entrar con el modelo en sí es un redondeo a tres decimales de los valores float, puesto que daba problemas el hecho de no acotarlos. Por otro lado, se ha realizado un escalado mediante la función “StandarSacler” de la librería “sklearn.preprocessing”, con el objetivo de escalar todos los valores entre -1 y 1, para que ninguno tenga más peso sobre el valor de la decisión final por el simple hecho de ser diferente al resto en cuantía.

Interfaz de usuario gráfica, Texto, Aplicación

Descripción generada automáticamente

El último paso de esta pregunta consta de la realización del modelo en sí. Para ello, se han probado tres clasificadores diferentes para comprobar cuál de ellos realiza la predicción más precisa y en menos tiempo. De esta forma, los algoritmos probados han sido un “Decision Tree Classifier”, un “Logistic Regressor”, un modelo “Random Forest” y un modelo “SVM Classifier”.

Los resultados se analizarán en el apartado “RESUMEN DE RESULTADOS”, pero el modelo que mejor resultados en cuanto a métricas ha sido el modelo “Random Forest”, mientras que el segundo más rápido y el segundo también en mejores métricas ha sido el modelo “Decision Tree Classifier”.

## “Enzaprost”

El objetivo de este apartado es comprobar el efecto del medicamento denominado “Enzaprost” sobre los verracos inyectados.

Este medicamento provoca en los verracos una estimulación puntual de la libido, con la posibilidad de producir una disminución en la producción de dosis a corto-medio plazo. El reto consiste en analizar el efecto de este medicamento sobre la producción de dosis del verraco y la calidad seminal de las mismas.

Interfaz de usuario gráfica, Gráfico, Gráfico de dispersión

Descripción generada automáticamente

La foto adjunta anteriormente es un ejemplo del estudio que se realizará para todos los verracos de la base de datos. En este caso, se grafica la información sobre el verraco con número de identificación 2221, el cual servirá de ejemplo para explicar el desarrollo de la cuestión.

Como introducción, se define la línea azul como la representación de las dosis reales producida por el verraco 2221 a lo largo de su vida. Por otro lado, con marcas rojas, se resaltan los días en los que se administró una inyección de ENZAPROST sobre el verraco. Finalmente, la gráfica de color verde muestra el nivel de excitación del verraco, con el objetivo de poder analizar qué ocurre en niveles más altos de excitación con las dosis reales y la calidad seminal.

Al existir un gran número de verracos (1453 verracos) es imposible hacer esta gráfica para cada uno de ellos porque se obtendría un resultado muy difícil de comprender. El grueso de la solución se engloba en la ejecución de la celda de la imagen adjunta a continuación, la cual se procede justamente después a explicar los bloques que la componen.

Texto

Descripción generada automáticamente con confianza baja

En primer lugar, se crean una listas vacías en las que se almacenará la información de valor extraída del análisis. Así mismo, se elabora un bucle el cual ejecutará el código encargado de extraer dicha información tantas veces como verracos existan en la base de datos de la empresa.

Texto

Descripción generada automáticamente con confianza media

A continuación, se procede a filtrar tanto la base de datos global de los verracos como la base de datos de las inyecciones en ellos según el verraco a analizar en cada iteración del bucle.

Interfaz de usuario gráfica, Texto, Aplicación, Correo electrónico

Descripción generada automáticamente

El siguiente bloque trata de la elaboración del modelo creado para definir teóricamente el nivel de excitación de los verracos. Antes de entrar en detalles, se quiere hacer hincapié en que dicho modelo es una representación gráfica aproximada de lo que a priori se supone que es el comportamiento de las hormonas excitadas por la inyección en el cuerpo de un verraco. No se han encontrado estudios científicos que detallen exactamente la representación matemática que explica el comportamiento de las hormonas excitadas en los cuerpos de los verracos, así que dicha representación es una aproximación.

Una vez se ha aclarado lo necesario, es importante mencionar que para la descripción del comportamiento de excitación en los sementales se ha seguido una de las muchas representaciones de un filtro paso bajo, cuya expresión matemática seguida en esta aplicación es la siguiente:

Inicialmente se redefine la base de datos de los tratamientos según el orden de la columna denominada “FECHA”, con el objetivo de poder representar posteriormente de una forma cronológica. Así mismo, se almacena la diferencia en días entre una inyección y la siguiente. De esta forma, la variable adquirirá valores de 1 o 0 si existe inyección o no, mientras que la variable “prev\_value” almacena la información anterior, teniendo así “memoria” pudiendo valorar los niveles de excitación del verraco de una forma numérica aproximada.

Por último, es importante mencionar que los valores de las variables “tau” y “step” son valores establecidos basados en una experimentación visual de la representación gráfica del primer verraco, pero que son usados por igual para el resto de los verracos, por lo que hay unificación de resultados.

Texto

Descripción generada automáticamente

Finalmente, se procede a almacenar la información obtenida en la base de datos denominada “x\_model” para posteriormente almacenar en la lista denominada “model\_inyection” los diferentes valores de excitación de cada verraco, así como filtrar por fechas para almacenar la información de las dosis reales producidas únicamente durante la etapa de excitación de los diferentes verracos.

Texto

Descripción generada automáticamente con confianza baja

Las conclusiones serán analizadas en apartados posteriores.

## Analizadores

En lo respectivo al objetivo de este apartado, consiste en realizar un estudio de los diferentes analizadores contratados por la empresa.

Así mismo, estos “analizadores” son los encargados de realizar un estudio detallado de la muestra extraída del verraco, analizando parámetros como motilidad, calidad de movimiento, cantidad de esperma, formas anormales, etc.

Este proceso es uno de los pocos que queda sin automatizar de todo el proceso de elaboración de las dosis de eyaculados (producto final de la empresa). Estos analizadores realizan un proceso de aprendizaje minucioso con el objetivo de minimizar posibles errores de interpretación de la muestra. Aun así, realizan una clasificación muy subjetiva muy interesante a analizar.

En primer lugar, se procede a realizar un análisis inicial de las bases de datos en las que se oculta la información a extraer y analizar. El objetivo es realizar una primera toma de contacto en la

De esta forma, se ha elaborado una tabla informativa acerca de todas las muestras tratadas por los distintos analizadores contratados por la empresa

Tabla

Descripción generada automáticamente

En la tabla adjunta anteriormente se han creado nuevas columnas con el objetivo de analizar y comparar los valores de las mediciones entre analizadores, siendo de importancia destacar la detección de formas anormales ya que, al igual que la calidad de movimiento y motilidad, son valores subjetivos a cada analizador. Es muy importante para la empresa encontrar aquellos analizadores más propensos a detectar formas anormales para, por ejemplo, reentrenarlos o cambiar de entrenamiento/entrenador. El código elaborado para la correcta obtención de la tabla anteriormente adjuntada es el siguiente:

Texto

Descripción generada automáticamente

El funcionamiento general es el mismo que el funcionamiento del bloque de código del apartado anterior; se crea una lista con los identificadores de todos los empleados (lista2) y se elabora un bucle que ejecute determinadas líneas de código por cada uno de los analizadores.

Como se puede apreciar en la imagen anteriormente adjuntada, se calculan parámetros con el objetivo de comprobar la eficiencia de los analizadores realizando estudios sobre las muestras extraídas de los sementales. De esta forma, se pueden destacar parámetros como dosis admitidas, dosis rechazadas, dosis totales analizadas y otros parámetros que comprobaran las notas asignadas en forma de calidad seminal, así como las detecciones rechazadas por formas anormales.

Mediante la interpretación de los resultados se obtendrán las conclusiones que se adjuntarán en el apartado CONCLUSIONES.

## Diluyentes

El objetivo de este apartado es analizar la influencia de los diferentes diluyentes sobre la calidad seminal, tanto en la base de datos de las extracciones analizadas al instante como la base de datos de estas después de cierto tiempo de conservación. Estos diluyentes son líquidos preparados para realizar una conservación mejor de las muestras extraídas a los verracos, siendo de gran importancia su composición química puesto que estos son una clara diferencia competitiva entre rivales dentro del sector. De esta forma, los diluyentes serán definidos con los números uno, dos y tres.

En primera instancia, es importante destacar que no existe la información de qué tipo de diluyente se está usando en cada una de las extracciones. De esta forma, mi tutor en la empresa me adjunto la siguiente información:

* El Diluyente 1 fue usado entre las fechas 03/11/2020 y 04/12/2020, así como entre 17/01/2021 y el 12/02/2021.
* El Diluyente 2 fue utilizado entre las fechas 31/05/2020 y 07/07/2020, y entre 21/03/2021 hasta la actualidad.
* El Diluyente 3 fue utilizado en el resto de las fechas.

El código elaborado para la distinción de cada una de los diferentes diluyentes es adjuntado a continuación, en el que se puede apreciar la definición de las fechas límite para cada uno de ellos y el establecimiento del tipo de diluyente en función a estas fechas.

Imagen que contiene Calendario

Descripción generada automáticamente

Una ves realizada la distinción anteriormente nombrada, se puede agrupar la base de datos en función a estos diluyentes y analizar parámetros estadísticos de sus valores de motilidad, calidad de movimiento y formas anormales, que son tres de los diferentes parámetros que la empresa pidió a estudiar. Además, se adjunta también la creación de una base de datos temporal con el objetivo de ser exportada a formato Excel para ser manipulada en Tableau.

Interfaz de usuario gráfica, Aplicación

Descripción generada automáticamente

Para el análisis de la calidad de movimiento se ha realizado un estudio similar, mientras que para el análisis de las formas anormales se ha optado por un diagrama de tarta en el que se representen todos los motivos de desecho por los cuales se han anulado las muestras analizadas por cada diluyente. En la imagen que se adjunta a continuación se puede apreciar lo anteriormente explicado, así como una separación de la representación del motivo de desecho “B”, el cual representa las formas anormales según el código establecido por los trabajadores de la empresa.

Gráfico, Gráfico circular

Descripción generada automáticamente

En el análisis de los valores de la bases de datos de “TOMAS-CONSERVACION” se ha incluido (como es obvio) una nueva dimensión; los diferentes días de conservación a los que han sido sometidas las muestras extraídas de los verracos. De esta forma, se ha procedido a agrupar los datos según los días de conservación de las extracciones y según el tipo de diluyente utilizado.

Texto, Tabla

Descripción generada automáticamente

Como se ha comentado anteriormente, de los datos de las conservaciones de las muestras se han obtenido conclusiones acerca de la motilidad y la calidad de movimiento. Para ello, se ha procedido a almacenar las informaciones de todos los diluyentes en listas anteriormente creadas vacías, teniendo en cuenta el siguiente punto de interés; al ser representados los tres diluyentes en la misma gráfica es necesario que ellos tengan la misma dimensión. Si se analiza la tabla anteriormente adjuntada, se puede apreciar que no todos los tipos de diluyentes aparecen en todos los días de conservación, siendo esto un problema a resolver.

Texto

Descripción generada automáticamente con confianza media

Así mismo y como se puede apreciar en la imagen superior, se han rellenado los huecos con NaN, creando así vectores de información con las mismas longitudes.

A la hora de mostrar por pantalla el gráfico de la información obtenida, se ha usado la función “fill\_between” la cual permite la representación de la desviación estándar de los valores de la media para cada diluyente. La representación, para el estudio de la motilidad, es la siguiente:

Gráfico, Gráfico de líneas

Descripción generada automáticamente

Mediante la interpretación de los resultados se obtendrán las conclusiones que se adjuntarán en el apartado CONCLUSIONES.

## Fecha de nacimiento vs Calidad Seminal y Supervivencia

En este apartado se pedía realizar un estudio entre la fecha de nacimiento de los verracos y la calidad seminal, así como con la supervivencia de estos en sus respectivos centros.

Para ello, en primera instancia se ha separado las fechas por meses y años, así como calculado los días de estancia del verraco en el centro. Para esto último se han utilizado la base de datos denominada “VERRACOS.xslx”, la cual posee la información de la fecha de llegada y salida de cada verraco en sus respectivos centros.

Para la matriz de correlación se ha elaborado otra base de datos con la información sobre la que se quiere saber la relación; mes, año, días de estancia en el centro, motilidad y calidad de movimiento. Se adjunta a continuación el resultado obtenido.

Gráfico

Descripción generada automáticamente

# RESUMEN DE RESULTADOS Y CONCLUSIONES

Este apartado consistirá en adjuntar los resultados obtenidos en cada uno de los estudios independientes realizados en este proyecto de Data Science.

## Modelo Predictivo

A continuación, se adjuntan los resultados obtenidos de los diferentes modelos de machine Learning para clasificación que se han empleado para predecir si un semental será productivo o no.

Tabla

Descripción generada automáticamente

En la tabla se pueden apreciar los valores de los diferentes medidores de la calidad de un modelo, pudiendo llegar a la conclusión de que los modelos “Decision Tree Classifier” y “Random Forest” son aquellos cuyas métricas son las más elevadas. Debido al tiempo de ejecución del modelo “Random Forest”, se ha elegido el modelo “Decision Tree Classifier” como el modelo más optimo en cuanto a tiempo de ejecución y métricas, de los cuatro modelos diferentes probados para este conjunto de datos. Además, se adjunta la representación de su curva ROC.

Gráfico, Gráfico de líneas

Descripción generada automáticamente

## “Enzaprost”

Patrón de fondo

Descripción generada automáticamente

Esta primera representación gráfica la distribución de dosis reales de todos los verracos registrados en la base de datos de la empresa (más de 1500) en función a los valores de excitación provocados por las inyecciones del medicamento en cuestión. La conclusión que se obtiene del gráfico resultado es que no hay diferencias significativas entre las dosis reales obtenidas si el verraco está muy excitado. Al inicio de la gráfica se concentran más puntos, pero únicamente porque el muestreo de verracos poco excitados es mayor al número de verracos muy excitados. El hecho de excitar al verraco a niveles altos no implica una producción de dosis reales mayor.

Gráfico, Gráfico de cajas y bigotes

Descripción generada automáticamente

En esta gráfica se puede apreciar las distribuciones para cada una de las notas asignadas en cuanto a Calidad de Movimiento para diferentes niveles de excitación. Como conclusión interesante, se puede afirmar que las notas que más niveles distintos de excitación abarcan son las calificaciones de 5 y 6, así como se puede apreciar que las notas 8 y 9 poseen gran cantidad de “outlayers”, es decir, valores fuera de la distribución más frecuente. Para valores de excitación altos, hay una gran cantidad de muestras fuera del rango Intercuartil para ambas calificaciones. Lo mismo ocurre con los valores de Motilidad.

Gráfico, Gráfico de cajas y bigotes

Descripción generada automáticamente

## Analizadores

En cuanto a los analizadores, se han obtenido diferentes conclusiones.

Por un lado, se ha estudiado la proporción de dosis admitidas y dosis rechazadas del total de dosis analizadas por cada empleado, obteniendo así las siguientes gráficas:

Gráfico, Gráfico de barras

Descripción generada automáticamenteGráfico, Gráfico de barras

Descripción generada automáticamente

A continuación, se han obtenido dos gráficas relacionadas con la detección de las formas anormales. Por un lado, en la gráfica superior, se muestran el porcentaje de dosis rechazadas por formas anormales entre todas las dosis rechazadas por los diferentes analizadores. En la gráfica inferior se puede apreciar los valores para cada analizador según su ratio de formas anormales entre todas las dosis analizadas.

Gráfico, Gráfico de barras

Descripción generada automáticamenteGráfico, Gráfico de barras

Descripción generada automáticamente

Por último, se adjuntarán a continuación las gráficas correspondientes a las notas asignadas en Calidad de Movimiento y Motilidad por cada uno de los analizadores.

Gráfico, Gráfico de barras

Descripción generada automáticamente

Gráfico, Gráfico de barras

Descripción generada automáticamente

En cuanto a las conclusiones de este apartado, se adjunta a continuación un listado de detecciones realizadas sobre cada gráfica adjunta anteriormente:

* En las gráficas de barras apiladas que mostraban la proporción de dosis admitidas y dosis rechazadas en cuanto al global de dosis analizadas por cada trabajador, destacan los analizadores con identificadores 59, 71 y 79, los cuales han desarrollado la mayor parte de sus análisis en el centro denominado “CAL”, con un 39%, 34% y un 85% de dosis rechazadas.
* En las conclusiones relativas al porcentaje de formas anormales en las dosis rechazadas, destacan los analizadores 60, 57 y 43, con valores por encima del 40% de rechazos por formas anormales entre todas sus dosis rechazadas.
* En cuanto al ratio de formas anormales por dosis analizada, solo se resalta la actuación del analizador 57, el cual rechaza 1 dosis por formas anormales de 100 que analiza.
* Y, por último, se ha concluido que la dispersión de las calificaciones en cuanto a Motilidad y Calidad de Movimiento es mínima, existiendo así escasa diferencia entre los valores asignados por los diferentes analizadores.

## Diluyentes

Para el análisis y las conclusiones de este apartado se han dividido en dos partes claramente diferenciables. Por un lado, se adjunta a continuación las conclusiones obtenidas durante el análisis de las muestras recién obtenidas del verraco:

Gráfico, Gráfico de barras

Descripción generada automáticamenteGráfico, Gráfico de barras

Descripción generada automáticamente

Gráfico, Gráfico circular

Descripción generada automáticamenteGráfico, Gráfico circular

Descripción generada automáticamenteGráfico, Gráfico circular

Descripción generada automáticamente

Como se puede observar en las gráficas anteriormente adjuntas, el diluyente 1 es el mejor de los tres utilizados por la empresa en los análisis de las muestras recién extraídas del verraco. Dicho diluyente es el que mejores valores presenta de motilidad y calidad de movimiento, así como el valor más bajo de rechazos por formas anormales.

Finalmente, se adjuntan las gráficas correspondientes a los valores de media de la motilidad y de la calidad de movimiento, junto con sus desviaciones estándar, todo en función a los días de conservación.

Gráfico, Gráfico de líneas

Descripción generada automáticamente

Gráfico, Gráfico de líneas

Descripción generada automáticamente

El objetivo de estas dos últimas gráficas es que estas sirvan como apoyo a la hora de decidir el tipo de diluyente en función a los días que se pretende dejar en conservación la muestra deseada. Una apreciación común a las dos es que, por un lado, no existen registros de los diluyentes 1 y 2 a partir de los cinco días de conservación y, además, el diluyente número 3 presenta demasiada desviación en sus valores de motilidad y calidad de movimiento.

## Fecha de nacimiento vs Calidad Seminal y Supervivencia

A continuación, se adjunta la matriz de correlación obtenida como respuesta a esta última cuestión planteada por la empresa de inseminación porcina.

Gráfico

Descripción generada automáticamente

En dicha matriz de correlación se obtiene como conclusión la inexistencia de una relación solida entre fecha de nacimiento (AÑO, MES) con la calidad seminal (MOTILIDAD, CALIDAD MOVIMIENTO), así como con la supervivencia en el centro (DIAS ESTANCIA).

# MANUAL DE USUARIO

Como instrucciones para el correcto funcionamiento del proyecto, se necesita abrir el código en un notebook de jupyter en Google y abrir desde allí los diferentes archivos correspondientes a cada ejercicio. Dicho código está subido al repositorio de Github que se adjunta a continuación (*https://github.com/jesusparedes/TFM*)

Por otro lado, será necesario descargar las correspondientes bases de datos utilizadas, teniendo en cuenta que la concesión de los datos está regulada por un contrato de confidencialidad. Las bases de datos se encuentran subidas a Google Drive, a las cuales se puede acceder entrando en el siguiente enlace (*https://drive.google.com/drive/u/0/folders/1a9d2D90ArCfgtUde4SuwmpXUkxiYIhBH*)

Finalmente, para poder apreciar el “frontend” elaborado para la resolución de los ejercicios planteados por la empresa, es necesario contar con el software de Tableau instalado en el ordenador, así como contener una licencia en activo del mismo software. En el caso de no contener dicha licencia, es posible obtener una de forma gratuita en Tableau Public. Se puede acceder a los documentos generados en Tableau desde el mismo enlace a Github.