UNIVERSIDAD NACIONAL DE EDUCACIÓN A DISTANCIA

ESCUELA TÉCNICA SUPERIOR DE INGENIERÍA INFORMÁTICA

Trabajo Fin de Máster del

Máster Universitario en Ingeniería y Ciencia de Datos

Estudio de Modelos de Aprendizaje Automático Probabilístico Para la Predicción de Casos de Covid-19 en España

Pablo Marcos Alarcón

Dirigido por: José Luis Aznarte

Rafael Pastor Vargas

Curso: 2020-2021: 1ª Convocatoria

Estudio de Modelos de Aprendizaje Automático Probabilístico Para la Predicción de Casos de Covid-19 en España

TFM – Master en Ingeniería y Ciencia de Datos

Alumno: Pablo Marcos Alarcón Fecha: 27 de septiembre de 2021

Contenido

1. Introducción

En el presente estudio se va a analizar la efectividad de diferentes modelos de aprendizaje automático con salida probabilística e inferencia Bayesiana, aplicados a la predicción de casos de Covid-19 en España. Para ello, se utilizarán los datos y las métricas de evaluación utilizadas en el European Covid-19 Forecast Hub (a partir de ahora, Hub). Dicho Hub está coordinado por el Centro Europeo de Predicción y Control de Enfermedades y se encarga de coleccionar y combinar predicciones a corto plazo de casos, hospitalizaciones y muertes por Covid-19 en toda Europa (Unión Europea, países EFTA y UK), generados por diferentes equipos de modelado independientes, utilizando diferentes tipos de metodologías [1].

Durante el estudio se analizará el rendimiento de modelos predictivos que utilizan diferentes metodologías, siguiendo un enfoque incremental en cuanto a la complejidad. El análisis se llevará a cabo analizando, tanto los resultados del entrenamiento de los modelos, como la comparación de las predicciones generadas por los mismos con aquellas predicciones de los modelos existentes en el Hub, utilizando para ello las medidas de evaluación con las que se evalúan semanalmente las predicciones enviadas al Hub.

Los modelos estudiados utilizan únicamente datos de casos producidos en el pasado, focalizando el estudio, por tanto, en el componente autoregresivo y de dependencia temporal del modelado. El estudio pone especial atención en la evaluación y selección de modelos de regresión.

Por tanto, el alcance del estudio es el de analizar hasta qué punto los modelos de regresión probabilística e inferencia bayesiana pueden alcanzar resultados fiables con respecto a modelos predictivos más sofisticados que utilizan conjuntos de datos ampliados con conceptos de epidemiología.

El presente documento se compone de las siguientes secciones:

- Introducción al European Covid-19 Forecast Hub (el Hub), donde se pone en contexto el estudio y los objetivos del Hub.
- Metodología de estudio. En esta sección se sientan las bases del estudio a realizar, describiendo en detalle los datos, el preprocesado realizado en los mismos, la generación de predicciones, la generación de los ficheros de texto que contienen las predicciones para su evaluación posterior, la descripción teórica de los modelos a estudiar, así como el razonamiento de su elección para el estudio, y, por último, las métricas de evaluación que se han utilizado.
- El estudio, analizando cada uno de los modelos por separado, estudiando en detalle los resultados del entrenamiento, así como la evaluación posterior de los resultados utilizando las métricas de evaluación definidas.
- Un apartado donde se definen las conclusiones finales del estudio.
- Áreas de mejora y estudio posteriores

Las herramientas utilizadas para el modelado, entrenamiento y generación de predicciones y ficheros ha sido Tensorflow Probability, mientras que las evaluaciones posteriores de las predicciones se han realizado con un programa en R proporcionado por el Hub.

Para una mayor comprensión del estudio, se recomienda seguir este documento en paralelo con el Notebook de Jupyter proporcionado.

Por último, los números entre corchetes, indican referencias a la bibliografía.

2. El Hub

El European Covid-19 Forecast Hub (Hub) se ejecuta en colaboración con el Centro de Modelado Matemático de Enfermedades Infecciosas de la Escuela de Higiene y Medicina Tropical de Londres, y el Centro Europeo para el Control y la Prevención de Enfermedades (ECDC).

El Hub recopila y combina pronósticos a corto plazo de Covid-19 en Europa (países de la UE y EFTA y el Reino Unido) generados por diferentes equipos de modelado independientes que utilizan una amplia gama de enfoques. El enfoque subyacente fue iniciado por el laboratorio Reich y sigue proyectos similares en los EE. UU., Alemania y Polonia.

El objetivo principal del Hub es proporcionar a los responsables de la toma de decisiones y al público en general información fiable sobre la trayectoria futura a corto plazo de la pandemia. Esto se logra mediante la recopilación de pronósticos de diferentes modelos en un conjunto (ensemble), un enfoque que en el pasado ha demostrado proporcionar un rendimiento consistentemente mejor que cualquier enfoque de modelado individual.

Los objetivos secundarios son obtener información sobre el rendimiento predictivo de diferentes enfoques de modelado, evaluar la calidad de los pronósticos con respecto a diferentes medidas de gravedad de la enfermedad (por ejemplo, casos o muertes) y mantener una comunidad de modeladores de enfermedades infecciosas respaldada por una ética de ciencia abierta [1].

3. Metodología del Estudio

3.1. Datos

Para realizar una comparación lo más objetiva posible con las predicciones disponibles en el Hub, los datos de origen que se han utilizado para el análisis han sido, únicamente, los datos (proporcionados por el Hub), que contienen los datos oficiales de la Universidad Johns Hopkins. Estos datos se han utilizado como la base para la generación de datos de entrada y etiquetas, y, por tanto, como fuente de verdad para su evaluación durante su entrenamiento.

Los datos de origen proporcionados tienen como fecha inicial el día 23 de enero de 2020, y se actualizan a diario con nuevas cifras. Para nuestro análisis, el último día del que se dispone de datos es el 20 de septiembre de 2021, pero el Hub impone como requisito la generación de predicciones durante al menos 4 semanas para realizar correctamente el cálculo de todas las métricas de evaluación. Para cumplir con este requisito, y con el objetivo de no introducir sesgo en los modelos, el entrenamiento se ha realizado con datos hasta el día 31 de Julio de 2021, de donde se han obtenido los conjuntos de entrenamiento, validación y test.

Los datos de origen tienen frecuencia diaria, pero las predicciones, de acuerdo con los requisitos del Hub, deben ser semanales, por lo que es necesaria una agregación de los datos de origen previa al entrenamiento. Esta agregación supone una disminución de los datos de origen disponibles para el entrenamiento, lo que se suma a la escasa cantidad de datos disponible inicialmente, problema debido a que la pandemia producida por el Covid-19 tiene un origen relativamente reciente. Esta limitación en los datos supone uno de los principales retos de generación de predicciones a corto plazo para cualquiera de los modelos predictivos.

Los países para los que se proporcionan datos son los siguientes: Austria, Bélgica, Bulgaria, Croacia, Chipre, República Checa, Dinamarca, Estonia, Finlandia, Francia, Alemania, Grecia, Hungría, Islandia, Irlanda, Italia, Letonia, Liechtenstein, Lituania, Luxemburgo, Malta, Países Bajos, Noruega, Polonia, Portugal, Rumania, Eslovaquia, Eslovenia, España, Suecia, Suiza y Reino Unido.

Para el presente estudio se han utilizado únicamente los datos de España. Esto ha permitido acotar el alcance del análisis, evitando el ruido inherente a los datos generado por las diferencias en escala en los datos de los diferentes países (debidos a las diferentes cifras de población) y en la distribución de los datos (debidos a la diferente velocidad de propagación del virus en los diferentes países y al diferente ritmo de aplicación de medidas de contención no farmacéuticas [2]), y ha evitado añadir complejidad al análisis, permitiendo a su vez la observación de conclusiones al final del estudio.

Es importante destacar que los datos utilizados en el estudio, como es lógico, de acuerdo al tema tratado (personas infectadas por Covid-19), se corresponden con datos discretos (no continuos) mayores que 0. Además, los datos tienen un componente de auto regresión, ya que el valor de cada día depende en cierta medida, del valor del día anterior, aunque no sea el único parámetro determinante. De esta forma: $P(x_t|x_{t-1}, x_{t-2}, \dots, x_0)$, donde x_t es el valor del momento actual, y x_{t-i} son los valores de las semanas anteriores. Este estudio se centra, por tanto, en la predicción de dicho componente autoregresivo del modelado.

Ambos factores se han tenido en cuenta a lo largo del presente estudio.

3.2. Preprocesado de datos

Los datos facilitados por el Hub, provenientes de la Universidad Johns Hopkins, se componen de dos ficheros en formato ".csv" correspondientes a los datos de casos de Covid-19 y de muertes producidas por Covid-19 respectivamente. Para el presente estudio se ha utilizado únicamente el fichero correspondiente a casos de Covid-19, que se utilizará como fuente de verdad única para el entrenamiento y validación de los modelos.

Los datos de casos de Covid-19 suponen un reto para la predicción debido a que se ven afectados por las diferentes políticas de los gobiernos, el comportamiento de la población y las prácticas de testeo. Los datos de muertes son datos más retrasados en el tiempo con respecto al número de casos, y por tanto son más fáciles de predecir [1].

El Hub, además, facilita otras fuentes de datos adicionales, que no se han tenido en cuenta para el presente estudio.

El fichero de casos de Covid-19 proporcionado por el Hub dispone de los siguientes campos:

- Location: Código de país al que corresponden los datos, en formato ISO de 2 caracteres.
- Location_name: Nombre completo del país al que corresponden los datos.
- Date: Fecha del dato representado, con frecuencia diaria.
- Value: valor numérico de los casos de Covid-19 reportados el día correspondiente.

Para el propósito del estudio, se han filtrado los datos obtenidos para trabajar únicamente con los datos de España, para posteriormente realizar una agrupación semanal de los mismos. Además, para alinear las semanas de los datos con las semanas utilizadas en el Hub, la agregación se ha realizado en torno a los sábados de cada semana.

Adicionalmente, y con el objetivo de realizar un análisis alineado con los requisitos del Hub, los datos se limitarán para utilizar solamente los datos desde el día 23 de enero de 2020 hasta el día 31 de Julio de 2021, incluyendo en ese rango de fechas los conjuntos de entrenamiento, validación y test. Esto permitirá realizar predicciones de hasta 4 semanas en el futuro, utilizando para cada predicción datos más recientes no utilizados en el entrenamiento, simulando de esta forma el paso del tiempo sin caer en el sesgo de "look-ahead" en los datos (para más detalle, consultar la sección "Predicciones" de este documento).

Por último, el 70% de los datos corresponden al conjunto de entrenamiento, el 20% a los datos de validación y el 10% restante a los datos de test.

Para obtener los inputs y las etiquetas que servirán para entrenar los diferentes modelos y dado que los datos tienen dependencia temporal entre ellos, estos se han estructurado de acuerdo a la función Wavenet [3], en tensores unidimensionales de longitud 1 desplazados hacia la derecha. De esta forma, para cada valor de entrada, se asigna una etiqueta que se corresponde con un único valor posterior en el tiempo, estableciéndose un salto entre el valor de input y el valor de la etiqueta que variará dependiendo del horizonte temporal de la predicción. De esta forma:

$$Etiqueta_t = Input_{t+n} \; ; \; n = 1, 2, 3, 4$$

Donde n es el horizonte temporal de la predicción.

También cabe destacar que para el preprocesado de los datos no se ha realizado ningún escalado de los datos, dado que ninguno de los modelos estudiados se beneficia del mismo.

3.3. Predicciones

Los objetivos de las predicciones del Hub se centran en el número de casos, hospitalizaciones y muertes. Es posible enviar predicciones de cualquiera de los objetivos o combinaciones de estos.

El Hub Europeo se centra en las predicciones de cada uno de los países con horizonte temporal de 1 a 4 semanas en el futuro, sin embargo, es posible enviar predicciones con cualquier combinación de horizontes hasta un total de 20 semanas. Debe tenerse en cuenta que las predicciones serán comparadas con los datos originales, y es previsible que los modelos pierdan precisión a medida que las predicciones realizadas sean más lejanas en el tiempo, por lo que se recomienda que las predicciones se centren en un horizonte temporal de entre 1 y 4 semanas [4].

El envío de los datos al Hub puede realizarse en cualquier momento, pero solo se puede realizar una vez por semana. Se recomienda que se realice el envío cada lunes, ya que se considera que la semana termina en sábado, y los resultados enviados entre el viernes y el lunes serán tomados como predicciones de la semana que termina el mismo sábado, mientras que los resultados enviados en los días posteriores al lunes se tomarán como predicciones de la semana que termina el sábado posterior a la actual [4].

También es posible enviar predicciones con cualquier combinación de países de los anteriormente mencionados.

Para la correcta evaluación de las predicciones y el cálculo de todas las métricas relevantes para la evaluación, es necesario que las predicciones especifiquen los siguientes cuantiles: 0.010, 0.025, 0.050, 0.100, 0.150, 0.200, 0.250, 0.300, 0.350, 0.400, 0.450, 0.500, 0.550, 0.600, 0.650, 0.700, 0.750, 0.800, 0.850, 0.900, 0.950, 0.975, 0.990. Además, es posible enviar una predicción puntual que representará el

valor con mayor probabilidad. En el caso del presente estudio, se ha considerado que el valor con mayor probabilidad es el correspondiente al cuantil 0.500.

Se han analizado las predicciones con un horizonte temporal de 1, 2, 3 y 4 semanas, de acuerdo con la recomendación del Hub. El único país que se ha analizado ha sido España, pero el estudio es extensible a cualquiera de los países de los que se disponen datos. Todas las predicciones realizadas proporcionarán todos los cuantiles requeridos por el Hub, de forma que sea posible el análisis de todas las métricas.

Para el presente estudio, se adaptarán los datos para ajustarse a los requisitos del Hub en lo relativo al formato de las semanas, generando las predicciones cada lunes, con fecha de predicción el sábado correspondiente al horizonte temporal correspondiente a la predicción. En el caso de España, y debido a la forma de reportar datos de los diferentes organismos gubernamentales, los fines de semana (sábado y domingo), los datos reportados son siempre 0, mientras que los datos del lunes agregan los datos de sábado, domingo y lunes.

Dado que los modelos a estudiar tienen un enfoque probabilístico, las predicciones reportadas se basan en un muestreo de 1000 predicciones, a partir de las cuales se calculan la media y los diferentes cuantiles.

3.4. Generación de ficheros

Una vez generadas las predicciones para todos los horizontes temporales, se generan ficheros con cada uno de los cuantiles del intervalo de predicción, especificando la fecha de origen de los datos, el horizonte temporal de la predicción y la fecha del final de la predicción, que variará dependiendo del horizonte temporal.

Dado que para la correcta evaluación de las predicciones y la visualización de todas las métricas asociadas es necesario generar al menos 4 semanas de predicciones, la generación de ficheros se ha llevado a cabo mediante una simulación que realiza predicciones sobre datos pasados. Para esto, se debe tener en cuenta el sesgo de “look-ahead”, y evitar que el modelo haga predicciones sobre datos con los que ha entrenado. A su vez, para reflejar el paso del tiempo, se deben realizar las predicciones teniendo en cuenta los últimos datos disponibles en el momento de la predicción.

El siguiente ejemplo ilustra cómo se lleva a cabo la simulación de predicciones sobre datos pasados:

- Primera iteración:
  - Se entrena el modelo con datos hasta el 31 de julio de 2021
  - Se generan predicciones para la primera semana, correspondiente al intervalo entre el 2 de agosto y el 7 de agosto de 2021, y para el resto de semanas, para los intervalos cuyo final es el 14, 21 y 28 de agosto de 2021 respectivamente.
  - Se genera el fichero correspondiente, con las predicciones de las 4 semanas y los cuantiles correspondientes a cada una
- Segunda iteración:
  - Se obtienen los datos hasta 7 días después de los datos disponibles en la iteración anterior, en este caso, el 7 de agosto de 2021.
  - Se generan predicciones para los intervalos de tiempo que empiezan el 9 de agosto de 2021 y terminan el 14, 21, 28 de agosto y el 4 de septiembre de 2021, respectivamente
  - Se genera el fichero con las predicciones de las 4 semanas y los cuantiles correspondientes.

- El resto de las iteraciones son equivalentes a la segunda iteración.

Para cada uno de los modelos estudiados se han generado 8 ficheros con predicciones, siendo la fecha del primer fichero el 2 de agosto de 2021 y la última fecha el 20 de septiembre de 2021, lo que implica que las predicciones tienen un rango que abarca desde el 7 de agosto de 2021 hasta el 16 de octubre de 2021. Esto garantiza que la evaluación de los modelos tiene un rango suficiente sobre el que operar.

3.5. Modelos

Los modelos predictivos que se van a estudiar son modelos de regresión mediante aprendizaje automático con salida probabilística, en algunos casos, e inferencia bayesiana en otros. Los modelos que se van a analizar serán los siguientes:

- Modelo de regresión lineal con desviación estándar variable. Este modelo servirá como línea base para el resto de los modelos a estudiar. Se implementará mediante la modificación de una distribución normal, de forma que la salida del modelo sea probabilística.
- Modelo de regresión con distribución de Poisson. Esta distribución es apropiada para datos contables, como los que se quieren estudiar. Se trata de una distribución con un solo parámetro (λ), por lo que supone un buen punto de partida para el análisis por su sencillez de modelado.
- Modelo de regresión con mezcla de distribuciones logísticas discretizadas. Este modelo aplica una mezcla de varias distribuciones logísticas, y discretiza la salida para que no sea continua, lo que se ajusta a la tipología de los datos a predecir.
- Modelo de regresión mediante inferencia variacional. Este modelo utiliza un enfoque de aproximación de inferencia Bayesiana y aplica una distribución logística discretizada a la salida.
- Modelo de regresión mediante inferencia variacional con mezcla de distribuciones logísticas discretizadas. Este modelo mezcla los dos anteriores, y utiliza el enfoque de aproximación de inferencia Bayesiana de la inferencia variacional junto con la mezcla de distribuciones logísticas con salida discretizada.

La arquitectura de los diferentes modelos no varía independientemente del horizonte temporal de la predicción, aunque sí se han entrenado diferentes modelos para cada uno de los horizontes temporales, con el objetivo de ajustar los pesos adecuadamente a cada uno de ellos. En el caso de que se hubieran introducido cambios específicos a un horizonte temporal de predicción, se indicará en el experimento correspondiente, aportando la justificación adecuada.

Quedan fuera del ámbito del estudio modelos expertos que incluyen conceptos complejos de epidemiología para sus predicciones. Se espera, por tanto, que los modelos analizados tengan una capacidad predictiva más baja que dichos modelos expertos.

Además, cabe puntualizar que se desconocen los detalles de los demás modelos que reportan predicciones al Hub, debido a que:

- Las implementaciones de los diferentes modelos de los colaboradores no se encuentran publicadas en el Hub, solo las predicciones que realizan.
- Los detalles de los datos utilizados por los modelos del resto de colaboradores del Hub para el entrenamiento tampoco se encuentran publicados en el Hub, y son desconocidos, siendo posible que se utilicen datos de fuentes adicionales a las proporcionadas por el Hub, ya que el Hub no establece limitaciones al respecto.

- Como consecuencia de los puntos anteriores, se desconoce si el resto de los modelos incluyen conceptos de epidemiología para el cálculo de sus predicciones.

Por todo lo anterior, es de esperar que los modelos analizados en el presente estudio no mejoren los resultados de los modelos existentes en el Hub. Por tanto, el alcance del estudio será doble:

1. Comparar el rendimiento entre los diferentes modelos propuestos, estudiando hasta que punto los desarrollos más complejos aportan mejores resultados predictivos que desarrollos más sencillos.
2. Analizar hasta qué punto los modelos de regresión probabilística y bayesiana pueden alcanzar resultados fiables con respecto a modelos predictivos más sofisticados, como los que colaboran en las predicciones reportadas por el Hub.

Cada uno de los modelos ha sido probado con diferentes arquitecturas antes de su evaluación con las métricas del Hub, hasta que se ha encontrado una variante que minimizaba el resultado de NLL (Negative Log Likelihood) en el conjunto de validación. Como resultado de este análisis, se han probado diferentes números de neuronas en las diferentes capas, así como diferentes números de capas, y se ha observado que, debido al bajo número de datos con los que se entrenan los modelos, estos no se benefician de:

- Tener un alto número de neuronas, ya que esto aumenta considerablemente el número de parámetros a entrenar, y en muchos casos no se disponen de datos suficientes para entrenarlos todos.
- Tener un alto número de capas. Un mayor número de capas genera curvas más complejas y por lo general, produce resultados más extremos, lo que no se adapta a la forma de los datos de entrenamiento.

Por estas razones, los modelos a estudiar son de arquitectura bastante sencilla, lo que se ha observado que mejora los resultados de las predicciones.

A continuación, se describen en detalle los modelos a estudiar.

3.5.1. Modelo de Regresión Lineal

El modelo de regresión lineal servirá como punto de partida del estudio y se tomará como línea base para el resto de los modelos, siendo el objetivo de estos superar el rendimiento de este modelo.

El modelo de regresión lineal representa el modelo de regresión más simple posible. Su objetivo es trazar una línea recta entre todos los puntos, de forma que se minimice la suma de las distancias al cuadrado de todos los puntos hasta la línea.

La línea viene dada por la fórmula $y = ax + b$, por lo que minimizar la función de pérdida supone encontrar valores a y b tales que la función de pérdida sea mínima, lo que se consigue durante el entrenamiento del modelo. Una vez encontrados estos valores, se divide el resultado entre el número de puntos de datos, de forma que el resultado de la función de pérdida se ve penalizado cuanto mayor es el número de datos en la muestra. Esta función de pérdida se conoce como el MSE (mean squared error):

$$MSE = \frac{1}{n}\sum_{i=1}^{n}(y_i - \hat{y}_i)^2 = \frac{1}{n}\sum_{i=1}^{n}(y_i - (a \cdot x_i + b))^2$$

Para el objetivo del estudio es necesario que el modelo disponga de una salida probabilística que cuantifique la incertidumbre del modelo, lo cual se consigue mediante el modelado una distribución normal. Esta distribución tiene dos parámetros, μ, la media (centro) de la distribución, y σ, la varianza de la distribución.

De esta forma, se tiene que, la incertidumbre del modelo viene dada por la densidad de probabilidad $f(y|x,w) = f(y|x,\mu,\sigma)$. La distribución de probabilidad de las etiquetas y viene dada por los inputs x y los parámetros w. Para etiquetas continuas, la fórmula de NLL deriva en MSE, que viene dada por una distribución normal con varianza (σ) fija.

Sin embargo, en el modelo utilizado no se ha considerado homocedasticidad, es decir, una varianza (σ) fija, por lo que la salida del modelo tendrá dos parámetros. En este caso, la función de pérdida sigue siendo NLL, que viene dada por $NLL = \sum_{i=1}^{n} -\log\left(\frac{1}{\sqrt{2\pi\sigma x_i{}^2}}\right) + \frac{(\mu_{x_i}-y_i)^2}{2\sigma x_i{}^2}$ en el caso de la distribución normal.

Con los dos parámetros de la salida del modelo, podremos controlar el centro de la distribución (media) y la varianza de la misma, y de esta forma podremos capturar la incertidumbre aleatórica, referente a la variabilidad inherente al problema. [5][6]

Para controlar que la desviación estándar no proporcione valores negativos, se ha utilizado una función de activación softplus, que asigna valores en el rango [0, infinito) a partir de cualquier valor de entrada.

Por último, dado que se calcularán la media y la desviación típica de la distribución, la red tendrá únicamente 4 parámetros a entrenar.

3.5.2. Modelo con Distribución de Poisson

El segundo modelo estudiado es una red neuronal sencilla que alimenta una distribución de Poisson. Esta distribución es apropiada para trabajar con datos contables, como es el caso de los datos del estudio (casos de Covid-19 en España).

La distribución de Poisson solo recibe un parámetro, λ, que representa el número medio de eventos por unidad. Dado que es una media, no siempre es un número entero, y representa tanto el centro de la distribución como su varianza.

En la distribución de Poisson obtenemos una probabilidad, no una función de densidad como en la distribución Gaussiana, y se denota como función de probabilidad de masa, en lugar de función de probabilidad acumulada (CPD). Esta probabilidad viene dada por la fórmula [8]:

$$P(y = k) = \frac{\lambda^k \cdot e^{-\lambda}}{k!}, donde\ k! = 1 \cdot 2 \cdot 3 \cdot \ldots \cdot k$$

En el modelo utilizado para el estudio, el parámetro λ viene determinado por una red neuronal de una única capa. λ puede tomar cualquier valor entre -∞ y +∞. Puesto que en nuestro caso solo nos interesan valores mayores que 0, la capa densa de la red neuronal está activada por una función Softplus, lo que asegura valores en el rango [0, +∞). La función Softplus tiene una forma similar a la función exponencial, pero crece de forma lineal para valores muy grandes de x, como se puede ver en la figura, [7]:

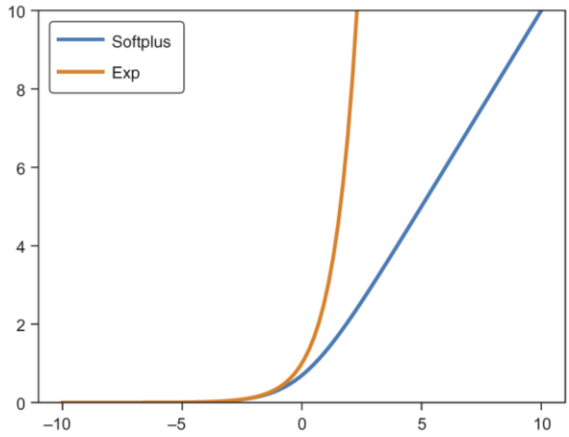


Figura 1 - Función Softplus comparada con la función exponencial [7]

3.5.3. Modelo con Mezcla de Distribuciones Logísticas Discretizadas

Este modelo permite modelar distribuciones complejas mediante la mezcla de varias distribuciones logísticas, a las que se asignan pesos para determinar las proporciones de cada una de las distribuciones en la mezcla. El uso de distribuciones logísticas garantiza que no se generen valores negativos, como puede verse es los gráficos de su función de densidad de probabilidad (similar al de la distribución normal), y su función de densidad de probabilidad acumulada, en el siguiente gráfico:

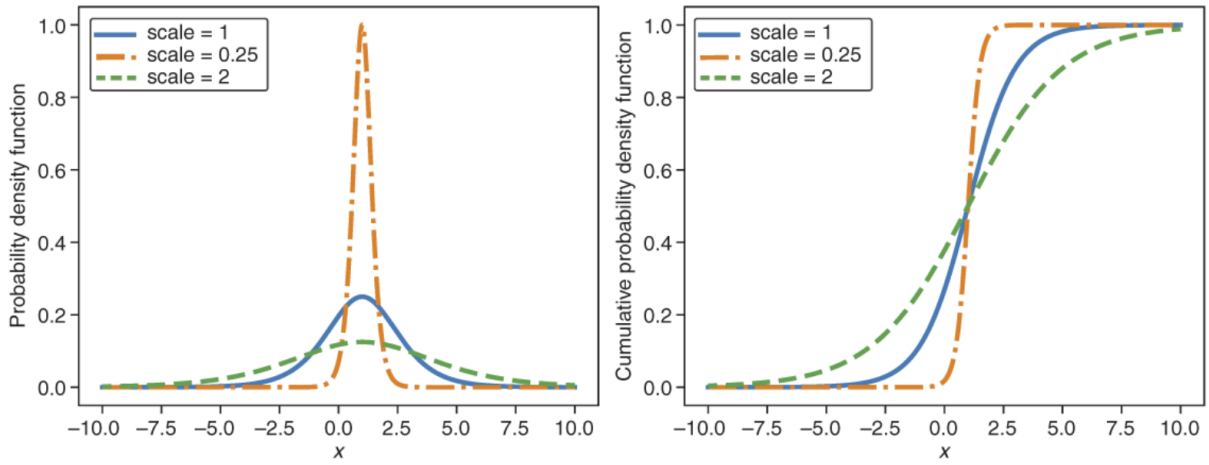


Figura 2 - Función de densidad de probabilidad y función de densidad de probabilidad acumulada para la distribución logística, con diferentes valores de varianza.

Adicionalmente, es necesario que la salida de la mezcla sean valores discretos (no continuos), lo que en el modelo se consigue utilizando la función QuantizedDistribution de Tensorflow Probability. Un ejemplo de la aplicación de esta función se ilustra en la siguiente figura, donde se muestra la versión cuantizada de la distribución con varianza 0,25 de la figura anterior:

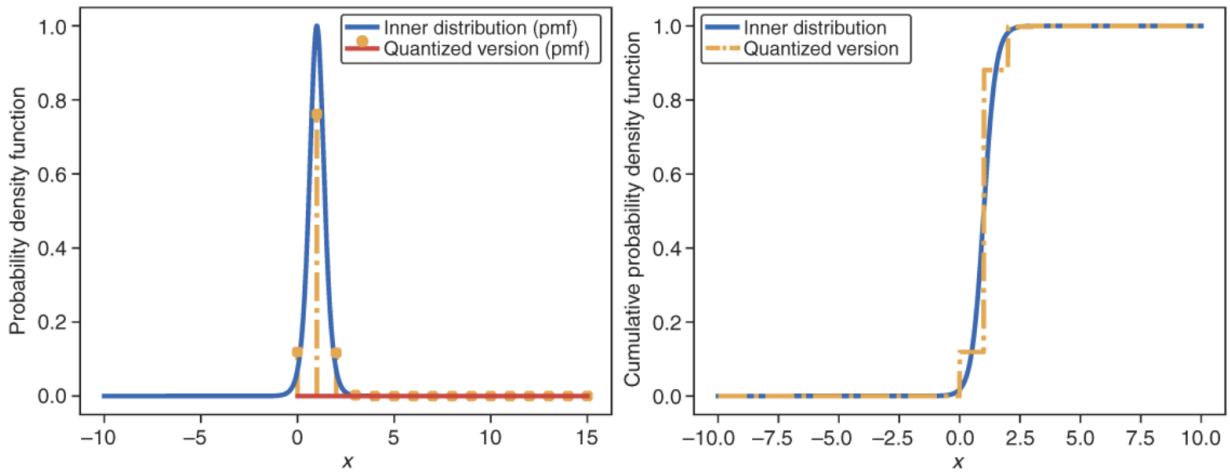


Figura 3 - Versión cuantizada de la distribución logística con varianza 0,25 y media 1 [8]

El modelado de la mezcla de logísticas necesita controlar 3 parámetros para cada uno de los componentes, la localización (media), la varianza y el peso. Estos parámetros se controlan mediante una capa densa, cuyo número de neuronas debe ser múltiplo de 3. En este caso se han generado los modelos con 15 neuronas, de forma que se genere una mezcla de 5 distribuciones logísticas [8]. Durante las pruebas realizadas a este modelo, se ha comprobado que aumentar el número de neuronas en la capa densa no produce mejores resultados durante el entrenamiento. A su vez, utilizar menos neuronas sin afectar el rendimiento del modelo es posible para el entrenamiento del modelo que produce predicciones con horizonte temporal de 1 semana, pero degrada considerablemente el rendimiento del resto de modelos con horizontes temporales de 2, 3 y 4 semanas.

3.5.4. Modelo de Inferencia Variacional con Distribución Logística

Debido a la complejidad de encontrar soluciones analíticas a modelos Bayesianos de aprendizaje automático, es necesario realizar aproximaciones al resultado. La inferencia variacional es una técnica de aproximación aplicable a cualquier modelo de aprendizaje automático. Por tanto, este modelo se diferencia de los anteriores en que implementa inferencia Bayesiana, convirtiendo la red neuronal en una red neuronal Bayesiana (BNN). Esta tipología de red sustituye los valores de los pesos por distribuciones, lo que genera una distribución muy compleja y que no es independiente entre los diferentes pesos.

La idea principal de la inferencia variacional es que la distribución compleja de la posterior (probabilidad de los pesos), se puede aproximar mediante una distribución sencilla llamada distribución variacional [9].

La inferencia Bayesiana toma como punto de partida el teorema de Bayes: $P(A|B) = \frac{P(B|A)\,P(A)}{P(B)}$, que junta 4 probabilidades diferentes:

- P(A|B): la probabilidad condicional de A, dado B, también llamada posterior, y representa la probabilidad de un valor de un parámetro dados los datos.
- P(B|A): La inversa o probabilidad (Likelihood).
- P(A): La previa (prior), representa la probabilidad de cada valor de los parámetros y debe ser definida antes de conocer los datos. Representa la "creencia" de cómo se distribuyen los datos, antes de conocerlos [10].

- P(B): La cantidad o evidencia (también llamada probabilidad marginal). Representa la probabilidad de los datos.

La filosofía de la inferencia bayesiana es la de incorporar evidencia a medida que se obtienen más datos sobre el problema. De esta forma, siempre se mantiene un grado de incertidumbre sobre la predicción, que será menor a medida que se obtienen más evidencias. Por tanto, al comienzo de las predicciones, todo el peso de la predicción lo tendrá la probabilidad previa P(A), establecida antes de ver ningún dato, y a medida que se vayan viendo datos, esta "creencia" inicial se verá actualizada por las observaciones, y haciendo que tenga menos peso en la predicción [10].

Este modelo se ha implementado utilizando dos capas del tipo "DenseFlipout", que son las capas que aplican inferencia variacional en Tensorflow Probability. La salida de esas capas alimenta una distribución logística discretizada.

3.5.5. Modelo de Inferencia Variacional con Mezcla de Distribuciones Logísticas Discretizadas

Este último modelo mezcla el modelo de inferencia variacional descrito en el apartado anterior con la mezcla de distribuciones logísticas con salida discretizada implementado también en otro modelo anterior.

El objetivo de este modelo es comprobar si la inferencia variacional se beneficia de utilizar una mezcla de distribuciones logísticas y si los resultados de la evaluación de este modelo superan los de los dos modelos anteriores por separado.

3.6. Métricas de Evaluación

El entrenamiento de los diferentes modelos se ha realizado minimizando la función NLL (Negative Log Likelihood), presentada anteriormente, para todos los modelos analizados. Esta métrica es una función de pérdida comúnmente utilizada en problemas de clasificación y regresión, y minimizándola se maximiza la probabilidad de las estimaciones (principio de Maximum Likelihood o MaxLike).

Una vez entrenados los modelos, y con el objetivo de realizar la evaluación de sus predicciones y su comparación con el resto de predicciones existentes en el Hub, se han utilizado las medidas oficiales de este, prestando especial atención al WIS relativo (Relative Weighted Score Interval o Puntuación de Intervalo Ponderado Relativo) y al Error Absoluto Relativo. Las métricas utilizadas por el Hub para la evaluación de modelos son las siguientes:

- La puntuación de intervalo ponderado (WIS) [13] es una regla de puntuación adecuada (es decir, no se puede "engañar") que es adecuada para puntuar intervalos de pronósticos. Generaliza el error absoluto (es decir, los valores más bajos son mejores) y tiene tres componentes: dispersión, predicción insuficiente y predicción excesiva. La dispersión es un promedio ponderado de los anchos de los intervalos de predicción enviados. Se agregan penalizaciones por predicción excesiva y falta de predicción cada vez que una observación cae fuera de un intervalo de predicción central informado, y la fuerza de la penalización depende del nivel nominal del intervalo y de como de lejos del intervalo cayó la observación. Se debe tener en cuenta que el WIS promedio puede referirse a diferentes objetivos en diferentes modelos y, por lo tanto, no siempre se puede comparar entre modelos. Estas comparaciones deben realizarse en función de la habilidad relativa.

El WIS relativo es una medida relativa del rendimiento de la predicción que tiene en cuenta que es posible que diferentes modelos no cubran exactamente el mismo conjunto de objetivos de predicción (en el presente estudio, se refiere a semanas y ubicaciones). En términos generales, un WIS relativo de X significa que, promediado sobre los objetivos que abordó un equipo determinado, su WIS fue X veces mayor / menor que el rendimiento del modelo base descrito en Cramer et al. (2021). Los valores más pequeños son, por tanto, mejores y un valor por debajo de uno significa que el modelo tiene un rendimiento por encima de la media. El WIS relativo se calcula mediante un "torneo de comparación por pares" en el que para cada par de modelos se calcula una proporción de puntuación media en función del conjunto de objetivos compartidos. El WIS relativo es la media geométrica de estas relaciones. Los detalles sobre el cálculo se pueden encontrar en Cramer et al. (2021). Esta métrica se calcula solo si se proporciona un conjunto completo de cuantiles.

- El error absoluto relativo se calcula a partir de los pronósticos puntuales predictivos, que es el valor pronosticado individual que los modelos consideran más probable. En todos los modelos estudiados, se ha considerado como el pronóstico individual más probable, el valor correspondiente al cuantil 50%. El EA relativo se calcula mediante un "torneo de comparación por parejas", en el que para cada par de modelos se calcula una relación de puntuación media en función del conjunto de objetivos compartidos (semanas y ubicaciones). El EA relativo es la media geométrica de estas relaciones. Los detalles sobre el cálculo se pueden encontrar en Cramer et al. (2021).
- La cobertura es la proporción de observaciones de datos reales que cayeron dentro de un intervalo de predicción determinado. Idealmente, un modelo de pronóstico alcanzaría una cobertura del 50% de 0,50 (es decir, el 50% de las observaciones caen dentro del intervalo de predicción del 50%) y una cobertura del 95% de 0,95 (es decir, el 95% de las observaciones caen dentro del intervalo de predicción del 95%). Los valores de cobertura superiores a estos valores nominales indican que los pronósticos son poco fiables, es decir, los intervalos de predicción tienden a ser demasiado amplios, mientras que los valores de cobertura más pequeños que estos valores nominales indican que los pronósticos son demasiado confiados, es decir, los intervalos de predicción tienden a ser demasiado estrechos. Esta métrica se calcula para todos los modelos que proporcionan los cuantiles relevantes.
- El sesgo (sesgo) es una medida entre -1 y 1 que expresa la tendencia a predecir menos (-1) o predecir en exceso (1), ver la descripción en Funk et al. (2019). Esta métrica se calcula para todos los modelos que proporcionan cuantiles. [14]

4. Estudio

4.1. Modelo de regresión lineal

4.1.1. Entrenamiento del modelo

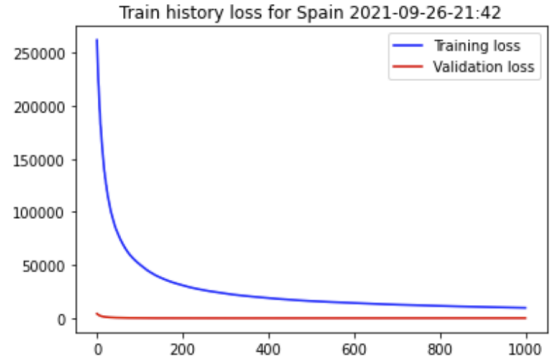
Como se ha mencionado anteriormente en la descripción del modelo, la regresión lineal solo tiene 4 parámetros a entrenar, que representan la media y la desviación típica de la distribución normal con la que se ha modelado la regresión lineal.

El modelo de regresión lineal muestra unos resultados mejores de los esperado durante el entrenamiento para la predicción con horizonte temporal de 1 semana, ya que obtiene un NLL mejor en los datos de validación que en los datos de entrenamiento, y aún mejores en el conjunto de test, lo que ilustra la capacidad predictiva del modelo, e indica que no se ha producido “overfit” sobre los datos de

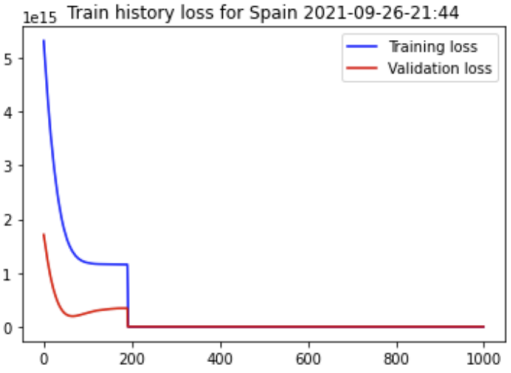
entrenamiento, aunque este resultado puede deberse al escaso número de puntos de datos disponibles en el conjunto de test. Este resultado se repite en la predicción con horizonte temporal a 2 semanas, aunque en este caso el rendimiento ya se degrada notablemente. Para los horizontes temporales de 3 y 4 semanas, el resultado del entrenamiento es muy pobre, por lo que las predicciones que realizará el modelo serán completamente aleatorias.

La siguiente figura ilustra los resultados del entrenamiento:

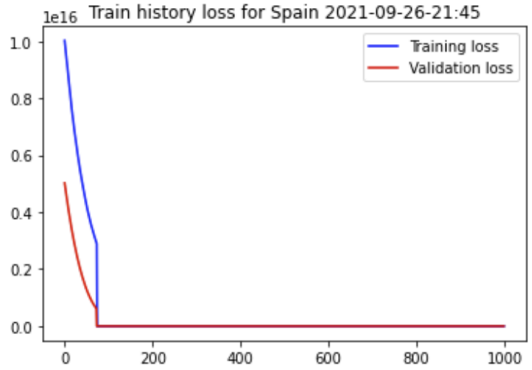

```
Linear Regression NLL on training 1 week:  105.6368637084961
Linear Regression NLL on validation 1 week:  18.34954833984375
Linear Regression NLL on test 1 week:  15.977630615234375
```




```
Linear Regression NLL on training 2 weeks:  9709.693359375
Linear Regression NLL on validation 2 weeks:  94.28871154785156
Linear Regression NLL on test 2 weeks:  52.6368408203125
```




```
Linear Regression NLL on training 3 weeks:  5592227.0
Linear Regression NLL on validation 3 weeks:  893.8184814453125
Linear Regression NLL on test 3 weeks:  51942.9296875
```




```
Linear Regression NLL on training 4 weeks:  35718288.0
Linear Regression NLL on validation 4 weeks:  5236.6767578125
Linear Regression NLL on test 4 weeks:  209267.875
```

Figura 4 - Resultado del entrenamiento del modelo de regresión lineal con predicciones de 1 a 4 semanas (de derecha a izquierda y de arriba abajo), con detalle del resultado del NLL en los 3 conjuntos de datos (entrenamiento, validación y test).

Todo lo anterior se encuentra respaldado si analizamos a la representación gráfica de las predicciones frente a las observaciones en cada uno de los horizontes temporales, así como los CPD correspondientes a cada uno de los conjuntos.

Se puede observar que el modelo realiza un ajuste correcto, aunque mejorable, en los datos de entrenamiento y validación con horizonte temporal de 1 semana, donde además, es capaz de capturar cierto nivel de incertidumbre de la predicción (representado por las líneas grises y azules), pero rápidamente empeora a medida que se aumenta el horizonte temporal de la predicción, dando resultados aleatorios en los horizontes temporales más altos, donde el modelo tampoco es capaz de capturar ningún tipo de incertidumbre en el modelo.

En los CPD se puede observar que muchas de las predicciones se observan fuera de los cuantiles 50% y 95% para cualquiera de los horizontes temporales.

Las siguientes figuras ilustran este hecho:

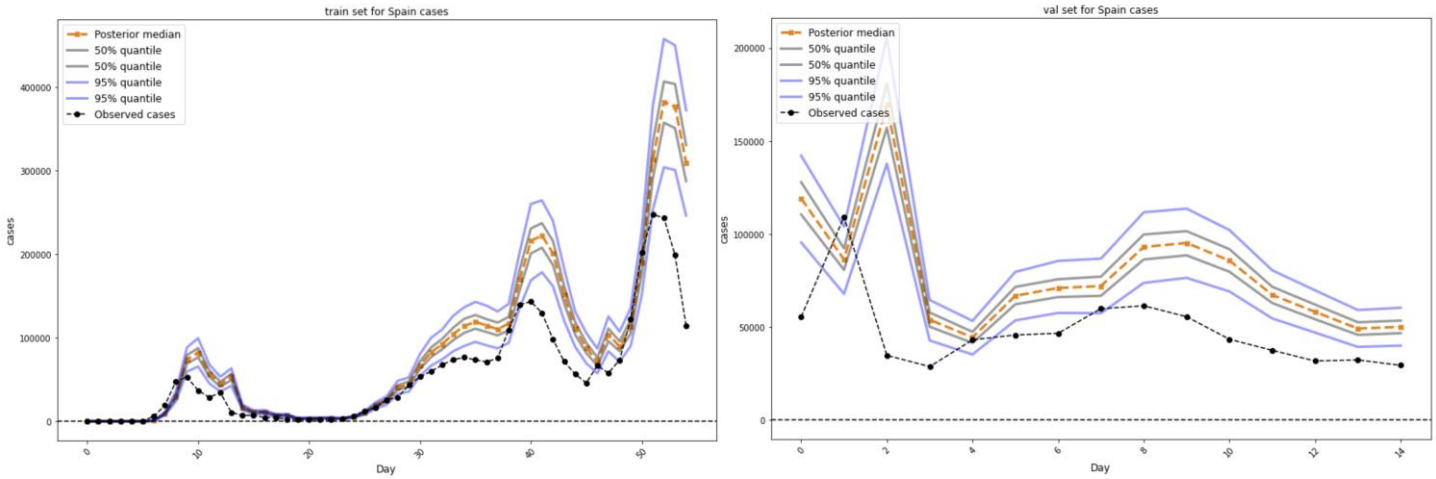


Figura 5 - Predicciones sobre los datos de entrenamiento (izquierda) y validación (derecha) del modelo de regresión lineal con horizonte temporal de predicción de 1 semana

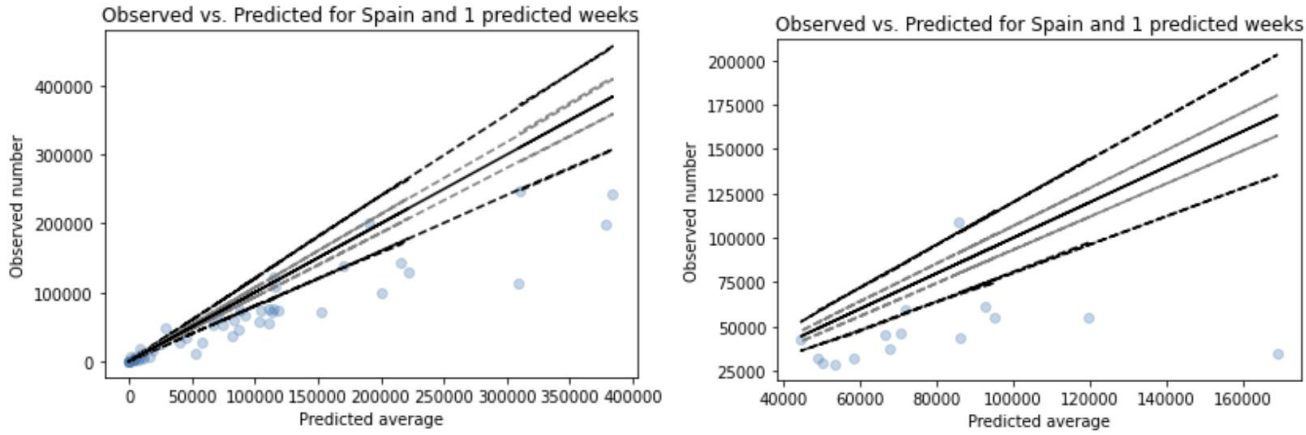


Figura 6 - CPDs de las predicciones sobre los datos de entrenamiento (izquierda) y validación (derecha) del modelo de regresión lineal con horizonte temporal de predicción de 1 semana

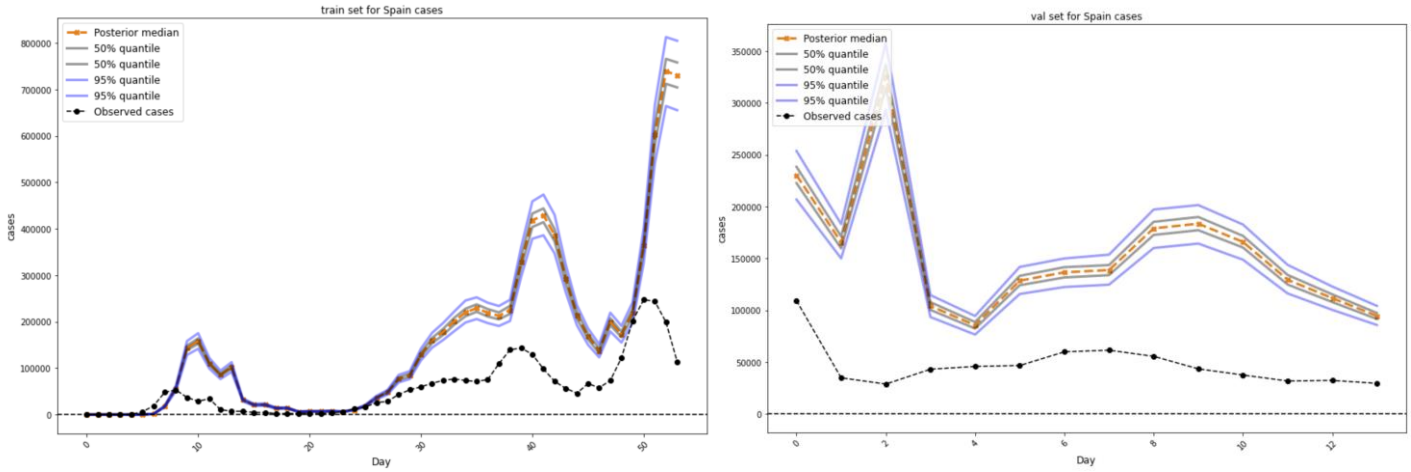


Figura 7 - Predicciones sobre los datos de entrenamiento (izquierda) y validación (derecha) del modelo de regresión lineal con horizonte temporal de predicción de 2 semanas

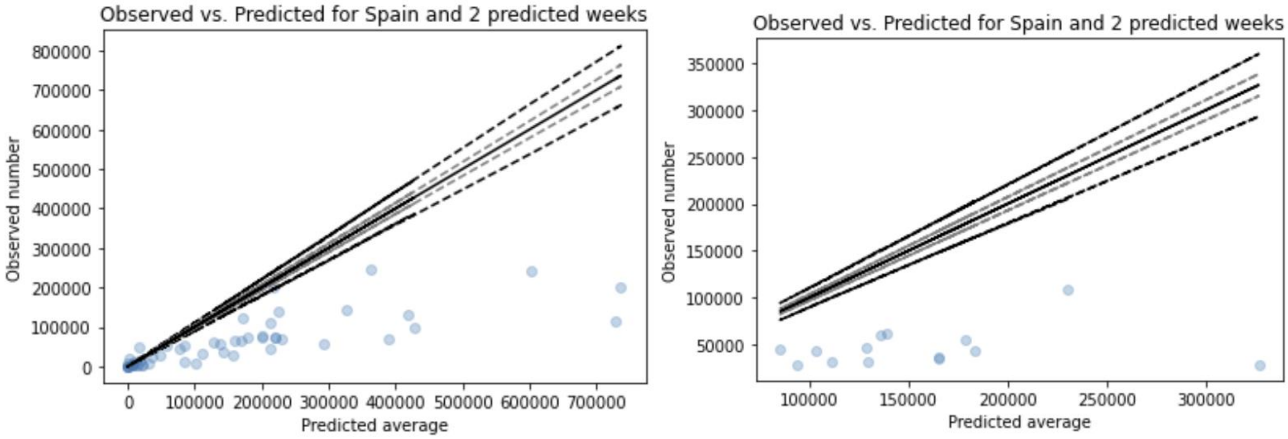


Figura 8 - CPDs de las predicciones sobre los datos de entrenamiento (izquierda) y validación (derecha) del modelo de regresión lineal con horizonte temporal de predicción de 2 semanas

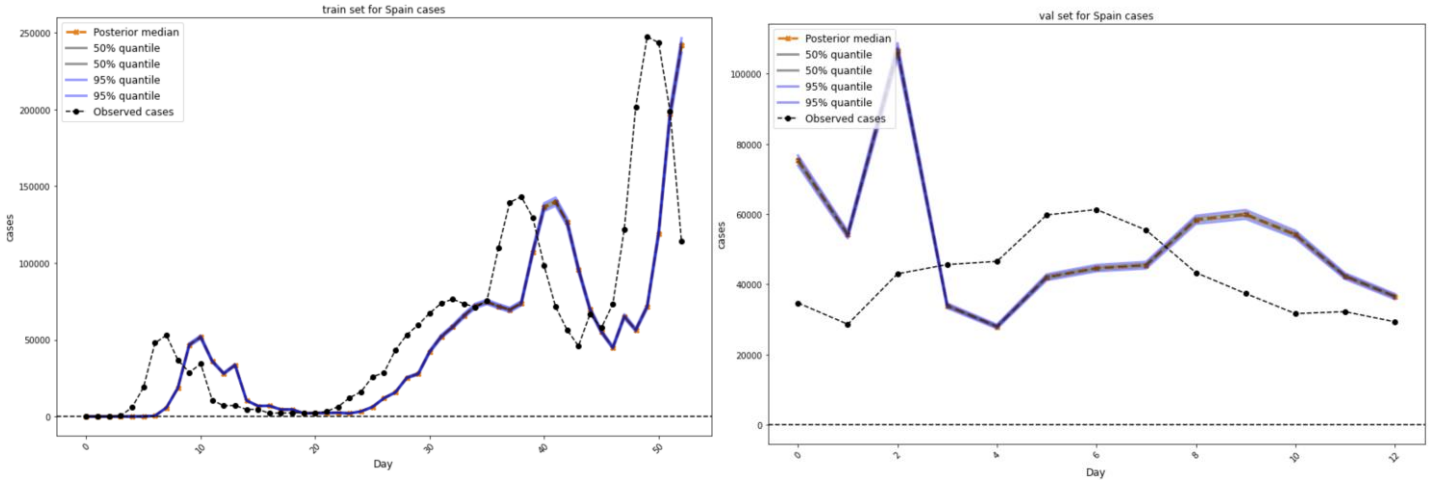


Figura 9 - Predicciones sobre los datos de entrenamiento (izquierda) y validación (derecha) del modelo de regresión lineal con horizonte temporal de predicción de 3 semanas

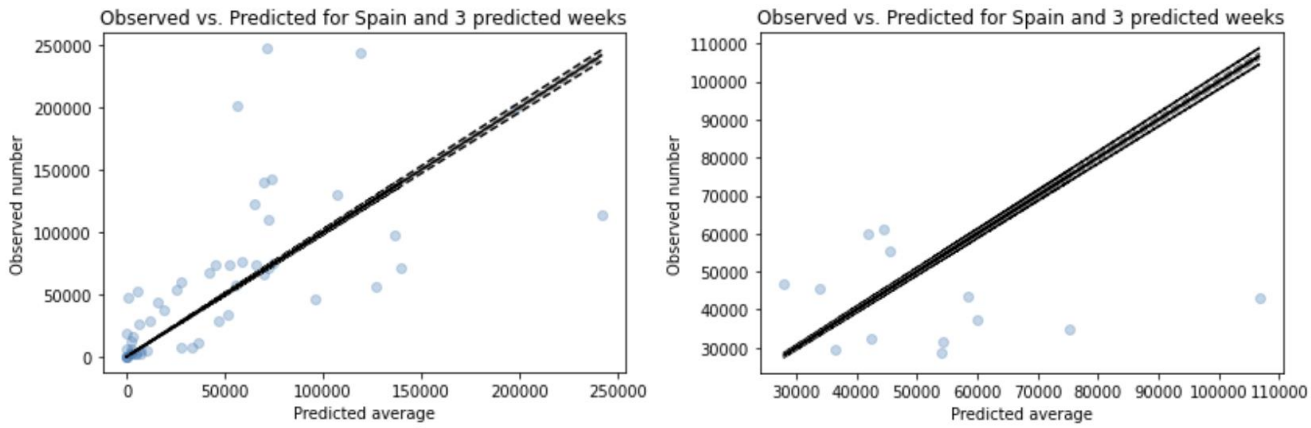


Figura 10 - CPDs de las predicciones sobre los datos de entrenamiento (izquierda) y validación (derecha) del modelo de regresión lineal con horizonte temporal de predicción de 3 semanas

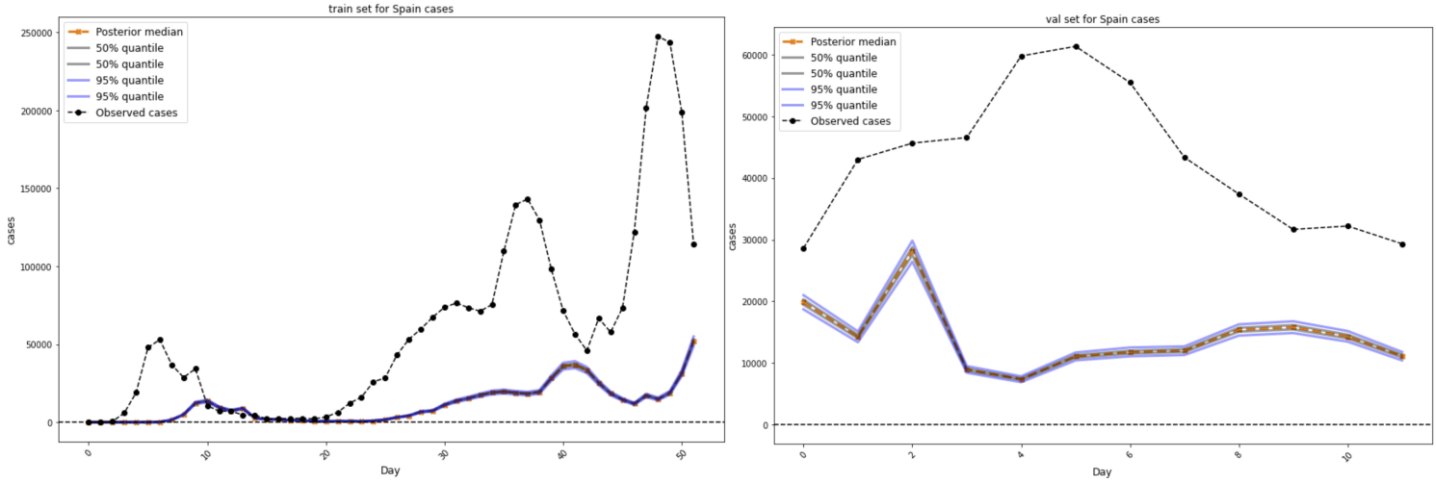


Figura 11 - Predicciones sobre los datos de entrenamiento (izquierda) y validación (derecha) del modelo de regresión lineal con horizonte temporal de predicción de 4 semanas

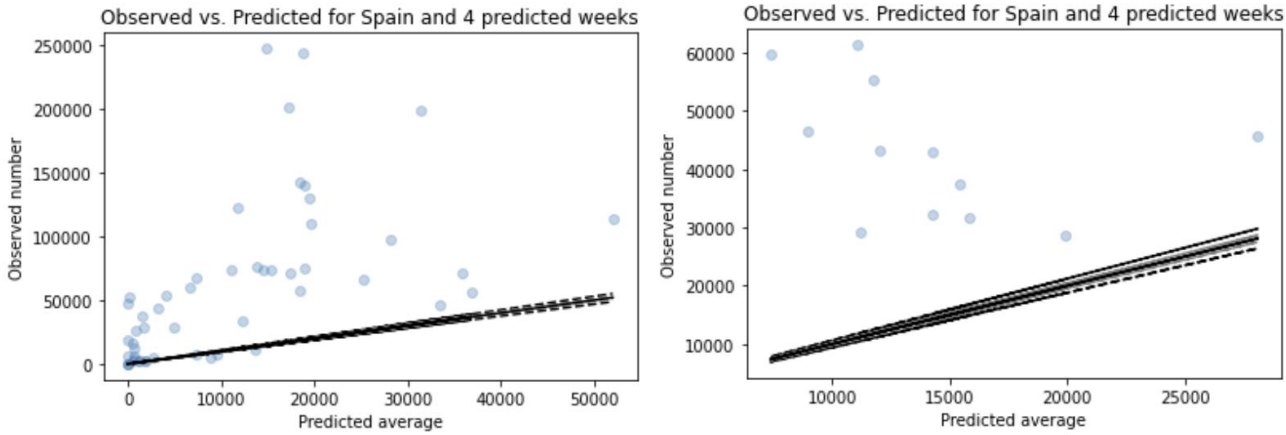


Figura 12 - CPDs de las predicciones sobre los datos de entrenamiento (izquierda) y validación (derecha) del modelo de regresión lineal con horizonte temporal de predicción de 4 semanas

4.1.2. Evaluación del Modelo

Tras la evaluación del modelo de regresión lineal con las métricas de evaluación, se obtiene la siguiente tabla, que es una muestra, centrada en este modelo, de la tabla adjunta a este documento:

	model	target_variable	horizon	location	interval_score	sharpness	underprediction	overprediction	rel_wis	ae	rel_ae	cov_50	cov_95	bias	n_loc	n	location_name
	UNED-CovidPredPMA																
946	UNED-CovidPredPMA	inc case	1	ES	45797,0	13006,0	3659,00	29133,0	5,76000	73615,0	5,97000	0,430000	0,570000	0,510000	1	7	Spain
961	UNED-CovidPredPMA	inc case	4	ES	212271	49368,0	0,00000	162903	10,4700	291219	8,70000	0,00000	0,00000	1,00000	1	4	Spain
956	UNED-CovidPredPMA	inc case	3	ES	354077	56281,0	0,00000	297796	19,6400	500810	17,0800	0,00000	0,00000	1,00000	1	5	Spain
951	UNED-CovidPredPMA	inc case	2	ES	300826	50782,0	0,00000	250045	22,0600	430648	20,1700	0,00000	0,170000	0,930000	1	6	Spain

En la tabla anterior se observa lo siguiente:

- El valor del WIS relativo para todos los horizontes temporales de predicción es muy superior a 1, lo que indica que el modelo tiene un rendimiento muy por debajo de la media del resto de modelos. Además, este valor se incrementa a medida que aumenta el horizonte temporal de la predicción.
- De la misma forma que ocurre con el WIS relativo, el error absoluto relativo (rel_ae) también es muy alto en todos los horizontes temporales y aumenta a medida que aumenta el horizonte temporal.
- La cobertura del 50% (cov_50) para el horizonte temporal de 1 semana es cercano a 50% (43%), lo que indica que las observaciones de datos reales están dentro del cuantil 50% de la predicción en un número de casos cercano al ideal para esta medida (50%). Además, el 57% de las observaciones entran en el cuantil 95% (cov_95). Este dato, en combinación con el obtenido en cov_50, indica que el 100% de las observaciones se encuentran dentro del rango de predicción generado por el modelo.
- Los horizontes temporales de 2, 3 y 4 semanas tienen una cobertura, tanto del 50% como del 95% de 0% o cercana al 0%, lo que aporta datos claros acerca de la aleatoriedad de las predicciones que realiza este modelo para estos horizontes temporales, tal como se observó durante el entrenamiento.
- Todos los horizontes temporales tienen un sesgo alto (mayor del 50%) y tienden a hacer predicciones por encima del valor real observado. Únicamente el horizonte temporal de 1 semana realiza predicciones por debajo del valor real y por encima del mismo, como se puede observar en las columnas “underprediction” y “overprediction”.

En la comparación de los resultados obtenidos, tomando el WIS relativo, con aquellos obtenidos por el resto de los modelos del Hub que generan predicciones para España y con los mismos horizontes temporales estudiados, este modelo tiene un rendimiento que es el peor de todos los modelos que colaboran con el Hub. Únicamente la predicción con horizonte temporal de 1 semana es capaz de superar el rendimiento de algunos modelos del presente estudio (no del Hub), como se verá más adelante.

Por todo lo anterior, el resultado del modelo de regresión lineal se puede considerar muy bajo, y, por tanto, supone una línea base fácil de mejorar, en teoría, por el resto de los modelos del estudio.

4.2. Modelo con Distribución de Poisson

4.2.1. Entrenamiento del Modelo

Puesto que la distribución de Poisson únicamente tiene un parámetro (λ), y dado que únicamente se ha añadido una capa densa para generar los valores de λ, el modelo solamente tiene 2 parámetros que entrenar. Este hecho hace que el resultado de la función de pérdida NLL al final del entrenamiento sea

muy alto, dado que el modelo no consigue ajustarse todo lo necesario a las observaciones reales. Esto es debido a que, a pesar de que el parámetro λ de la distribución represente la media y la desviación típica, este no expresa de forma precisa la incertidumbre del modelo, como se puede comprobar en los gráficos de representación de las predicciones frente a las observaciones y en los CPD correspondientes, más adelante. Este hecho se observa en todos los horizontes temporales de predicción.

Tal y cómo ocurría en el modelo de regresión lineal, en este caso también se degrada el rendimiento del modelo a medida que el horizonte temporal de la predicción aumenta. En este caso, es habitual ver que el NLL del conjunto de validación es más bajo que el NLL del conjunto de entrenamiento, lo que es una prueba positiva del poder predictivo del modelo. Sin embargo, los altos resultados del NLL, unidos a la degradación del modelo en horizontes temporales altos, hace ver que los resultados de la evaluación no serán buenos.

La siguiente figura muestra los resultados del entrenamiento para todos los horizontes temporales:

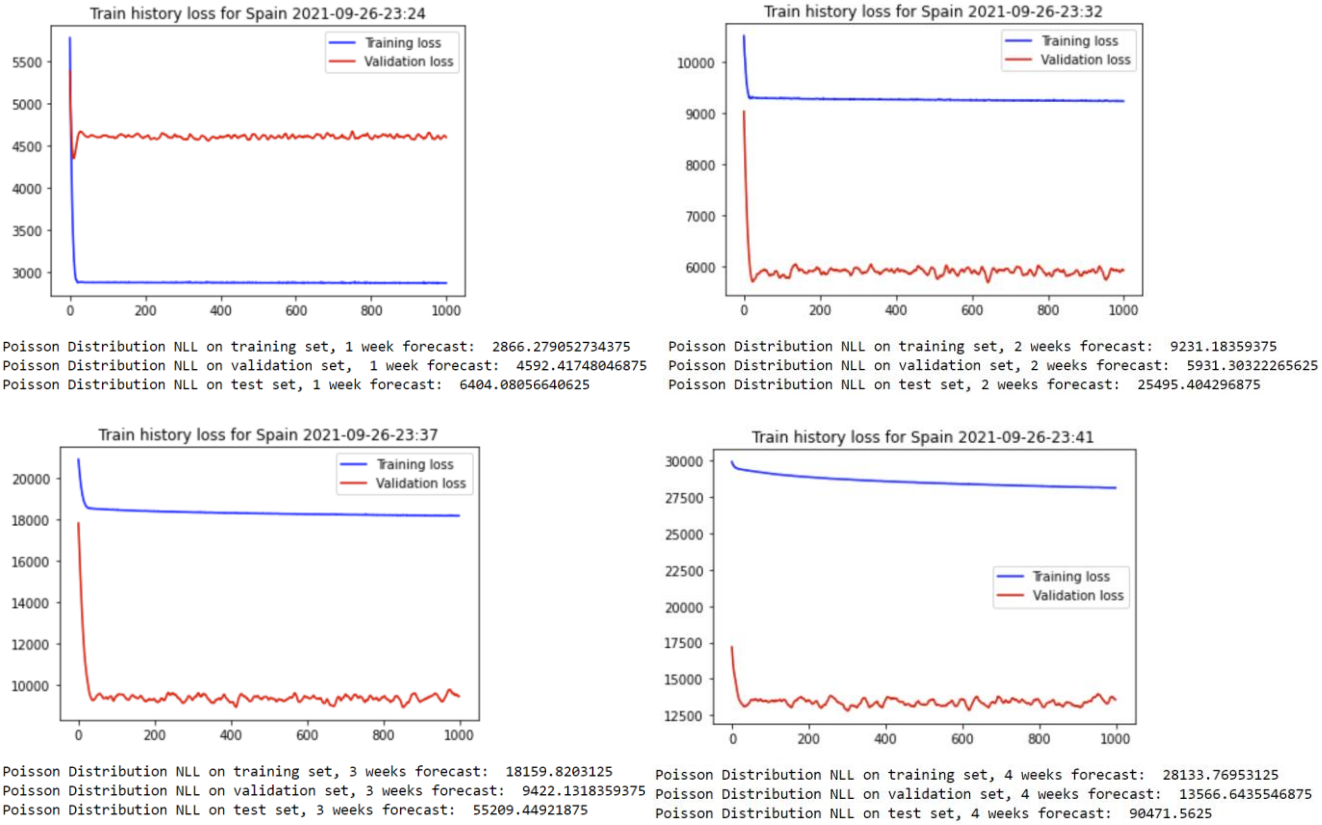


Figura 13 - Resultado del entrenamiento del modelo con distribución de Poisson con predicciones de 1 a 4 semanas (de derecha a izquierda y de arriba abajo), con detalle del resultado del NLL en los 3 conjuntos de datos (entrenamiento, validación y test).

En las siguientes figuras, correspondientes a las visualizaciones de las predicciones frente a las observaciones y a los CPD de todos los horizontes temporales, se puede observar que el modelo traza unas predicciones que se aproximan a los valores de las observaciones, pero es incapaz de capturar la incertidumbre del modelo, y no aporta ningún tipo de información acerca de la misma. Además, las predicciones pierden precisión a medida que se aumenta el horizonte temporal de la predicción.

Las siguientes figuras permiten observar este hecho:

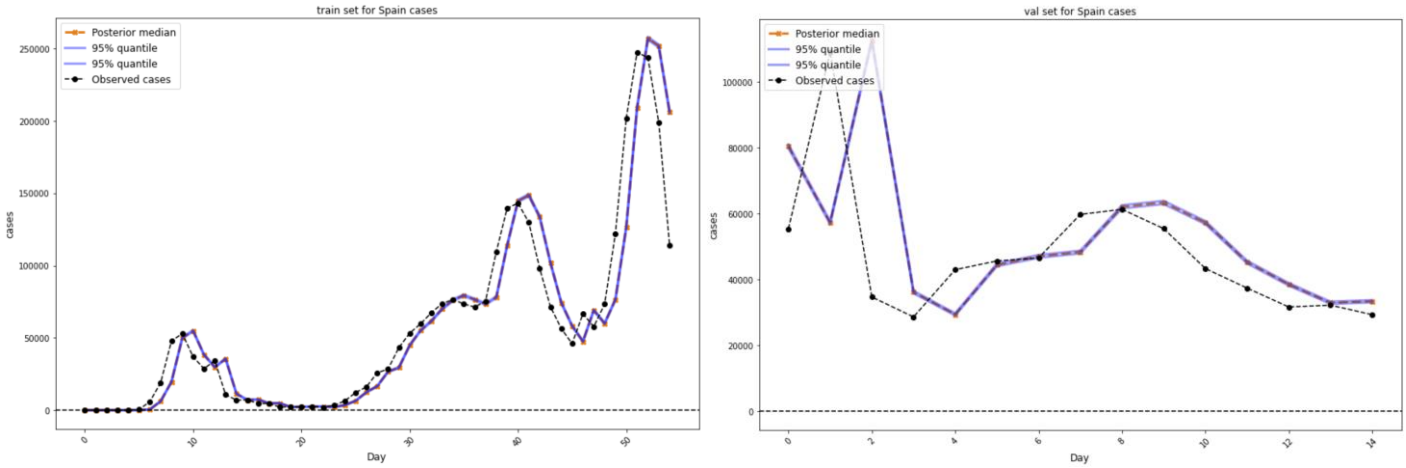


Figura 14 - Predicciones sobre los datos de entrenamiento (izquierda) y validación (derecha) del modelo con distribución de Poisson con horizonte temporal de predicción de 1 semana

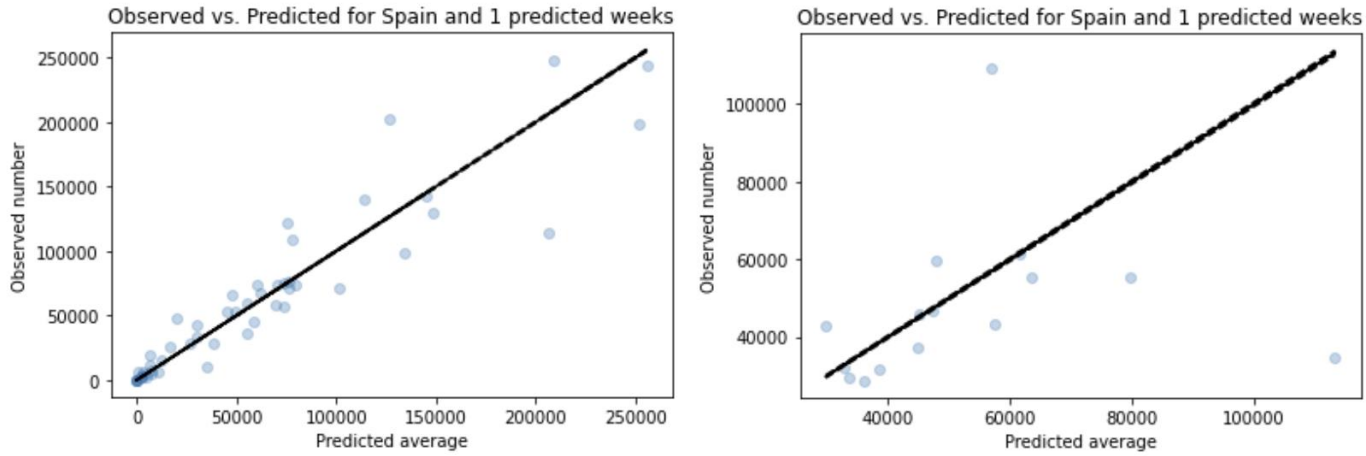


Figura 15 - CPDs de las predicciones sobre los datos de entrenamiento (izquierda) y validación (derecha) del modelo con distribución de Poisson con horizonte temporal de predicción de 1 semana

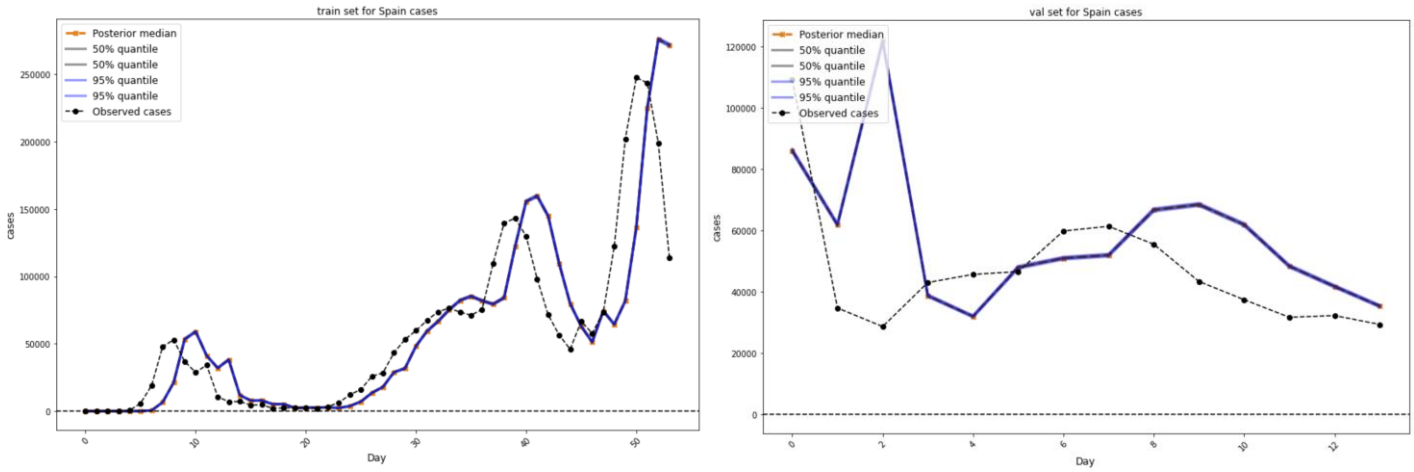


Figura 16 - Predicciones sobre los datos de entrenamiento (izquierda) y validación (derecha) del modelo con distribución de Poisson con horizonte temporal de predicción de 2 semanas

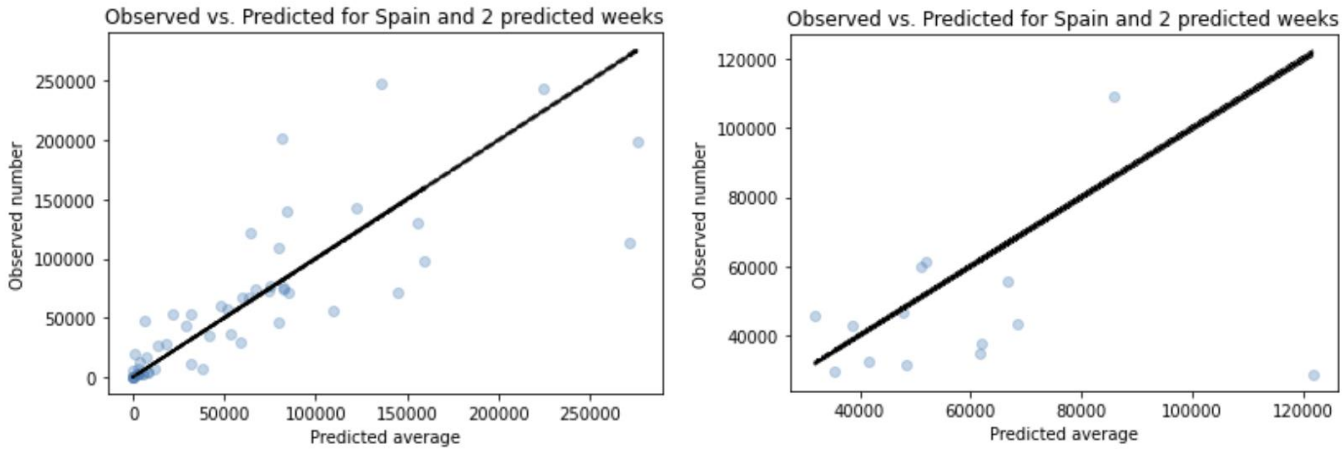


Figura 17 - CPDs de las predicciones sobre los datos de entrenamiento (izquierda) y validación (derecha) del modelo con distribución de Poisson con horizonte temporal de predicción de 2 semanas

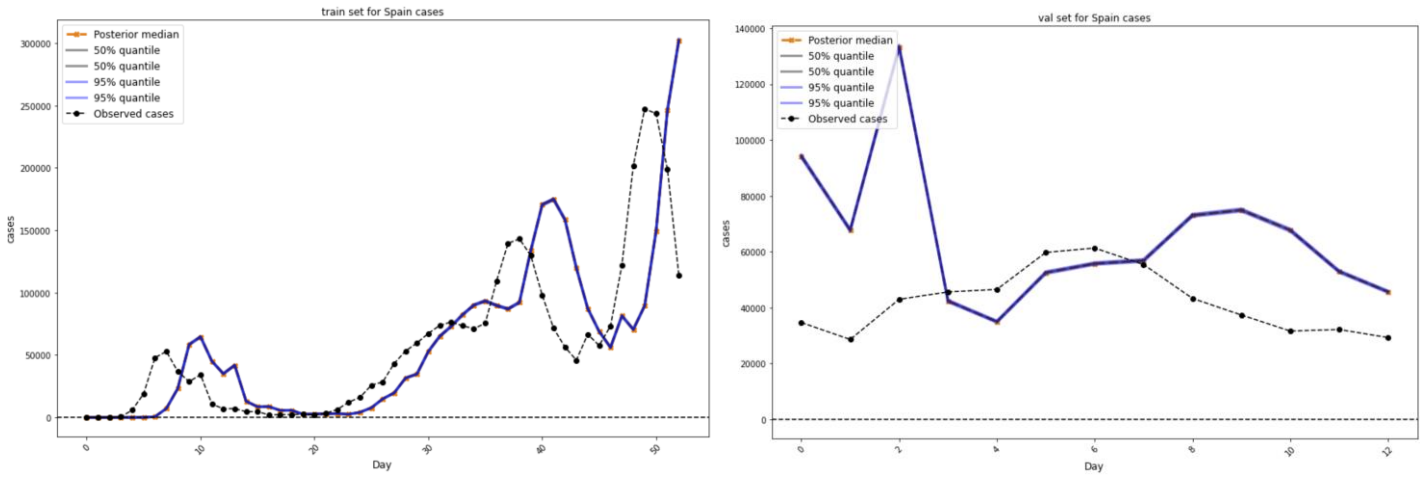


Figura 18 - Predicciones sobre los datos de entrenamiento (izquierda) y validación (derecha) del modelo con distribución de Poisson con horizonte temporal de predicción de 3 semanas

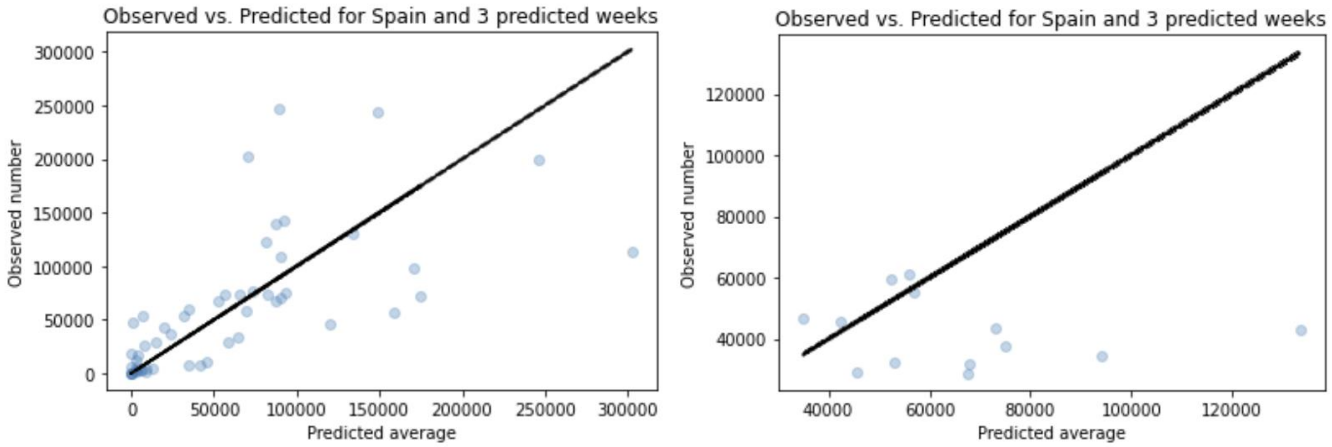


Figura 19 - CPDs de las predicciones sobre los datos de entrenamiento (izquierda) y validación (derecha) del modelo con distribución de Poisson con horizonte temporal de predicción de 3 semanas

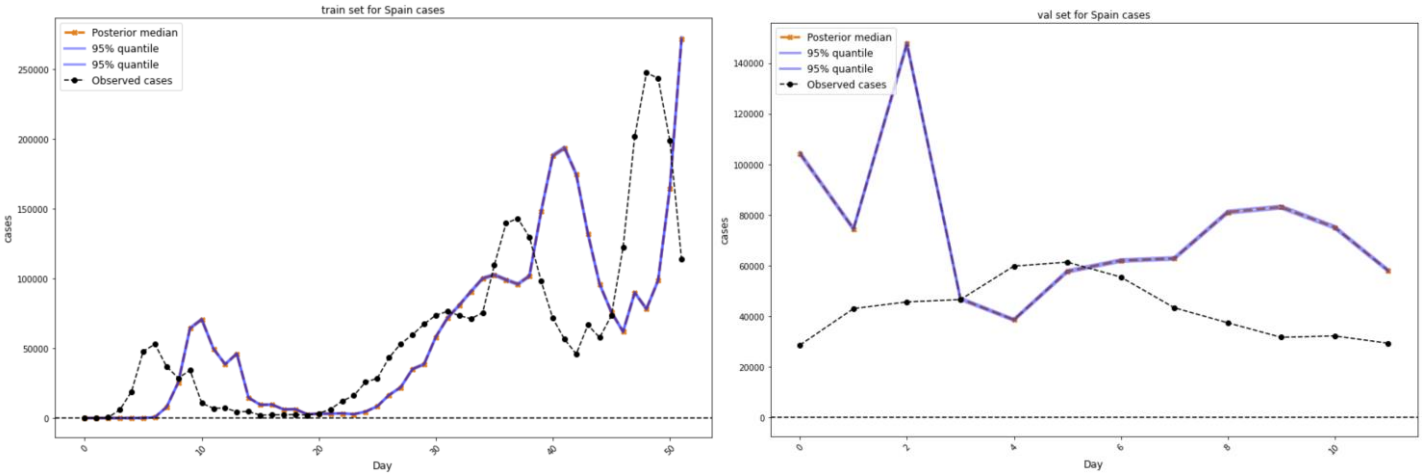


Figura 20 - Predicciones sobre los datos de entrenamiento (izquierda) y validación (derecha) del modelo con distribución de Poisson con horizonte temporal de predicción de 4 semanas

4.2.2. Evaluación del Modelo

Tras la evaluación del modelo con distribución de Poisson con las métricas de evaluación, se obtiene la siguiente tabla, que es una muestra, centrada en este modelo, de la tabla adjunta a este documento:

	model	target_variable	horizon	location	interval_score	sharpness	underprediction	overprediction	rel_wis	ae	rel_ae	cov_50	cov_95	bias	n_loc	n	location_name
	UNED-CovidPredPMA-P																
948	UNED-CovidPredPMA-Poisson	inc case	1	ES	46661,0	67,0000	0,00000	46594,0	5,87000	46899,0	3,80000	0,00000	0,00000	1,00000	1	7	Spain
963	UNED-CovidPredPMA-Poisson	inc case	4	ES	126596	82,0000	0,00000	126514	6,25000	126881	3,79000	0,00000	0,00000	1,00000	1	4	Spain
953	UNED-CovidPredPMA-Poisson	inc case	2	ES	97937,0	81,0000	0,00000	97857,0	7,18000	98221,0	4,60000	0,00000	0,00000	1,00000	1	6	Spain
958	UNED-CovidPredPMA-Poisson	inc case	3	ES	134207	86,0000	0,00000	134121	7,45000	134500	4,59000	0,00000	0,00000	1,00000	1	5	Spain

En la tabla anterior se observa lo siguiente:

- En todos los horizontes temporales, el WIS relativo es muy alto, lo que indica que el modelo tiene un rendimiento inferior al de la media de los modelos del Hub. Sin embargo, este supone una mejora considerable con respecto al modelo de regresión lineal, en todos los horizontes temporales excepto en el de 1 semana, donde no se aprecia mejora. El cálculo del WIS relativo considera 3 componentes diferentes, la definición/dispersión (sharpness), la predicción excesiva (overprediction) y la predicción insuficiente (underprediction). Si observamos estos 3 componentes por separado en la tabla, podemos ver que el cálculo del WIS relativo se ve penalizado notablemente por los resultados de la predicción excesiva si la comparamos con los demás modelos, hecho al que ayuda que el modelo no sea capaz de capturar la incertidumbre de la predicción. Esta lectura también se extrae de que todas las predicciones tengan un sesgo igual a 1.
- La cobertura, tanto del 50% como del 95%, es 0 en todos los horizontes temporales, de acuerdo con lo mencionado en el punto anterior. Dado que el modelo no es capaz de capturar la incertidumbre de la predicción, todas las observaciones se encuentran fuera del percentil 95%, lo que se debe a la estrechez del rango de predicciones.
- El error absoluto relativo es bajo en comparación con el modelo de regresión lineal, lo que junto con el resultado del WIS relativo, ilustra la mejora con respecto a aquel modelo.

Por todo lo observado anteriormente, se puede concluir que el modelo supone una mejora de rendimiento con respecto a la línea base (regresión lineal), pero todavía tiene mucho margen de mejora, especialmente en lo relativo a la captura de la incertidumbre de la predicción.

4.3. Modelo con Mezcla de distribuciones Logísticas Discretizadas

4.3.1. Entrenamiento del Modelo

La arquitectura del modelo de distribuciones logísticas discretizadas, descrita en la sección 3.5.3 de este documento, produce que el modelo tenga 30 parámetros que entrenar, debido a las 15 neuronas de su capa densa, muchos más que el modelo de regresión lineal (4 parámetros) o el modelo con distribución de Poisson (2 parámetros). Esto le aporta mayor flexibilidad para producir funciones complejas que permitan ajustarse mejora los datos de entrenamiento, lo que, como puede observarse en los gráficos de la función de pérdida a continuación, reduce considerablemente el resultado final de NLL para todos los horizontes temporales. Además, se observa que, en todos los casos, el conjunto de validación tiene un resultado mucho mejor que el del conjunto de entrenamiento, lo que muestra que el modelo no hace "overfit" de los datos de entrenamiento y tiene capacidad predictiva. Este hecho se ve respaldado también por los resultados del conjunto de test, aunque tengan un resultado algo peor que el del conjunto de validación.

Además, no se observa una pérdida de rendimiento destacable en el modelo a medida que el horizonte temporal de las predicciones aumenta, lo que supone un progreso con respecto al modelo de regresión lineal y el modelo con distribución de Poisson.

La siguiente figura ilustra los resultados del entrenamiento para cada uno de los horizontes temporales de la predicción:


```
Discretized Logistic Mixture NLL on training set, 1 week forecast:  1225.0836181640625   Discretized Logistic Mixture NLL on training set, 2 weeks forecast:  70.10052490234375
Discretized Logistic Mixture NLL on validation set,  1 week forecast:  11.560486793518066 Discretized Logistic Mixture NLL on validation set, 2 weeks forecast:  13.246403694152832
Discretized Logistic Mixture NLL on test set, 1 week forecast:  15.10047721862793        Discretized Logistic Mixture NLL on test set, 2 weeks forecast:  13.36078929901123
```

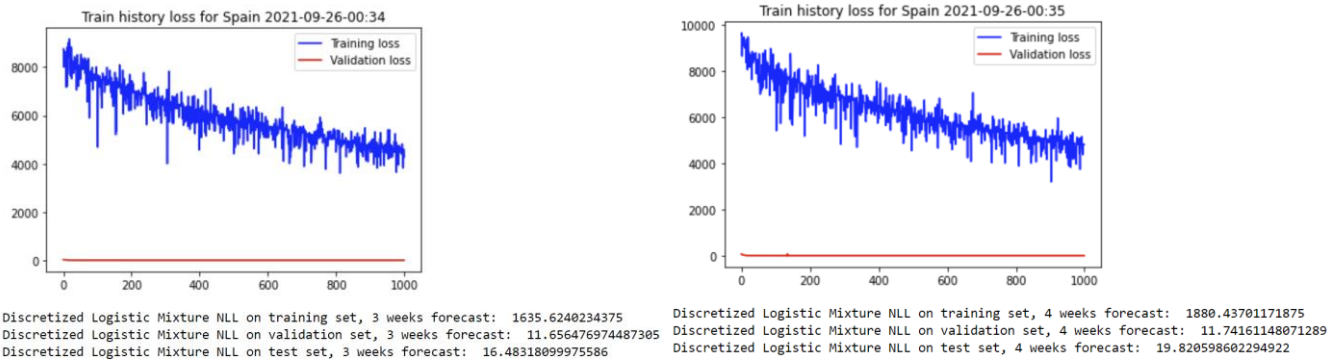


Figura 21 - Resultado del entrenamiento del modelo con mezcla de distribuciones logísticas discretizadas con predicciones de 1 a 4 semanas (de derecha a izquierda y de arriba abajo), con detalle del resultado del NLL en los 3 conjuntos de datos (entrenamiento, validación y test).

En las siguientes figuras, correspondientes a la comparación de las predicciones frente a las observaciones en los conjuntos de entrenamiento y validación, así como a los CPD de cada uno de los conjuntos, para todos los horizontes temporales, se observa que el modelo es capaz de realizar unas predicciones mucho mejores que los modelos anteriores, así como de capturar la incertidumbre del modelo correctamente.

Esto se puede apreciar especialmente en el modelo con horizonte temporal de 1 semana, donde el modelo muestra un alto grado de certidumbre en las predicciones cuyos valores se encuentran en un rango de valores conocido (días 0 a 35, aproximadamente), mostrando un rango de predicción estrecho, mientras que muestra un rango de predicción ancho, y por tanto alta incertidumbre, en cuando el valor del dato de entrada no se encuentra en un rango de valores conocido (días 35, aproximadamente, en adelante). Esto se ve reflejado en los CPD de horizonte temporal de 1 semana, donde prácticamente todas las predicciones se encuentran dentro del cuantil 50%.

Analizando estas figuras también se puede apreciar una cierta degradación del rendimiento del modelo a medida que aumenta el horizonte temporal de la predicción, algo que no se reflejaba notablemente analizando las gráficas de la función de pérdida durante el entrenamiento. Esto se ve reflejado en los CPD, donde se puede ver que hay más predicciones fuera del cuantil 50% y que el mismo es más ancho que en el horizonte temporal de 1 semana.

También cabe destacar que, en las gráficas de las predicciones con horizonte temporal de 2 semanas, se puede observar que el modelo ha hecho overfit de los datos de entrenamiento. Este es el horizonte temporal que mejor resultado de NLL tuvo durante el entrenamiento, pero se puede observar que las predicciones realizadas son las que más alejadas se encuentran de las observaciones, y que muestra mayor incertidumbre en las predicciones.

Además, en todos los casos se puede observar que tanto el cuantil 50% como el cuantil 95% establecen su límite inferior en 0 en todas las predicciones, excepto en algunos casos aislados en las predicciones con horizonte temporal de 4 semanas. Esto se debe a la utilización de las distribuciones logísticas, cuyo límite inferior es 0, pero también representa la baja flexibilidad del modelo para ajustar este límite inferior.

En las siguientes figuras se puede observar todo lo comentado:

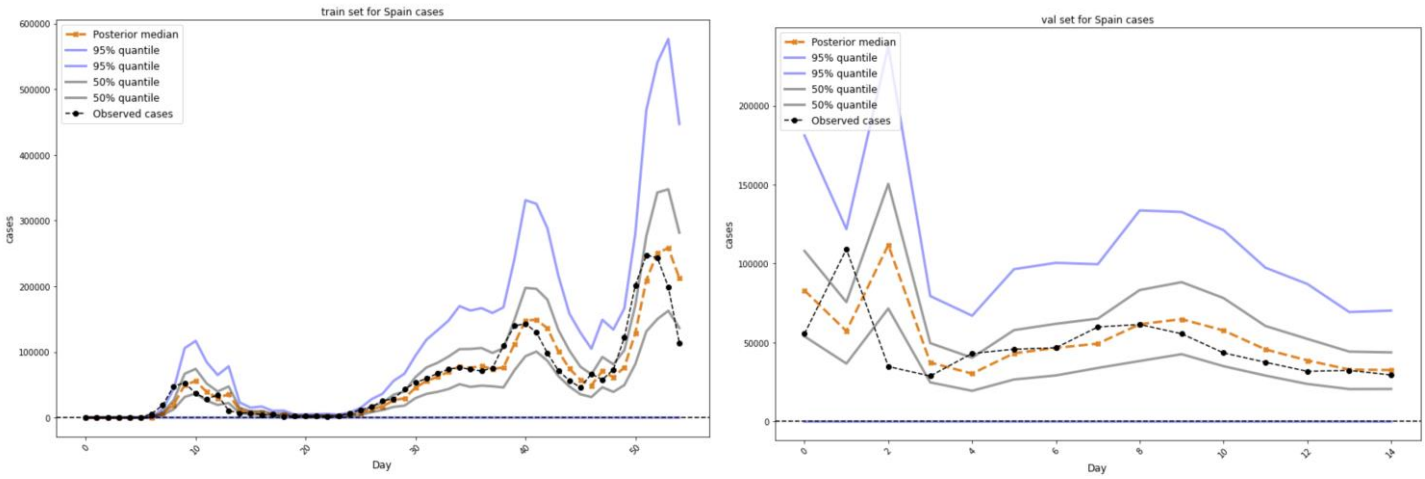


Figura 22 - Predicciones sobre los datos de entrenamiento (izquierda) y validación (derecha) del modelo con mezcla de distribuciones logísticas discretizadas, con horizonte temporal de predicción de 1 semana

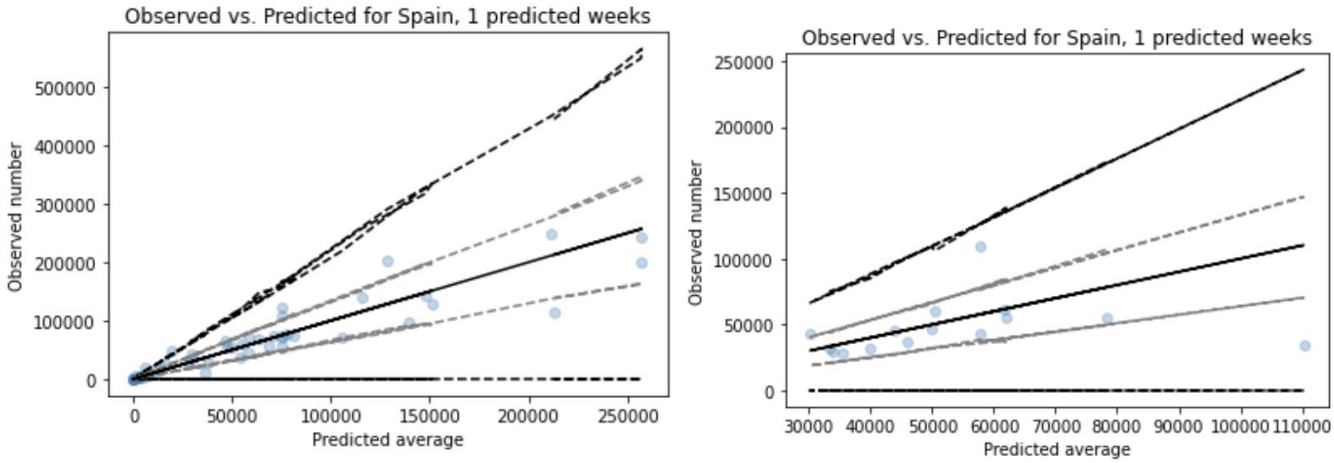


Figura 23 - CPDs de las predicciones sobre los datos de entrenamiento (izquierda) y validación (derecha) del modelo con mezcla de distribuciones logísticas discretizadas, con horizonte temporal de predicción de 1 semana

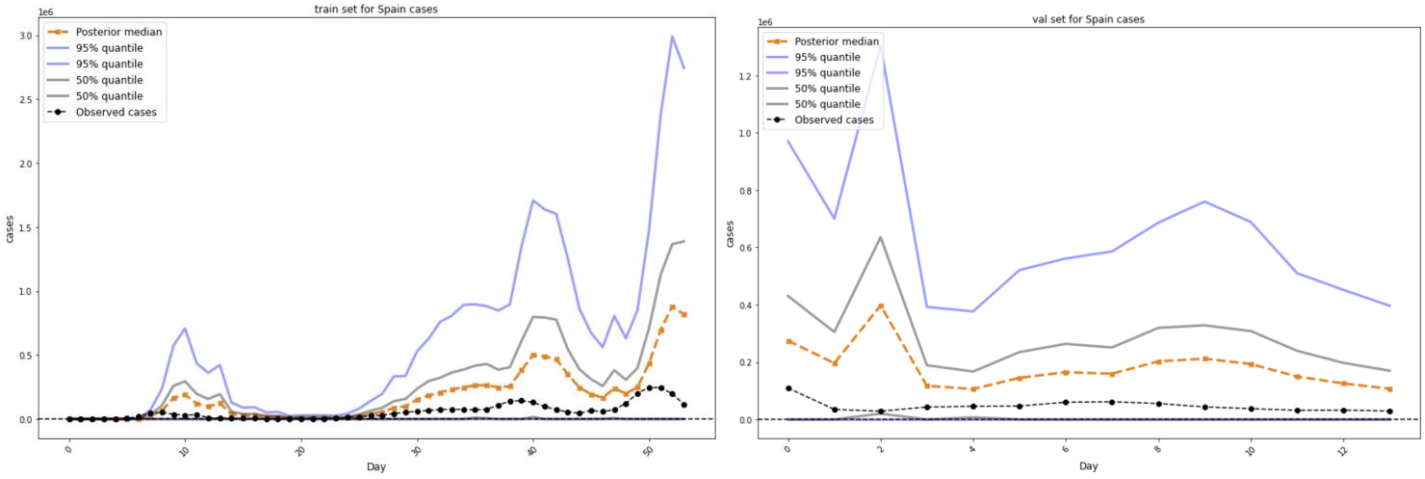


Figura 24 - Predicciones sobre los datos de entrenamiento (izquierda) y validación (derecha) del modelo con mezcla de distribuciones logísticas discretizadas, con horizonte temporal de predicción de 2 semanas

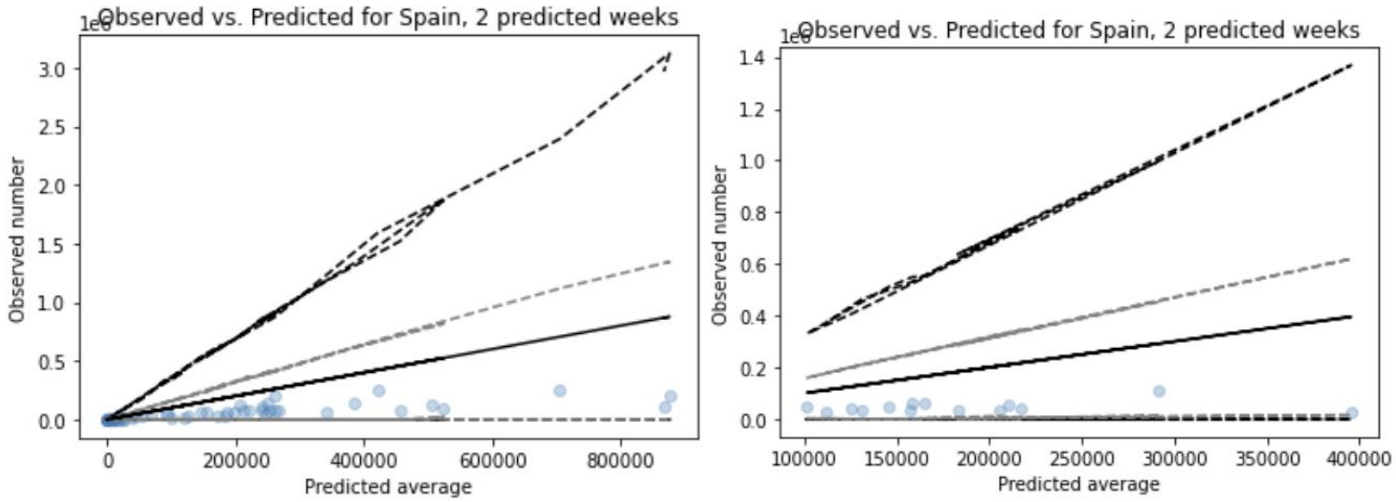


Figura 25 - CPDs de las predicciones sobre los datos de entrenamiento (izquierda) y validación (derecha) del modelo con mezcla de distribuciones logísticas discretizadas, con horizonte temporal de predicción de 2 semanas

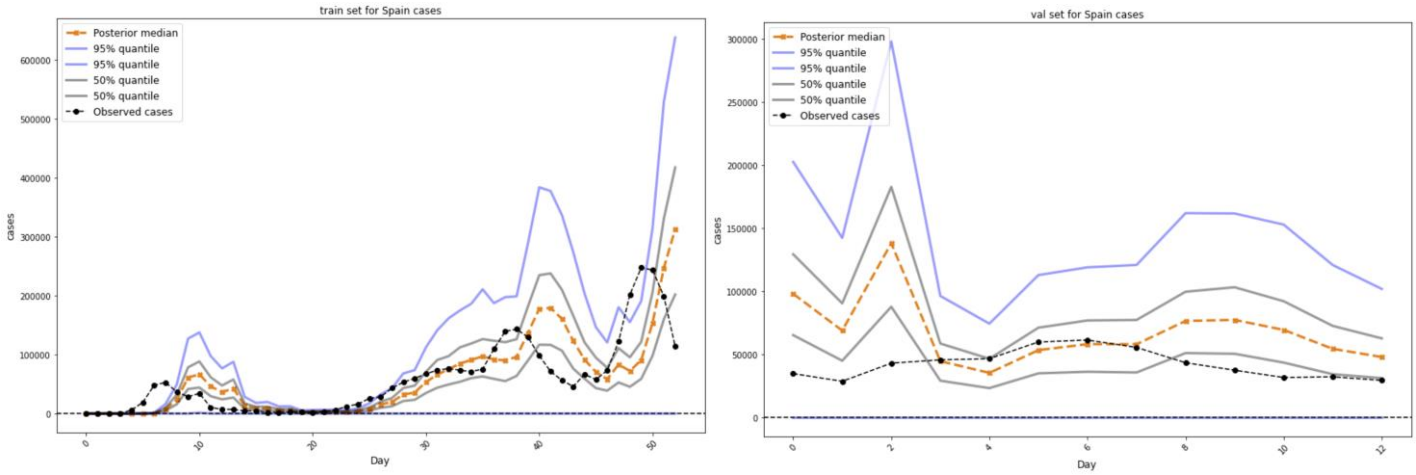


Figura 26 - Predicciones sobre los datos de entrenamiento (izquierda) y validación (derecha) del modelo con mezcla de distribuciones logísticas discretizadas, con horizonte temporal de predicción de 3 semanas

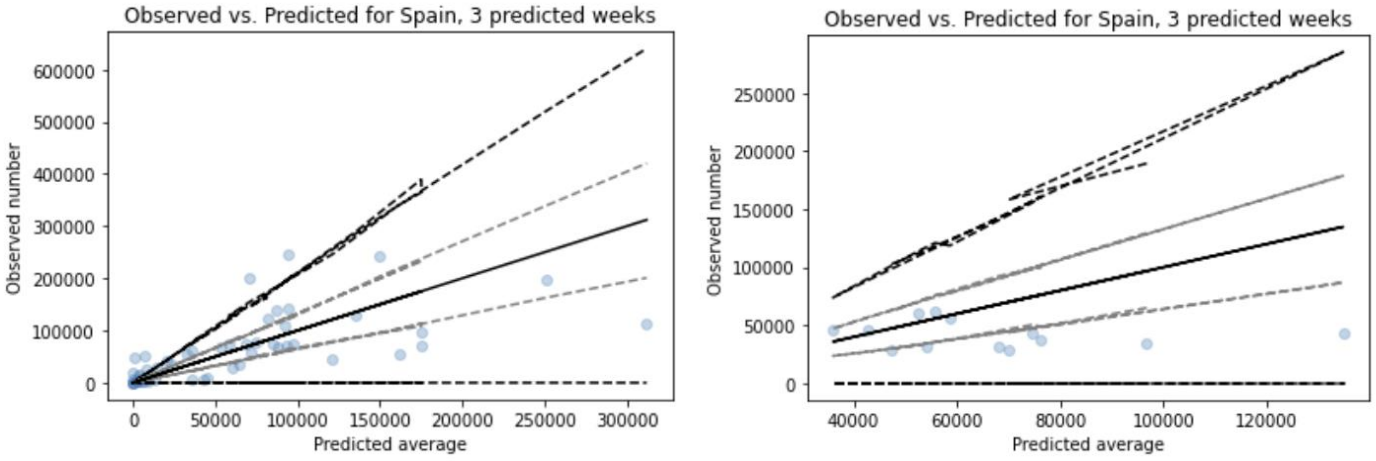


Figura 27 - CPDs de las predicciones sobre los datos de entrenamiento (izquierda) y validación (derecha) del modelo con mezcla de distribuciones logísticas discretizadas, con horizonte temporal de predicción de 3 semanas

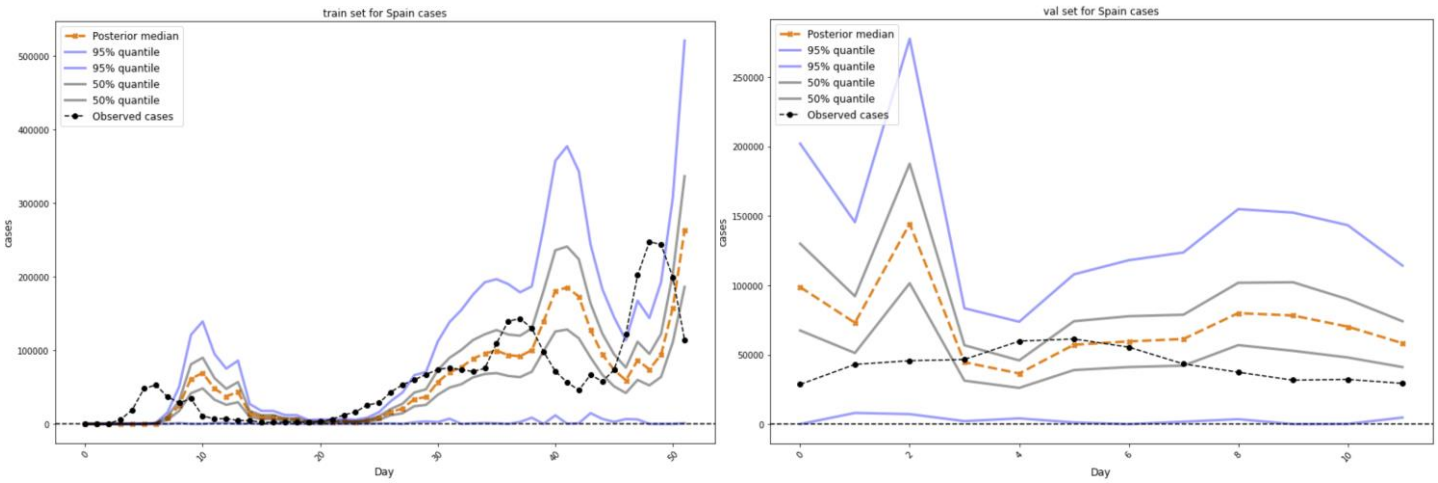


Figura 28 - Predicciones sobre los datos de entrenamiento (izquierda) y validación (derecha) del modelo con mezcla de distribuciones logísticas discretizadas, con horizonte temporal de predicción de 4 semanas

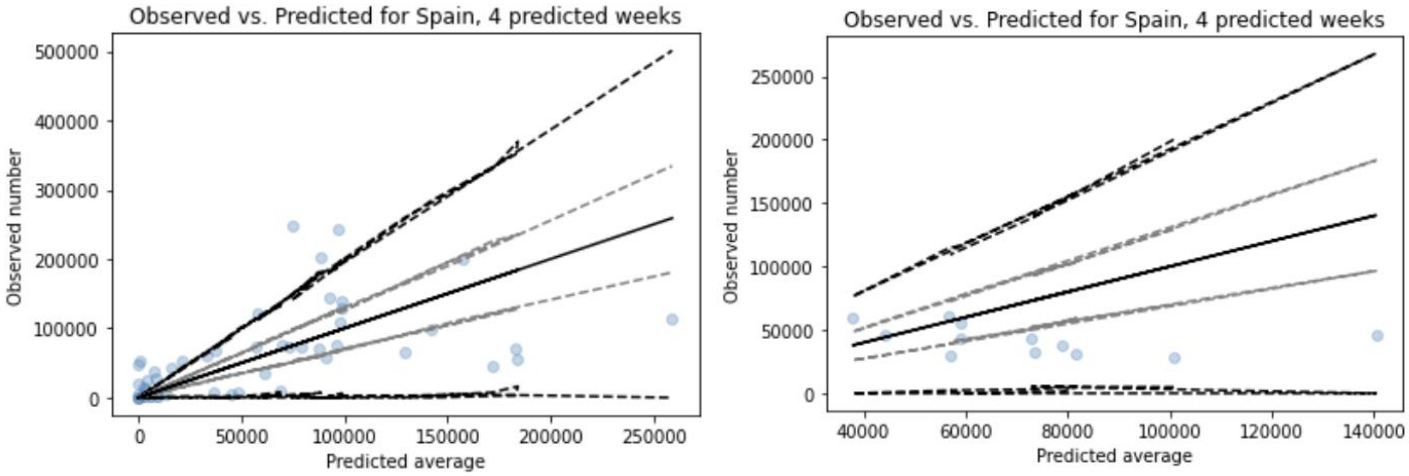


Figura 29 - CPDs de las predicciones sobre los datos de entrenamiento (izquierda) y validación (derecha) del modelo con mezcla de distribuciones logísticas discretizadas, con horizonte temporal de predicción de 4 semanas

4.3.2. Evaluación del Modelo

Tras la evaluación del modelo con las métricas de evaluación descritas, se obtiene la siguiente tabla, que es una muestra, centrada en este modelo, de la tabla adjunta a este documento:

	model	target_variable	horizon	location	interval_score	sharpness	underprediction	overprediction	rel_wis	ae	rel_ae	cov_50	cov_95	bias	n_loc	n	location_name
	UNED-CovidPredPMA-M																
947	UNED-CovidPredPMA-Mix	inc case	1	ES	25556,0	12668,0	0,00000	12888,0	3,21000	44776,0	3,63000	0,140000	1,00000	0,610000	1	7	Spain
962	UNED-CovidPredPMA-Mix	inc case	4	ES	80610,0	14759,0	0,00000	65851,0	3,98000	122800	3,67000	0,250000	1,00000	0,750000	1	4	Spain
952	UNED-CovidPredPMA-Mix	inc case	2	ES	56902,0	0,00000	56902,0	0,00000	4,17000	56902,0	2,66000	0,00000	0,00000	-1,00000	1	6	Spain
957	UNED-CovidPredPMA-Mix	inc case	3	ES	82276,0	19775,0	0,00000	62501,0	4,56000	137525	4,69000	0,00000	1,00000	0,830000	1	5	Spain

En la tabla anterior se observa lo siguiente:

- La puntuación del WIS relativo (rel_wis) supone una mejora con respecto al modelo con distribución de Poisson (también con respecto al modelo de regresión lineal) en todos los horizontes temporales, pero continúa teniendo un rendimiento por debajo de la media en el cómputo global.
- Lo mismo sucede, aunque en menor medida, con el error absoluto relativo (rel_ae).

- A diferencia de lo observado en los datos de entrenamiento y validación, el modelo no captura correctamente la incertidumbre del modelo, dado que la cobertura de 50% (cov_50) es muy baja (menor que 0,5). Por el contrario, la cobertura de 95% (cov_95) es muy alta (superior a 0,95), mostrando una incertidumbre excesiva, como ya se pudo apreciar anteriormente.
- A excepción del horizonte temporal de 2 semanas, el modelo tiende a hacer predicciones excesivas, como puede verse en los campos "overprediction" y "bias", lo que penaliza el resultado del WIS relativo.

Si se compara el rendimiento del modelo, tomando como referencia el WIS relativo, con el de otros modelos del Hub, se observa que este modelo todavía está lejos del rendimiento obtenido por otros modelos. Este modelo continúa estando entre los modelos con peor puntuación de WIS relativo entre aquellos que realizan predicciones para España con los mismos horizontes temporales, y solo el modelo con horizonte temporal de 1 semana es capaz de batir a uno de los modelos del Hub.

Por todo lo anterior, los siguientes modelos deberían mejorar la capacidad de capturar la incertidumbre de las predicciones, para aportar mejores resultados.

4.4. Modelo de Inferencia Variacional con Distribución Logística

4.4.1. Entrenamiento del Modelo

En este modelo se ha implementado una red neuronal de dos capas del tipo "DenseFlipout" (tipología de capa de Tensorflow Probability que implementa la inferencia variacional) la primera de 10 neuronas, y la segunda de 2 neuronas, para alimentar los 2 parámetros de la distribución logística utilizada. Todas las funciones asociadas al sesgo y al kernel de la previa y la posterior son las funciones por defecto en ambas capas. Esta arquitectura implica un mayor número de parámetros, 84, no solo originados por el incremento en el número de capas y neuronas con respecto a los modelos anteriores, también porque la implementación de la inferencia Bayesiana duplica el número de parámetros necesarios en el modelo, debido a que sustituye los valores de los pesos por distribuciones.

Si observamos las gráficas de la función de pérdida NLL durante el entrenamiento y validación para todos los horizontes temporales, se observa que, en todos los casos, la pérdida del conjunto de entrenamiento, y en menor medida la del conjunto de validación, empiezan con valores muy altos, que rápidamente descienden hasta niveles más bajos a los vistos en el modelo de mezcla de distribuciones logísticas, donde se mantiene estable.

Se observa también que el modelo no presenta signos de overfit de los datos de entrenamiento, ya que el resultado del NLL del conjunto de evaluación es similar o menor que el NLL del conjunto de entrenamiento, lo que se mantiene, además, en el NLL del conjunto de test.

Además, como ha podido observarse también en el resto de los modelos estudiados anteriormente, puede verse una disminución del rendimiento del modelo a medida que aumenta el horizonte temporal de la predicción, siendo más pronunciado que en el modelo de mezcla de distribuciones logísticas.

La siguiente figura muestra los datos de la función de pérdida en los conjuntos de entrenamiento y validación para todos los horizontes temporales:

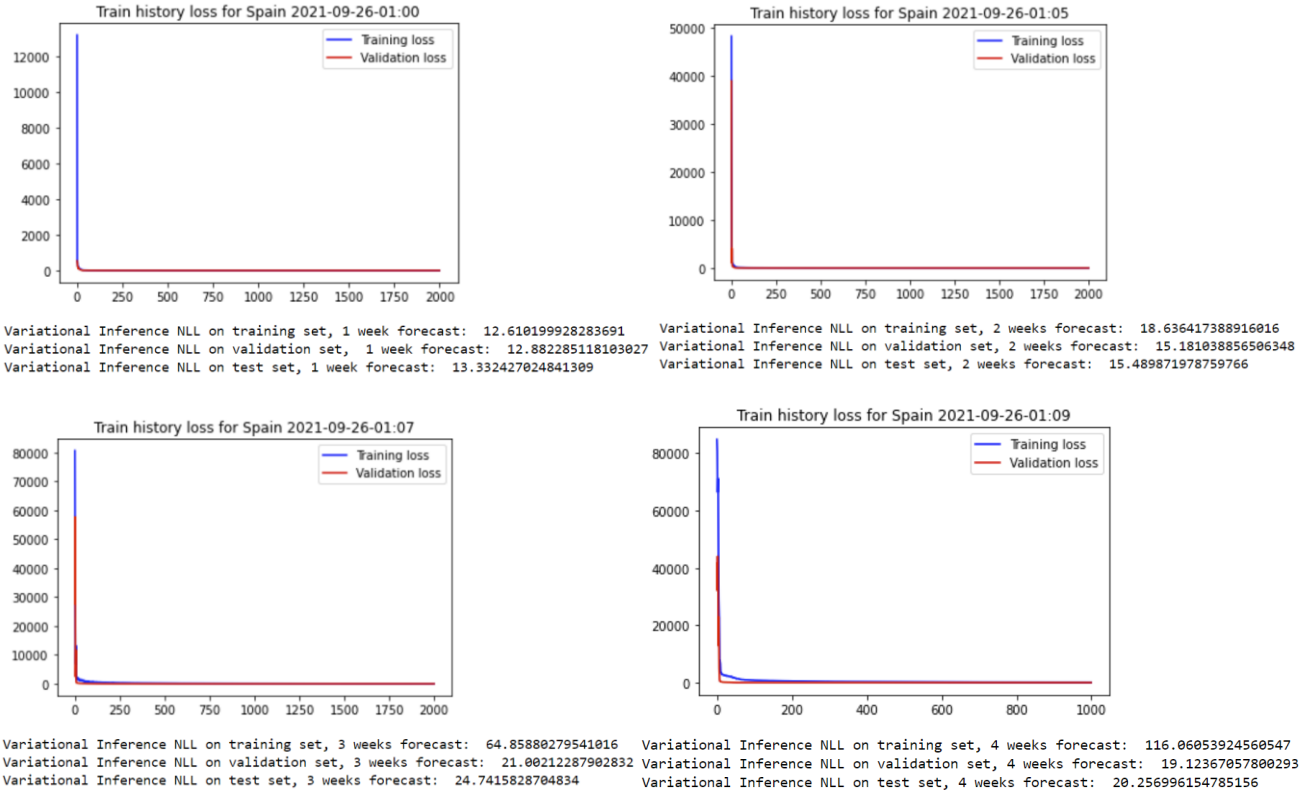


Figura 30 - Resultado del entrenamiento del modelo de inferencia variacional con predicciones de 1 a 4 semanas (de derecha a izquierda y de arriba abajo), con detalle del resultado del NLL en los 3 conjuntos de datos (entrenamiento, validación y test).

Si observamos los gráficos de comparación de las predicciones frente a las observaciones, así como los CPD de las predicciones, puede observarse que el modelo muestra un comportamiento similar al visto en la mezcla de distribuciones logísticas, realizando predicciones razonablemente precisas en el horizonte temporal de 1 semana, aunque algo peores que aquel, pero capturando correctamente la incertidumbre del modelo, dado que casi todas las predicciones se encuentran dentro del cuantil 50%, y ninguna fuera del cuantil 95%. Esto es algo que supone una mejora frente al modelo anterior, especialmente en lo relativo al cuantil 50%, y que se observa también en el resto de los horizontes temporales, dado que un mayor número de predicciones se observan en este cuantil.

Sin embargo, no se ha observado, con respecto al modelo anterior, una mejora en la cuantificación de la incertidumbre en el cuantil 95%, ya que el límite inferior es casi invariablemente 0 en todos los casos (a excepción del horizonte temporal de 3 semanas), y el límite superior se dispara hasta valores muy extremos cuando el valor de entrada no está en el rango de lo observado en el conjunto de entrenamiento. Este último punto no tiene por qué tomarse como un dato negativo, ya que el modelo es capaz de identificar cuando existe una alta incertidumbre en la predicción, pero esto afectará negativamente el resultado de las métricas de evaluación (WIS relativo).

En cuanto al resultado de las predicciones para los horizontes temporales de 2, 3 y 4 semanas, se puede apreciar un mejor resultado que la mezcla de distribuciones logísticas en el horizonte temporal de 2 semanas, mientras que, para los horizontes temporales más altos, el resultado se degrada considerablemente, de forma más acentuada de lo que se apreciaba en aquel modelo.

Las siguientes figuras ilustran las observaciones comentadas:

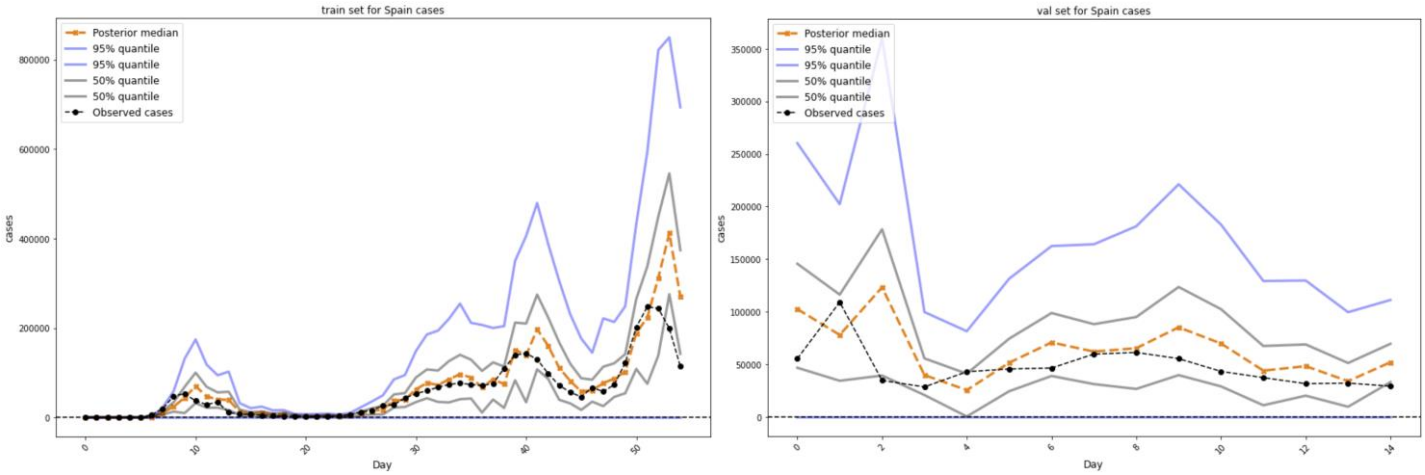


Figura 31 - Predicciones sobre los datos de entrenamiento (izquierda) y validación (derecha) del modelo de inferencia variacional con distribución logística, con horizonte temporal de predicción de 1 semana

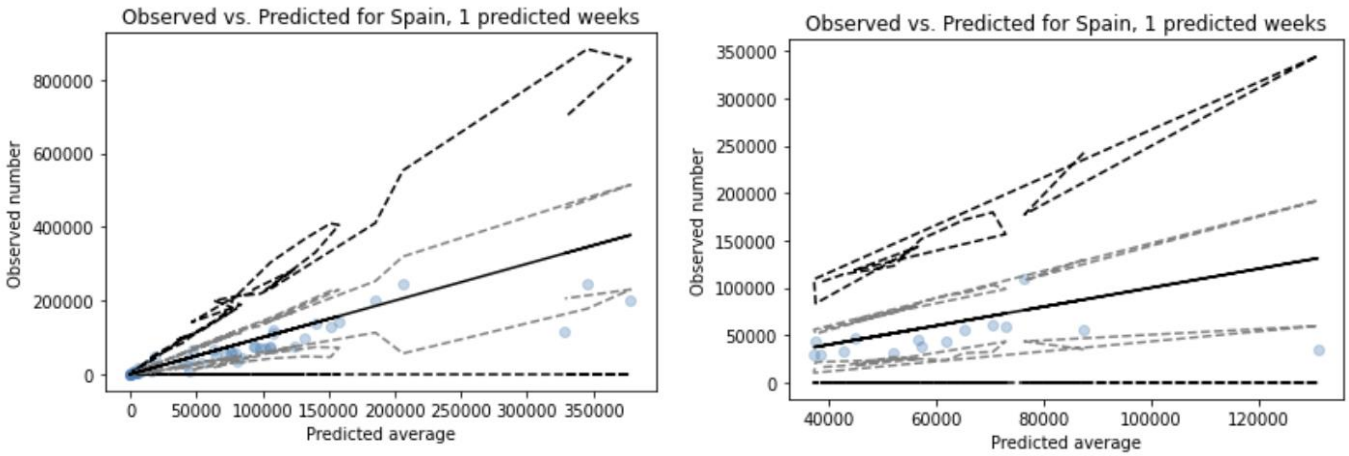


Figura 32 - CPDs de las predicciones sobre los datos de entrenamiento (izquierda) y validación (derecha) del modelo con mezcla de inferencia variacional con distribución logística, con horizonte temporal de predicción de 1 semana

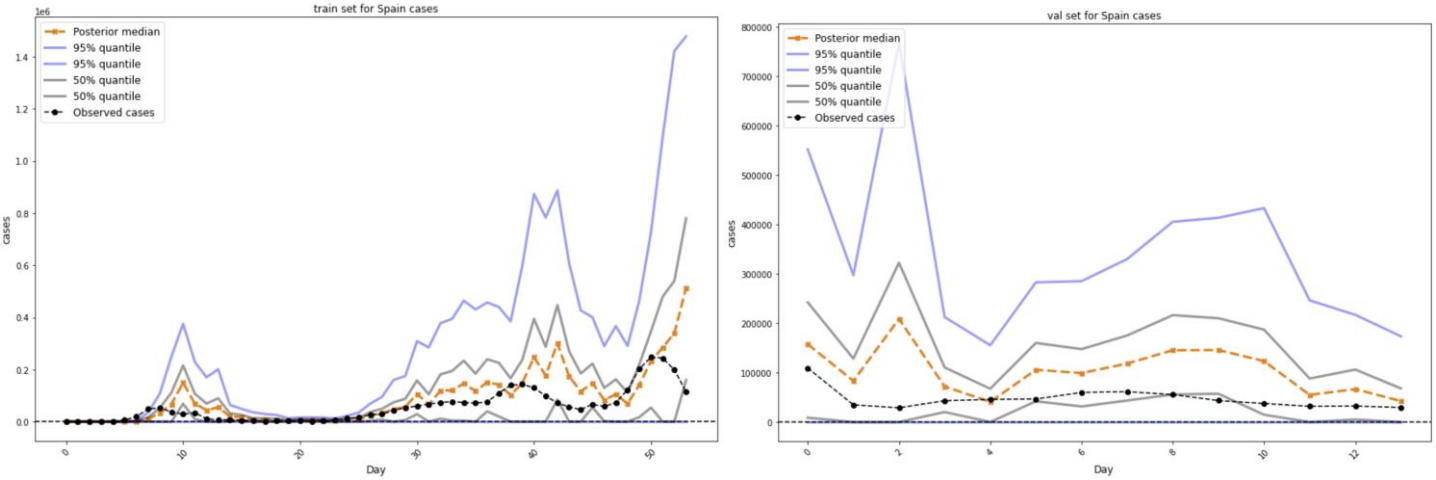


Figura 33 - Predicciones sobre los datos de entrenamiento (izquierda) y validación (derecha) del modelo de inferencia variacional con distribución logística, con horizonte temporal de predicción de 2 semanas

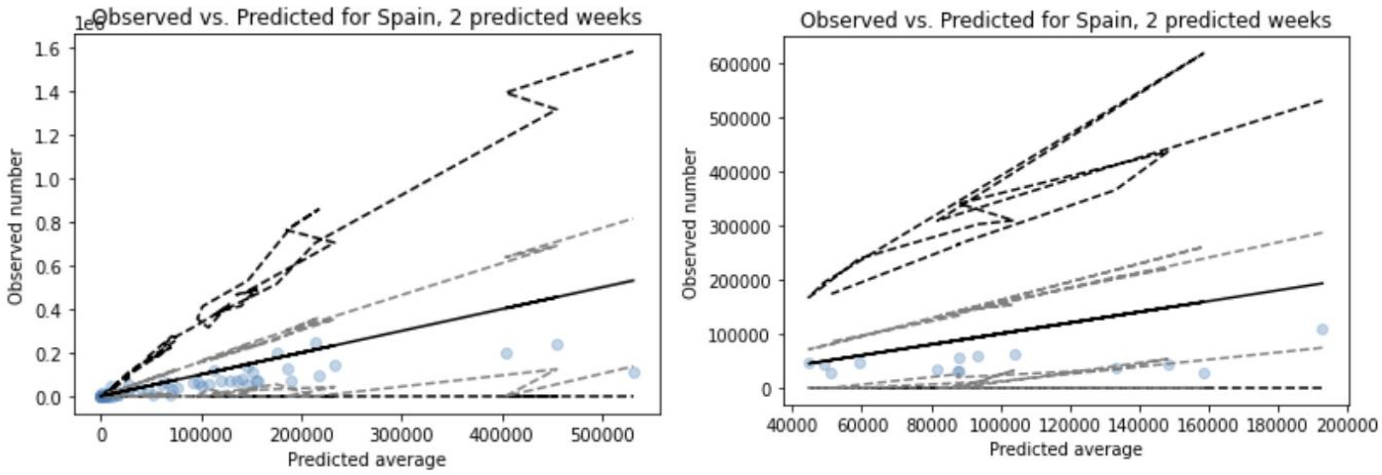


Figura 34 - CPDs de las predicciones sobre los datos de entrenamiento (izquierda) y validación (derecha) del modelo con mezcla de inferencia variacional con distribución logística, con horizonte temporal de predicción de 2 semanas

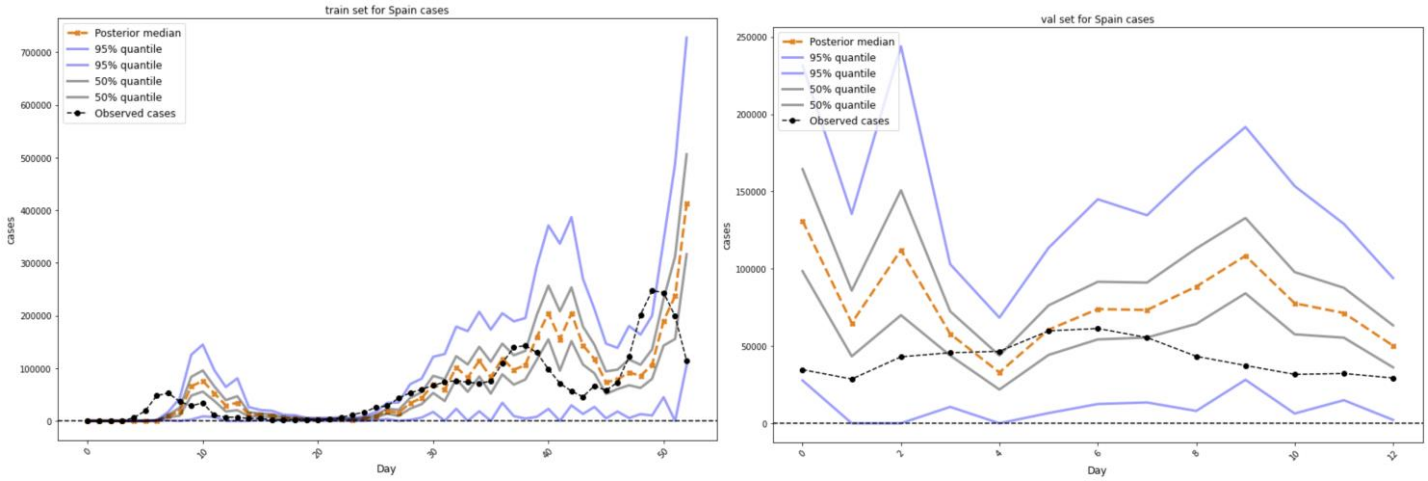


Figura 35 - Predicciones sobre los datos de entrenamiento (izquierda) y validación (derecha) del modelo de inferencia variacional con distribución logística, con horizonte temporal de predicción de 3 semanas

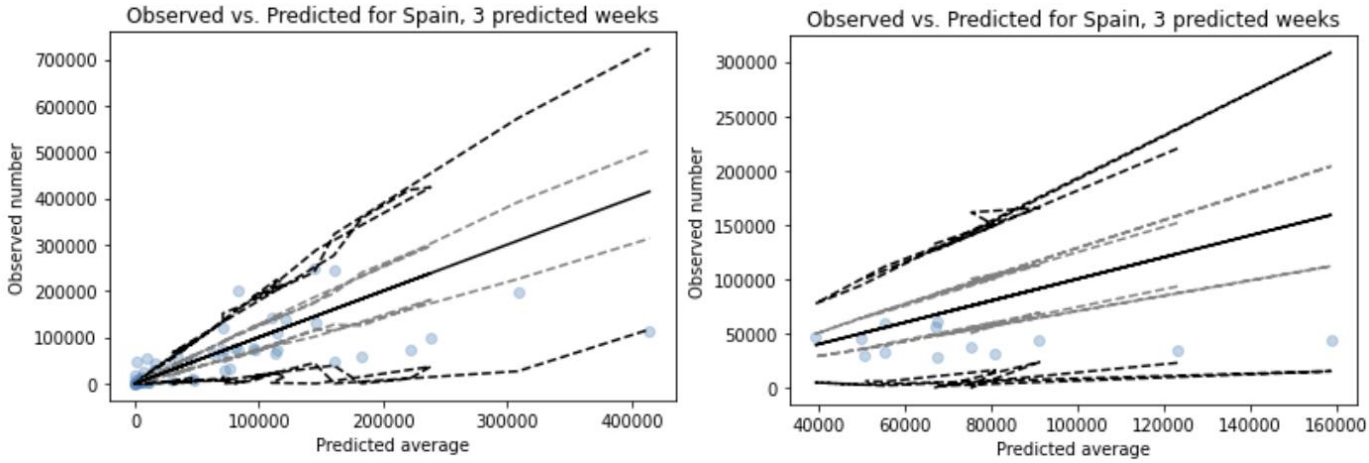


Figura 36 - CPDs de las predicciones sobre los datos de entrenamiento (izquierda) y validación (derecha) del modelo con mezcla de inferencia variacional con distribución logística, con horizonte temporal de predicción de 3 semanas

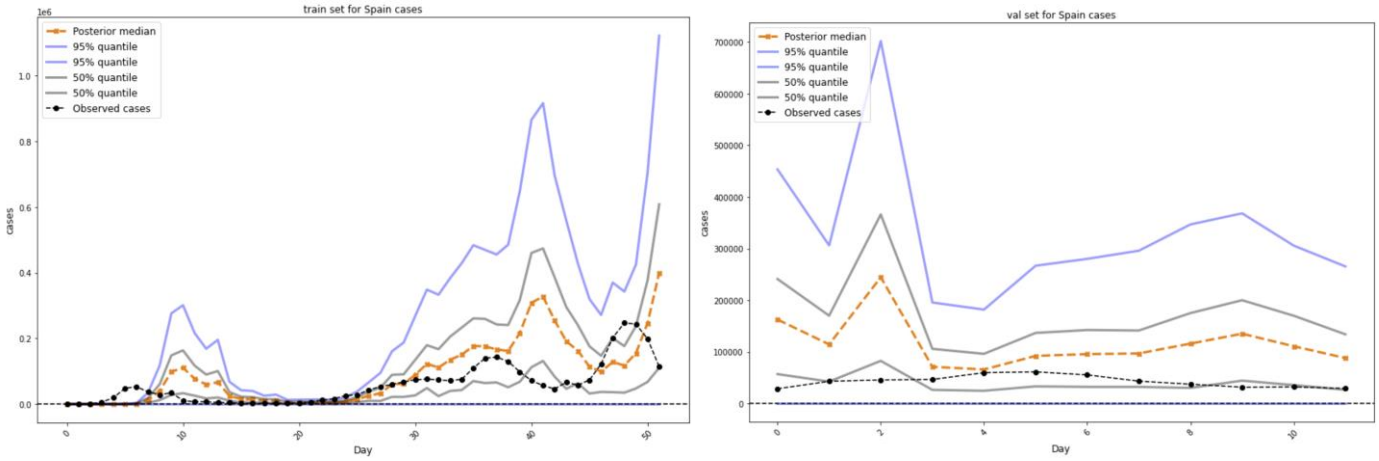


Figura 37 - Predicciones sobre los datos de entrenamiento (izquierda) y validación (derecha) del modelo de inferencia variacional con distribución logística, con horizonte temporal de predicción de 4 semanas

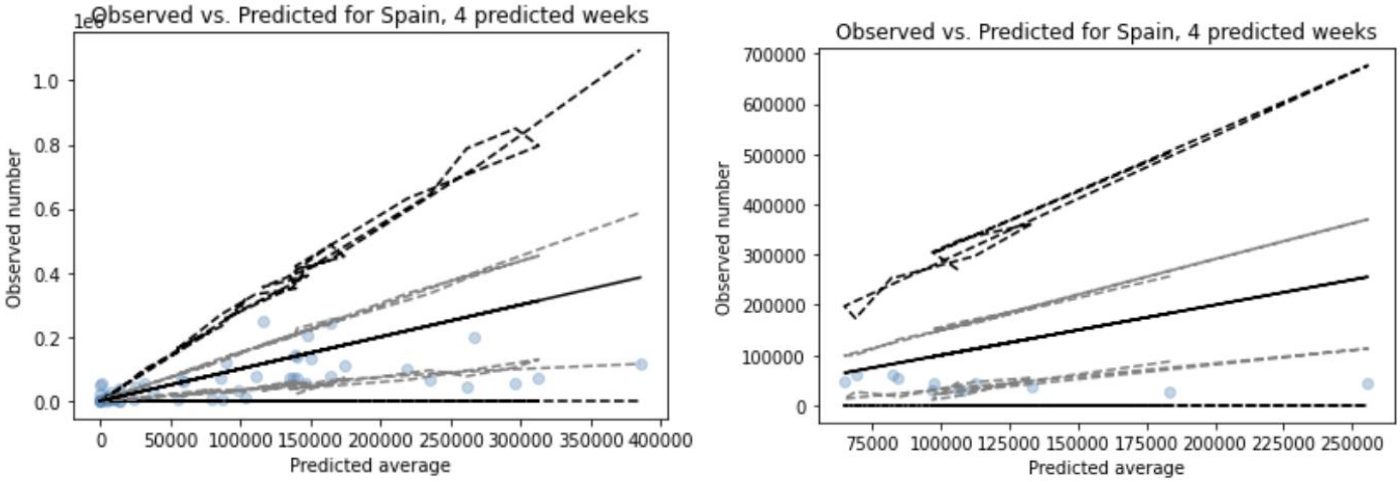


Figura 38 - CPDs de las predicciones sobre los datos de entrenamiento (izquierda) y validación (derecha) del modelo con mezcla de inferencia variacional con distribución logística, con horizonte temporal de predicción de 4 semanas

Por tanto, con la información observada hasta el momento, no es posible determinar si este modelo mejorará los resultados vistos en la mezcla de distribuciones logísticas.

4.4.2. Evaluación del Modelo

Tras evaluar el modelo con las métricas de evaluación descritas, se obtiene la siguiente tabla, que es una muestra, centrada en este modelo, de la tabla adjunta a este documento:

	model	target_variable	horizon	location	interval_score	sharpness	underprediction	overprediction	rel_wis	ae	rel_ae	cov_50	cov_95	bias	n_loc	n	location_name
	UNED-CovidPredPMA-V																
949	UNED-CovidPredPMA-VI	inc case	1	ES	37462,0	19726,0	0,00000	17736,0	4,71000	66078,0	5,36000	0,570000	1,00000	0,540000	1	7	Spain
964	UNED-CovidPredPMA-VI	inc case	4	ES	117988	44735,0	0,00000	73253,0	5,82000	210751	6,29000	0,250000	1,00000	0,650000	1	4	Spain
959	UNED-CovidPredPMA-VI	inc case	3	ES	117303	18342,0	0,00000	98961,0	6,51000	178981	6,11000	0,00000	0,600000	0,930000	1	5	Spain
954	UNED-CovidPredPMA-VI	inc case	2	ES	90542,0	53344,0	0,00000	37198,0	6,64000	146498	6,86000	0,670000	1,00000	0,470000	1	6	Spain

En la tabla anterior se observa lo siguiente:

- La puntuación del WIS relativo (rel_wis) no mejora la puntuación del modelo de mezcla de distribuciones logísticas en ninguno de los casos. Esto es debido a unos valores más altos de predicción excesiva que aquel modelo.

- Tal y como se ha observado en los datos de entrenamiento y validación, la cobertura del 50% (cov_50) es superior a la del modelo anterior, aunque la cobertura del 95% (cov_95) se mantiene igual que aquel, a excepción del horizonte temporal de 3 semanas, donde se ha incrementado notablemente (de 0% a 60%) el número de predicciones observadas en este cuantil.
- Sigue existiendo un sesgo hacia la predicción excesiva en el modelo, en este caso incluso superior al modelo anterior, dato que se puede observar en las columnas “overprediction” y “bias”.
- El error relativo del modelo también empeora los resultados vistos en el modelo de mezcla de distribuciones logísticas.

Si se comprara el resultado de la evaluación, tomando como referencia el WIS relativo, con los resultados del resto de modelos, se observa que este modelo tiene una puntuación superior a la de la regresión lineal y la distribución de Poisson en todos los casos, pero obtiene una puntuación peor que el modelo de mezcla de distribuciones logísticas en todos los casos. Con respecto al resto de modelos del Hub que realizan predicciones para España con los horizontes temporales de 1, 2, 3 y 4 semanas, este modelo continúa posicionándose entre los que peor puntuación obtiene en todos los horizontes temporales.

Por todo lo anterior, para mejorar los resultados de las predicciones, el siguiente modelo debería mejorar la capacidad de predicción del modelo de mezcla de distribuciones logísticas, pero mantener (o mejorar) la captura de la incertidumbre observada en el modelo de inferencia variacional.

4.5. Modelo de Inferencia Variacional con Mezcla de Distribuciones Logísticas Discretizadas

4.5.1. Entrenamiento del Modelo

Este modelo replica la arquitectura definida en el modelo de mezcla de distribuciones logísticas, con una capa densa de 15 neuronas, solo que, en este caso, dado que se implementa la inferencia variacional, se sustituye la capa densa tradicional por una capa del tipo “DenseFlipout”, tal y como se hizo en el modelo de inferencia variacional anterior. La salida de la capa densa, tal y como ocurría en el modelo de mezcla de distribuciones logísticas, alimenta una función que mezcla las distribuciones logísticas y discretiza la salida de la mezcla. Este cambio en la arquitectura produce que el modelo tenga 60 parámetros a entrenar, justo el doble que el modelo de mezcla de distribuciones logísticas, dado que en este caso se están sustituyendo los valores de los pesos por distribuciones, lo que dobla el número de parámetros.

Analizando las gráficas de la función de pérdida, se observa que la puntuación del NLL es peor que la del modelo de mezcla de distribuciones logísticas en todos los casos, aunque la puntuación se mantiene estable entre todos los horizontes temporales, algo que no se ha observado en ninguno de los modelos estudiados anteriormente.

También se observa que el resultado del NLL de validación es mejor que el resultado del NLL del entrenamiento, lo que implica que no existe overfit de los datos y que el modelo tiene, por tanto, capacidad predictiva sobre datos nuevos. Sin embargo, este dato no se ve respaldado por los resultados del NLL en el conjunto de test, ya que se observa un incremento en dicha puntuación.

Por tanto, a la vista de estos resultados de entrenamiento y validación, no se pueden apreciar indicios de que el modelo vaya a proporcionar mejores predicciones que el modelo de mezcla de distribuciones logísticas o que el modelo de inferencia variacional.

La siguiente figura ilustra las observaciones desarrolladas en este apartado:

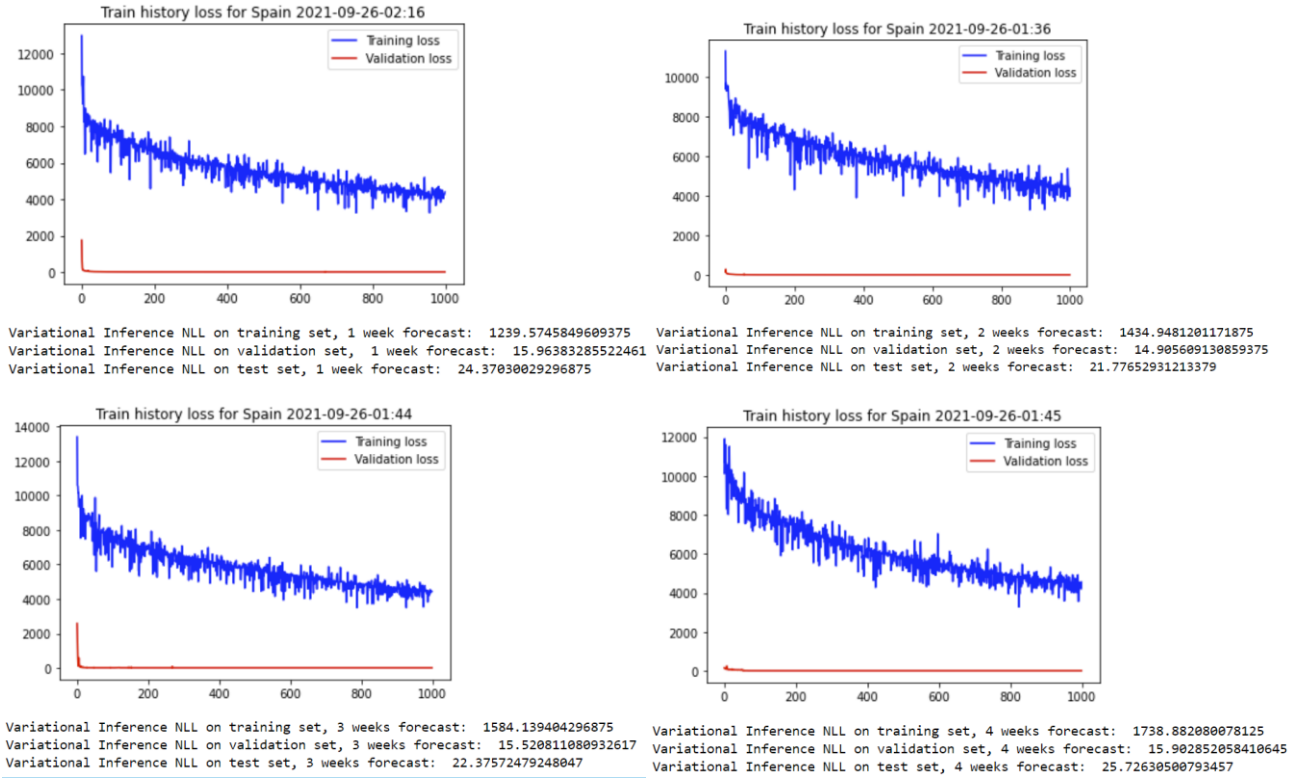


Figura 39 - Resultado del entrenamiento del modelo de inferencia variacional con mezcla de distribuciones logísticas discretizadas, con predicciones de 1 a 4 semanas (de derecha a izquierda y de arriba abajo), con detalle del resultado del NLL en los 3 conjuntos de datos (entrenamiento, validación y test).

Si atendemos a las gráficas de las predicciones frente a las observaciones, y a los CPD de dichas predicciones, se observa que el modelo hace un buen trabajo a la hora de realizar predicciones, aunque tiende a la predicción insuficiente, y captura correctamente la incertidumbre de las predicciones, con intervalos de predicción más estrechos cuando los valores de entrada son conocidos, y más anchos cuando son desconocidos. Este hecho se observa en todos los horizontes de predicción.

Además, se observa que el sistema mejora los modelos anteriores a la hora de definir un límite inferior más preciso en los cuantiles 50% y 95%, y, de hecho, las predicciones que no se observan dentro de ambos cuantiles se observan, en la mayoría de los casos, por encima de los límites superiores, consecuencia de que el modelo tiende a la predicción insuficiente.

En cuanto al límite superior de los cuantiles 50% y 95%, estos siguen alcanzando, especialmente en los horizontes temporales más altos, valores extremos, lo que penalizará la puntuación obtenida por el modelo en las métricas de evaluación.

Las siguientes figuras ilustran las observaciones comentadas:

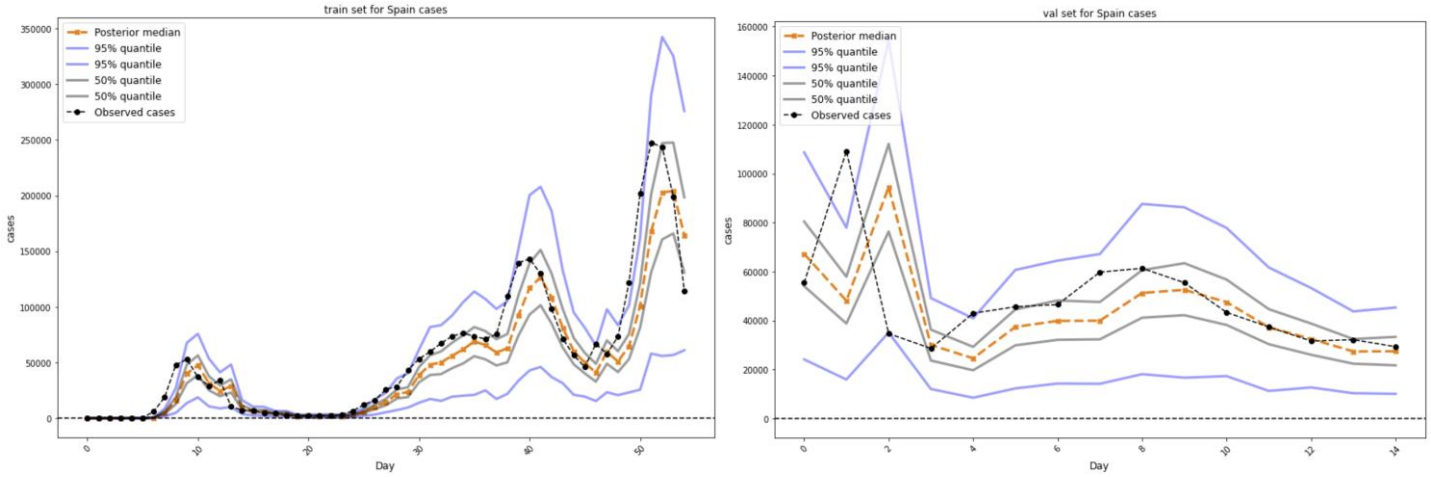


Figura 40 - Predicciones sobre los datos de entrenamiento (izquierda) y validación (derecha) del modelo de inferencia variacional con mezcla de distribuciones logísticas, con horizonte temporal de predicción de 1 semana

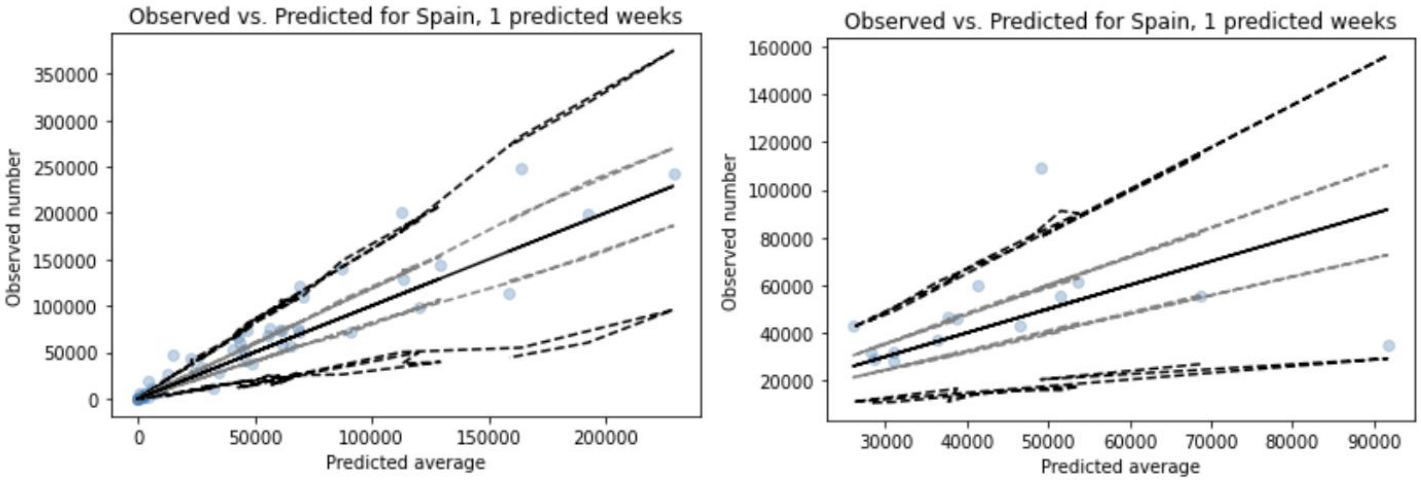


Figura 41 - CPDs de las predicciones sobre los datos de entrenamiento (izquierda) y validación (derecha) del modelo con mezcla de inferencia variacional con mezcla de distribuciones logísticas, con horizonte temporal de predicción de 1 semana

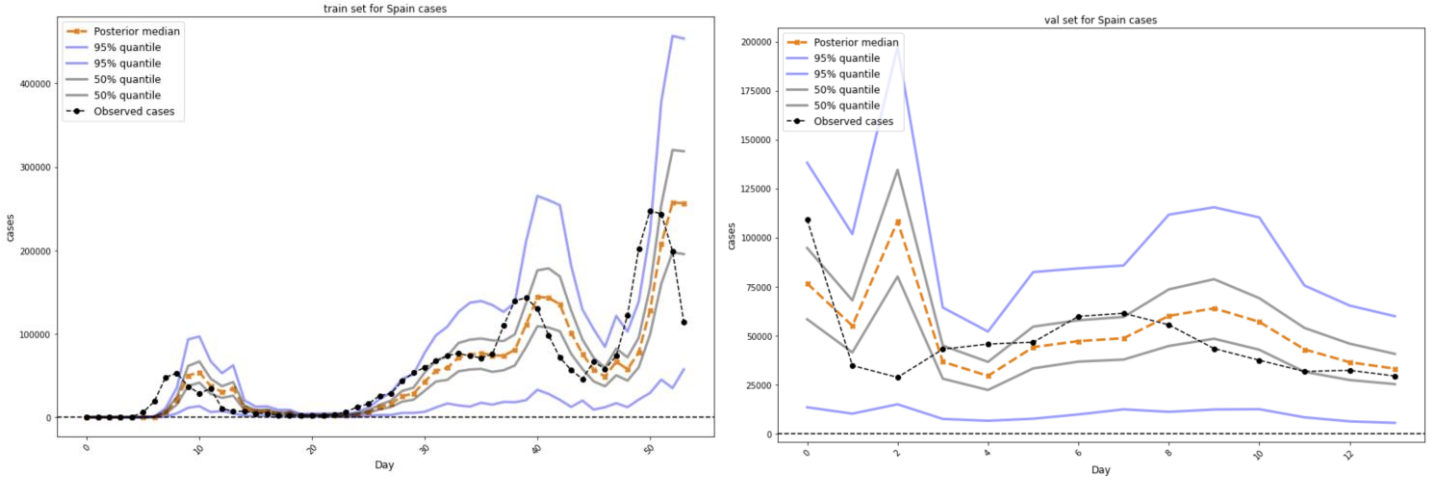


Figura 42 - Predicciones sobre los datos de entrenamiento (izquierda) y validación (derecha) del modelo de inferencia variacional con mezcla de distribuciones logísticas, con horizonte temporal de predicción de 2 semanas

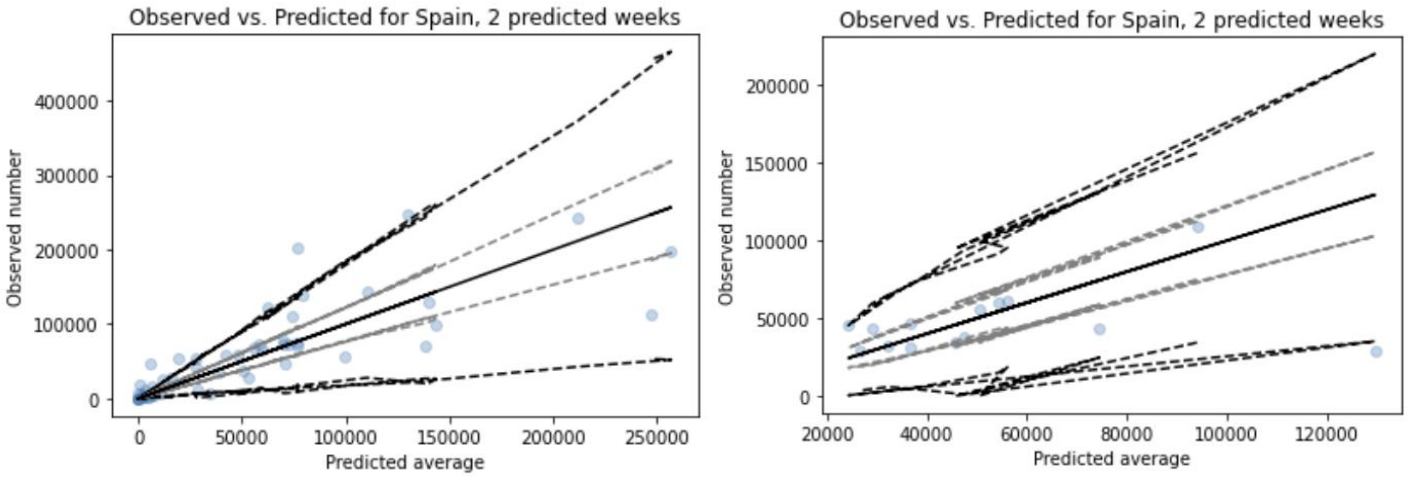


Figura 43 - CPDs de las predicciones sobre los datos de entrenamiento (izquierda) y validación (derecha) del modelo con mezcla de inferencia variacional con mezcla de distribuciones logísticas, con horizonte temporal de predicción de 2 semanas

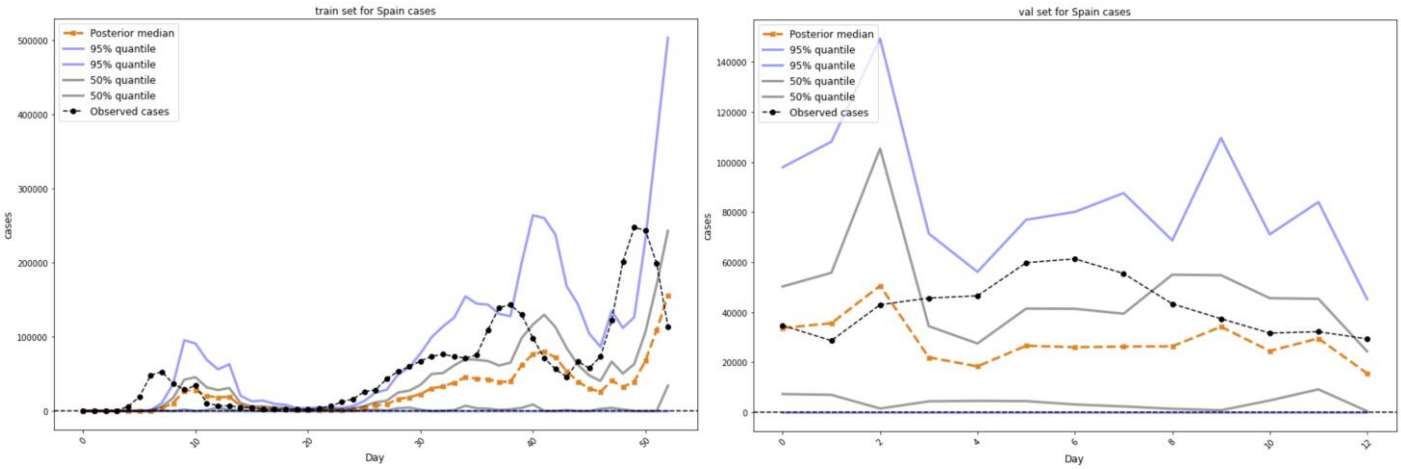


Figura 44 - Predicciones sobre los datos de entrenamiento (izquierda) y validación (derecha) del modelo de inferencia variacional con mezcla de distribuciones logísticas, con horizonte temporal de predicción de 3 semanas

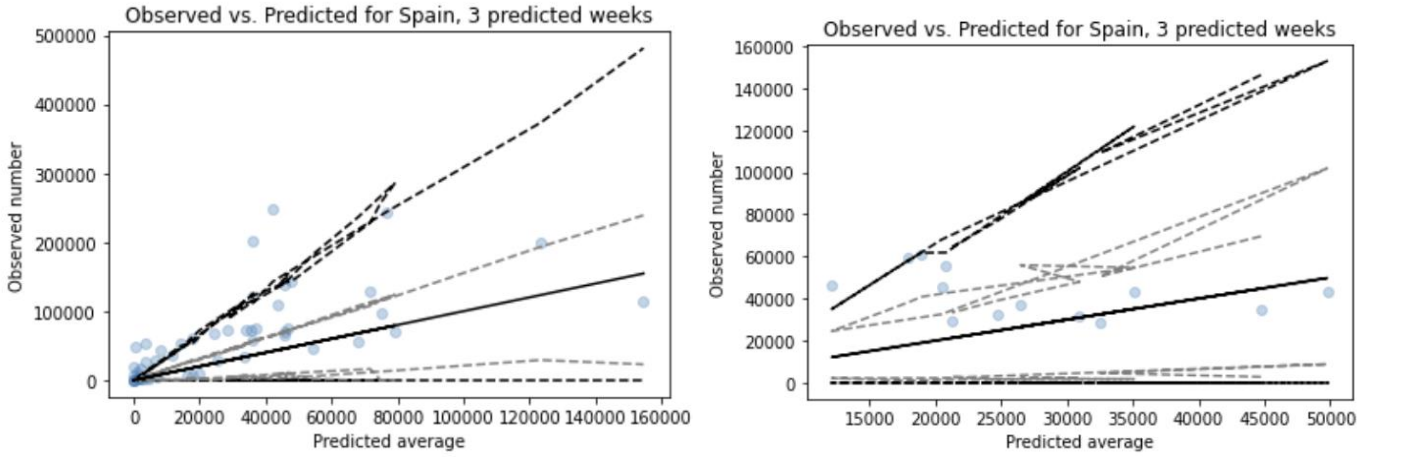


Figura 45 - CPDs de las predicciones sobre los datos de entrenamiento (izquierda) y validación (derecha) del modelo con mezcla de inferencia variacional con mezcla de distribuciones logísticas, con horizonte temporal de predicción de 3 semanas

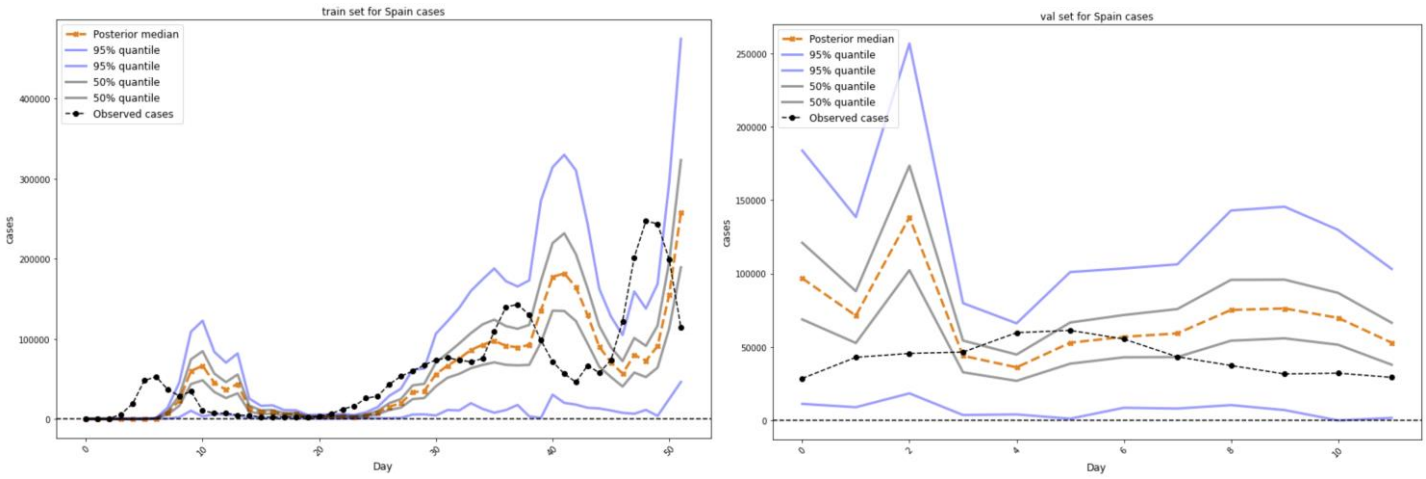


Figura 46 - Predicciones sobre los datos de entrenamiento (izquierda) y validación (derecha) del modelo de inferencia variacional con mezcla de distribuciones logísticas, con horizonte temporal de predicción de 4 semanas

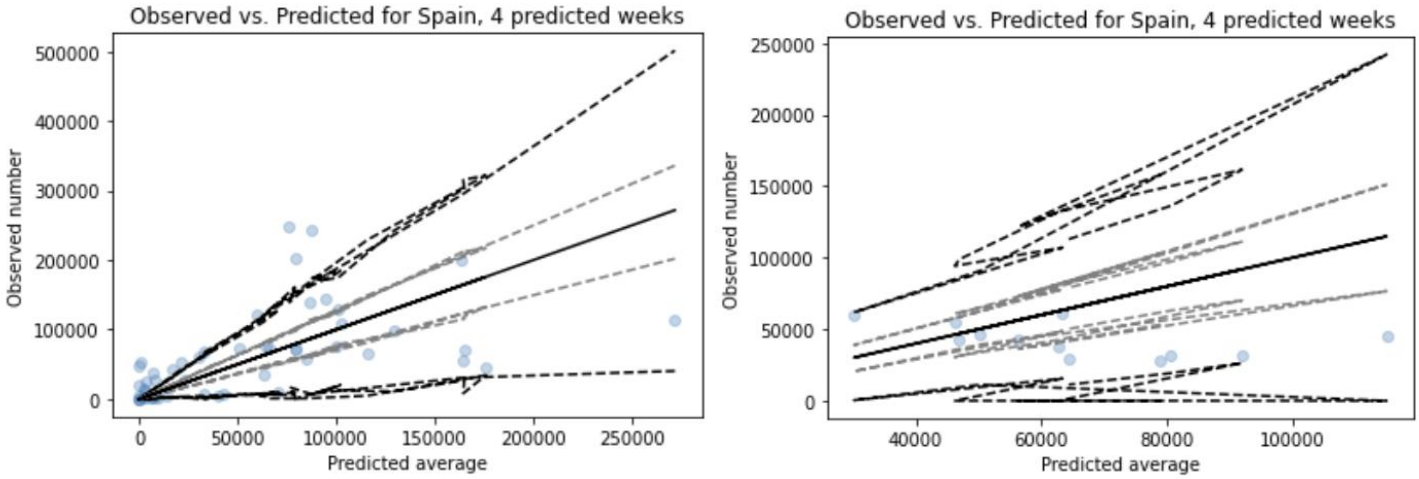


Figura 47 - CPDs de las predicciones sobre los datos de entrenamiento (izquierda) y validación (derecha) del modelo con mezcla de inferencia variacional con mezcla de distribuciones logísticas, con horizonte temporal de predicción de 4 semanas

Por todo lo anterior, es razonable esperar una cierta mejora en la puntuación de las métricas de evaluación en este modelo, con respecto a los observados anteriormente.

4.5.2. Evaluación del Modelo

Tras la evaluación del modelo utilizando las métricas de evaluación, se obtiene la tabla mostrada en la siguiente figura, que es un extracto de la tabla con todos los resultados, adjunta a este documento:

	model	target_variable	horizon	location	interval_score	sharpness	underprediction	overprediction	rel_wis	ae	rel_ae	cov_50	cov_95	bias	n_loc	n	location_name
	UNED-CovidPredPMA-VI-																
960	UNED-CovidPredPMA-VI-mix	inc case	3	ES	24765,0	17750,0	1645,00	5370,00	1,37000	35755,0	1,22000	1,00000	1,00000	0,120000	1	5	Spain
965	UNED-CovidPredPMA-VI-mix	inc case	4	ES	70270,0	12748,0	0,00000	57522,0	3,47000	108150	3,23000	0,250000	0,750000	0,780000	1	4	Spain
950	UNED-CovidPredPMA-VI-mix	inc case	1	ES	27829,0	15198,0	0,00000	12631,0	3,50000	45692,0	3,70000	0,430000	1,00000	0,560000	1	7	Spain
955	UNED-CovidPredPMA-VI-mix	inc case	2	ES	55659,0	10744,0	0,00000	44915,0	4,08000	87628,0	4,10000	0,00000	1,00000	0,900000	1	6	Spain

Analizando la tabla anterior, se observa lo siguiente:

- La puntuación del WIS relativo mejora notablemente las puntuaciones vistas en los modelos anteriores en todos los horizontes temporales, especialmente en el horizonte temporal de 3 semanas, donde la puntuación se acerca a 1, lo que significa que el modelo realiza predicciones

con una precisión similar a la media del resto de modelos del Hub, aunque todavía peor que dicha media.

- Al contrario de lo observado en los datos de entrenamiento y validación, el modelo tiende a la predicción excesiva en todos los casos, aunque aquí, los valores de predicción excesiva son más bajos que en el resto de los modelos, lo que contribuye a la mejora de la puntuación del WIS relativo. Esto se puede observar en las columnas “overprediction” y “bias”.
- La cobertura de 50% (cov_50) y de 95% (cov_95) mejora la de los modelos anteriores, aunque el presente modelo continúa, por lo general, produciendo intervalos de predicción demasiado anchos en el cuantil 95% y demasiado estrechos en el cuantil 50%, lo que a su vez penaliza la puntuación del WIS relativo.
- El error absoluto relativo también mejora lo observado anteriormente en otros modelos.

Al comparar los resultados de la evaluación, tomando como referencia el WIS relativo, del presente modelo con aquellas de los demás modelos del estudio, este modelo presenta una mejora sustancial del rendimiento con respecto el resto de los modelos para todos los horizontes temporales. Al compararlo con el resto de los modelos del Hub, este sigue estando entre los peores de la comparativa para todos los horizontes temporales. Sin embargo, el modelo con predicción de horizonte temporal de 3 semanas supera en rendimiento a algunos de los modelos que aportan predicciones al Hub, lo que supone un avance notable con respecto al resto de modelos.

Por todo lo anterior, se considera que este modelo supone un salto cualitativo considerable con respecto al resto de modelos del estudio, y pone de manifiesto la capacidad predictiva de los modelos probabilísticos y de inferencia Bayesiana teniendo en cuenta el escaso tamaño del conjunto de datos con el que se ha trabajado.

5. Conclusiones

Tras realizar el estudio de los diferentes modelos propuestos, se observa que los modelos más complejos obtienen resultados considerablemente más fiables, con una progresión clara en el rendimiento mostrado por los modelos estudiados, y compensando así el esfuerzo de implementación que supone el desarrollo de modelados más complejos. En cualquier caso, dado el alcance del estudio, los resultados de los modelos estudiados todavía se encuentran lejos de los resultados obtenidos por modelos que incorporen conceptos de epidemiología en sus predicciones.

En cualquier caso, los resultados obtenidos son notables y muestran la potencia de los modelos de aprendizaje automático probabilísticos y de inferencia bayesiana incluso cuando el conjunto de datos es pequeño y no se complementa con datos adicionales extraídos de otras fuentes, aunque, evidentemente, esto mejoraría los resultados.

Este estudio marca una línea clara de la tendencia que deben seguir los modelos de estudios posteriores aplicados a la predicción de casos de Covid-19 en un futuro estudio.

6. Áreas de Mejora y Estudio

De cara a mejorar los resultados del estudio propuesto en este documento, existen varias áreas a las que apunta el presente estudio en algunos casos, pero se han quedado fuera del alcance de este, y cuyo impacto debería analizarse convenientemente:

- El análisis de los mismos modelos estudiados con un conjunto de datos más amplio, o mejorado con datos adicionales, tales como los datos de población total, fechas de aplicación de las medidas no farmacéuticas por el gobierno, datos de hospitalizaciones o mortalidad, entre otros datos específicos de epidemiología aplicados a la pandemia de Covid-19.
- Análisis de la aplicación de los modelos estudiados para los datos de otros países, así como de su aplicación en la predicción de datos de hospitalizaciones y mortalidad.
- Estudio del impacto en los resultados de la aplicación de distribuciones más complejas en los modelos.
- Análisis del impacto en la capacidad predictiva de los diferentes modelos cuando se utilizan ventanas de tiempo (valores) como input para la predicción, en lugar de valores individuales.
- Estudio del impacto de la aplicación de otros métodos de aproximación de inferencia Bayesiana, o de mejoras sobre el mismo método de aproximación, tales como la implementación de dropout en las capas de la red neuronal.

7. Bibliografía

[1]: Background: https://covid19forecasthub.eu/background.html

[2]: Modeling COVID-19 spread in Europe and the effect of interventions: https://www.tensorflow.org/probability/examples/Estimating_COVID_19_in_11_European_countries

[3] Aaron van den Oord et al. Parallel WaveNet: Fast High-Fidelity Speech Synthesis. arXiv preprint arXiv:1711.10433, 2017

[4]: Targets and horizons: https://github.com/epiforecasts/covid19-forecast-hub-europe/wiki/Targets-and-horizons

[5]: Beate Sick, Oliver Duer : Probabilistic Deep Learning, capítulo 4 “Building loss functions with the likelihood approach”.

[6]: Regression with Probabilistic Layers in TensorFlow Probability: https://blog.tensorflow.org/2019/03/regression-with-probabilistic-layers-in.html

[7]: Beate Sick, Oliver Duer : Probabilistic Deep Learning, capítulo 5 “Probabilistic deep learning models with Tensorflow Probability”

[8]: Beate Sick, Oliver Duer : Probabilistic Deep Learning, capítulo 6 “Probabilistic deep learning models in the wild”

[9]: Beate Sick, Oliver Duer : Probabilistic Deep Learning, capítulo 8 “Bayesian Neural Networks”

[10]: Probabilistic Programming and Bayesian Methods for Hackers Chapter 1: https://github.com/CamDavidsonPilon/Probabilistic-Programming-and-Bayesian-Methods-for-Hackers/blob/master/Chapter1_Introduction/Ch1_Introduction_TFP.ipynb

[11]: Evaluating epidemic forecasts in an interval format: https://journals.plos.org/ploscompbiol/article?id=10.1371/journal.pcbi.1008618

[12]: Cramer et al.: Evaluation of individual and ensemble probabilistic forecasts of COVID-19 mortality in the US

[13]: Funk et al. (2019): Assessing the performance of real-time epidemic forecasts: A case study of Ebola in the Western Area region of Sierra Leone, 2014-15

[14]: Reports: https://covid19forecasthub.eu/reports.html