Universidad de Buenos Aires
Facultad de Ciencias Económicas
Escuela de Estudios de Posgrado

CARRERA DE ESPECIALIZACIÓN EN MÉTODOS CUANTITATIVOS PARA LA GESTIÓN Y ANÁLISIS DE DATOS EN ORGANIZACIONES

TRABAJO FINAL INTEGRADOR

Modelado de datos meteorológicos para la prevención de inundaciones en la Cuenca Matanza Riachuelo

AUTOR: MANUEL FLOCCO

MENTORA: VERÓNICA GARCÍA FRONTI

DICIEMBRE 2020

Contenido

RESUMEN

La cuenca Matanza Riachuelo es el curso fluvial más urbanizado, densamente poblado e industrializado del país. Sus aguas presentan riesgo de inundación y altos niveles de contaminación. La degradación del ambiente y el riesgo de desborde inciden en la condición de vida de los millones de habitantes de la cuenca. Más aún afecta, a aquellas personas que en condición de vulnerabilidad elevada viven en asentamientos informales emplazados en zonas inundables. Las causas de las crecidas encuentran origen tanto en factores medioambientales como sociales y económicos. Se trata de un problema complejo de gestión, solo parcialmente descomponible, cuyo origen puede rastrearse hasta la fundación de la Ciudad de Buenos Aires. Su solución requiere de un enfoque multidisciplinario a fin de cerrar la deuda social y ecológica vigente.

Para fortalecer el sistema de alerta temprana resulta útil pronosticar la altura del río a partir de los datos abiertos que provee ACUMAR. En el contexto de este trabajo se analiza la aplicabilidad de un modelo basado en redes neuronales artificiales para la predicción de series temporales que permitan modelar el nivel del río. El modelo que aquí se presenta permite pronosticar su altura con una semana de antelación.

PALABRAS CLAVE

Autoridad Cuenca Matanza - Riachuelo, Inundaciones, Aprendizaje automático supervisado, Redes neuronales artificiales LSTM, Series de tiempo multidimensionales.

INTRODUCCIÓN

Un reciente informe del Banco Mundial **(**2016) destaca que el 60 por ciento de los desastres naturales en Argentina se debe a inundaciones. Uno de los cursos fluviales con riesgo de desborde es el río Matanza Riachuelo. Su cuenca abarca aproximadamente 2047 kilómetros cuadrados en el noreste de la provincia de Buenos Aires. Según datos de la Autoridad Cuenca Matanza Riachuelo (ACUMAR, 2020) sus lindes son habitados por casi seis millones de personas. Esta cifra representa el quince por ciento de la población total del país. A su vez, es una de las vías acuáticas más contaminadas del mundo. El organismo señala que el proceso histórico de asentamiento urbano condujo a la explotación excesiva de los recursos naturales, la contaminación y ocupación de áreas vulnerables a inundaciones. El impacto de los desbordes y la contaminación afecta desproporcionalmente a las personas de menores recursos, sobre todo aquellas que por necesidad debieron instalarse en sectores naturalmente inundables. Los fenómenos que generan el desborde de las aguas se vinculan con múltiples causas. Por un lado, la crecida de los ríos que nutren el cauce principal. Otras causas encuentran origen en fenómenos meteorológicos tales como las sudestadas y las lluvias intensas. A estos eventos naturales, se les suma uno más asociado con el aumento ocasional de las napas freáticas. Además, puede distinguirse un último elemento que contribuye al desborde al considerar la combinación de dos o más de estos fenómenos.

Para gestionar la problemática de las inundaciones, restaurar el medioambiente y contribuir a la mejora de la calidad de vida de los habitantes, en el año 2006 se crea ACUMAR. Su origen es producto de una demanda colectiva iniciada en año 2004, conocida como *Causa Mendoza* por el apellido de la mujer que encabeza la demanda contra el Estando Nacional, los gobiernos de la Provincia de Buenos Aires y de la Ciudad Autónoma de Buenos Aires, así como 44 empresas privadas alojadas en el polo industrial de Dock Sud. La demanda colectiva es la cristalización de un problema cuyo origen es anterior y que puede rastrearse hasta la fundación misma de la Ciudad de Buenos Aires. Este problema se relaciona directamente con el tratamiento y gestión de los deshechos que la ciudad genera. No sólo se trata de una deuda social y ecológica, sino también de una deuda política. Durante los últimos cincuenta años distintos gobiernos de turno, democráticos y de facto, han prometido la implementación de un plan de saneamiento que mejore la calidad de vida de los millones de habitantes de la Cuenca. Finalmente para el año 2008, la Corte Suprema de Justicia de la Nación obliga a ACUMAR a diseñar y ejecutar un Plan Integral de Saneamiento

Ambiental. Este plan define tres objetivos fundamentales: Mejorar la calidad de vida de los habitantes de la Cuenca, recomponer el ambiente y prevenir daños con suficiente y razonable grado de predicción. La propuesta del organismo es integral, no sólo busca reparar el daño ambiental y prevenir eventuales crecidas, sino que tiene a su cargo el control de la contaminación ambiental, el monitoreo de la calidad del aire y agua, la tramitación de sitios contaminados, la concientización ambiental y la restauración ecológica (ACUMAR, 2020).

Desde la década de 1980, los métodos de pronóstico climático han mejorado en precisión y asertividad. Los avances tecnológicos, junto con el constante monitoreo y relevamiento de información atmosférica, han posibilitado crear modelos de pronósticos climáticos de alta fiabilidad y precisión. Estos modelos resultan de gran utilidad a la hora de prevenir catástrofes naturales. No obstante, no es posible conocer exactamente el estado de la atmósfera en todo momento del tiempo. Esta incertidumbre degrada la potencia del pronóstico en el tiempo, al depender de las condiciones iniciales (Alley, Emanuel & Zhang, 2019). A fin de sortear esta dificultad, se han implementado con éxito múltiples métodos de aprendizaje automático para predecir fenómenos meteorológicos[1]. En Maqsood, Khan y Abraham (2003) se comparan distintos modelos de redes neuronales artificiales (RNA) para regresar variables climáticas. Estas redes consisten en modelos computacionales inspirados en el comportamiento de las neuronas biológicas. A partir de una capa de entrada de datos se definen una serie de ponderadores aleatorios que transmiten señales a otras capas de redes. La transferencia se realiza al superar un umbral de activación. La precisión del resultado surge a partir de minimizar una función de error. El enfoque que propone Maqsood et al. busca determinar qué modelo de redes neuronales artificiales provee el mejor ajuste, a la vez que generaliza mejor frente a las variaciones provocadas por las distintas estaciones del año.[2] Para justificar la elección de este método los autores argumentan: "Las redes neuronales artificiales proveen una metodología para resolver muchos problemas de tipo no lineal ya que ellas tienen la capacidad de extraer relaciones a partir de un proceso sistemático de entradas y salidas" (Maqsood et al., 2003, p.34). Su implementación, por lo tanto, permite aproximar y

[1] Algunos casos de éxito en la implementación de modelos de aprendizaje automático para regresar variables climáticas pueden consultarse en: Culclasure (2013), Maqsood, Khan y Abraham (2003) Otros casos de éxito referidos específicamente a modelar niveles de agua fluvial a partir de datos meteorológicos pueden encontrarse en Mogrovejo (2018), Wica, Witkowski, Szumiec & Ziebura (2019) y en Paul & Das (2014).

[2] Muchos de los enfoques de pronósticos climáticos utilizan una combinación de técnicas empíricas para elaborar modelos. Para profundizar en estas técnicas puede consultarse Sarle(1994) Chollet (2018) y Wica et al. (2019).

resolver problemas complejos, allí donde las técnicas tradicionales de resolución presentan dificultades [3].

La aplicación de redes neuronales artificiales para pronosticar la altura de ríos a partir de datos meteorológicos ha sido estudiada también en Argentina. El trabajo de La Red Martínez y Crespo Fidalgo (2013) busca reconstruir las crecidas del río Paraná a partir de datos meteorológicos y datos del nivel del río. Aplicando una función de penalización a la red neuronal, los investigadores pudieron obtener resultados confiables respecto de su altura. Ellos justifican su trabajo como un método de prevención y alerta frente a las importantes consecuencias económicas y sociales que tuvieron las inundaciones en la provincia de Corrientes.

> Las inundaciones de las últimas décadas provocaron pérdidas millonarias; desde el año 1982 hasta el 2000 las inundaciones y vendavales en Corrientes provocaron pérdidas por un valor de 70 millones de dólares. A los desastres naturales se agregan situaciones de vulnerabilidad generadas por precaria infraestructura vial, falta de acceso a servicios básicos y cobertura social para grandes sectores de la población. (La Red Martínez y Crespo Fidalgo, 2013, p. 15)

Algo más de un millón de personas están actualmente expuestas a riesgos de inundación en 32 ciudades de todo el país. Bajo la problemática del cambio climático global, en los últimos sesenta años se ha registrado en Argentina un aumento en las precipitaciones de un 20 por ciento anual aproximadamente (Barros y Camilloni, 2016). Según definiciones del Ministerio de Salud de la Nación, las catástrofes asociadas a inundaciones provocan un numero inesperado de muertes, lesiones, aumento de enfermedades que exceden la capacidad de atención sanitaria instalada, así como el aumento potencial de enfermedades posteriores. Por otra parte, las inundaciones producen pérdidas materiales no solo en el momento en el que ocurren, sino que a su vez condicionan el desarrollo futuro. Por todas estas razones, la pregunta central que motiva esta investigación busca conocer si es posible pronosticar con precisión el nivel del río Matanza Riachuelo a partir de datos meteorológicos. Una respuesta afirmativa a esta pregunta resulta útil a la hora de considerar las

[3] Las técnicas tradicionales consisten en los métodos basados en el cálculo diferencial multivariado. Este enfoque busca precisar matemáticamente la relación subyacente que gobierna determinado fenómeno. La diferencia con el enfoque basado en aprendizaje automático radica en que da por supuesta dicha relación, en el sentido de que ésta se encuentra implícita en los datos. Para mayor profundidad en estos conceptos puede consultarse Chollet (20018) y también Zurada (1992) .

posibles pérdidas humanas y materiales, el surgimiento y expansión de enfermedades asociadas al agua, eventuales pérdidas de puestos de trabajo, entre otras. Para ello, se buscará obtener resultados precisos a fin de prevenir y anticipar una crecida con la antelación necesaria para gestionar mediadas que permitan atenuar el impacto. Se sostiene como hipótesis de trabajo que es posible aplicar un modelo basado en redes neuronales artificiales para la predicción de series temporales que permitan modelar el nivel la Cuenca Matanza Riachuelo a partir de datos meteorológicos y de la altura fluvial.

El presente informe contiene tres capítulos. El primero de ellos presenta la problemática de la Cuenca Matanza Riachuelo desde una perspectiva integral. Primeramente, se elabora un recorrido histórico que permita comprender la gestión de la Cuenca desde la fundación de Buenos Aires. Este recorrido histórico permitirá comprender la situación actual desde una perspectiva de largo alcance, en particular la vigencia de la deuda social y ecológica que dio origen a ACUMAR. Posteriormente, se describen los lineamientos institucionales de este ente autónomo, su misión y objetivos institucionales. También se repasa el actual Proyecto Integral de saneamiento Ambiental y los modos en los que la institución toma decisiones para asegurar el alcance de los objetivos de dicho proyecto. A continuación, se indaga en las características geográficas y políticas de la cuenca. Dada su extensión y múltiples paisajes, este recorrido permite comprender la dificultad en materia de gestión. En el último apartado de este primer capítulo se introduce la problemática de producción de datos y su gestión en el marco de Gobierno Abierto. Aquí, se describen qué datos se generan y con qué fines, a la vez que se reconstruye el proceso de recolección, procesamiento y producción de conocimiento a partir de ellos. El segundo capítulo centra el enfoque en la descripción metodológica. En el primer apartado se detalla formalmente el modelo de redes neuronales artificiales implementado y el mecanismo de *retro propagación*. Luego, se presentan las fuentes de información secundaria utilizada para la generación de modelos. El último momento de este capítulo presenta el análisis multivariado de datos y los preprocesamientos aplicados antes de ingresar al modelo. Finalmente en el capítulo tercero se indica cuáles son los resultados alcanzados. Un último apartado recupera los pasos necesarios para una eventual implementación del modelo elaborado en la organización.

CAPÍTULO I: DESCRIPCIÓN DE LA CUENCA MATANZA-RIACHUELO

1. Brevísima reseña histórica de la Cuenca Matanza Riachuelo.

El tándem hombre - naturaleza ha sido objeto de estudio de intelectuales, científicos de múltiples disciplinas y filósofos a lo largo de la historia de la humanidad. El interés en distintas épocas históricas indica que la pregunta encuentra respuestas parciales, siempre circunscriptas al contexto social vigente. En este sentido, las respuestas posibles tienen lugar en la sociedad en la que surgen. No son los individuos aislados los que hacen usos legítimos o ilegítimos de la naturaleza, sino que son las sociedades y sus modos de obrar los que habilitan o prohíben estos usos. Así, aquello que se identifique como un problema medioambiental, se enlaza con el uso social del sustrato natural. Cada sociedad establece vínculos con los recursos naturales. Su mediación es la tecnología, pero esta relación da cuenta de una profunda concepción de proyecto económico, social, cultural y de régimen político.

La historia argentina puede analizarse a partir de fases o etapas de desarrollo económico y social. Cada una de estas etapas implica un proyecto de país y un uso legítimo de los recursos naturales. Brailovsky y Foguelman (2009) reconocen cinco fases de desarrollo, delimitados a partir de una serie de sucesos políticos. Cada fase circunscribe una forma particular de producir la vida social, es decir de crear comunidad y de relacionarse con la naturaleza. En esta reseña histórica, se presentan dichas fases y se explora el uso del río Matanza Riachuelo en cada una de ellas. Este paso resulta necesario para comprender el contexto histórico de la Cuenca y la importancia que ha tenido en el desarrollo económico y social argentino. Actualmente sus lindes conforman una de las zonas más urbanizadas, densamente pobladas y altamente industrializadas del país. Esta perspectiva de largo alcance contribuye a entender cuál es la deuda social y ecológica vigente.

La primera fase de desarrollo económico y social refiere a la etapa colonial, la cual abarca desde la fundación de Buenos Aires hasta 1816. No fue casualidad que tanto Pedro de Mendoza como Juan de Garay fundaran la ciudad a orillas del Riachuelo. Los conquistadores españoles buscaban un puerto natural donde resguardar sus navíos de madera de las tormentas. El Riachuelo, por sus características geográficas, constituía un punto singular de asentamiento. En sus orillas se conjuga un puerto natural con una barranca elevada donde asentar a la población. Las Leyes de

Indias prohibían explícitamente construir edificaciones en terrenos inundables[4]. Sin embargo en la historia oficial de la ciudad se afirma que "Todo el valle del Matanza (por el río), en general, era – y sigue siéndolo en su parte superior – inundable y anegadizo" (Zabala y Gandia, 1980). El problema del asentamiento urbano y las inundaciones fue tratado en múltiples ocasiones en el Cabildo. Para el año 1772, la presencia de pestes y ratas impulsa a la ciudad a llevar adelante reformas entre las cuales se señala la necesidad de construir desagües pluviales, desalojar las viviendas ubicadas en áreas inundables, entre otras. Sin embargo, las prohibiciones para construir en las zonas del bajo inundable tendrán lugar recién en el primer gobierno Peronista. Desde la fundación de Buenos Aires, el Riachuelo se convirtió en la cloaca a cielo abierto de la cuidad. Las mismas Leyes de las Indias establecían que las industrias que producen contaminación debían instalarse aguas abajo de las ciudades. De este modo, el agua que consumía la población no se contaminaba. Con esta misma lógica los saladeros, curtiembres, lavaderos de lana y mataderos se ubicaron en estas orillas.

El grito de independencia en 1816 sienta las bases de una nueva sociedad y con ella un nuevo relacionamiento con los recursos naturales. Durante estos años, el recurso ganadero comienza a explotarse de manera intensiva, dando lugar a múltiples industrias. Algunas crónicas de la época[5] señalan que el Riachuelo ya presentaba un estado deplorable. Para 1800 la ciudad tenía quince saladeros que vertían la sangre de la faena al Riachuelo, junto con otros deshechos que tarde o temprano llegaban al río. Para calmar el malestar en la población, los gobiernos de Rivadavia y Martín Rodríguez expulsan de la ciudad a los depósitos de cueros, y la fundiciones de velas. En 1830 se prohíbe arrojar los desperdicios de los saladeros al Riachuelo. Sin embargo, dado que estas industrias conformaban el centro de la producción económica, las prohibiciones y decretos alcanzaron escasa efectividad y la actividad continúa.

[4] En el caso de la fundación de buenos aires parecen no haberse cumplido estas prohibiciones. Brailovsky y Foguelman señalan que "Buenos Aires fue insalubre casi desde el principio. Callejones, callejuelas y plazuelas, huecos y aceras, perduraron hasta fines del siglo XVIII y aún más, en un estado de absoluto abandono, invadidos por las aguas y lodazales (…) y sus vecinos fuertemente diezmados por las pestes durante los días más fuertes del estío" (Brailovsky y Foguelman, 2009, pp 72)

[5] La actividad de la industria de los saladeros llevó a Guillermo Hudson a llamar a Buenos Aires "la ciudad más pestilente del mundo". Martinez Estrada también señala la contaminación: "El pobre Riachuelo arrastra sus seculares de las curtiembres y los saladeros, lavándose constantemente en su misma suciedad." (Brailovsky y Foguelman, 2009, pp 137)

Alrededor de la década de 1860 y hasta 1930, Argentina ingresa en el proceso de inserción en la división internacional del trabajo. En esta fase, la creciente demanda de materias primas por parte del impero británico, impulsa al país a inscribirse en el mercado mundial como productor de carnes, cereales y lanas. La creciente demanda intensificó el uso del Riachuelo como vertedero de desperdicios industriales. A comienzos de 1871 la ciudad atravesó una peste de fiebre amarilla. Los olores que emanaban de sus aguas condujeron a los habitantes de la cuidad a afirmar que la razón de la peste era el río. Brailovsky y Foguelman recuperan una editorial del diario La Nación de febrero de aquel año:

> El lecho del Riachuelo es una inmensa capa de materias en putrefacción. Su corriente no tiene ni el color del agua. Unas veces sangrienta, otras veces verde y espesa, parece un torrente de pus que escapa a raudales de la herida abierta en el seno gangrenado de la tierra. Un foco tal de infección puede ser causa de todos los flagelos, el cólera y la fiebre. ¿Hasta cuándo inspiraremos el aliento y beberemos la podredumbre de ese gran cadáver tendido a espaldas de nuestra ciudad? (2009, pp. 220)

Posteriormente se llegó a la conclusión de que la epidemia de fiebre amarilla no se debía a la pestilencia. Sin embargo, existía un vínculo entre la enfermedad y el río. Las inundaciones provocadas por las lluvias generaban charcos costeros donde proliferaron los mosquitos, el agente transmisor de esta fiebre. Cuantiosas inundaciones se registraron en este período[6], todas ellas con grandes consecuencias económicas.

La etapa de sustitución de importaciones (1930-1976) aceleró vertiginosamente el proceso de urbanización e industrialización obligando a utilizar todos los espacios disponibles, incluso aquellos pasibles de inundarse. Entre estos años se produce un proceso de migraciones internas que conformarán el Gran Buenos Aires. El fenómeno de la contaminación continúa pero cambian las empresas contaminantes. Crecen las curtiembres, fábricas de pinturas, procesadoras de pulpa de papel, derivados del caucho y el petróleo, fertilizantes, plaguicidas. En líneas generales, disminuyen los residuos orgánicos y aumentan los residuos químicos, metales pesados y derivados del petróleo. A la contaminación del agua se le suma la contaminación del aire. Durante estos años

[6] Por ejemplo las crecidas de 1877, 1884 y 1889 todas con pérdidas de animales y evacuaciones. Las inundaciones de 1910 de los barrios de La Boca, Palermo, Belgrano y Núñez. Testimonios de época señalan que el desborde de los arroyos obligó a que las calles debieran transitarse caballo o en canoa. Véase Brailovsky y Foguelman (2009)

comienzan a instalarse equipos de medición de la altura del río en distintos puntos sobre todo en la zona alta de la cuenca[7], con el objetivo de prevenir desbordes.

La última fase comienza en 1976 y continúa hasta nuestros días. Para estos años, la Cuenca Matanza Riachuelo ya presentaba altos niveles de contaminación del agua y del aire. Durante los años 1970 algunas de las fábricas e industrias comenzaron a trasladarse al interior. Al vaciamiento le sobrevino la holgura de los espacios vacantes, síntoma de una industria marginal, desregulada y precaria. Durante estos años, gobiernos de facto y electos por el voto popular, impulsaron proyectos de saneamiento. Ninguno de ellos se ejecutó. Ante el fracaso de estos intentos, la conciencia pública no calla y continúa demandando una solución a múltiples organismos de gestión y gobierno. En el año 2004, un grupo de vecinos habitantes de la Cuenca presentó una demanda judicial colectiva contra los daños y perjuicios ocasionados por inundaciones, la contaminación y deficiencias sanitarias. Esta acción judicial se inició contra el Estado nacional, la Provincia de Buenos Aires, la Ciudad Autónoma de Buenos Aires, las catorce jurisdicciones provinciales que abarca la Cuenca, y las 44 empresas localizadas en el sector del Polo Petroquímico de Dock Sud. En la demanda colectiva, se exigió la recomposición del ambiente, saneamiento de las aguas, control de los efluentes vertidos en el río y otros elementos tendientes a reparar y evitar el deterioro medioambiental. En el año 2006, el Congreso de la Nación sanciona la ley nacional 26168 dando origen a la Autoridad de Cuenca Matanza Riachuelo (ACUMAR). El juicio, conocido popularmente como "Causa Mendoza", obtuvo sentencia por parte de la Corte Suprema de Justicia de la Nación (CSJN) en el año 2008. La sentencia colectiva contiene una condena general que recayó sobre el Estado nacional, la Provincia y la Ciudad de Buenos Aires, cuyo objeto sentó los lineamientos, objetivos y obligaciones del ente interjurisdiccional ACUMAR.

2. ACUMAR Lineamientos institucionales.

El fallo de la CSJN del año 2008 intimó a ACUMAR a efectuar un plan de saneamiento como respuesta a la demanda colectiva presentada cuatro años antes. Su ley de creación dota al ente de

[7] La cuenca Matanza Riachuelo se encuentra dividida en tres zonas denominadas Cuenca Alta, Media y Baja. La cuenca Alta hace referencia a los municipios de Marcos Paz, Merlo, San Vicente y Cañuelas. Posteriormente en el presente informe se analizan los límites políticos de cada una de estas divisiones.

facultades de regulación, control y fomento de actividades industriales, servicios públicos y toda otra actividad que tenga incidencia ambiental en todo el espacio de la Cuenca, pudiendo intervenir en materia de prevención, saneamiento, recomposición y utilización racional de los recursos naturales[8]. ACUMAR es un ente autónomo autárquico e interjurisdiccional, compuesto por ocho integrantes, cuya presidencia es encabezada por el titular de la Secretaría de Ambiente y Desarrollo Sustentable. Su actividad es coordinada a través del Consejo Municipal, con los Municipios de la Provincia de Buenos Aires que integran la Cuenca y el gobierno de la Ciudad Autónoma de Buenos Aires. Además de la coordinación de las actividades con los municipios, a fin de alcanzar los objetivos organizacionales, el Consejo Municipal asiste, asesora y coopera con ACUMAR en materia de ejecución presupuestaria, legal y técnica. Bajo los lineamientos de Gobierno Abierto[9] y los Objetivos de Desarrollo Sostenible de Naciones Unidas[10], ACUMAR incorpora el concepto de Visión Compartida, cuyo objeto consiste en generar un espacio de construcción común y colaborativo entre gobernantes, administraciones y la sociedad civil, a fin de producir una visión común de la Cuenca Matanza Riachuelo. Entre estas organizaciones se incluyen instituciones del ámbito académico, colegios profesionales de distintas disciplinas, Defensorías del Pueblo, Organizaciones No Gubernamentales, el Cuerpo Colegiado de ACUMAR y ciudadanos en general. Bajo la forma de Comisiones de Participación Social, se presentan diferentes instancias de canalización del diálogo y de gestión tales como audiencias públicas y mesas de seguimiento.

En connivencia con la ley que da origen al ente, la sentencia de la CSJN indica que plan de saneamiento debe tener como objetivo la mejora de la calidad de vida de los habitantes de la Cuenca, la recuperación del ambiente en lo que respecta al aire, agua y tierra, así como prevención daños con suficiente antelación y grado de predicción. A partir del año 2009, se implementa un Plan Integral de Saneamiento Ambiental (PISA) cuyas definiciones y lineamientos fundan el plan

[8] Artículo quinto de la ley nacional 26168

[9] En el año 2012 la CEPAL integra conceptualmente las bases de un Gobierno Abierto y los beneficios en términos de calidad de gestión, eficiencia de las políticas públicas y de transparencia. Argentina se integra a este proceso de apertura de las instituciones estatales y organismos públicos en el año 2013. Esto permitió la gestión de políticas públicas con base en evidencia, así como favorecer el proceso de rendición de cuentas de manera transparente. Actualmente se encuentra vigente el Cuarto Plan de Acción Nacional de Gobierno Abierto 2019-2021. Al respecto puede consultarse CEPAL (2012) y Secretaria de Modernización (2019)

[10] En el año 2015 los países integrantes de Naciones Unidas aceptaron una agenda de desarrollo sostenible. Dicha agenda presenta diecisiete objetivos globales con miras a erradicar la pobreza, proteger el planeta y asegurar la prosperidad. Véase Naciones Unidas (2018*), La Agenda 2030 y los Objetivos de Desarrollo Sostenible: una oportunidad para América Latina y el Caribe.*

de acción de ACUMAR en materia de control y gestión. Durante el año 2016, las directrices de este plan integral fueron actualizadas, con el objeto de crear nuevas acciones y verificar el cumplimento de los objetivos. El PISA afirma que:

> Toda la cuenca se articula sobre la base de tres componentes interrelacionados: físicos (agua, suelo, aire), biológicos (flora, fauna) y antropológicos (socioeconómicos, culturales, institucionales, normativos). Si uno o algunos de ellos resultan afectados por un factor de carácter interno o externo, el equilibrio de su constitución original se pierde y el conjunto del sistema primario es puesto en riesgo. Un modelo integral de cuenca debe ser capaz de hacer coexistir de forma coordinada y armónica todas las dimensiones involucradas en la realidad de su geografía, su biología, su sociología, su política, su economía y su administración de gobierno. Aun siendo independientes, con el fin de lograr la prevención de la contaminación y la recomposición de la cuenca, todas sus dimensiones deben poder ser evaluadas, monitoreadas y en caso de ser preciso, redireccionadas, bajo una visión compartida, de modalidad de conjunto, de compromisos, de transparencia, de eficiencia y de innovación. (ACUMAR, 2016, pp. 31)

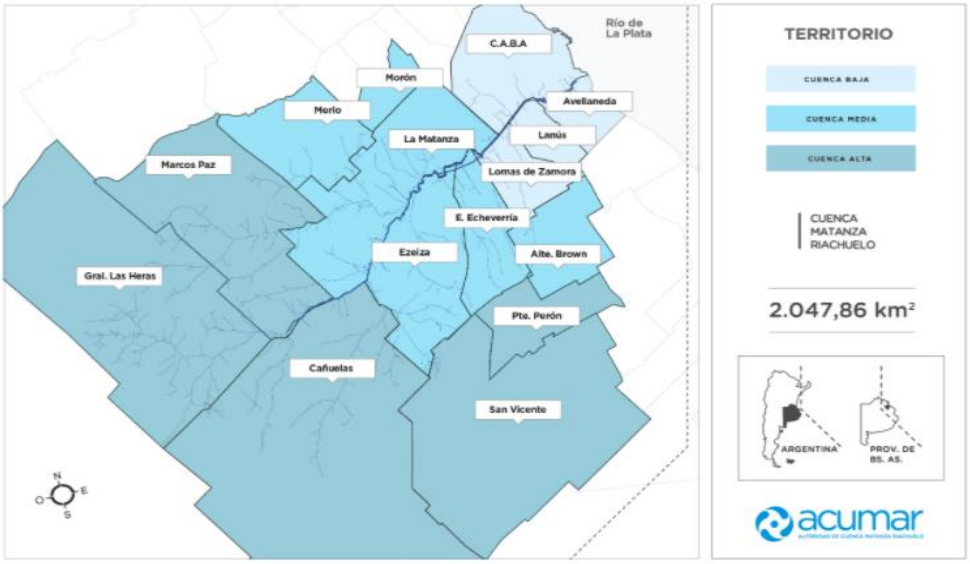
El enfoque institucional define un programa multidisciplinar e integral. Se trata de generar acciones y objetivos de corto, medio y largo alcance. El PISA reconoce que no es suficiente con proteger la biodiversidad de las doce reservas biológicas que integran la Cuenca, disminuir el daño y riesgo ambiental a partir del control y fiscalización o reubicar a las familias vulnerables asentadas en sectores con riesgo de inundación, sino que es necesaria una respuesta sistemática e integral que contribuya a *mejorar la calidad* de vida de los habitantes de la cuenca.

3. La Cuenca: características geográficas y límites políticos.

A modo de dimensionar la complejidad y variedad de situaciones, riesgos y problemas que la institución debe gestionar, se presenta en este apartado una descripción geográfica y política de la Cuenca Matanza Riachuelo. Desde un punto de vista territorial, la cuenca se ubica en el noreste de la Provincia de Buenos Aires. Sus aguas se vierten en el Rio de la Plata, en la Boca del Riachuelo. El tamaño de la cuenca se extiende por más de dos mil kilómetros cuadrados, cuyo emplazamiento abarca los municipios de Almirante Brown, Avellaneda, Cañuelas, Esteban Echeverría, Ezeiza,

General Las Heras, Lanús, La Matanza, Lomas de Zamora, Marcos Paz, Merlo, Morón, Presidente Perón, y San Vicente, en la Provincia de Buenos Aires. Además, atraviesa la totalidad de la comuna 8 en la Ciudad Autónoma de Buenos Aires y parcialmente las comunas 1, 3, 4, 5, 6, 7, 9, y 10 de la misma ciudad[11]. Debido a su extensión y a las características geográficas, políticas, económicas y sociales, la Cuenca divide su territorio en tres regiones: Cuenca Alta, Media y Baja. La primera presenta paisaje rural, principalmente ligado con actividades económicas relacionadas al sector agroindustrial. A medida que se avanza hacia la Ciudad Autónoma, el paisaje cambia de rural a

Figura 1: Mapa de la Cuenca Matanza Riachuelo



Fuente: ACUMAR

urbano con actividad industrial y de servicios. De esta manera, las actividades productivas que se desenvuelven actualmente en la Cuenca son amplias, siendo ésta la zona más industrializada del país[12].

Por su impacto ambiental, las industrias de mayor relevancia pertenecen al sector petroquímico, alimenticio, curtiembres, frigorífica y metalúrgico. A lo largo del tiempo, el tramo Bajo y Medio de la Cuenca sufrieron alteraciones físicas provocadas por la actividad humana, entre ellas la canalización y ensanche del Riachuelo. Producto de estas obras, el río dejó ver una serie de meandros, los cuales fueron ocupados por familias de escasos recursos económicos. Estas

[11] Resolución ACUMAR número 1113/13.
[12] Más de 13 mil industrias se encuentran presentes en la zona, principalmente emplazadas en los sectores Bajo y Medio de la Cuenca.

urbanizaciones espontáneas se realizaron sin planificación y en zonas naturalmente inundables del río. La Cuenca cuenta con muy baja pendiente a lo largo de toda su extensión. Tratándose de un río de llanura, su escurrimiento es lento lo cual dificulta la evacuación de agua durante las crecidas. Este fenómeno se potencia con las sudestadas[13] dado que cuando ellas ocurren, sube el nivel del Río de la Plata, ralentizando el escurrimiento. Por otra parte, desde la política pública históricamente se privilegió el drenaje pluvial como método para evitar las inundaciones. Este enfoque promueve la creación de micro desagües urbanos, típicamente cordones cuneta, canales y conductos cloacales entre otros, a fin de conducir el agua hacia su natural destino de evacuación. Sin embargo, no se han generado estructuras que permitan el almacenamiento temporal del agua a fin de contener la crecida y prevenir el desborde.

4. Gestión de datos en la Cuenca

Como parte de las directrices de Gobierno Abierto, ACUMAR dispone de una amplia variedad de datos para su consulta en línea. Según declara en el portal de acceso, la información es pública; 'toda persona puede acceder a los datos sin necesidad de aclarar para qué utilizará los datos o brindar otras explicaciones' (ACUMAR, 2020). De esta manera, ACUMAR autoriza explícitamente la difusión de los registros, documentos y datos producidos, a la vez que dispone una serie de repositorios de acceso público. El monitoreo y medición de parámetros ambientales es parte de la tarea llevada adelante por el ente a fin de alcanzar los objetivos delineados en el PISA. El organismo cuenta con estaciones de monitoreo ambiental, meteorológicas e hidrométricas a lo largo de toda la Cuenca. Estas estaciones recaban información de manera puntual y continua para su posterior modelado matemático. Los datos recopilados incluyen el relevamiento topo-batimétrico de toda la Cuenca, los cauces de los cuerpos de agua, incluyendo relieve, depresiones, niveles, velocidad de escurrimiento, monitoreo superficial y subterráneo de las subcuencas, características fisicoquímicas y microbiológicas de las aguas, series estadísticas de actividades industriales y volcamientos. A los fines de investigación de este informe, se

[13] La sudestada consiste en un fenómeno climático que tiene lugar cuando se produce el pasaje de un frente de aire frío o bien por efecto de dos sistemas de presión combinados, uno de alta y otro de baja presión. El fenómeno se caracteriza por la caída de intensas lluvias en cuestión de minutos. Su consecuencia directa es la elevación del Río de la Plata.

utilizaron datos obtenidos en las estaciones meteorológicas. Existen quince estaciones de este tipo en la Cuenca. Catorce de ellas se encuentran en los municipios de la Provincia de Buenos Aires y una en la Ciudad. Las estaciones de monitoreo recopilan y transmiten datos de forma automática respecto de múltiples variables tales como la temperatura ambiente, el porcentaje de humedad, la sensación térmica, el punto de rocío, la presión barométrica, la velocidad y dirección del viento, la intensidad y cantidad de lluvias, entre otras. Todos estos datos se reportan públicamente a partir de sus promedios diarios. Estas estaciones meteorológicas constituyen la base del funcionamiento del sistema de alerta temprana frente a eventuales catástrofes. El sistema permite gestionar acciones con antelación para mitigar los efectos negativos de cambios climáticos repentinos, como es el caso de las sudestadas[14]. En este sentido, la política de decisión que el organismo ejerce está basada en datos. Esta idea 'refiere a la práctica de tomar decisiones al analizar los datos antes que decidir en base a la intuición'[15] (Provost & Fawcett, 2013, pp. 53). Las estaciones de monitoreo constituyen un punto de recolección constante del estado del río y del medioambiente, los cuales deben ser transmitidos e integrados. LaValle, Lesser, Shocjley, Hopkins y Kruschwitz (2011) señalan que este es uno de los desafíos del Big Data[16]. El dato por sí solo no permite tomar decisiones, ellos deben ser capturados y almacenados para su posterior explotación. Es en este sentido, en el que la palabra misma encubre una neutralidad valorativa. El dato es lo que está *dado*, lo *incuestionable*, lo que *emerge*, lo *fáctico*, cuando el dato es siempre producido con un fin específico. Algo común a las definiciones de Big Data es la dimensión tecnológica. Lo que la tecnología permitió es generar y capturar más datos, pero estos no existen por fuera de su producción y de su interpretación. (Del Rio Riande, 2020, pp 3). En el caso de ACUMAR, el ente integra la información recabada en las estaciones meteorológicas en una Base de Datos

[14] Este fenómeno climático introduce una dificultad adicional a la hora de elaborar pronósticos. La tasa de incremento con la que aumenta el cauce del río cuando tiene lugar una sudestada provoca que las zonas vulnerables a una inundación se encuentren bajo agua en cuestión de minutos.

[15] Traducción propia del inglés.

[16] Boyd y Crawford (2012) señalan que el concepto de Big data es, en cierta medida, *vago*. Los autores definen Big data como un fenómeno cultural, tecnológico a la vez que erudito, en la interrelación de tres dimensiones: la tecnología que provee medios para analizar grandes conjuntos de datos, el análisis para identificar patrones en estos datos y la mitología, en el sentido de la creencia socialmente extendida de que en el volumen existe una forma superior de inteligencia y conocimiento que puede producir una visión disruptiva, imposible de alcanzar de otro modo, bajo el aura la verdad, objetividad y precisión. Numan y Di Domenico (2013) indican que la vaguedad del término quizás se deba a la velocidad en la que el uso del término se diseminó en la sociedad. Su definición también enlaza tres perspectivas: es una respuesta a los problemas tecnológicos relacionados el volumen, es un medio para producir valor comercial para las organizaciones, con amplio efecto social sobre todo en lo relacionado con los valores éticos y el uso comercial de dichos datos.

Hidrológicos[17] (BDH). El sistema de almacenamiento es de acceso público y permite obtener datos de series históricas sobre la dinámica y calidad de las aguas subterráneas y superficiales. La BDH también contiene otras series con datos meteorológicos en base a las variables presentadas anteriormente. Además, la base de datos comprende un repositorio con publicaciones científicas, estudios ambientales e imágenes satelitales. La BDH es una fuente secundaria de esta investigación. Se extrajeron de allí las series históricas que indican la altura del río. Esta información fue complementada con datos climáticos extraídos de manera automática de dos sitios distintos[18].

El siguiente capítulo tiene por objetivo presentar el enfoque metodológico de esta investigación. Primeramente se buscará presentar el concepto de Redes Neuronales Artificiales (RNA) e introducir las razones que fundamentan la utilización de este método para regresar variables climáticas. Luego se presentarán las fuentes de datos utilizados y las variables con las que se trabaja. Por último, se analizan los principales estadísticos descriptivos de las variables consideradas y las transformaciones aplicadas sobre las ellas.

CAPÍTULO II: METODOLOGÍA

1. ¿Qué es una Red Neuronal Artificial?

Una RNA consiste en un modelo computacional que se asemeja, con cierto grado de abstracción, a las neuronas biológicas (La Red Martínez y Crespo Fidalgo, 2013). La estructura básica de una red presenta capas y conexiones entre esas capas. La información se propaga desde las capas presinápticas hacia las capas sinápticas. La primer capa proporciona la entrada de datos. A partir de ella, se inician una serie de ponderadores aleatorios que indican la relevancia de cada conexión con respecto a todas las demás conexiones. Así, cada neurona está representada por una

17 El Sistema BDH permite el filtrado y búsqueda de datos hídricos, meteorológicos y de calidad de aguas. Cuenta con una interfaz gráfica en línea disponible en http://www.bdh.acumar.gov.ar/bdh3/index_contenido.php

18 Los datos meteorológicos fueron recuperados del portal *Power Data* (https://power.larc.nasa.gov) y del portal Wunderground (https://www.wunderground.com/). El detalle de estas fuentes y el procesamiento realizado puede consultarse en el apartado segundo del siguiente capítulo.

combinación lineal entre estos ponderadores aleatorios multiplicados por el respectivo valor de la capa de entrada. El resultado de esta sumatoria de estímulos es seguido por una función de activación no lineal[19] para determinar la salida de la neurona. El término *capa* indica el resultado de aplicar el producto punto entre la matriz de ponderadores y la matriz con los coeficientes de entrada. Como complemento, a esta operación matricial puede adicionársele un coeficiente denominado sesgo. Desde un punto de vista matemático, una RNA es un grafo orientado y puede ser comprendida como una composición de funciones, en donde el resultado de la cadena de cálculos de la capa anterior es transmitido a la siguiente capa.

Formalmente:

$$Y = f(\sum_{i=1}^{n} \omega_{\mathrm{i}} X_{\mathrm{i}})$$

Esta ecuación reconstruye el resultado de salida de cada neurona hacia una nueva capa, hasta alcanzar la capa final. La matriz X_{i} contiene los datos de entrada, ω_i representa la matriz de ponderadores y f es una función no lineal de activación. Existen distintos modos de determinar la efectividad de una RNA, según el problema con el que se trabaje. El objetivo radica en encontrar los ponderadores adecuados para alcanzar un representante fiel de los datos. Para controlar el resultado obtenido, puede medirse la distancia que hay en el resultado alcanzado respecto del valor esperado. Esto permite introducir el concepto de función de pérdida. Una métrica útil para calcular la función de pérdida y por lo tanto medir la efectividad de la red es la raíz del error medio cuadrático[20] o RMSE por sus siglas en inglés. En cada iteración, la selección de los ω_i pesos varía con el objetivo de hacer mínimo este error. Los pesos pueden ser positivos o negativos a fin de favorecer o inhibir factores que permitan alcanzar el resultado esperado. Chollet (2018) señala que el *truco* fundamental del aprendizaje automático consiste en utilizar este resultado como una señal de retroalimentación y control. Por lo tanto, el proceso de ajuste de los pesos se convierte en un problema de optimización, más precisamente en minimizar una función de pérdida. Como se

[19] Algunos ejemplos de estas funciones de activación no lineales son la tangente hiperbólica, la función sigmoidea o la función lineal rectificada, entre muchas otras. Esta última es la utilizada en el modelo presentado en este trabajo.

[20] La raíz del error medio cuadrático presenta la siguiente fórmula $\sqrt{\sum \frac{(\widehat{y_l}-y_l)}{N}}$. Donde $\widehat{y_l}$ representa el valor predicho, y_l el valor observado, y N la cantidad de registros a predecir.

señaló, inicialmente los ponderadores son asignados aleatoriamente. El resultado de esta primera iteración no necesariamente resulta útil a los fines de representar los datos en el sentido buscado. Sin embargo, esto constituye el punto inicial. El objetivo consiste en encontrar la combinación de pesos que retorne el menor valor para la función de pérdida, es decir encontrar los ω_i parámetros tal que el gradiente de $f(\omega_i)$ sea igual a cero. En la sintonización de los hiperparámetros del modelo es donde la heurística del investigador tiene lugar. Por ejemplo, el algoritmo incorpora un hiperparámetro denominado tasa de aprendizaje, el cual equivale al tamaño del paso en el descenso del gradiente en cada iteración. Una tasa de aprendizaje muy baja presenta el riesgo de que la red converja en un mínimo local, mientras que una tasa de aprendizaje muy grande presenta el riesgo de que se esquive sistemáticamente el mínimo.

En la práctica una RNA está compuesta por un vector multidimensional de datos, también llamado tensor[21], encadenados entre sí. Aplicando la regla de la cadena[22] para el cálculo del descenso del gradiente se obtiene el algoritmo de retropropagación, del inglés *backpropagation.* Su nombre representa gráficamente la idea: retropropagación implica propagar el error que se obtiene producto de la inicialización aleatoria de los ponderadores, hacia atrás (Wica et al., 2019). A partir de conocer el valor de la función de pérdida, se introduce dicho valor *hacia atrás,* es decir, desde las capas superiores de la red o salidas, hasta las capas de entrada. La regla de la cadena permite calcular con precisión la contribución de cada parámetro al valor final de la función de pérdida. Al conocer cada contribución es posible sintonizar con un criterio los pesos iniciados aleatoriamente.

El concepto de RNA surge a mediados del siglo XX. Reconstruir su historia desde el desarrollo conceptual y el lugar que este método ocupa en el campo de la inteligencia artificial[23] requiere

[21] Un tensor es una generalización del concepto de matriz. Un tensor se caracteriza por su rango y forma. Por rango se entiende la cantidad de ejes que presenta. Por forma, se indica la cantidad de dimensiones que el tensor presenta alrededor de cada eje. Por ejemplo, un escalar es un tensor de 0 dimensiones, mientras que un vector y una matriz son tensores de 1 y 2 dimensiones respectivamente.

[22] La regla de la cadena implica que la derivada de una función compuesta puede ser calculada apelando a la siguiente identidad $f(g(x)) = f'(g(x)) * g'(x)$.

[23] El campo de la inteligencia artificial abarca múltiples disciplinas y puede ser estudiado desde diversos enfoques. Uno de sus subcampos es el aprendizaje automático. Desde un punto de vista exclusivamente técnico, los métodos de aprendizaje automático implican la búsqueda de representaciones útiles a partir de una serie de datos iniciales, en el contexto de un espacio de posibilidades predefinido. Su guía u orientación se define a partir de una señal de retroalimentación. Aquí, *aprendizaje* implica un método sistemático de búsqueda automatizada a fin de alcanzar representaciones eficientes desde los datos, es decir que permitan alcanzar el resultado esperado. El aprendizaje

mucho más espacio que este apartado. El recorrido que aquí se propone no pretende ser exhaustivo sino presentar algunos de los principales conceptos en el desarrollo de esta técnica de aprendizaje automático, caro a múltiples disciplinas. Estas ideas tienen como objetivo facilitar la comprensión de la relevancia metodológica del enfoque presentado para la resolución del problema de esta investigación.

El concepto de RNA tiene origen en 1943, con el trabajo de McCulloch y Pitts (Zurada, 1992). El modelo reconstruido por los autores incluía los elementos teóricos necesarios para que el sistema preforme operaciones lógicas. Sin embargo, la tecnología de la época no posibilitaba implementar el modelo en la práctica. A inicios de los años 1950 comienzan a realizarse las primeras implementaciones de modelos. En 1958, Rosenblatt presentó teóricamente la unidad mínima de una RNA, el perceptrón. Este psicólogo estadounidense presenta al perceptrón como el mecanismo capaz de aprender a clasificar una serie de datos. A inicios de la década de 1960 se introduce el dispositivo ADALINE, o combinador lineal adaptativo por sus siglas en inglés. La novedad es introducida a partir del ajuste adaptativo de los pesos aleatorios, a fin de obtener el resultado esperado[24]. El éxito en la implementación de este dispositivo se verifica tanto en la búsqueda de patrones para clasificación y regresión. Un caso de uso célebre es la elaboración de pronósticos climáticos. El entusiasmo de la época llevó a formular nuevos sistemas de RNA. Surge en estos años el concepto de redes en capas (del inglés, *layered networks*) aunque su implementación requerirá nuevos avances tecnológicos tanto en procesamiento como poder de cómputo. Durante la década de los 70, se elaboran los primeros resultados teóricos del concepto de memoria asociativa en RNA. Este constructo teórico, recuperado de la psicología, refiere a la habilidad de aprender y recordar relaciones entre elementos. Esto permite retener contextos, en

corresponde a la internalización de los cambios en los parámetros a partir de los datos. Al respecto puede consultarse Chollet (2018) y Zurada (1992). Para un enfoque más global sobre el concepto de inteligencia artificial puede consultarse el texto elaborado por el Grupo de expertos de alto nivel sobre inteligencia artificial de la Comisión Europea. Disponible en https://ec.europa.eu/digital-single-market/en/news/definition-artificial-intelligence-main-capabilities-and-scientific-disciplines

[24] Puede observarse aquí la diferencia entre paradigmas en programación. El paradigma simbólico requiere que el programador ingrese una capa de reglas que se aplicarán a los datos con el fin de obtener un resultado determinado. En el paradigma basado en aprendizaje automático, el programador ingresa los datos y los resultados deseables para obtener las reglas que permitan clasificar o regresar las variables ingresadas. Por ello, a menudo se indica que un sistema de aprendizaje automático se *entrena* más que se *programa*. El modelo encuentra una estructura subyacente en los datos, es decir encuentra fundamentos estadísticos en los mismos que permiten representarlos y proyectarlos hacia adelante. Véase Chollet (2018).

base a experiencias pasadas, y obrar en consecuencia cuando un contexto similar se presenta. Los avances alcanzados en este plano permitieron introducir un nuevo modelo de RNA, las redes neuronales recurrentes (RNR). En esta clase particular de redes, las conexiones entre las distintas capas describen un grafo a partir de una secuencia dinámica temporal. Así, una red neuronal recurrente permite procesar secuencias de datos a partir de ingresar cada elemento y retener una memoria, denominada *estado,* de la secuencia anterior (Chollet, 2018, pp. 196). El concepto de recurrencia implica que el resultado obtenido en el momento actual se convierte en el input del siguiente período. Es decir, el modelo considera el dato actual más lo que retuvo de todos los períodos anteriores. Esto permite que el modelo internalice secuencias de dependencias en el tiempo a fin de comprender el contexto en el cual el valor de la variable es considerado. Este enfoque se ha implementado con éxito en problemas donde la secuencia temporal en la que los elementos son presentados es relevante[25]. La particularidad de las RNR consiste en que presentan una tercer matriz de ponderadores que media entre el estado actual y los estados anteriores. Mientras que las RNA simples se alimentan con los datos de entrada y las capas ocultas son actualizadas en cada paso del proceso, las RNR se alimentan no sólo con el input sino también con los valores que se han calculado anteriormente. Esta es la particularidad que le permite a este tipo de redes neuronales retener una secuencia temporal de datos.

Formalmente:

$$h(t) = f(h(t-1)\,, x(t))$$

Donde el estado actual de la capa oculta $h(t)$ es el resultado de la aplicación de una función no lineal f, que toma como parámetros, la historia anterior, representada en el estado oculto $h(t-1)$ y el estado actual $x(t)$. La optimización de los pesos en este tipo de redes también se realiza con el algoritmo de retropropagación. Como se enunció más arriba, consiste en calcular la derivada parcial del error con respecto a la matriz de pesos de manera recursiva para cada neurona. No obstante, a medida que se incrementan las cantidades de capas ocultas el método de propagar el error hacia atrás pierde potencia. Al tratarse de operaciones encadenadas, cambia la sensibilidad con la cual el mecanismo de corrección afecta a las capas. Concretamente, el punto central del

[25] Ejemplos de casos de éxito en la implementación de este tipo de red pueden observarse en la aproximación de series temporales, en la reconstrucción de cuadros de imágenes de video, en la predicción de la próxima palabra en una oración o la siguiente nota musical en una canción, entre otros.

método del descenso de gradiente consiste en mejorar de manera recursiva la matriz de ponderadores tal que el valor esperado de salida coincida con el valor retornado por el modelo. No obstante, a medida que la red se vuelve más profunda, la fuerza del gradiente se desvanece. Es decir, la magnitud de la corrección que ejerce la retropropagación del error resulta menor en las capas iniciales de la red. Esto implica que la tasa de actualización sea menor a medida que se retrocede en las capas, ya que el gradiente es cada vez más cercano a cero. Como señala Chollet, para una RNR simple, las dependencias de largo plazo son imposibles de aprender (2018, pp. 202). Este fenómeno se lo conoce como el problema del desvanecimiento del gradiente[26], el cual sucede en los casos en los que los coeficientes de la matriz de pesos son menores a 1 y cuando se utilizan funciones de activación cuyo resultado se encuentra entre 0 y 1. Para otros casos, el riesgo que se presenta es el del gradiente explosivo, es decir un gradiente que tienda a infinito. Una solución a este problema, son las redes de memoria corto-largo plazo o LSTM[27], por sus siglas en inglés. Estas son redes recursivas más complejas cuya ventaja algorítmica radica en poder retropropagar el gradiente a través de las capas sin pérdida. Su característica consiste en que es capaz de retener información del pasado para ser reinyectada en un momento posterior del entrenamiento. En su versión más simplificada, una LSTM se representa como una célula compuesta por tres compuertas. Dos de ellas representan el ingreso y salida de datos. La tercera compuerta representa la memoria, también denominada *forget gate* o compuerta de olvido, por su traducción del inglés. El objetivo de esta última es retener información para utilizarla en un momento posterior del entrenamiento. Dicha información busca dar cuenta del contexto, para que el modelo alcance una representación fidedigna de los datos. Lo que articula a estas tres compuertas en la célula LSTM es el estado actual. A riesgo de sobre simplificar la explicación, una compuerta consiste en una serie de operaciones matriciales tal como en las RNA más sencillas. Esto implica que cada compuerta presenta una matriz de pesos diferenciable con lo cual es posible utilizar el método de retropropagación para actualizar los ponderadores. Lo interesante es que el modelo debe ser capaz de discernir qué olvidar y qué conservar de cada estado a fin de retener el contexto de manera efectiva. El modelo que se desarrolla aquí está basado en redes recursivas del tipo células LSTM.

[26] Las razones teóricas de este problema fueron estudiadas a principios de 1990. Por ejemplo, puede consultarse Bengio, Y; Simard, P; & Frasconi, P; (1994).Learning Long-Term Dependencies with Gradient Descent Is Difficult, *IEEE Transactions on Neural Networks* 5, no. 2.

[27] El algoritmo fue desarrollado por Hochreiter y Schmidhuber en 1997. Al respecto véase: Hochreiter y Schmidhuber (1997).

Además, se realiza una adaptación para la ingesta de los datos por parte del modelo. Presentar estos elementos es el objeto de los siguientes apartados.

2. Fuentes de datos: Descripción y procesamiento

ACUMAR tiene a su cargo una serie de estaciones de monitoreo continuo de datos climáticos, ambientales y del nivel del río. Estas estaciones conforman una red a lo largo de toda la cuenca. La captura de datos y series históricas es integrada por la institución en la BDH. A partir de estos indicadores ACUMAR, en conjunto con la Universidad Nacional de La Plata, la Universidad de Buenos Aires y la Universidad de La Matanza, elaboran modelos hidrológicos para la predicción de crecidas. Estos modelos fundan un sistema de alerta temprana que permite reducir el impacto frente a las crecidas. A su vez, distintos departamentos de estas universidades y otros entes gubernamentales y no gubernamentales elaboran informes de medición ambiental y de contaminación del río.

Diversos informes[28] señalan que el riesgo de inundación no es el mismo a lo largo de toda la cuenca. Dada esta característica y con el objetivo de reducir la dificultad del problema planteado, se consideran aquí los datos provenientes de punto de recolección ubicado en la comuna número 4 de la Ciudad de Buenos Aires. La disponibilidad de datos públicos del nivel del río permitió obtener un período de 18 meses, entre ellos meses de abril de 2019 y septiembre de 2020[29]. Esta variable presenta el nivel de altura en intervalos cada quince minutos, totalizando 53782 mediciones para dicho período. La extracción se realizó de manera manual desde el portal en línea de la BDH. Para la obtención de los datos meteorológicos, se recuperaron las variables temperatura media, punto de rocío, humedad, velocidad del viento promedio, presión atmosférica y el nivel de precipitación, del portal Wunderground.com. La frecuencia disponible de estos datos es promedio hora. Esta extracción se realizó de manera automática. En el anexo del presente informe se encuentra el código a tal fin elaborado. No obstante, esta fuente de datos presenta días sin registros,

[28] ACUMAR dispone estos informes en línea. También puede consultarse el trabajo de Miller (2014). Esta tesis de maestría reconstruye zonas de vulnerabilidad frente a inundaciones a partir de imágenes obtenidas por satélite. Otra fuente de consulta es la organización *Anticipando la Crecida*, un proyecto de extensión universitaria la Facultad de Ciencias Exactas y Naturales de la UBA. http://anticipandolacrecida.cima.fcen.uba.ar/

[29] ACUMAR tiene series históricas de mayor antigüedad. Para obtener estos datos puede solicitarse al ente que los facilite. Para la elaboración de este informe se consideraron sólo los datos disponibles directamente en el portal BDH.

sobre todo para los días del año 2019. Para completar los espacios faltantes se utilizó otra fuente de datos proveniente del portal *POWER data*. Este portal, dependiente de la Administración Nacional de Aeronáutica y Espacio[30] estadounidense, integra y dispone datos satelitales sobre indicadores climáticos para todo el mundo. El portal presenta datos de agregados en promedios diarios. De esta fuente secundaria de información se adicionaron las variables temperatura máxima y mínima diaria y velocidad del viento mínima y máxima diaria.

La dimensión temporal disponible, en término de frecuencia de registro de los datos climáticos, no coincide con la disponibilidad de alturas del río Matanza Riachuelo. Por ello fue necesario aplicar una serie de transformaciones y agregación de datos, previos a la generación del modelo. Estas tres fuentes de datos se integraron a nivel diario completando un conjunto de 574 observaciones para el nivel del río (variable a predecir) y los indicadores meteorológicos. Para este nivel de agregación, los datos no presentaban faltantes. Cada registro en esta base está representado por su promedio diario. Por otra parte, se crearon variables que cuantifiquen la amplitud térmica y la amplitud en la intensidad del viento. Además, se crea una variable que mide el desvío estándar diario de la altura del río. El objetivo de esta variable es representar variaciones intra diarias fuertes en la altura, para aprovechar el nivel de detalle del datos. La variable a predecir presentaba valores faltantes, los cuales fueron interpolados tomando la semisuma entre el valor anterior y el siguiente. De esta manera se subsana la serie antes de su agregación. Finalmente, la ausencia de valores faltantes en todas las variables permite garantizar la ejecución del modelo basado en RNA.

3. Análisis de los datos.

Para el análisis preliminar de los datos se utilizó la librería Pandas[31] en Python. El código generado para el análisis puede consultarse en el anexo. A continuación, se detalla la base de datos a partir de sus estadísticos descriptivos más relevantes.

[30] NASA por sus siglas en inglés. El portal está disponible en línea en: https://power.larc.nasa.gov
[31] Versión de la librería utilizada: 1.1.1. https://pandas.pydata.org/

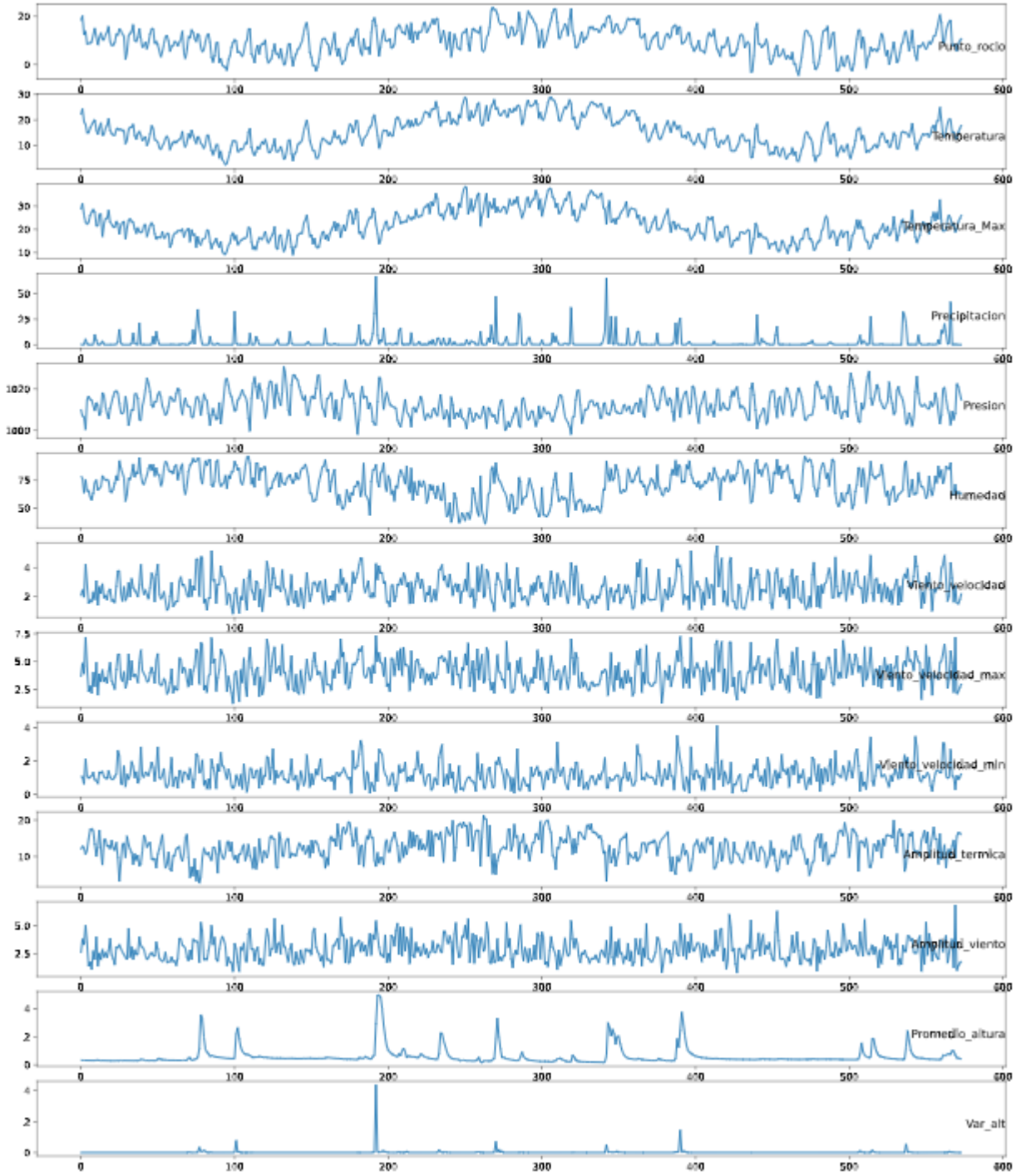
Tabla 1 : Resumen de las variables

Estadísticos	Punto_rocio	Temperatura	Temperatura_Max	Temperatura_Min	Precipitación	Presión	Humedad
Recuento	574	574	574	574	574	574	574
Faltantes	0	0	0	0	0	0	0
Media	9,77	15,12	21,81	9,30	2,55	1.012,32	71,66
Desvío estandar	5,44	5,92	6,39	5,58	7,29	5,87	13,40
Asimetría	-0,05	0,28	0,38	0,21	4,62	0,31	-0,38
Curtosis	-0,32	-0,68	-0,52	-0,50	27,31	-0,26	-0,46
Mínimo	-4,44	2,22	9,04	-4,04	0,00	998,10	35,37
Percentil_25	6,29	10,64	16,86	5,99	0,00	1.008,00	62,78
Percentil_50	9,60	14,45	20,94	8,89	0,00	1.011,80	72,84
Percentil_75	13,67	19,45	26,18	13,26	0,67	1.016,27	81,90
Máximo	23,76	29,00	38,30	23,67	66,94	1.013,90	96,90

Estadísticos	Viento_velocidad	Viento_velocidad_max	Viento_velocidad_min	Amplitud_térmica	Amplitud_viento	Promedio_altura	Var_alt
Recuento	574	574	574	574	574	574	574
Faltantes	0	0	0	0	0	0	0
Media	2,55	4,04	1,18	12,50	2,86	0,59	0,02
Desvío estandar	0,89	1,27	0,65	3,43	1,03	0,61	0,20
Asimetría	0,55	0,34	0,98	-0,10	0,67	4,09	18,86
Curtosis	-0,16	-0,50	1,43	-0,38	0,29	19,80	396,69
Mínimo	0,82	1,23	0,05	2,68	0,74	0,14	0,00
Percentil_25	1,83	3,04	0,78	10,03	2,07	0,33	0,00
Percentil_50	2,47	3,99	1,08	12,55	2,72	0,41	0,00
Percentil_75	3,10	4,89	1,49	15,01	3,51	0,53	0,00
Máximo	5,48	7,32	4,13	21,30	6,88	4,99	4,39

Se observa en la tabla que el conjunto de datos está compuesto por catorce variables más un índice temporal. Las observaciones totalizan 574 registros con promedios diarios de los indicadores. Se observa, que algunas de estas variables presentan bastante amplitud respecto de su valor medio. A continuación se presenta un gráfico de la serie de tiempo para cada variable.

Gráfico 1. Series de los indicadores



Como estamos considerando promedios diarios no emerge a simple vista la estacionalidad de los datos. Podrían exceptuarse de esta observación las variables que hacen referencia a la temperatura. Más allá de eso, el algoritmo LSTM no requiere de la corrección de la estacionalidad para su ejecución, aunque el modelo podría verse beneficiado si se la realiza.

Al analizar la variable a regresar, se observan fuertes picos en el nivel de altura del río. Si se recupera de la tabla 1 el coeficiente de asimetría se observa que la variable presenta asimetría positiva, lo cual indica que su distribución presenta una cola pesada a derecha. El gráfico 2 permite observar esto mismo:

Gráfico 2: Y = Distribución de la altura del río

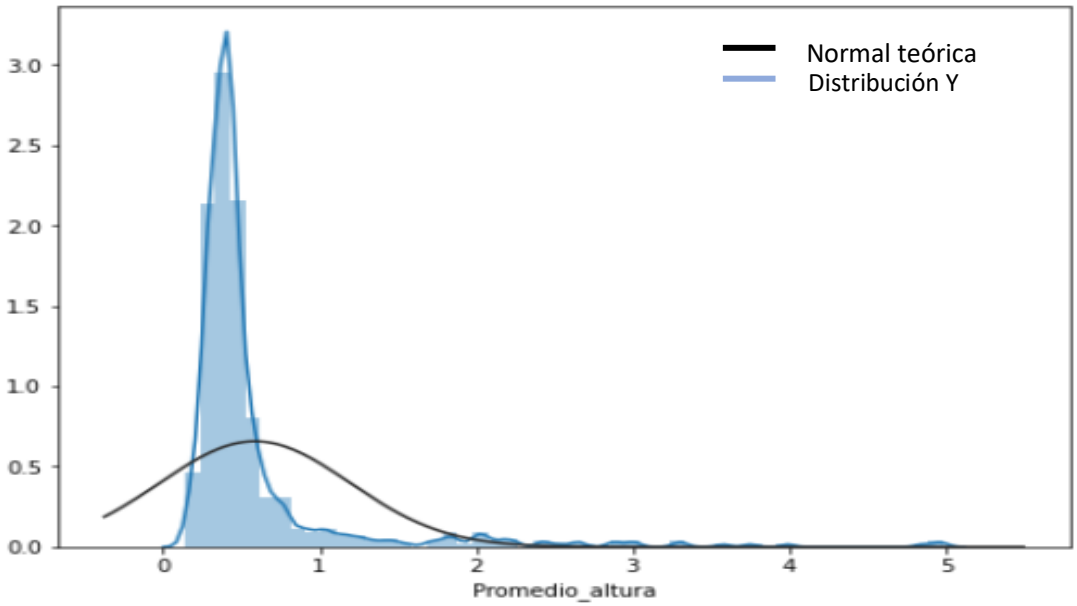
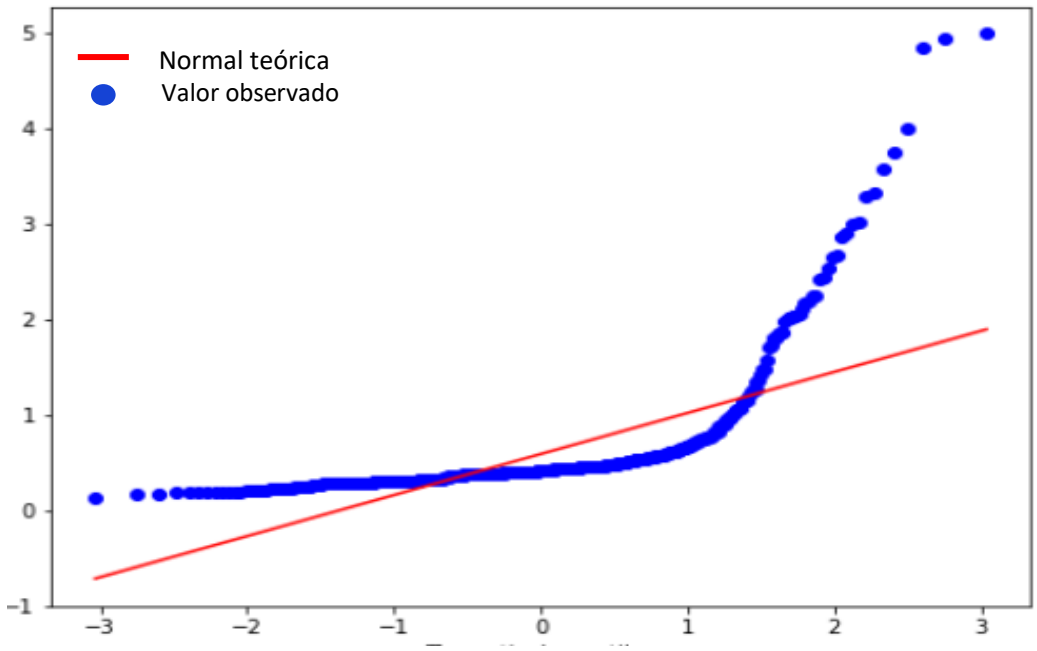


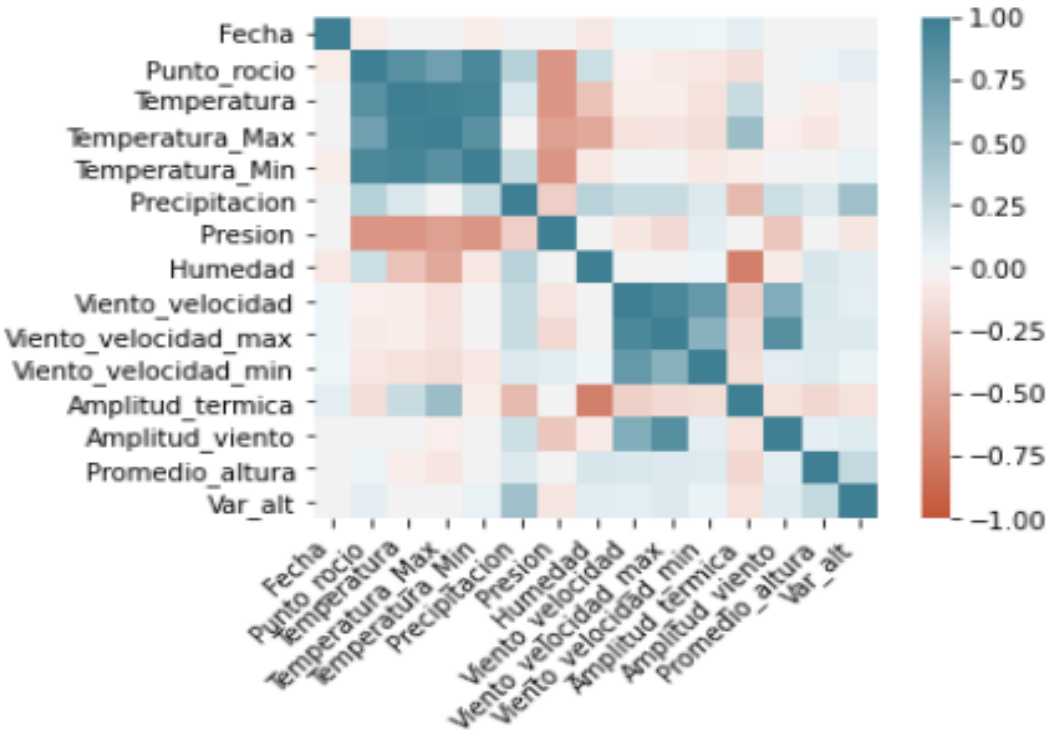
Gráfico 3: Altura del río



Mientras que la altura del río está representada por la curva celeste, en negro se indica la distribución de una normal teórica. Se observa que más del 97% de la curva se encuentra entre cero y un metro de altura. En el gráfico 3 se reconstruye un esquema Q-Q, el cual permite comparar los valores observados de la variable a predecir (puntos azules), contra la ubicación que tendrían dichos puntos de encontrarse normalmente distribuidos. Ambos gráficos permiten intuir que no se trata de una variable normalmente distribuida. Sin embargo, un nivel formal de rigurosidad exige la elaboración de un test de hipótesis. Las conclusiones extraídas de estos test permiten afirmar que ninguna de las variables consideradas en este trabajo se encuentra normalmente distribuida. En el anexo puede consultarse el detalle de los test elaborados para cada variable.

Para cerrar con el análisis preliminar, se elabora una matriz de correlación a fin de identificar relaciones subyacentes entre las variables. Se utiliza como medida el coeficiente de correlación de Pearson. Se observa una débil correlación negativa entre la altura del rio y la amplitud térmica. A la vez, se verifica correlación débil positiva de la variable a predecir con las variables referidas al viento, las precipitaciones y a la humedad del ambiente.

Gráfico 4: Matriz de correlaciones



En la matriz se observa que existe una fuerte correlación lineal entre las variables referidas a la temperatura y, por otra parte, entre los indicadores del viento. Esto se debe a que algunas de estas variables son derivadas a partir de otras, es decir están midiendo el mismo fenómeno. Sin embargo, dado que los elementos que influyen en una eventual crecida que dé origen a inundaciones tienen relación con variaciones bruscas en los indicadores climáticos, la decisión metodológica en este caso es conservar estas variables.

A modo de conclusión, el análisis preliminar de los datos permite señalar que las variables no están normalmente distribuidas, algunas de ellas presentan amplia variabilidad y sesgo. Esto indica que el modelo a construir deberá considerar relaciones no lineales entre los datos. Respecto de los valores atípicos observados, aquí no pueden descartarse. Metodológicamente estos son los valores que interesa predecir con mayor precisión.

En el siguiente capítulo se presentará el modelo desarrollado con estos datos. Primeramente se presenta el argumento utilizado para convertir el análisis multivariado de series temporales a un modelo de aprendizaje supervisado. Luego se reconstruyen los resultados del modelo. Finalmente, se presenta un esquema metodológico para una eventual incorporación del modelo de aprendizaje automático en la institución.

CAPÍTULO III: ANÁLISIS DE LOS RESULTADOS Y PROPUESTA DE IMPLEMENTACIÓN

1. De series temporales a aprendizaje supervisado

Las proyecciones de series temporales multidimensionales son un problema complejo dentro de los modelos analíticos predictivos. Dicha dificultad queda representada en los diferentes algoritmos de RNA desarrollados, cuyo objetivo busca superar las proyecciones vigentes. Pero además, las series temporales agregan un factor adicional de complejidad a la hora de elaborar pronósticos. En estos casos, los datos no pueden tomarse al azar ya que existe una secuencia de dependencia de las variables que ingresan al modelo. En el capítulo anterior, se revisó que existe un tipo específico de redes neuronales cuya ventaja radica precisamente en poder manejar esta

dependencia de secuencias temporales. Estas son las RNR y en particular la arquitectura LSTM es la que aquí se implementa.

El primer paso para implementar un modelo LSTM es preparar los datos para que puedan ser considerados por el algoritmo. Una vez capturados, limpiados y ordenados, la matriz de datos debe ser enmarcada como un problema de aprendizaje supervisado. La idea es la siguiente: el conjunto de datos se indexa por fecha para establecer una secuencia temporal. De esta manera, dada una fecha, el estado del sistema queda representado por las variables climáticas más la altura del río. Por lo tanto, para elaborar la proyección de la altura actual Y(t), el modelo tomará las variables climáticas del período anterior X(t-1). Para el ejercicio aquí propuesto, el período anterior equivale al día anterior. Sin embargo, otros caminos de análisis pueden considerar diferentes variaciones en el tiempo, por ejemplo períodos semanales.

Para el entrenamiento se definió un modelo inicial con una sola capa oculta. Se prueba este el modelo indicando como grupo de validación la última semana en el conjunto de datos. La cantidad de neuronas de esta capa surge a partir de un proceso de prueba y ajuste. Para decidir su número se prueban iterativamente múltiplos de dos. La primer capa oculta queda constituida por 128 neuronas. Se decide incorporar una segunda capa, bajo la misma metodología. Mejores resultados se obtienen al incorporar esta capa adicional compuesta por 64 neuronas. El resumen del modelo indicando la cantidad de parámetros por capa puede consultarse en el anexo.

Por otra parte, se ejecutó un proceso de optimización de los hiperparámetros de la red. En este caso se optimizan a la cantidad de épocas a entrenar, la tasa de aprendizaje del modelo y la decadencia. Por épocas se entiende la cantidad de veces que se ejecuta el ciclo completo de la red. En cada iteración se obtienen resultados que permiten observar si el modelo mejora con sucesivos entrenamientos. La tasa de aprendizaje permite controlar la reducción de la perdida en el modelo de aprendizaje automático. Como se señaló en el primer apartado del capítulo anterior, el vector gradiente presenta una dirección y una magnitud. La tasa de aprendizaje se multiplica por dicho vector a fin de determinar el siguiente punto a evaluar. Si ese punto es menor que el anterior entonces el modelo alcanza mejoras, es decir el ajuste de la matriz de ponderadores resulta eficiente. El parámetro de decadencia está relacionado con el de tasa de aprendizaje. Este parámetro permite modificar dicha tasa a medida que suceden las iteraciones. La idea conceptual

radica en explorar espacios de la superficie que han sido descartados en ciclos anteriores. Por último, para evitar el sobreajuste del modelo se establece una tasa de abandono del 0.2 por ciento.

2. Resultados del modelo

Para evaluar la performance del modelo se utiliza la métrica RMSE. El juego de hiperparámetros que surge de la optimización del modelo permitió alcanzar un resultado para esta métrica de 0.076. Previamente la base fue normalizada entre 0 y 1 por lo que el resultado alcanzado resulta bueno.

En el grafico 5 se observa el historial de entrenamiento a lo largo de las 250 épocas. Se nota la mejora en la performance a lo largo de ellas. Alrededor de los 150 ciclos la mejora comienza a estacionarse. Se probó una serie de ejecuciones con este valor para el hiperparámetro correspondiente pero los resultados obtenidos fueron inferiores.

Gráfico 5: Entrenamiento del modelo

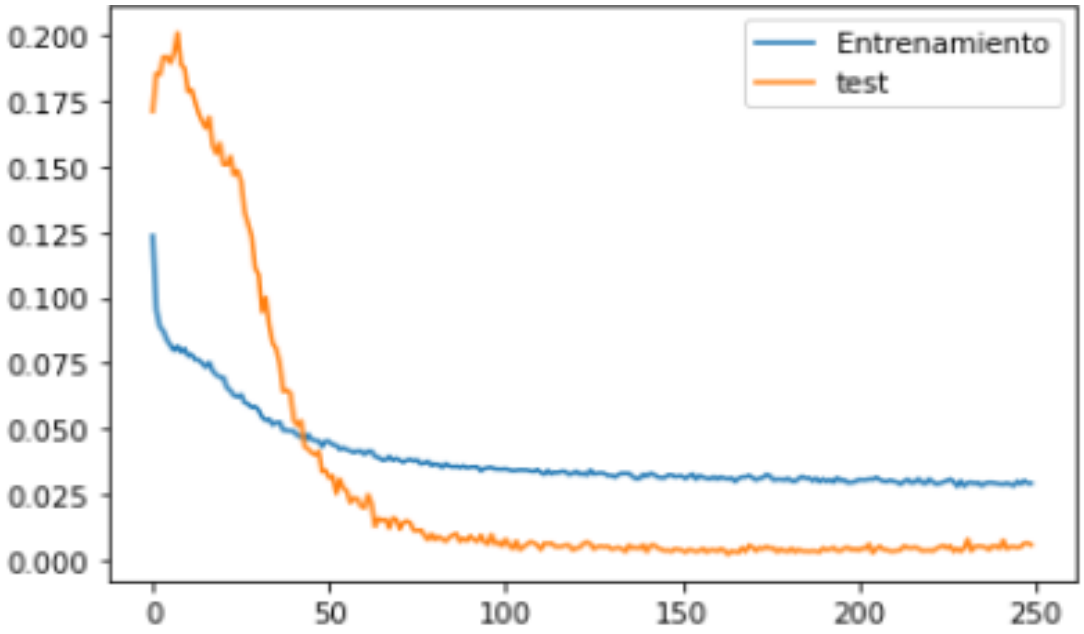
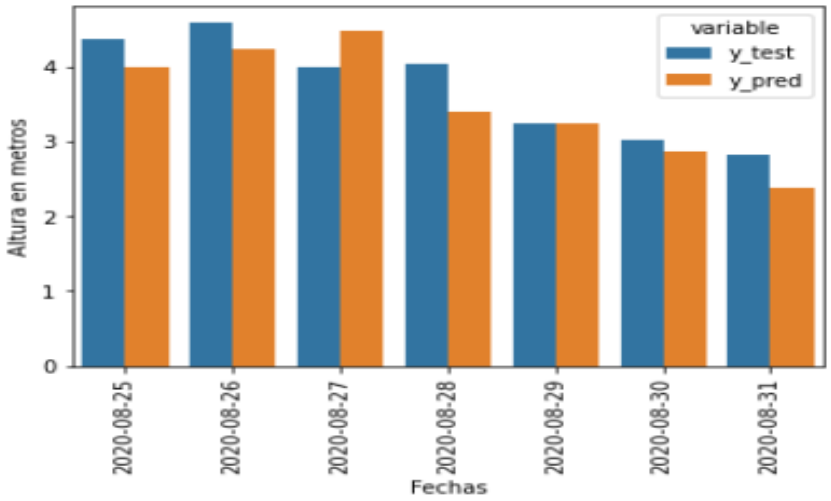
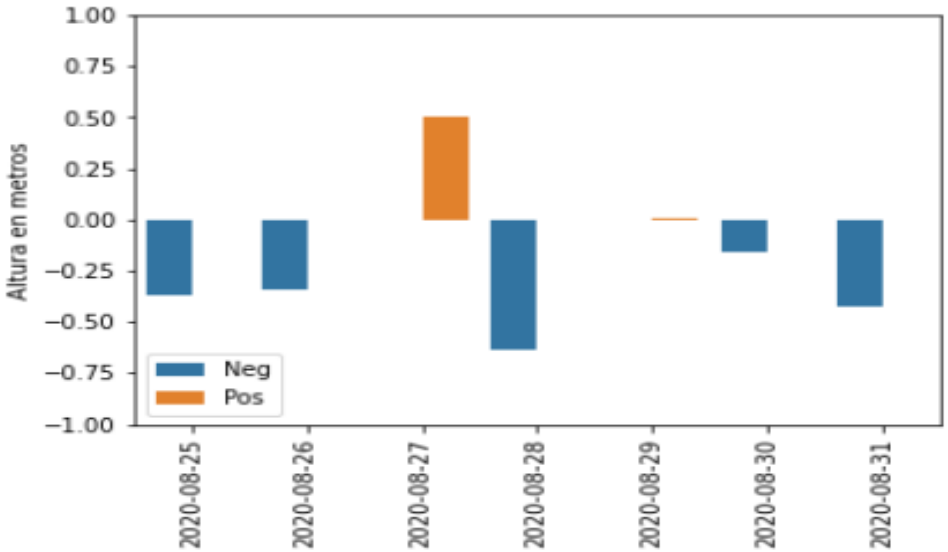


Gráfico 6: Pronósticos del modelo



A continuación se le pide al modelo que pronostique la altura del río para los siguientes siete días. Para la presentación se discretiza la variable dado que la unidad de medición de los estadísticos son promedios diarios. Los resultados se evalúan en los gráficos 6 y 7. El error medio en el pronóstico es de 0.39 metros de altura. La tabla con el detalle de la conversión para cada día puede consultarse en el anexo. El primer gráfico compara el valor predicho contra el observado, mientras que el segundo recoge las diferencias entre ambos. Un elemento a señalar es que el modelo tiende a subestimar la altura. Este punto puede constituirse en una vía de mejora del modelo ya que el riesgo es mayor en una subestimación que en una sobre estimación. Los errores en las estimaciones pueden mitigarse al considerar un período de tiempo que abarque más años. Un conjunto de datos más extenso que el presente podría favorecer al pronóstico ya que el algoritmo podría captar otras dependencias en el tiempo.

Gráfico 7: Error de las predicciones



3. **Acercamiento a una eventual implementación del modelo**[32]

El objetivo de este último apartado consiste en presentar los pasos a seguir en el caso de que el modelo desarrollado se quiera implementar de manera productiva, como contribución a un sistema alerta temprana. Entre las incumbencias profesionales ACUMAR, se señala la protección y alerta en la población respecto de variaciones en el clima que pudieran ocasionar daños materiales y a la vida. Como se señaló, para alcanzar este y otros objetivos institucionales, el organismo presenta una serie de estaciones de medición toda la Cuenca. Dichas estaciones recolectan y transmiten datos que se integran en la BDH para su posterior explotación. Además, el fenómeno a pronosticar es parte de un sistema dinámico, solo parcialmente descomponible, cuyas condiciones iniciales no pueden ser precisadas totalmente, en el sentido de que no es posible conocer exactamente el estado de la atmósfera en todo momento. Sin embargo, el riesgo de una catástrofe persiste. Por lo tanto, resulta necesario, no sólo elaborar un sistema de alerta temprana sino que además es indispensable que el modelo se retroalimente con los datos todo el tiempo.

Con un fuerte foco en los usuarios finales, se propone implementar la metodología *Desing Thinking* para llevar a cabo el proyecto. Partiendo de su definición conceptual, esta metodología permite alcanzar soluciones basadas en las necesidades del usuario final, en este caso la prevención de una inundación. La primer fase de esta metodología consiste en la *empatía*. Resulta clave en esta primer etapa, comprender las circunstancias, los problemas y necesidades de los usuarios finales. Esto implica valorizar el conocimiento que aporta el usuario final desde el inicio del proyecto, entendiendo sus motivaciones y haciéndolas propias. Las preguntas que operacionalizan este primer punto buscan comprender cómo se conectan los usuarios con el servicio brindado, en qué circunstancias y contextos. A posterior se introduce la etapa de *definición* a fin de comprender la dimensión estratégica del reto que se enfrenta. En esta fase, se busca sintetizar el conocimiento actual para alcanzar nuevas perspectivas. Pasado el momento de la definición sigue la fase que refiere a *idear* opciones de solución. Se busca generar múltiples ideas que den solución al problema, a fin de generar *prototipos* en un siguiente momento. El prototipo es la materialización de la idea. Consiste en dar forma, en otorgar una definición de alto nivel de abstracción pero que

[32] ACUMAR presenta un esquema de modelado e integración de datos para simular el comportamiento de múltiples variables en la Cuenca. Este detalle está disponible en: https://www.acumar.gob.ar/monitoreo-ambiental/modelacion-matematica/

permita transmitir el concepto subyacente. La última etapa de estos cinco momentos consiste en *evaluar* los prototipos. Es la fase empírica de validación que permite identificar qué elementos del problema alcanzan solución y cuáles pueden mejorarse en sucesivas iteraciones. Este elemento permite vincular a *Design Thinking* con la metodología *Agile* ya que ambos se basan en iteraciones y desarrollo incremental. Además, ambos enfoques promueven equipos multidisciplinares de trabajo. La capacidad de aportar la singularidad de cada individuo en un proceso de trabajo colaborativo permite estar más cerca del usuario final y pensar soluciones cercanas a la realidad del usuario. Algunos de los puntos que motivan el maridaje de ambos enfoques son los múltiples beneficios que se alcanzan. Por ejemplo, *Backlogs* más asertivos e innovadores, historias de usuarios reales, es decir enfocadas en el problema del usuario final, requisitos específicos y con menor dificultad de comprensión, inversiones en tiempo y dinero más consistentes y confiables, entre otras En conclusión, esta combinación de enfoques permite alcanzar soluciones de punta a punta. *Design Thinking* centra su aporte en el pensamiento disruptivo y *Agile* posibilita que estas ideas innovadoras alcancen carácter tangible de manera productiva.

Las fases del *Design Thinking* se operacionalizarán de la siguiente manera. Para la fase de empatía se emplean herramientas que recaben testimonio en los usuarios. Algunos ejemplos son encuestas, entrevistas, estadísticas, *focus group*, entre otros. En particular para la prevención de inundaciones resulta útil considerar estadísticas de inundaciones en zonas urbanas a la vez que recopilar el testimonio de personas afectadas. Este paso ya permite definir un rol dentro del proyecto, el usuario final del servicio. La fase de definición permitirá construir un "*service blueprint*", en base a la información obtenida en el paso anterior. Este informe consiste en un diagrama que permite visualizar las relaciones entre los diferentes componentes del servicio y los usuarios finales. Este enfoque resulta de mucha utilidad ya que la prevención de una catástrofe es por definición un problema complejo que requiere el esfuerzo coordinado de muchos actores involucrados. Al momento de innovar, todas las ideas que permitan ir más allá de lo obvio son bienvenidas. Por ello, se propone un enfoque que considere múltiples ideas también conocido como *brainstorming*, pero que a la vez considere el acontecimiento de errores y fallas ("*Worst possible idea*") y cómo obrar cuando sucedan. Dado que de lo que se trata es prevenir, se buscar privilegiar un método de pensamiento que valore ampliamente la diversidad de enfoques y propuestas. Para la fase de prototipado la metodología propone que su construcción no demore a

la vez que no presente un alto costo económico. En principio para un sistema de alertas el requisito mínimo es la alerta en sí. El prototipo deberá considerar cómo los usuarios harán uso de esa alerta. Por ejemplo, se podrá realizar un maqueteado web para considerar cómo se verá esta alerta en el portal de ACUMAR o del Servicio Meteorológico Nacional. Un *backlog* posible para este desarrollo requiere generar alertas precisas para los próximos días. Cuanto mayor sea el espacio entre el evento y la proyección, mayor será la oportunidad de gestionar le riesgo al tomar medidas preventivas. Además, de que la alerta esté disponible en la web, en una siguiente etapa se buscará disponer de este resultado como un servicio automático para ser consumido por terceros. Para la evaluación del consumo por parte del usuario final, puede considerarse cómo se verá la alerta en el portal. Luego esta representación puede ser evaluada por un grupo de usuarios exteriores al proyecto a fin de validar la calidad e intuitividad de la alerta, así como también los pasos a seguir.

El plan de proyecto permitirá definir un proceso de ciencia de datos en equipo. Sumado al rol del usuario final para quien se construye este servicio, se definen los siguientes roles: un jefe de proyecto (*Project Manager*) cuyo objetivos consiste en coordinar el trabajo del equipo. Su labor produce el acta de constitución ("*Project charter*") que consiste en un documento que autoriza formalmente la existencia del proyecto. El rol del jefe de proyectos debe garantizar que el proyecto se ejecute en los tiempos acordados. Es el responsable de que los objetivos institucionales se alcancen en tiempo y forma. Una vez implementado el modelo produce el reporte de resultados, el cual resume los logros alcanzados. Otro perfil requerido es el ingeniero de datos. Su rol consiste en garantizar la disponibilidad de los datos para la implementación de los modelos. Su actividad está coordinada con el científico de datos, el cual será encargado de limpiar los datos para que puedan ser utilizados en el proceso de aprendizaje supervisado. A su vez, este rol es responsable de la comunicación a través de visualizaciones, reportes y tableros de los resultados alcanzados. Además, el equipo debe estar conformado por un Arquitecto de soluciones ("*Solution Architect*") cuya misión es diseñar la solución que dará respuesta a la necesidad detectada en la definición del problema. Tiene conocimiento no solo en el área técnica, es decir en los requerimientos funcionales necesarios para implementar la solución, sino que además tiene responsabilidad por la gestión de requisitos no funcionales, esto es las necesidades del usuario final.

El éxito del proyecto se medirá en función de los pronósticos elaborados. El objetivo final de largo plazo podrá consistir en un sistema de alerta en tiempo real que notifique a los habitantes

frente a riesgos de inundaciones. Periódicamente deberá reentrenarse el modelo en producción con nuevos datos para garantizar que el modelo implementado es una representación eficiente de los datos. Esta revisión periódica resulta esencial para la correcta gestión del riesgo ya que podría cambiar el comportamiento de algún parámetro, por ejemplo por factores asociados al cambio climático o por la intervención del hombre, a fin de garantizar la vigencia del modelo implementado en producción. Todo el código y los documentos que se generen a partir del proyecto deberán almacenarse bajo un sistema de control de versiones a fin de facilitar la colaboración al interior del equipo. Este proyecto deberá alojarse en un repositorio independiente de otros proyectos institucionales.

CONCLUSIONES

En el presente informe se elabora un pronóstico aproximado del valor del nivel del río Matanza Riachuelo a partir de su serie histórica y un conjunto de indicadores climáticos tales como temperatura, viento, humedad, punto de rocío y otros, los cuales describen el estado del ambiente en cada momento. El modelo aquí implementado permite anticipar el nivel de altura del río con una semana de distancia, lo cual verifica la hipótesis de trabajo. El error promedio de la predicción es de 0.4 metros de diferencia. La disponibilidad de datos posibilitó describir el estado medio por día respecto del conjunto de variables climáticas y el nivel del río. El análisis busca proporcionar una herramienta adicional a la toma de decisiones, por ejemplo a la hora de prevenir con suficiente antelación eventuales crecidas. Dicha gestión del riesgo permitirá tomar acciones preventivas a fin de disminuir o evitar daños materiales y a la vida en general.

En cuanto al modelo, primero se reconstruyó la serie histórica para las variables mencionadas. Se tomaron datos del período abril 2019 septiembre 2020. Para crear el modelo se utilizaron redes neuronales recurrentes del tipo LSTM, las cuales son ampliamente utilizadas en problemas de predicción con series temporales multivariantes. El diseño algorítmico de este tipo de redes permite retener factores importantes en los datos de manera que se conserva una memoria de los datos que puede ser reinyectada en otro momento del entrenamiento. Esta característica es la que permite a este tipo particular de redes alcanzar estimaciones precisas para series históricas de largos períodos con comportamientos complejos no lineales. Para el entrenamiento se definió un modelo con dos capas ocultas de 128 y 64 neuronas respectivamente. Los datos debieron transformarse a fin de que el pronóstico de las series temporales pueda ser entrenado como un modelo de aprendizaje supervisado. El proceso recursivo de pronóstico itera de la siguiente manera: los datos de las series históricas ingresan al modelo, previa transformación, y retornan como resultado la predicción para el siguiente período. Para elaborar el pronóstico se toman los valores anteriores de la serie con una ventana fija, en este caso el día anterior. A partir del segundo período el algoritmo incorpora una memoria que le permite contextualizar dependencias de largo plazo. El proceso se reitera hasta concluir el entrenamiento.

El presente trabajo no pretende ser concluyente. Su aporte puede ser considerado como una colaboración a los modelos hidrológicos y de alerta temprana que ACUMAR, en colaboración con distintas universidades del país, organizaciones no gubernamentales y otras instituciones, realiza.

Existen múltiples caminos por donde continuar. Algunas vías de mejora para este modelo pueden generarse si se considera un período de tiempo mas largo, o bien si en lugar de tomar promedios diarios se dispusiera de indicadores horarios de las variables. Otra vía de exploración para la mejora puede consistir en considerar otras variables climáticas tales como la nubosidad, la dirección del viento o la humedad en distintas alturas. Por otra parte, podrían incorporarse imágenes satelitales y de radares.

Fenómenos climáticos extremos tales como las inundaciones, están influenciados por múltiples factores. El desarrollo económico, el uso del sustrato natural por parte del hombre, el cambio climático y la imprevisibilidad natural del clima son algunos de ellos. El proceso histórico de asentamiento urbano en la Cuenca condujo a la explotación excesiva de los recursos naturales, la contaminación y ocupación de áreas vulnerables a inundaciones. La naturaleza y el accionar del hombre contribuyen al acontecimiento de desastres naturales. Sin embargo, los riesgos de desastre surgen de la combinación de fenómenos meteorológicos y sociales. Se trata de un problema complejo de gestión, solo parcialmente descomponible, cuyo origen puede rastrearse hasta la fundación de la Ciudad de Buenos Aires y que requiere de un enfoque multidisciplinario a fin de cerrar la deuda social y ecológica vigente. En esta línea es que surge en los últimos años ACUMAR, un ente que por definición busca tener una visión compartida de la gestión pública de la Cuenca. Si bien conserva una estructura vertical de gobierno, la institución presenta componentes horizontales ya que en las Comisiones de Participación Social cualquier persona puede participar por el solo hecho de ser ciudadano.

La elaboración de un modelo de aprendizaje automático implica la inmersión en un proceso heurístico y creativo. El proceso de generación de modelos involucra prueba, error y ajuste. La iteración potencia el razonamiento deductivo del investigador implicado en la gestión del riesgo. Su compromiso requiere un marcado sesgo en favor del cuidado de la vida.

BIBLIOGRAFÍA

Alley R. B., Emanuel K. A., Zhang, F., (2019) Advances in weather prediction. *Science*, *363*(6425) 342–344. Recuperado de: https://science.sciencemag.org/content/363/6425/342/tab-figures-data

Autoridad de Cuenca Matanza Riachuelo. (2020). *Definiciones institucionales*. Recuperado de https://www.acumar.gob.ar/institucional/

Autoridad de Cuenca Matanza Riachuelo (2016) *Plan Integral de Saneamiento Ambiental*. Actualización PISA 2016. Ciudad Autónoma de Buenos Aires, Buenos aires. Recuperado de: https://www.acumar.gob.ar/wp-content/uploads/2016/12/PISA-2016.pdf

Banco Mundial. (2016). *Argentina. Análisis ambiental de País* (9). Serie de informes técnicos del Banco Mundial en Argentina, Paraguay y Uruguay. Oficina Regional de América Latina y Caribe. Recuperado de: http://documents1.worldbank.org /curated/es/552861477562038992/pdf/109527-REVISED-PUBLIC-AR-CEA-An%C3%A1lisis-Ambiental-de-Pa%C3%ADs-Segunda-Edici%C3%B3n.pdf

Barrios V. , Camilloni, I. (2016) *La Argentina y el cambio climático. De la física a la política.* Buenos Aires, Argentina: EUDEBA.

Boyd, D., Crawford, K., (2012) Critical Questions for Big Data. *Information, Communication & Society*, 15(5), 662-679. Disponible en: http://dx.doi.org/10.1080/1369118X.2012.678878.

Brailovsky A. E., Foguelman D., (2009) *Memoria Verde. Historia ecológica de la Argentina* Buenos Aires, Argentina: Debolsillo.

Chollet, F., (2018). *Deep Learning with Python*. Nueva York Estados Unidos: Manning Publications Company.

Comisión Económica para América Latina y el Caribe (2012) Colección Documentos de proyectos*. El desafío hacia el gobierno abierto en la hora de la igualdad*. Santiago de Chile, Chile. Naciones Unidas. Disponible en línea en: https:// repositorio.cepal.org/bitstream/handle/11362/3969/1/S2012004_es.pdf

Culclasure, A. (2013) *Using Neural Networks to Provide Local Weather Forecast* (tesis de maestría). Georgia Southern University, Georgia, Estados Unidos.

Del Rio Riande, G. (2020) Marcos teóricos y prácticas críticas para entender la gestión, el procesamiento y análisis de grandes datos: lectura distante y macroanálisis. Material de la cátedra: *Gestión de datos en contextos organizacionales*.

Hochreiter, S., & Schmidhuber, J. (1997). Long short-term memory. *Neural Computation, 9*(8), 1735–1780. https://doi.org/10.1162/neco.1997.9.8.1735

La Red Martínez Martínez, M., Crespo Fidalgo, J., (2013). *Sistemas Inteligentes para el ajuste de Modelos Hidrológicos. Aplicación al Río Paraná* (tesis doctoral). Universidad de Cantabria, Santander, España.

LaValle, S., Lesser, E., Shockley, R., Hopkins, M., S., Kruschwitz, N., (2011) Big Data, Analytics and the path from insights to value. *MIT Sloan Management Review*, 2(52) 20-31.

Maqsood, I., Riaz Khan, M., Abraham A., (2003). Weather Forecasting Models Using Ensembles of Neural Networks. En: Abraham A., Franke K., Köppen M. (eds) *Intelligent Systems Design and Applications. Advances in Soft Computing* (33-45), Berlín, Alemania: Springer.

Miller, M., (2014). *Delimitación de Áreas de Riesgo Ambiental para la Salud en la Cuenca Matanza Riachuelo, a partir de Técnicas de Análisis Espacial e Inteligencia Artificial* (tesis de maestría). Universidad Nacional de Córdoba, Córdoba, Argentina.

Mogrovejo, A. (2018). *Integración de un sistema de alerta temprana mediante modelación hidrodinámica y predicción de flujos con redes neuronales. Caso de estudio: río Tomebamba* (tesis doctoral). Universidad de Cuenca, Cuenca, Ecuador.

Numan, D., Di Domenico, M., (2013) Market research the ethics of big data. *International Journal of Market Research* 55 (4) 505-520. Disponible en: https://eprints.bbk.ac.uk/id/ eprint/16229/

Paul, A. Das, P. (2014). Flood prediction model using artificial neural network. *International Journal of Computer Applications Technology and Research, 3(7),* 473-478. doi:10.7753/IJCATR0307.1016

Provost, F., Fawcett T., (2013). Data Science and its relationship to big data and data driven decisions. En: Liebert M., A., inc. Big Data (51-59). DOI: 10.1089/big.2013.1508

Sarle W. (Abril, 1994). Neural Networks and Statistical Models, *Proceedings of the Nineteenth Annual SAS Users Group International Conference*, 1-13. Simposio llevado a cabo en el Instituto SAS, Cary, Carolina del Norte, Estados Unidos Recuperado de https://people.orie.cornell.edu/davidr/or474/nn_sas.pdf

Secretaría de Modernización de la Nación (2019) *Cuarto plan de acción nacional de gobierno abierto 2019-2021*. (Publicación vigesimotercera) Dirección de Gobierno Abierto, Secretaria de Innovación pública y Gobierno Abierto, Presidencia de la Nación. Recuperado de: https://www.argentina.gob.ar/sites/default/files/cuarto_plan_de_ accion_nacional_de_gobierno_abierto_-_ argentina_-_v4.pdf

Wica, M. , Witkowski, M. , Szumiec, A., Ziebura, (2019). *Weather forecasting system with the use of Neural Network and Backpropagation Algorithm*. CEUR Workshop Proceedings Silesian University of Technology, Gliwice, Polonia. Recuperado de: http://ceur-ws.org/Vol-2468/p8.pdf

Zabala R., Gandia E., (1980), capítulo VII Buenos Aires durante el gobierno de Don Pedro de Mendoza. En: *Historia de la Ciudad de Buenos Aires* capítulo VII(I) pp. 135-153. Buenos Aires, Argentina: Municipalidad de la Ciudad de Buenos Aires.

Zurada, J. M. (1992). *Introduction to Artificial Neural Systems*, Saint Paul, Minnesota, Estados Unidos: West Publishing Company. Recuperado de: https://anuradhasrinivas.files.wordpress.com/20 13/08/29721562-zurada-introduction-to-artificial-neural-systems-wpc-1992.pdf

BIBILIOTECAS UTILIZADAS

A continuación se listan los artículos que refieren a las bibliotecas utilizadas para el modelado de datos en Python. Todas las licencias son de código abierto. La versión de Python utilizada es la 3.7.

Abadi, M. y otros, (2015) Tensorflow: Large-Scale Machine Learning on Heterogenous system (2.3.1) [Tensorflow]. Recuperado de: https://tensorflow.org/

Chollet, F. y otros. (2015) Keras (2.1.5) [Keras]. Recuperado de: https://keras.io

Harris, C.R., Millman, K.J., van der Walt, S.J. et al. *Array programming with NumPy*.(1.18.3) Nature 585, 357–362 (2020). DOI: 0.1038/s41586-020-2649-2

Hunter J.D. (2007), . Matplotlib: A 2D Graphics Environment, Computing in Science & Engineering, 9, 90-95. DOI:10.1109/MCSE.2007.55

Pedregosa F., Varoquaux, G., Gramfort A., y otros.(2011) Scikit-learn: Machine Learning in Python (0.22) [Scikit-learn] Recuperado de:https://scikit-learn.org/

Selenium, (2020). Selenium, a portable framework for testing web applications. (3.141.0) [Selenium]. Disponible en: https://www.selenium.dev/

The Pandas development team. (2020). Pandas(1.1.1)[Pandas]. Recuperado de https://doi.org/10.5281/zenodo.3509134.

Virtanen P., Gommers R., Oliphant T., Haberland M., y otros(2020) SciPy 1.0: Fundamental Algorithms for Scientific Computing in Python (1.4.1) [SciPy]. Recuperado de: https://www.scipy.org/

Waskom, M. (2020) Seaborn (0.10.1) [Seaborn]. Recuperado de https://doi.org/10.5281/zenodo.592845

ANEXO

Código desarrollado para el informe. Software Python versión 3.7. La versión de las bibliotecas utilizadas puede consultarse en el apartado anterior.

Importante: Si bien se ha incluido una semilla, los resultados obtenidos de la ejecución de los modelos pueden variar dada la naturaleza estocástica del algoritmo LSTM, y por diferencias en el cómputo debido a la configuración de la precisión numérica de la computadora donde se ejecute.

#Busqueda de datos

```
from selenium import webdriver
from selenium.webdriver.common.by import By
from selenium.webdriver.support.ui import WebDriverWait
from selenium.webdriver.support import expected_conditions as EC
from datetime import timedelta
import datetime as dt
import pandas as pd
import numpy as np
import seaborn as sns
from scipy.stats import norm
from scipy.stats import normaltest
from  matplotlib.pyplot as plt
import tensorflow as ts
from tensorflow.keras.models import Sequential
from tensorflow.keras.layers import Dense, LSTM, concatenate
from sklearn.preprocessing import LabelEncoder, MinMaxScaler
from sklearn.metrics import mean_squared_error
import os, time, math, random

np.random.seed(4)

def unificar_links(link, fecha_ini, fecha_fin):
    """Funcion q  devuelve una lista con los links"""
    delta = fecha_fin - fecha_ini
    fechas = []
    for i in range(delta.days+1):
        day = fecha_ini + timedelta(days=i)
        fechas.append(str(day))
    links = len(fechas)*link
    links_ = [i+ j for i, j in zip(links, fechas)]
    return links_
def buscar(links):
```

```
    """Funcion q devuelve las tablas del link"""
    for i in links:
        driver.get(i)
        tables = WebDriverWait(driver,20).until(EC.presence_of_all_elements_located((By.CSS_S
ELECTOR, "table")))
        yield(tables)
driver = webdriver.Chrome("C:\Program Files (x86)\chromedriver.exe")
link = ["https://www.wunderground.com/history/daily/ar/palermo/SABE/date/"]
fecha_ini = date(2019,4,1)
fecha_fin = date(2020,8,30)
links= unificar_links(link,fecha_ini, fecha_fin)
tables = buscar(links)
df = pd.DataFrame()
df_ = pd.DataFrame()
for i in tables:
    print(driver.current_url)
    time.sleep(3)
    try:
        df = pd.read_html(driver.page_source)[1]
        df["Fecha"] =  driver.current_url[64:]
        df_ = df_.append(df, ignore_index=False)
        df = pd.DataFrame()
    except:
        print("Revisar: "+driver.current_url)
        next
```

#limpieza de datos climaticos

```
df[["Hora", "Huso_horario"]]= df['Time'].str.split(" ",expand=True)
df[["Temperatura", "unidad_t"]]= df['Temperature'].str.split("F",expand=True)
df[["Punto_rocio", "unidad_pr"]]= df['Dew Point'].str.split("F",expand=True)
df[["Humedad", "ptj"]]= df['Humidity'].str.split("%",expand=True)
df[["Viento_velocidad", "unidad_v"]]= df['Wind Speed'].str.split("mph",expand=True)
df[["Viento_rafagas", "unidad_vr"]]= df['Wind Gust'].str.split("mph",expand=True)
df[["Presion", "unidad_p"]]= df['Pressure'].str.split("in",expand=True)
df[["Precipitacion", "unidad_prec"]]= df['Precip.'].str.split("in",expand=True)
df[["hr", "min"]]= df['Hora'].str.split(":",expand=True)
df["Fecha_Hora"] =pd.to_datetime(df.Hora)
df= df[df['min'] == '00']
conservar = ["Fecha_Hora", "Huso_horario", "Temperatura", "Punto_rocio", "Humedad","Wind"
, "Viento_velocidad","Viento_rafagas", "Presion", "Precipitacion" ,"Condition"]
df = df[conservar]
convertir = {'Temperatura': int, 'Punto_rocio': int, 'Humedad': int, 'Viento_velocidad': int, 'Viento
_rafagas': int, 'Presion': float, 'Precipitacion': float }
```

```
df = df.astype(convertir)
df['Temperatura'] = round((df['Temperatura'] -32) * 5/9, 0)
df['Punto_rocio'] = round((df['Punto_rocio'] -32) * 5/9, 0)
df['Humedad'] = round(df['Humedad']/100, 2)
df['Viento_velocidad'] = round(df['Viento_velocidad']*1.60934, 2)
df['Viento_rafagas'] = round(df['Viento_rafagas']*1.60934, 2)
df['Presion'] = round(df['Presion']*33.8639,2)
```

#datos del río

```
df = pd.DataFrame()
for i in os.listdir():
    if i.endswith(".xlsx"):
        data = pd.read_excel(i)
        df = df.append(data)
#limpieza de datos del río
df = df.rename(columns={'Fecha y Hora': 'Fecha_Hora', 'Valor [m]': 'Valor'})
df =df.drop_duplicates(['Fecha_Hora'], ignore_index=True)
df[["day", "hr"]]= df['Fecha_Hora'].str.split(" ",expand=True)
df[["hr", "min"]]= df['hr'].str.split(":",expand=True)
df =df.groupby(["day"]).agg(['mean', "var"])
df = pd.DataFrame(df.to_records())
df["Fecha"] = pd.to_datetime(df["day"], dayfirst=True)
df = df.sort_values(by=['Fecha'], ignore_index=True)
df.rename(columns={"('Valor', 'mean')": 'Promedio_altura', "('Valor', 'var')":'Var_alt'}, inplace=T
rue)
keep = ['Fecha', 'Promedio_altura', 'Var_alt']
df = df[keep]
clima["Amplitud_termica"] =clima['Temperatura_Max'] -clima['Temperatura_Min']
clima["Amplitud_viento"] =clima['Viento_velocidad_max'] -clima['Viento_velocidad_min']
clima['Presion'] = clima['Presion']*10
df = pd.merge(clima, df, how="left", on= "Fecha")
```

#Análisis y visualización

```
df = df.replace([np.inf, -np.inf], np.nan).dropna()
valores = df.values
variables = [1, 2, 3, 5, 6, 7,8,9,10,11,12,13,14]
i = 1
pyplot.figure(figsize=(18, 22))
for v in variables:
    pyplot.subplot(len(variables), 1, i)
    pyplot.plot(valores[:, v])
    pyplot.title(df.columns[v], y=0.3, loc='right')
```

```
    i += 1
a = pyplot.show()
df.describe()
df.skew()
df.kurt()
dims = (8, 6)
fig, ax = pyplot.subplots(figsize=dims)
sns.distplot(df['Promedio_altura'], fit=stats.norm)
fig = plt.figure(figsize=(8, 6))
res = stats.probplot(df['Promedio_altura'], plot=plt)
correlaciones = df.corr()
ax = sns.heatmap(
    correlaciones,
    vmin=-1, vmax=1, center=0,
    cmap=sns.diverging_palette(20, 220, n=400),
    square=True)
ax.set_xticklabels(
    ax.get_xticklabels(),
    rotation=45,
    horizontalalignment='right')
ax.figsize=(12, 9)
```

#test de hipótesis sobre las variables

```
for i in df.columns:
        estat, pvalor = normaltest(i)

        print('Variable: %.3f. Estadistico de prueba=%.3f, p=%.3f' % (i, estat, pvalor))
        alfa = 0.05
        if p > alfa:
                print('La muestra se distribuye normalmente (No rechaza H_0)')
        else:
                print('La muestra no se distribuye normalmente (Rechaza H_0)')
```

Variable: Temperatura. Estadístico de prueba=33.194, p=0.000 La muestra no se distribuye normalmente (Rechaza H_0)

Variable: Temperatura_Max. Estadístico de prueba=25.330, p=0.000 La muestra no se distribuye normalmente (Rechaza H_0)

Variable: Temperatura_Min. Estadístico de prueba=14.615, p=0.001 La muestra no se distribuye normalmente (Rechaza H_0)

Variable: Precipitacion. Estadístico de prueba=572.995, p=0.000 La muestra no se distribuye normalmente (Rechaza H_0)

Variable: Presion. Estadístico de prueba=10.715, p=0.005 La muestra no se distribuye normalmente (Rechaza H_0)

Variable: Humedad. Estadístico de prueba=21.179, p=0.000 La muestra no se distribuye normalmente (Rechaza H_0)

Variable: Viento_velocidad. Estadístico de prueba=26.999, p=0.000 La muestra no se distribuye normalmente (Rechaza H_0)

Variable: Viento_velocidad_max. Estadístico de prueba=21.501, p=0.000 La muestra no se distribuye normalmente (Rechaza H_0)

Variable: Viento_velocidad_min. Estadístico de prueba=88.064, p=0.000 La muestra no se distribuye normalmente (Rechaza H_0)

Variable: Amplitud_termica. Estadístico de prueba=6.179, p=0.046 La muestra no se distribuye normalmente (Rechaza H_0)

Variable: Amplitud_viento. Estadístico de prueba=38.906, p=0.000 La muestra no se distribuye normalmente (Rechaza H_0)

Variable: Promedio_altura. Estadístico de prueba=513.559, p=0.000 La muestra no se distribuye normalmente (Rechaza H_0)

Variable: var_alt. Estadístico de prueba=1301.993, p=0.000 La muestra no se distribuye normalmente (Rechaza H_0)

#Modelado

```
df.Fecha = df.Fecha.values.astype(np.int64) // 10 ** 9
df.index.name = 'Fecha'

def series_to_supervised(data, n_in=1, n_out=1, dropnan=True):
    n_vars = 1 if type(data) is list else data.shape[1]
    df = pd.DataFrame(data)
    cols, names = list(), list()
    for i in range(n_in, 0, -1):
        cols.append(df.shift(i))
        names += [('var%d(t-%d)' % (j+1, i)) for j in range(n_vars)]
    for i in range(0, n_out):
        cols.append(df.shift(-i))
        if i == 0:
            names += [('var%d(t)' % (j+1)) for j in range(n_vars)]
```

```
        else:
            names += [('var%d(t+%d)' % (j+1, i)) for j in range(n_vars)]
    agg = pd.concat(cols, axis=1)
    agg.columns = names
    if dropnan:
        agg.dropna(inplace=True)
    return agg

values = df.values
encoder = LabelEncoder()
values[:,14] = encoder.fit_transform(values[:,14])
values = values.astype('float32')
scaler = MinMaxScaler(feature_range=(0, 1))
scaled = scaler.fit_transform(values)
reframed = series_to_supervised(scaled, 1, 1)
reframed.drop(reframed.columns[[15,16,17,18,19,20,21,22,23,24,25,26,27,28,29]], axis=1, inpla
ce=True)
values = reframed.values
t_step = 566
train = values[:t_step, :]
test = values[t_step:, :]
train_X, train_y = train[:, :-1], train[:, -1]
test_X, test_y = test[:, :-1], test[:, -1]
train_X = train_X.reshape((train_X.shape[0], 1, train_X.shape[1]))
test_X = test_X.reshape((test_X.shape[0], 1, test_X.shape[1]))

# Red LSTM
model = Sequential()

model.add(LSTM(128, activation='relu',input_shape=(train_X.shape[1], train_X.shape[2]),return
_sequences=True))

model.add(Dropout(0.2))

model.add(LSTM(64, input_shape=(train_X.shape[1], train_X.shape[2]),return_sequences=True
))

model.add(Dropout(0.2))

model.add(Dense(1))

opt = tf.keras.optimizers.Adam(lr=0.001, decay =1e-5)

model.compile(loss="mean_squared_error", optimizer=opt)

history = model.fit(train_X, train_y, epochs=250, batch_size=7, validation_data=(test_X, test_y),
 verbose=3, shuffle=False)
```

```
model.summary()
model.compile(loss='mae', optimizer='adam')
history = model.fit(train_X, train_y, epochs=200, batch_size=10, validation_data=(test_X, test_y
), verbose=0, shuffle=False)
pyplot.plot(history.history['loss'], label='train')
pyplot.plot(history.history['val_loss'], label='test')
pyplot.legend()
pyplot.show()
yhat = model.predict(test_X)
mse = mean_squared_error(test_y, yhat)
print("MSE: %.2f" % mse)
print("RMSE: %.2f" % (mse**(1/2.0)))
yhat = model.predict(test_X)
test_X = test_X.reshape((test_X.shape[0], test_X.shape[2]))
inv_yhat = concatenate((yhat, test_X[:, 1:]), axis=1)
inv_yhat = scaler.inverse_transform(inv_yhat)[:, [0]]
inv_yhat = inv_yhat[:,0]
test_y = test_y.reshape((len(test_y), 1))
inv_y = concatenate((test_y, test_X[:, 1:]), axis=1)
inv_y = scaler.inverse_transform(inv_y)
inv_y = inv_y[:,0]
rmse = numpy.sqrt(mean_squared_error(inv_y, inv_yhat))
print('Test RMSE: %.3f' % rmse)
```

#Modelo

```
Model: "sequential"
_________________________________________________________________
Layer (type)                 Output Shape              Param #
=================================================================
lstm (LSTM)                  (None, 1, 128)            73216
_________________________________________________________________
dropout (Dropout)            (None, 1, 128)            0
_________________________________________________________________
lstm_1 (LSTM)                (None, 1, 64)             49408
_________________________________________________________________
dropout_1 (Dropout)          (None, 1, 64)             0
_________________________________________________________________
dense (Dense)                (None, 1, 1)              65
=================================================================
Total params: 122,689
Trainable params: 122,689
Non-trainable params: 0
_________________________________________________________________
```

#Optimizacion de hiperparámetros

```
def model_LSTM(optimizer= tf.keras.optimizers.Adam(lr=0.001, decay =1e-
6), activation = 'relu', dropout_rate =0.0, weight_constraint = 0):
    model = Sequential()
    model.add(LSTM(128, activation='relu',input_shape=(train_X.shape[1], train_X.shape[2]),ret
urn_sequences=True))

model.add(Dropout(0.2))

model.add(LSTM(64, input_shape=(train_X.shape[1], train_X.shape[2]),return_sequences=True
))

   model.add(Dropout(0.2))

    model.add(Dense(1))

    model.add(Dense(1))
    model.compile(loss='mean_squared_error', optimizer=optimizer)
    return model

model = KerasRegressor(build_fn=model_LSTM, verbose=0)
batch_size = [x for x in np.arange(10,100,10)]
epochs = [x for x in np.arange(10,200,20)]
param_grid = dict(batch_size=batch_size, epochs=epochs)
grid = GridSearchCV(estimator=model, param_grid=param_grid, n_jobs=-1, cv=5)
grid_result = grid.fit(train_X, train_y)
print("Best: %f using %s" % (grid_result.best_score_, grid_result.best_params_))
best1=grid_result.best_params_
means = grid_result.cv_results_['mean_test_score']
stds = grid_result.cv_results_['std_test_score']
params = grid_result.cv_results_['params']
for mean, stdev, param in zip(means, stds, params):
    print("%f (%f) with: %r" % (mean, stdev, param))
```

#Proyecciones del modelo

Medición del RMSE (Datos normalizados)				
Y_pred	**Y_test**	**Y_pred - Y_test**	**(Y_pred - Y_test)^2**	**(Y_pred - Y_test)^2/N**
0,79993	0,87525	-0,07533	0,00567	0,00081
0,84840	0,91952	-0,07112	0,00506	0,00072
0,90032	0,83300	0,06733	0,00453	0,00065
0,67551	0,80684	-0,13133	0,01725	0,00246
0,64344	0,64185	0,00159	0,00000	0,00000
0,56546	0,59759	-0,03212	0,00103	0,00015
0,46511	0,55332	-0,08821	0,00778	0,00111

Conversión a los valores de la variable original				
Y_pred	**Y_test**	**Y_pred - Y_test**	**(Y_pred - Y_test)^2**	**(Y_pred - Y_test)^2/N**
4,01500	4,38023	-0,36523	0,13340	0,01906
4,25003	4,59486	-0,34483	0,11891	0,01699
4,50181	4,00000	0,50181	0,25181	0,03597
3,41172	4,04853	-0,63681	0,40552	0,05793
3,25626	3,24854	0,00772	0,00006	0,00001
2,87814	3,03391	-0,15576	0,02426	0,00347
2,39157	2,81928	-0,42771	0,18293	0,02613

	RMSE	MSE
Error en metros	0,3994	0,1596
Datos normalizados	0,0768	0,0059