

UNIVERSIDAD DISTRITAL FRANCISCO JOSÉ DE CALDAS



FACULTAD DE INGENIERÍA

ESPECIALIZACIÓN EN INGENIERÍA DE SOFTWARE

PROTOTIPO DE USO DE MACHINE LEARNING APLICADA EN EL
ANÁLISIS DE LOS FACTORES QUE INFLUYEN EN LA CALIDAD DEL
AIRE EN LA CIUDAD DE BOGOTÁ

Director:

SANDRO JAVIER BOLAÑOS
CASTRO

Revisora:

JORGE MARIO CALVO

Autores:

EDGAR JUNIOR CASTRO
ESCORCIA, CÉSAR ALFONSO
SOLANO RUIZ

ABRIL DE 2021

Índice general

0.1. Introducción	3
I Contextualización de la investigación	4
1. Descripción de la investigación	5
1.1. Título y definición del tema de investigación	5
1.2. Estudio del problema de investigación	5
1.2.1. Planteamiento del problema	5
1.2.2. Formulación del problema	6
1.2.3. Sistematización del problema	6
1.3. Objetivos de la investigación	6
1.3.1. Objetivo general	6
1.3.2. Objetivos específicos	6
1.4. Justificación	7
1.4.1. Justificación práctica	7
1.5. Hipótesis del trabajo	7
1.6. Marco referencial	7
1.6.1. Marco teórico	7
1.6.2. Marco conceptual	8
1.7. Aspectos metodológicos	15
1.7.1. Tipo de estudio	15
1.7.2. Metodología de investigación	15
1.7.3. Fuentes y técnicas para la recolección de información	16
II Desarrollo de la investigación	18
2. Desarrollo de la investigación	19
3. Arquitectura de la aplicación	20
4. Modelado de los datos	21
III Cierre de la investigación	22
5. Resultados y discusión	23
6. Conclusiones	24
6.1. Verificación, contraste y evaluación de los objetivos	24
6.2. Síntesis del modelo propuesto	24
6.3. Aportes originales	24
6.4. Trabajos o Publicaciones derivadas	24

7. Prospectiva del trabajo de grado	25
7.1. Líneas de investigación futuras	25
7.2. Trabajos de Investigación futuros	25

0.1. Introducción

Desde hace algunos años y con el gran crecimiento de las tecnologías de la información aplicadas a diferentes aspectos de nuestro diario vivir, se produce información conseguida desde los diferentes dispositivos empleados en las actividades, estas fuentes de información que al ser manejada de forma masiva se convierte en base de partida para establecer las diferentes comportamientos, tendencias y necesidades. Es una tendencia que se está extendiendo globalmente a las ciudades y que es conocido como "Smart City" consiste en la aplicación del big data y sus diferentes técnicas a la gobernabilidad, una visión futurista para que las ciudades, recursos y necesidades pueden ser administradas de forma tecnológica y con sentido social, para ello las ciudades hacen uso de las diferentes fuentes de información que siempre han existido pero que ahora están disponibles de forma digital y que puede ser cruzada con otras instituciones de forma rápida y eficaz, cosa que en el pasado era imposible realizar. En Colombia desde el año 2018 se estableció la política de explotación de datos Big data en el documento CONPES 3920 la cual tiene como objetivo "La presente política tiene por objetivo aumentar el aprovechamiento de datos, mediante el desarrollo de las condiciones para que sean gestionados como activos para generar valor social y económico. En lo que se refiere a las actividades de las entidades públicas, esta generación de valor es entendida como la provisión de bienes públicos para brindar respuestas efectivas y útiles frente a las necesidades sociales". El Distrito Capital cuenta con la Red de Monitoreo de Calidad del Aire de Bogotá – RMCAB, la cual permite recolectar información sobre la concentración de contaminantes y variables meteorológicas de forma continua en diferentes lugares de la ciudad. El proceso inicia con la medición de los parámetros en cada estación y el almacenamiento de estos en un datalogger, luego cada hora se actualizan los datos y son transmitidos hacia el servidor central, donde se almacena la información en una base de datos. Esta información es validada desde su captura en los equipos, hasta su revisión en la sede central, con el fin de obtener los datos para la elaboración de los informes periódicos de calidad del aire. En dichos reportes se evalúa el cumplimiento de los niveles máximos permisibles definidos por la Resolución 2254 del 2017 del Ministerio de Ambiente y Desarrollo Sostenible (MADS).

Bajo esta normativa y con las diferentes fuentes de datos abiertas "open data" que son ofrecidas por parte de la alcaldía mayor de Bogotá y las entidades gubernamentales sobre diferentes aspectos de socio-económicos se busca desarrollar un prototipo que permita el aprovechamiento de estos datos para obtener la información suficiente para definir planes de acción y toma de decisiones que permitan tomar acciones preventivas que beneficien la calidad del aire en la ciudad de Bogotá.

Parte I

Contextualización de la investigación

Capítulo 1

Descripción de la investigación

1.1. Título y definición del tema de investigación

El título propuesto para el trabajo es "PROTOTIPO DE USO DE MACHINE LEARNING APLICADA EN EL ANÁLISIS DE LOS FACTORES QUE INFLUYEN EN LA CALIDAD DEL AIRE EN LA CIUDAD DE BOGOTÁ".

1.2. Estudio del problema de investigación

1.2.1. Planteamiento del problema

El número de datos ambientales recolectados por las estaciones de monitoreo en la ciudad de Bogotá a cargo de La Red de Monitoreo de Calidad del Aire de Bogotá - RMCAB que hace parte de la Subdirección de Calidad del Aire, Auditiva y Visual de la Secretaría Distrital de Ambiente y está conformada por catorce (14) estaciones de las cuales trece (13) son fijas y una (1) estación móvil y cuyas mediciones son realizadas en tiempo real y en periodos cortos entre una medición y otra, garantizan la precisión de los indicadores de la calidad del aire, las concentraciones de material particulado (PM10 y PM2.5), la concentración de ozono (O3), la concentración de dióxido de nitrógeno (NO2), la concentración de dióxido de azufre (SO2) y la concentración de monóxido de carbono (CO) en el aire, permitiéndonos conocer el comportamiento de la calidad del aire en las distintas zonas de la ciudad.[1] Pero para registrar los cambios ocurridos, descubrir relaciones entre los datos y patrones en la dinámica de los fenómenos a través del volumen de datos obtenidos, exige que se usen herramientas informáticas para su almacenamiento y gestión que permitan tener un mecanismo eficiente de análisis de datos y extracción de información y conocimiento relevante para la toma de decisiones y la generación de alertas tempranas por medio de la predicción de comportamientos. Es decir, es necesario contar con un modelo de gestión de datos para el análisis y extracción de información concreta que ayude, tanto a la toma de decisiones como a la construcción de mecanismos de alerta temprana, basados en predicción de eventos a partir del estudio de los datos históricos calculados a partir de estos. Lo anterior se convierte en una oportunidad para la aplicación de técnicas de análisis de datos machine learning, el uso de datos masivos y sus técnicas analíticas para el diseño e implementación de políticas públicas que redunden en el beneficio de la sociedad. Es por esto que las técnicas de machine learning han estado inmersos en muchos ámbitos de nuestra vida sin siquiera saberlo, Marketing y ventas son quizá las áreas de mayor aplicación en la actualidad, pero también se ve en otros ámbitos, como en los procesos de negocios, cuantificación y optimización del rendimiento personal, mejoramiento de la salud pública y un uso sorprendente es el de análisis de datos masivos para predecir la evolución de las epidemias y brotes de enfermedades. Integrando datos de historiales clínicos con análisis de datos de redes sociales pueden detectar brotes de enfermedades en tiempo real simplemente escuchando lo que la gente publica en sus perfiles públicos etc. Todo esto nos lleva a cuestionarnos cómo los gobiernos llevan a cabo el aprovechamiento de los grandes volúmenes de información de sus datos para su gobernabilidad y toma de decisiones críticas, a la hora de invertir, ejecutar y proyectar

soluciones en materia medioambiental, que si no son vistas desde una óptica propositiva no llegan a tener el impacto esperado al no generar las soluciones necesarias y esperadas.

1.2.2. Formulación del problema

¿La aplicación de machine learning en fuentes de datos abiertas de las diferentes entidades gubernamentales puede producir resultados específicos para la toma de decisiones en el ámbito medioambiental de la ciudad de Bogotá?

1.2.3. Sistematización del problema

- ¿Se pueden encontrar dataset de tipo público que ofrecen las entidades gubernamentales de la ciudad de Bogotá para la obtención de información relevante?
- ¿Se pueden determinar las problemáticas de medio ambiente que suceden en la ciudad de Bogotá con la aplicación de machine learning sobre dataset de uso público?
- ¿Qué tecnologías asociadas a machine learning se pueden adaptar mejor para el tratamiento de grandes volúmenes de datos ambientales?
- ¿Se puede evaluar la eficacia del prototipo y determinar medidas de mejora para versiones posteriores?
- ¿Es posible aplicar el enfoque machine learning a los datos ambientales referentes a la calidad del aire y resolver los problemas presentes en la contaminación del aire?

1.3. Objetivos de la investigación

1.3.1. Objetivo general

Diseñar un prototipo de machine learning que permita crear un modelo de predicción mediante la utilización de datos históricos asociados a la contaminación del aire de la ciudad de Bogotá haciendo uso de dataset públicos ofrecidos por las instituciones distritales.

1.3.2. Objetivos específicos

- Identificar los dataset de las entidades estatales y nacionales que puedan ser usados para el desarrollo de nuestro prototipo como fuente de información para establecer las principales problemáticas del sector de medio ambiente en la ciudad de Bogotá.
- Identificar las fuentes de datos ambientales generados por estaciones de monitoreo ambiental que permitan mediante técnicas de inteligencia artificial el análisis predictivo de la calidad del aire.
- Realizar un análisis de los resultados obtenidos por parte del prototipo para determinar su nivel de eficacia.
- Determinar los modelos predictivos que puedan ser usados con mayor eficacia en la predicción de los niveles de crecimiento de los agentes contaminantes del aire en la ciudad.
- Caracterizar los modelos predictivos por su eficacia en la identificación de aumentos de agentes contaminantes.
- Determinar si es posible establecer una predicción del índice de los agentes contaminantes al año 2030.

1.4. Justificación

1.4.1. Justificación práctica

A lo largo de la última década se han visibilizado nuevas disciplinas en el campo tecnológico esto debido al crecimiento exponencial del uso de recursos tecnológicos en los diferentes ámbitos de nuestras vidas. A nivel mundial se ha dado la tendencia de automatización de procesos y análisis de datos, donde el objetivo es la identificación de patrones y tendencias en muchas diferentes fuentes de datos, una de las más destacada es el machine learning mediante el cual se puede explotar de forma más profunda las diferentes fuentes de datos para obtener la información más importante a la hora de tomar decisiones. En Colombia desde el 17 de abril de 2018 se estableció la política de explotación de datos Big data en el documento CONPES 3920 la cual tiene como objetivo “La presente política tiene por objetivo aumentar el aprovechamiento de datos, mediante el desarrollo de las condiciones para que sean gestionados como activos para generar valor social y económico. En lo que se refiere a las actividades de las entidades públicas, esta generación de valor es entendida como la provisión de bienes públicos para brindar respuestas efectivas y útiles frente a las necesidades sociales” [2], de acuerdo a esto se busca crear un prototipo o modelo que permita realizar la aplicación de machine learning en las diferentes fuentes de datos abiertas ofrecidas por las entidades estatales y nacionales, las que más tienen avances en la preparación de sus datos son: Red de Monitoreo de Calidad del Aire de Bogotá – RMCAB.[1] Esta información sumada a la problemática de La calidad del aire es uno de los factores de importancia en la determinación del índice de calidad de vida de un centro urbano. Una ciudad con buena calidad del aire es preferible para vivir y más atractiva para las inversiones al ser comparada con otras ciudades con condiciones similares de ingreso, acceso a bienes y servicios y oportunidades de empleo, pero con aire contaminado. Un aire de baja calidad es aquel que produce una evidencia perceptible o medida de poco bienestar, v.g.: visibilidad reducida, suciedad en edificaciones, afectaciones a la naturaleza o perjuicios sobre la salud. En centros urbanos con altas concentraciones de población y la alta ocurrencia de procesos productivos, la afectación a la salud resulta ser la consecuencia más importante de la contaminación del aire. [1] Es por lo que un análisis utilizando técnicas IA de machine learning permitirá catalizar la información obtenida y generar índices de valor para la toma de decisiones en materia ambiental a fin de solventar la problemática actual.

1.5. Hipótesis del trabajo

El adecuado aprovechamiento de los datos recolectados por las estaciones de monitoreo llevados a cabo por la Red de monitoreo de la calidad del aire de la ciudad de Bogotá permite tener un conocimiento en tiempo real de lo que está sucediendo con el aire en la ciudad, es por ello que contar con las herramientas y los métodos adecuados de análisis que permitan tomar acciones preventivas a través del aprendizaje automático la IA descarta por sí misma las acciones más arriesgadas y aquellas que pueden poner en riesgo el desarrollo normal de los factores económicos, sociales y culturales de la ciudad. Debido a esto y con base en el planteamiento del problema definido antes, se define la siguiente hipótesis: “Con la elaboración de un prototipo usando tecnologías de machine learning aplicado en el análisis de los factores que influyen en la calidad del aire, se aportará favorablemente en el mejoramiento de la calidad del aire de la ciudad a través de resultados de información que permitirán tener herramientas más confiables para la toma de decisiones de índole medioambiental.”

1.6. Marco referencial

1.6.1. Marco teórico

Los datos abiertos son información pública dispuesta en formatos que permiten su uso y reutilización bajo licencia abierta y sin restricciones legales para su aprovechamiento. En Colombia, la Ley 1712 de 2014 [3] sobre Transparencia y Acceso a la Información Pública Nacional, define los datos abiertos en el numeral sexto como “todos aquellos datos primarios o

sin procesar, que se encuentran en formatos estándar e interoperables que facilitan su acceso y reutilización, los cuales están bajo la custodia de las entidades públicas o privadas que cumplen con funciones públicas y que son puestos a disposición de cualquier ciudadano, de forma libre y sin restricciones, con el fin de que terceros puedan reutilizarlos y crear servicios derivados de los mismos”.

Por medio de una revisión preliminar se determinaron los principales conceptos asociados al tema planteado en esta propuesta, los cuales se presentan a continuación.

1.6.2. Marco conceptual

Ciencia de los datos

La ciencia de datos es el cuarto paradigma de la ciencia, después del experimento, la teoría y ciencias computacionales, abreviado como ciencia de datos, se refiere a la realización del análisis de datos como una ciencia empírica, aprendiendo directamente a partir de los datos mismos. Esto puede tomar la forma de recopilación de datos seguida de un análisis abierto sin preconcebidas hipótesis (a veces denominada descubrimiento o exploración de datos). El segundo método empírico se refiere a la formulación de una hipótesis, la recopilación de datos nuevos o preexistentes, para abordar la hipótesis, y la confirmación analítica o negación de la hipótesis (o la determinación de que se requiere información o estudio adicional). En ambos métodos, las conclusiones se basan en los datos.

En muchos proyectos de ciencia de datos, los datos originales son examinados primero, lo que informa una hipótesis, que luego se investiga. Como en cualquier ciencia experimental, el resultado Podría ser que la hipótesis original en sí necesita ser reformulada.

El concepto clave es que la ciencia de datos es una ciencia empírica, realizando el proceso científico directamente sobre los datos. Hay que tener en cuenta que la hipótesis puede ser por una necesidad, o puede ser la reafirmación de una necesidad en términos de una hipótesis técnica. La ciencia de datos es la extracción de conocimiento útil directamente de los datos a través de un proceso de descubrimiento, o de formulación de hipótesis y prueba de hipótesis.

La ciencia de datos está estrechamente vinculada al análisis de Big Data y se refiere a la gestión y ejecución de los procesos de datos de un extremo a otro, incluidos los comportamientos de los componentes del sistema de datos. Como tal, la ciencia de datos incluye toda la analítica, pero la analítica no incluye toda la ciencia de datos. La ciencia de datos contiene diferentes enfoques para aprovechar los datos para resolver las necesidades de la misión. Si bien el término ciencia de datos puede entenderse como las actividades en cualquier canal de análisis que produce conocimiento a partir de datos, el término se usa típicamente en el contexto de Big Data.

Componente ambiental

Las dinámicas de los recursos naturales y el medio ambiente generan una cantidad de datos considerable, los cuales son medidos, almacenados y analizados. Tratar estos datos exige de un gran esfuerzo, pero pueden contribuir de manera significativa a la generación de conocimiento, de patrones de comportamiento, de alertas de prevención y de políticas ambientales para apoyo a la toma de decisiones. Dentro de las variables medioambientales que se pueden monitorear y que generan datos de forma continua se encuentran las de la medición de la calidad del aire. El aire es una mezcla de gases que constituye la atmósfera que envuelve la tierra y es un factor indispensable para la vida. La presencia en el aire de una sustancia extraña o una variación significativa en la proporción de sus constituyentes susceptibles de provocar un efecto perjudicial, de origen antrópico, se conoce como contaminación atmosférica (Arnedo, et. al., 2005) fenómeno que se ha asociado con efectos nocivos para el entorno en general y para la salud humana en particular.

Calidad del aire

La quema de combustibles de biomasa y carbón para la preparación de alimentos, el uso de combustibles fósiles para la movilización de vehículos y para la producción industrial y el

desarrollo exponencial de compuestos químicos, se identifican como algunas de las causas de la contaminación del aire (Enkerlin, Cano, Garza y Vogel. 1997). Dentro de todos los contaminantes que existen en la atmósfera, se identificaron 5 contaminantes criterio que afectan a la salud inmediatamente desde su inhalación: monóxido de carbono (CO), dióxido de azufre (SO₂), dióxido de nitrógeno (NO₂), ozono troposférico (O₃) y material particulado con diámetro aerodinámico menor a 10 µm (PM₁₀). Además de éstos, se incluye al CO₂ (dióxido de carbono) por su aporte al efecto invernadero.

En los últimos diez años, las partículas totales en suspensión más importantes son las conocidas como PM₁₀ y PM_{2.5}, respectivamente. La razón fundamental de esta especificación se debe a que las partículas más pequeñas son más peligrosas para la salud de los seres humanos porque son capaces de alcanzar la zona inferior de los pulmones (OPS, 2005a). A continuación, se listan algunos de los efectos más importantes en la salud (OPS, 2005b) para los contaminantes como son material particulado, óxidos de azufre, óxidos de nitrógeno y ozono.

- Dióxido de azufre (SO₂). La exposición a SO₂ puede disminuir la función pulmonar, agravar enfermedades respiratorias preexistentes (especialmente bronquitis) y reducir la habilidad de los pulmones para liberar partículas extrañas.
- Dióxido de nitrógeno (NO₂). El monóxido de nitrógeno (NO) es relativamente inofensivo, sin embargo, el NO₂ puede causar problemas respiratorios principalmente en asmáticos y niños.
- Ozono (O₃). Cabe anotar que el ozono producido naturalmente en la estratosfera es beneficioso porque protege a la tierra y sus habitantes de la nociva radiación ultravioleta del sol.
- Material particulado (MP). Son las partículas sólidas o líquidas suspendidas en el aire, tienen una composición química diversa y su tamaño varía de 0.005 a 100 µm de diámetro aerodinámico; sus fuentes principales son la quema incompleta del combustible para motores diésel y los combustibles sólidos, como la madera y el carbón, también se puede producir por la condensación de vapores ácidos y compuestos orgánicos semi-volátiles, mediante una serie de complejas reacciones del NO₂ y SO₂ en la atmósfera que finalmente forman sales de nitratos y sulfatos.

Las partículas que más afectan la salud son aquellas con diámetro aerodinámico menor de 10 µm (PM₁₀) y, más aún, aquellas con diámetro aerodinámico menor de 2,5 µm (PM_{2.5}).

Aire y contaminación

El aire es una mezcla de gases que constituye la atmósfera que envuelve la tierra y es un factor indispensable para la vida. La presencia en el aire de una sustancia extraña o una variación significativa en la proporción de sus constituyentes susceptibles de provocar un efecto perjudicial, de origen antrópico, se conoce como contaminación atmosférica (Arnedo, et. al., 2005) fenómeno que se ha asociado con efectos nocivos para el entorno en general y para la salud humana en particular.

Red de monitoreo de calidad del aire de Bogotá

El Distrito Capital cuenta con la Red de Monitoreo de Calidad del Aire de Bogotá – RMCAB, que permite recolectar información sobre la concentración de contaminantes de origen antropogénico y natural y el comportamiento de las variables meteorológicas que regulan la distribución de los mismos en la atmósfera bogotana. Los datos recolectados en distintos sitios de la ciudad se reciben en una estación central donde se someten a un proceso de validación final y posterior análisis con el fin de evaluar el cumplimiento de los estándares de calidad de aire en Bogotá dados por la Resolución 610 del 24 de marzo de 2010 expedida por el entonces Ministerio de Ambiente, Vivienda y Desarrollo Territorial (MAVDT). Además resulta información base para la definición de las políticas de control de la contaminación y de la gestión ambiental. Para el año 2019 la RMCAB estaba conformada por catorce (14) estaciones de monitoreo de las cuales

doce (12) son fijas, una (1) meteorológica y una (1) móvil. Están ubicadas en sitios estratégicos de la ciudad, dotadas con equipos automáticos que permiten realizar un monitoreo continuo de las concentraciones de contaminantes criterio: material particulado (PM10, PM2.5), y gases contaminantes (SO₂, NO₂, CO, O₃), así como de las variables meteorológicas de Precipitación, Velocidad y Dirección del Viento, Temperatura, Radiación Solar, Humedad Relativa y Presión Barométrica. Los métodos de medición utilizados por la RMCAB se encuentran descritos en el CFR (Code of Federal Regulations) Título 40 que están aprobados por la Environmental Protection Agency (EPA) de Estados Unidos (U.S. Government Printing Office, 2014). El comportamiento del contaminante PM10 se puede observar en el gráfico de boxplot presentado en la Figura 2, el cual se ha usado históricamente como referencia para definir los niveles de contaminación y definir políticas de control ambiental.

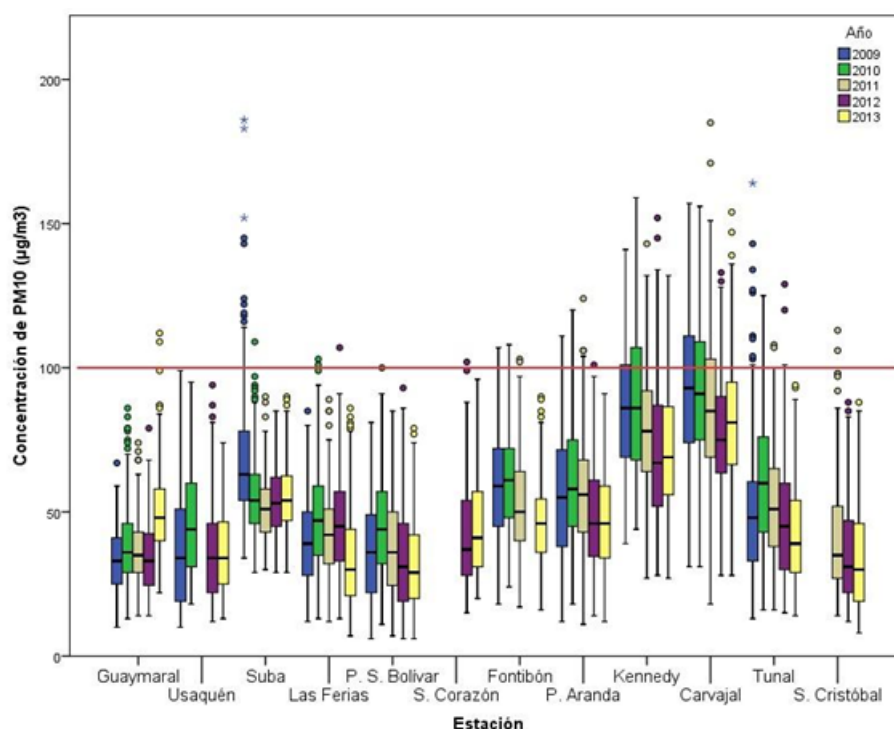


Figura 1.1: Gráfico boxplot

Se puede observar que en términos generales existe una tendencia a la reducción progresiva en los niveles de concentración en los últimos años y para la mayoría de estaciones. Sin embargo y comparando los dos últimos años 2012 y 2013, se puede apreciar que para las estaciones de San Cristóbal, Tunal, Fontibón, Parque Simón Bolívar y Las Ferias existe una reducción en los niveles de concentración; mientras que otras estaciones como Puente Aranda, Suba y Usaquén han permanecido constantes. Por su parte las estaciones Carvajal, Kennedy, Sagrado Corazón y Guaymaral presentaron aumentos en las concentraciones de 2013 en comparación a 2012 y además para el caso de Carvajal y Kennedy resultan los sitios de monitoreo en la ciudad con los mayores niveles de concentración con valores promedio de 81 y 71 µg/m³ y también en donde mayores excedencias a la norma diaria existen.

Información de contaminantes

SO₂ El dióxido de azufre es un gas incoloro y no inflamable, de olor fuerte e irritante. Su vida media en la atmósfera es corta de unos 2 a 4 días, y casi la mitad de las emisiones vuelven a depositarse en la superficie, mientras que el resto se transforma en iones sulfato (SO₄²⁻). Se trata de una sustancia reductora, que con el tiempo y en contacto con el aire y la humedad,

se convierte en trióxido de azufre. Es soluble en agua, formando una disolución ácida, y aún siendo inestable en estas condiciones, es capaz de formar sales como los sulfitos y bisulfitos.

PM10 Las PM10 se pueden definir como aquellas partículas sólidas o líquidas de polvo, cenizas, hollín, partículas metálicas, cemento o polen, dispersas en la atmósfera, y cuyo diámetro varía entre 2,5 y 10 μm (1 micrómetro corresponde la milésima parte de 1 milímetro).

PM2.5 Son partículas que tienen diámetro aerodinámico inferior o igual a los 2,5 micrómetros, es decir, son 100 veces más delgadas que un cabello humano. Además, el tamaño no es la única diferencia. Cada tipo de partículas está compuesto de diferente material y puede provenir de diferentes fuentes. En el caso de las PM2.5, su origen está principalmente en fuentes de carácter antropogénico como las emisiones de los vehículos diésel.

El monóxido de carbono, cuya fórmula química es CO, es un gas incoloro y altamente tóxico. Puede causar la muerte cuando se respira en niveles elevados. Se produce por la combustión deficiente de sustancias como gas, gasolina, queroseno, carbón, petróleo, tabaco o madera. Las chimeneas, las calderas, los calentadores de agua o calefactores y los aparatos domésticos que queman combustible, como las estufas u hornillas de la cocina o los calentadores a queroseno, también pueden producirlo si no están funcionando bien. Los vehículos con el motor encendido también lo expulsan. Grandes cantidades de CO se forman como subproducto durante los procesos oxidativos para la producción de productos químicos, lo que hace necesaria la purificación de los gases residuales.

Ozono El ozono es un gas incoloro e inestable de tres átomos de oxígeno (su fórmula química es O₃), además, es un oxidante fuerte, muy fácil de producir pero a la vez muy frágil y fácil de destruir. Este gas reacciona fácilmente con muchos compuestos químicos y es explosivo en pequeñas cantidades. En 1840 el gas fue bautizado como "ozono" (el olor) por el químico Christian Friedrich Schönbein, quien descubrió que esta sustancia se formaba durante descargas eléctricas. Muy pronto se descubrió que el ozono era un componente natural del aire. Se caracteriza por su olor peculiar, el cual puede ser detectado durante los episodios de tormentas eléctricas y en las proximidades de equipos eléctricos. Las descargas eléctricas son generalmente usadas en la producción de ozono para procesos industriales, tales como, la purificación de agua y aire y el blanqueamiento de textiles y productos alimenticios.

Dióxido de nitrógeno El dióxido de nitrógeno u óxido de nitrógeno (IV)₂ (NO₂), es un compuesto químico formado por los elementos nitrógeno y oxígeno, uno de los principales contaminantes entre los varios óxidos de nitrógeno. El dióxido de nitrógeno es de color marrón-amarillento. Se forma como subproducto en los procesos de combustión a altas temperaturas, como en los vehículos motorizados y las plantas eléctricas. Por ello es un contaminante frecuente en zonas urbanas.

Benceno El benceno es un hidrocarburo aromático de fórmula molecular C₆H₆, (originariamente a él y sus derivados se le denominaban compuestos aromáticos debido a la forma característica que poseen) también es conocido como benzol. En el benceno cada átomo de carbono ocupa el vértice de un hexágono regular, aparentemente tres de las cuatro valencias de los átomos de carbono se utilizan para unir átomos de carbono contiguos entre sí, y la cuarta valencia con un átomo de hidrógeno. Según las teorías modernas sobre los enlaces químicos, tres de los cuatro electrones de la capa de valencia del átomo de carbono se utilizan directamente para formar los enlaces covalentes típicos (2C-C y C-H) y el cuarto se comparte con los de los otros cinco átomos de carbono, obteniéndose lo que se denomina "la nube (π)" que contiene en diversos orbitales los seis electrones. El benceno es un líquido incoloro y muy inflamable de aroma dulce (que debe manejarse con sumo cuidado debido a su carácter cancerígeno), con un punto de ebullición relativamente alto.

Niveles máximos permisibles de contaminantes criterio en el aire

Resolución 2254 de 2017

Artículo 2°. Niveles máximos permisibles de contaminantes criterio. En la Tabla número 1 se establecen los niveles máximos permisibles a condiciones de referencia para contaminantes criterio que regirán a partir del primero de enero del año 2018:"

"Parágrafo 1°. A partir del 1° de julio de 2018, los niveles máximos permisibles de PM10 y PM2.5 para un tiempo de exposición 24 horas serán de 75 $\mu\text{g}/\text{m}^3$ y 37 $\mu\text{g}/\text{m}^3$ respectivamente."

Contaminante	Nivel máximo permisible ($\mu\text{g}/\text{m}^3$)	Tiempo de exposición
PM10	50	Anual
	100	24 horas
PM2.5	25	Anual
	50	24 horas
SO_2	50	24 horas
	100	1 hora
NO_2	60	Anual
	200	1 hora
O_3	100	8 horas
CO	5000	8 horas
	35000	1 hora

Cuadro 1.1: Niveles máximos permisibles de contaminantes criterio en el aire

"Parágrafo 2°. Para verificar el cumplimiento de los niveles máximos permisibles establecidos en la Tabla número 1, la concentración de los contaminantes del aire deberá evaluarse por cada punto de monitoreo. El promedio de concentraciones de diferentes puntos de monitoreo no será válido para evaluar el cumplimiento de dichos niveles."

"Parágrafo 3°. Las autoridades ambientales competentes deben realizar las mediciones de los contaminantes criterio establecidos en el presente artículo, de acuerdo con los procedimientos, frecuencias y metodologías establecidas en el Protocolo para el Monitoreo y Seguimiento de la Calidad del Aire."

"Parágrafo 4°. Con el objetivo de identificar fuentes de emisión, soportar investigaciones en salud ambiental y sus efectos sobre el clima, las autoridades ambientales competentes podrán monitorear las concentraciones en la atmósfera de carbono negro y otros contaminantes climáticos. Los resultados que se generen por parte de las autoridades ambientales deberán reportarse al Subsistema de Información sobre Calidad del Aire (SISAIRE)."

"Artículo 3°. Niveles Máximos Permisibles a 2030. En la Tabla número 2 se establecen los niveles máximos permisibles a condiciones de referencia para contaminantes criterio que regirán a partir del 1° de enero del año 2030"

Contaminante	Nivel máximo permisible ($\mu\text{g}/\text{m}^3$)	Tiempo de exposición
PM10	30	Anual
PM2.5	15	Anual
SO_2	20	24 horas
NO_2	40	Anual

Cuadro 1.2: Niveles máximos permisibles de contaminantes en el aire para el año 2030

"Lo dispuesto en los parágrafos 1° y 2° del artículo 2° será aplicable al presente artículo."

"Parágrafo. Para los contaminantes y tiempos de exposición que no se encuentren en la Tabla número 2 se mantendrán los establecidos en la Tabla No. 1 de la presente resolución."

"Artículo 4°. Niveles máximos permisibles de contaminantes tóxicos del aire. En la Tabla número 3 se establecen los niveles máximos permisibles a condiciones de referencia para contaminantes tóxicos del aire, teniendo en cuenta sus efectos adversos en la salud humana y el ambiente. Estos niveles regirán a partir del 1° de enero de 2018."

Machine Learning

Pero, ¿qué es exactamente Machine Learning?. El Machine Learning es el diseño y estudio de las herramientas informáticas que utilizan la experiencia pasada para tomar decisiones futuras; es el estudio de programas que pueden aprender de los datos. El objetivo fundamental del Machine Learning es generalizar, o inducir una regla desconocida a partir de ejemplos donde

Contaminantes tóxicos	Nivel máximo permisible ($\mu\text{g}/\text{m}^3$)	Tiempo de exposición
Plomo y sus compuestos	0.5	Anual
Cadmio	0.005	Anual
Benceno	5	Anual
Mercurio inorgánico(vapores)	1	Anual
Tolueno	260 1000	1 Semana 30 Minutos
Níquel y sus compuestos	260	Anual
Hidrocarburos aromáticos Policíclicos expresados como benzo (a) pireno	0.001	Anual

Cuadro 1.3: Niveles máximos permisibles de contaminantes tóxicos en el aire

esa regla es aplicada. El ejemplo más típico donde podemos ver el uso del Machine Learning es en el filtrado de los correos basura o spam. Mediante la observación de miles de correos electrónicos que han sido marcados previamente como basura, los filtros de spam aprenden a clasificar los mensajes nuevos. El Machine Learning combina conceptos y técnicas de diferentes áreas del conocimiento, como las matemáticas, estadísticas y las ciencias de la computación; por tal motivo, hay muchas maneras de aprender la disciplina.

Tipos de machine learning

El Machine Learning tiene una amplia gama de aplicaciones, incluyendo motores de búsqueda, diagnósticos médicos, detección de fraude en el uso de tarjetas de crédito, análisis del mercado de valores, clasificación de secuencias de ADN, reconocimiento del habla y del lenguaje escrito, juegos y robótica. Pero para poder abordar cada uno de estos temas es crucial en primer lugar distinguir los distintos tipos de problemas de Machine Learning con los que nos podemos encontrar.

Aprendizaje supervisado En los problemas de aprendizaje supervisado se enseña o entrena al algoritmo a partir de datos que ya vienen etiquetados con la respuesta correcta. Cuanto mayor es el conjunto de datos más el algoritmo puede aprender sobre el tema. Una vez concluido el entrenamiento, se le brindan nuevos datos, ya sin las etiquetas de las respuestas correctas, y el algoritmo de aprendizaje utiliza la experiencia pasada que adquirió durante la etapa de entrenamiento para predecir un resultado. Esto es similar al método de aprendizaje que se utiliza en las escuelas, donde se nos enseñan problemas y las formas de resolverlos, para que luego podamos aplicar los mismos métodos en situaciones similares.

Aprendizaje no supervisado En los problemas de aprendizaje no supervisado el algoritmo es entrenado usando un conjunto de datos que no tiene ninguna etiqueta; en este caso, nunca se le dice al algoritmo lo que representan los datos. La idea es que el algoritmo pueda encontrar por si solo patrones que ayuden a entender el conjunto de datos. El aprendizaje no supervisado es similar al método que utilizamos para aprender a hablar cuando somos bebés, en un principio escuchamos hablar a nuestros padres y no entendemos nada; pero a medida que vamos escuchando miles de conversaciones, nuestro cerebro comenzará a formar un modelo sobre cómo funciona el lenguaje y comenzaremos a reconocer patrones y a esperar ciertos sonidos.

Aprendizaje por refuerzo En los problemas de aprendizaje por refuerzo, el algoritmo aprende observando el mundo que le rodea. Su información de entrada es el feedback o retroalimentación que obtiene del mundo exterior como respuesta a sus acciones. Por lo tanto, el sistema aprende a base de ensayo-error. Un buen ejemplo de este tipo de aprendizaje lo podemos encontrar en los juegos, donde vamos probando nuevas estrategias y vamos seleccionando y perfeccionando aquellas que nos ayudan a ganar el juego. A medida que vamos adquiriendo más práctica, el efecto acumulativo del refuerzo a nuestras acciones victoriosas terminará creando una estrategia ganadora.

Sobreajuste Como mencionamos cuando definimos al Machine Learning, la idea fundamen-

tal es encontrar patrones que podamos generalizar para luego poder aplicar esta generalización sobre los casos que todavía no hemos observado y realizar predicciones. Pero también puede ocurrir que durante el entrenamiento solo descubramos casualidades en los datos que se parecen a patrones interesantes, pero que no generalicen. Esto es lo que se conoce con el nombre de sobreajuste o sobreentrenamiento.

El sobreajuste es la tendencia que tienen la mayoría de los algoritmos de Machine Learning a ajustarse a unas características muy específicas de los datos de entrenamiento que no tienen relación causal con la función objetivo que estamos buscando para generalizar. El ejemplo más extremo de un modelo sobreajustado es un modelo que solo memoriza las respuestas correctas; este modelo al ser utilizado con datos que nunca antes ha visto va a tener un rendimiento azaroso, ya que nunca logró generalizar un patrón para predecir.

Como evitar el sobreajuste Como mencionamos anteriormente, todos los modelos de Machine Learning tienen tendencia al sobreajuste; es por esto que debemos aprender a convivir con el mismo y tratar de tomar medidas preventivas para reducirlo lo más posible. Las dos principales estrategias para lidiar con el sobreajuste son: la retención de datos y la validación cruzada.

En el primer caso, la idea es dividir nuestro conjunto de datos, en uno o varios conjuntos de entrenamiento y otro/s conjuntos de evaluación. Es decir, que no le vamos a pasar todos nuestros datos al algoritmo durante el entrenamiento, sino que vamos a retener una parte de los datos de entrenamiento para realizar una evaluación de la efectividad del modelo. Con esto lo que buscamos es evitar que los mismos datos que usamos para entrenar sean los mismos que utilizamos para evaluar. De esta forma vamos a poder analizar con más precisión como el modelo se va comportando a medida que más lo vamos entrenando y poder detectar el punto crítico en el que el modelo deja de generalizar y comienza a sobreajustarse a los datos de entrenamiento.

La validación cruzada es un procedimiento más sofisticado que el anterior. En lugar de solo obtener una simple estimación de la efectividad de la generalización; la idea es realizar un análisis estadístico para obtener otras medidas del rendimiento estimado, como la media y la varianza, y así poder entender cómo se espera que el rendimiento varíe a través de los distintos conjuntos de datos. Esta variación es fundamental para la evaluación de la confianza en la estimación del rendimiento. La validación cruzada también hace un mejor uso de un conjunto de datos limitado; ya que a diferencia de la simple división de los datos en uno de entrenamiento y otro de evaluación; la validación cruzada calcula sus estimaciones sobre todo el conjunto de datos mediante la realización de múltiples divisiones e intercambios sistemáticos entre datos de entrenamiento y datos de evaluación.

Pasos para construir un modelo de Machine Learning

Construir un modelo de Machine Learning, no se reduce solo a utilizar un algoritmo de aprendizaje o utilizar una librería de Machine Learning; sino que es todo un proceso que suele involucrar los siguientes pasos:

Recolectar los datos. Podemos recolectar los datos desde muchas fuentes, podemos por ejemplo extraer los datos de un sitio web o obtener los datos utilizando una API o desde una base de datos. Podemos también utilizar otros dispositivos que recolectan los datos por nosotros; o utilizar datos que son de dominio público. ¡El número de opciones que tenemos para recolectar datos no tiene fin! Este paso parece obvio, pero es uno de los que más complicaciones trae y más tiempo consume.

Preprocesar los datos. Una vez que tenemos los datos, tenemos que asegurarnos que tiene el formato correcto para nutrir nuestro algoritmo de aprendizaje. Es prácticamente inevitable tener que realizar varias tareas de preprocesamiento antes de poder utilizar los datos. Igualmente, este punto suele ser mucho más sencillo que el paso anterior.

Explorar los datos. Una vez que ya tenemos los datos y están con el formato correcto, podemos realizar un pre análisis para corregir los casos de valores faltantes o intentar encontrar a simple vista algún patrón en los mismos que nos facilite la construcción del modelo. En esta etapa suelen ser de mucha utilidad las medidas estadísticas y los gráficos en 2 y 3 dimensiones para tener una idea visual de cómo se comportan nuestros datos. En este punto podemos detectar valores atípicos que debamos descartar; o encontrar las características que más influencia tienen para realizar una predicción.

Entrenar el algoritmo. Aquí es donde comenzamos a utilizar las técnicas de Machine Learning

realmente. En esta etapa nutrimos al o los algoritmos de aprendizaje con los datos que venimos procesando en las etapas anteriores. La idea es que los algoritmos puedan extraer información útil de los datos que le pasamos para luego poder hacer predicciones.

Evaluar el algoritmo. En esta etapa ponemos a prueba la información o conocimiento que el algoritmo obtuvo del entrenamiento del paso anterior. Evaluamos que tan preciso es el algoritmo en sus predicciones y si no estamos muy conforme con su rendimiento, podemos volver a la etapa anterior y continuar entrenando el algoritmo cambiando algunos parámetros hasta lograr un rendimiento aceptable.

Utilizar el modelo. En esta última etapa, ya ponemos a nuestro modelo a enfrentarse al problema real. Aquí también podemos medir su rendimiento, lo que tal vez nos obligue a revisar todos los pasos anteriores.

Librerías de Python para machine Learning

Como siempre me gusta comentar, una de las grandes ventajas que ofrece Python sobre otros lenguajes de programación; es lo grande y prolifera que es la comunidad de desarrolladores que lo rodean; comunidad que ha contribuido con una gran variedad de librerías de primer nivel que extienden las funcionalidades del lenguaje. Para el caso de Machine Learning, las principales librerías que podemos utilizar son:

SCIKIT-LEARN Es la principal librería que existe para trabajar con Machine Learning, incluye la implementación de un gran número de algoritmos de aprendizaje. La podemos utilizar para clasificaciones, extracción de características, regresiones, agrupaciones, reducción de dimensiones, selección de modelos, o preprocesamiento. Posee una API que es consistente en todos los modelos y se integra muy bien con el resto de los paquetes científicos que ofrece Python. Esta librería también nos facilita las tareas de evaluación, diagnóstico y validaciones cruzadas ya que nos proporciona varios métodos de fábrica para poder realizar estas tareas en forma muy simple.

STATSMODELS Es otra gran librería que hace foco en modelos estadísticos y se utiliza principalmente para análisis predictivos y exploratorios. Al igual que Scikit-learn, también se integra muy bien con el resto de los paquetes científicos de Python. Si deseamos ajustar modelos lineales, hacer un análisis estadístico, o tal vez un poco de modelado predictivo, entonces Statsmodels es la librería ideal. Las pruebas estadísticas que ofrece son bastante amplias y abarcan tareas de validación para la mayoría de los casos.

PYMC Es un módulo de Python que implementa modelos estadísticos bayesianos, incluyendo la cadena de Markov Monte Carlo(MCMC). pyMC ofrece funcionalidades para hacer el análisis bayesiano lo más simple posible. Incluye los modelos bayesianos, distribuciones estadísticas y herramientas de diagnóstico para la covarianza de los modelos. Si queremos realizar un análisis bayesiano esta es sin duda la librería a utilizar.

NLTK Es la librería líder para el procesamiento del lenguaje natural o NLP por sus siglas en inglés. Proporciona interfaces fáciles de usar a más de 50 cuerpos y recursos léxicos, como WordNet, junto con un conjunto de bibliotecas de procesamiento de texto para la clasificación, tokenización, el etiquetado, el análisis y el razonamiento semántico.

Obviamente, aquí solo estoy listando unas pocas de las muchas librerías que existen en Python para trabajar con problemas de Machine Learning, los invito a realizar su propia investigación sobre el tema

1.7. Aspectos metodológicos

1.7.1. Tipo de estudio

La naturaleza de este trabajo es de tipo propositivo ya que se pretende proponer un patrón de aprendizaje para contrastar con otros patrones.

1.7.2. Metodología de investigación

Para la ejecución del proyecto se plantean cuatro ciclos, cada ciclo está asociado al proceso del desarrollo de los objetivos, estos son a su vez caracterizados por una serie de actividades. Los ciclos y actividades de la metodología se resumen a continuación.

1.7.3. Fuentes y técnicas para la recolección de información

Las principales fuentes de información identificadas son los datos e indicadores publicados por el observatorio ambiental de Bogotá, donde está la información clasificada por zonas rurales y urbanas, también por localidades; otra fuente identificada es la plataforma de datos abiertos de Bogotá donde se encuentra la categoría de ambiente y desarrollo sostenible y donde hay diferentes dataset que pueden llegar a ser útiles para el desarrollo del proyecto.

Ciclo	Objeto asociado	Tiempo de exposición
Ciclo 1: Obtención, revisión y refinamiento del marco conceptual y preparación de fuentes de datos para el modelo.	Identificar los data set de las entidades estatales y nacionales que puedan ser usados para el desarrollo de nuestro prototipo como fuente de información para establecer las principales problemáticas del sector de medio ambiente en la ciudad de Bogotá. Identificar las fuentes de datos ambientales generados por estaciones de monitoreo ambiental que permitan mediante técnicas de inteligencia artificial el análisis predictivo de la calidad del aire.	Extracción de los data sets de calidad del aire por hora de la localidad de Kennedy del 1 de enero de 2010 a 6 de abril de 2021. Análisis exploratorio de los datos.
Ciclo 2: Estudio, evaluación y comparación de las metodologías de machine learning para el desarrollo de pronósticos, análisis predictivos de información climática.	Determinar los modelos predictivos que puedan ser usados con mayor eficacia en la predicción de los niveles de crecimiento de los agentes contaminantes del aire en la ciudad.	Realizar un estudio de los diferentes métodos de machine learning que permitan realizar un pronóstico más exacto.
Ciclo 3: Diseño del modelo para análisis predictivo de datos de la calidad del aire	Caracterizar los modelos predictivos por su eficacia en la identificación de aumentos de agentes contaminantes.	Dividir la serie en conjuntos de entrenamiento y prueba. para poder comparar los resultados entre los modelos a probar. Realizar prueba de estacionaridad, toda vez es necesaria para la utilización del modelo VAR Transformación de la serie de entrenamiento de ser necesaria. Construir un modelo VAR sobre las series transformadas. Emplear la causalidad de Granger. Realizar un diagnóstico del modelo.
Ciclo 4: Construcción y validación del prototipo.	Determinar si es posible establecer una predicción del índice de los agentes contaminantes al año 2030.	Realizar un pronóstico utilizando el modelo finalmente elegido. Realizar una transformación inversa del pronóstico a la escala original. Realizar una evaluación del pronóstico

Cuadro 1.4: Ciclos de actividades

Parte II

Desarrollo de la investigación

Capítulo 2

Desarrollo de la investigación

Capítulo 3

Arquitectura de la aplicación

Capítulo 4

Modelado de los datos

Parte III

Cierre de la investigación

Capítulo 5

Resultados y discusión

Capítulo 6

Conclusiones

- 6.1. Verificación, contraste y evaluación de los objetivos
- 6.2. Síntesis del modelo propuesto
- 6.3. Aportes originales
- 6.4. Trabajos o Publicaciones derivadas

Capítulo 7

Prospectiva del trabajo de grado

7.1. Líneas de investigación futuras

7.2. Trabajos de Investigación futuros

Bibliografía

- [1] Manuel Fernandez. “Augmented-Virtual Reality: How to improve education systems”. En: *Higher Learning Research Communications* (2017). ISSN: 2157-6254. DOI: 10.18870/hlrc.v7i1.373.
- [2] Chris Dede. *Immersive interfaces for engagement and learning*. 2009. DOI: 10.1126/science.1167311.
- [3] Zhigeng Pan y col. “Virtual reality and mixed reality for virtual learning environments”. En: *Computers and Graphics (Pergamon)* (2006). ISSN: 00978493. DOI: 10.1016/j.cag.2005.10.004.
- [4] Nikoletta Zampeta Legaki y col. “The effect of challenge-based gamification on learning: An experiment in the context of statistics education”. En: *International Journal of Human Computer Studies* (2020). ISSN: 10959300. DOI: 10.1016/j.ijhcs.2020.102496.
- [5] Jorge Francisco Figueroa Flores. *Using Gamification to enhance second language learning*. 2015. DOI: 10.1344/der.2015.27.32-54.
- [6] Christian Hederich Martínez y Angela Camargo Uribe. “Revisión crítica del modelo de patrones de aprendizaje de J. Vermunt”. En: *Revista Colombiana de Educación* 1.77 SE - (mar. de 2019). DOI: 10.17227/rce.num77-9469. URL: <https://revistas.pedagogica.edu.co/index.php/RCE/article/view/9469>.
- [7] Ferdinand Rivera. “Introduction”. En: *Teaching and Learning Patterns in School Mathematics: Psychological and Pedagogical Considerations*. Dordrecht: Springer Netherlands, 2013, págs. 1-24. ISBN: 978-94-007-2712-0. DOI: 10.1007/978-94-007-2712-0_1. URL: https://doi.org/10.1007/978-94-007-2712-0_1.
- [8] Douglas; Clements y Julie. Sarama. *Learning and Teaching Early Math. The Learning Trajectories Approach*. 2014.
- [9] Alexandra González Aguña. “Patrones en aprendizaje: Concepto, aplicación y diseño de un patrón”. En: *Revista de Educación a Distancia (RED)* 0.31 SE - (jul. de 2015). URL: <https://revistas.um.es/red/article/view/232811>.
- [10] Paloma Sainz de Vicuña. “Didáctica de la Educación Infantil”. En: *Ciclo formativo de Grado Superior en Educación Infantil*. 2000. ISBN: 84-369-3375-3.
- [11] Elizabeth Andrea Tabla Chaves. *CIBERCULTURA Y EDUCACIÓN VIRTUAL*. 2020.
- [12] K.M. Kapp. *The Gamification of Learning and Instruction: Game-based Methods and Strategies for Training and Education*. Wiley, 2012. ISBN: 9781118191989. URL: <https://books.google.com.co/books?id=GLr81qqtELcC>.
- [13] Gabriela Kiryakova, Nadezhda Angelova y Lina Yordanova. “Gamification in education”. En: *Proceedings of 9th International Balkan Education y Science Conference*. 2014.
- [14] Gregory Kipper. “Chapter 1 - What Is Augmented Reality?” En: ed. por Gregory B T - *Augmented Reality* Kipper. Boston: Syngress, 2013, págs. 1-27. ISBN: 978-1-59749-733-6. DOI: <https://doi.org/10.1016/B978-1-59-749733-6.00001-2>. URL: <http://www.sciencedirect.com/science/article/pii/B9781597497336000012>.

Bibliografía

- [1] Liliana Patricia Santacruz Valencia. “Automatización de los procesos para la generación, ensamblaje y reutilización de objetos de aprendizaje”. En: (2005).