Escuela de Administración
Escuela de Ingeniería, Ciencia y Tecnología

Maestría en Business Analytics

Desarrollo de un sistema de analíticas para identificar y analizar patrones de comportamientos irregulares en actividades académicas en plataformas virtuales

Presentado por:

Daniel Hernández Gómez y Ángel David Murgano Cáceres

Bogotá, D.C. 28 de mayo de 2023

Escuela de Administración
Escuela de Ingeniería, Ciencia y Tecnología

Maestría en Business Analytics

Desarrollo de un sistema de analíticas para identificar y analizar patrones de comportamientos irregulares en actividades académicas en plataformas virtuales

Presentado por:

Daniel Hernández Gómez y Ángel David Murgano Cáceres

Bajo la dirección de:
Sandra Liliana Sánchez Castañeda

Bogotá, D.C. 28 de mayo de 2023

Contenido

Preliminares

Declaración de originalidad y autonomía

Declaramos bajo la gravedad del juramento, que hemos escrito el presente Proyecto Aplicado Empresarial (PAE), en la propuesta de solución a una problemática en el campo de conocimientos del programa de Maestría por nuestra propia cuenta y que, por lo tanto, su contenido es original.

Declaramos que hemos indicado clara y precisamente todas las fuentes directas e indirectas de información y que este PAE no ha sido entregado a ninguna otra institución con fines de calificación o publicación.

Daniel Hernández Gómez

Ángel David Murgano Cáceres

Firmado en Bogotá, D.C. el 28 de mayo de 2023

Declaración de exoneración de responsabilidad

Declaramos que la responsabilidad intelectual del presente trabajo es exclusivamente de sus autores. La Universidad del Rosario no se hace responsable de contenidos, opiniones o ideologías expresadas total o parcialmente en él.

Daniel Hernández Gómez

Ángel David Murgano Cáceres

Firmado en Bogotá, D.C. el 28 de mayo de 2023

Lista de figuras

Lista de tablas

Abreviaturas

ACP	Análisis de Componentes Principales
IES	Institución de Educación Superior
IESoe	Institución de Educación Superior objeto de estudio
KPI	Key Performance Indicator o Indicador Clave de Rendimiento
LMS	Learning Management System o Sistema de Gestión de Aprendizaje

Resumen ejecutivo

Actualmente la digitalización de las actividades académicas permite acceder a herramientas que agilizan los procesos de evaluación de estudiantes, e igualmente ponen a disposición de éstos servicios, software o aplicativos que cuando son utilizados de manera inapropiada brindan resultados de manera ventajosa e incluso fraudulenta. La suplantación, entendida como uno de los tipos específicos de fraude académico (García Villegas et al., 2009), puede generar un impacto no deseado en la adecuada validación de las competencias del estudiante en formación, y para la institución de educación superior correr el riesgo de entregarle a la sociedad un profesional titulado y acreditado que en la práctica evidencia incompetencia, con todos los riesgos y las consecuencias que la aplicación de un saber bajo esta condición puede generar. Estructurar una herramienta basada en una serie de algoritmos, alimentada con datos resultantes de actividades académicas, que identifique patrones de comportamiento por usuario, exponga Indicadores Claves de Procesos (KPI's) y ayude a identificar anomalías que puedan surgir en los procesos de evaluación, puede permitir a las instituciones educativas contar con mecanismos para la mejora continua en la calidad de la formación de pregrado en la educación superior robusteciendo de ese modo la importancia que tiene para un perfil profesional una esperada actitud ética por parte del estudiante universitario; y que la IESoe, como entidad con la que esperamos nos brinde su apoyo a este proyecto, finalmente espera de sus alumnos tal y como lo tiene expresado en sus estatutos.

Palabras clave

Educación virtual, fraude académico, suplantación, indicadores, anomalías.

Abstract

Currently, the digitization of academic activities allows access to tools that speed up student evaluation processes, and make available to these services, software, or applications that, when used inappropriately, provide results in an advantageous and even fraudulent manner. Impersonation, understood as one of the specific types of academic fraud (García Villegas et al., 2009), can generate an unwanted impact on the adequate validation of the skills of the student in training, and for the higher education institution to run the risk of giving society a qualified professional and accredited that in practice it evidences incompetence, with all the risks and consequences that the application of knowledge under this condition can generate. By structuring a tool based on a series of algorithms, fed with data resulting from academic activities, that identifies behavior patterns by user, exposes Key Process Indicators (KPI's) and helps identify anomalies that may arise in the evaluation processes; allow educational institution to have mechanisms for continuous improvement in the quality of undergraduate training in higher education, thus strengthening the importance of an expected ethical attitude on the part of the university student for a professional profile; and that the IESoe, as an entity with which we hope to give us its support for this project, finally expects from its students as it has expressed in its organic statute.

Keywords

Virtual education, academic fraud, impersonation, indicators, anomalies.

1. Introducción

A mediados del 2018 el periódico El Heraldo de Barranquilla presentó una noticia titulada “Estudiantes en líos por fraude: ¿qué pasa con la ética de los jóvenes?” (Patiño M. & Iguarán, 2018). La noticia mencionaba el caso de jóvenes que suplantaron a aspirantes a ingresar a un programa de pregrado en la Universidad del Magdalena durante la realización del examen de admisión. A este grupo de jóvenes las autoridades los denominaron “Los Intelectuales” debido a que estos jóvenes eran beneficiarios de lo que en su momento se denominó el programa “Ser Pilo Paga”, donde jóvenes con buen rendimiento académico eran beneficiarios de becas y facilidades para el acceso a la educación superior. En este caso, fueron cuatro personas que pertenecían a ese programa las que participaron de la suplantación del examen de admisión gracias a su capacidad académica. Al parecer a cada uno de ellos se les pagó entre 20 y 24 millones de pesos por cometer el fraude (Patiño M. & Iguarán, 2018).

La noticia, más que describir la manera en que actuaba este grupo de jóvenes, planteaba el interrogante sobre el por qué estudiantes que ya se encontraban beneficiados por parte del estado para estudiar toda su carrera se exponen a perder un importante beneficio al cometer este delito. Lo destacado en ese momento de esa noticia no sólo fue el acto de suplantación en sí, sino la carrera a la cual aspiraban quienes pagaron por cometer ese ilícito: Medicina.

Este hecho fue detectado gracias a que se hizo seguimiento a la manera en que quienes presentaban las identificaciones en los exámenes, adulteraban la foto de las cédulas de ciudadanía de los aspirantes, algo que una persona encargada de supervisar la prueba pudo detectar al momento de validar la identidad de cada aspirante.

Pero ¿Qué hubiera sucedido si esta prueba fuera presentada de manera virtual como consecuencia del aislamiento que generó la pandemia de COVID-19?, ¿Cómo se detectaría este tipo de fraude cuando la persona que va a ser evaluada se encuentra detrás de un computador sin una supervisión presencial?, ¿Es posible hacer seguimiento al comportamiento de la persona que va a ser evaluada por medios virtuales?

Actualmente instituciones de educación superior ofrecen formación universitaria a través de plataformas virtuales. Este tipo de formación está soportada en tecnologías LMS, dentro de las que se destaca el aplicativo de sistema libre Moodle [1], utilizado por varias instituciones de educación superior en Colombia. Este aplicativo tiene múltiples herramientas para realizar la labor de creación de contenido, generar actividades de evaluación y calificarlas de manera automática. Y a la par de la aparición de las necesidades en el sector académico, la comunidad adscrita a este aplicativo a nivel mundial ofrece soluciones mediante programas complementarios (Plug-Ins) para la prevención del fraude académico. Sin embargo, en lo hasta ahora investigado, gran parte de las soluciones están dirigidas al fraude académico del tipo plagio. Es probable que ese énfasis en combatir el plagio esté dado debido a la insistente observación mundial por el respeto de los derechos de autor, como también a los diferentes escándalos mediáticos a nivel mundial donde figuras públicas has sido señaladas de cometer este tipo de acción para obtener títulos académicos.

Es claro que el plagio no es el único tipo de fraude académico, pero si el más observado. Esto hace que sea necesario conocer otros tipos de fraude académico que existen y determinar si existen herramientas en los sistemas LMS que permitan detectar esos otros tipos de fraude. Para determinar lo anterior, es necesario indagar si existe un estudio respecto

[1] https://moodle.org/

al fraude académico en Colombia y si en ese estudio se ha determinado una clasificación de los tipos de fraude.

Mauricio García Villegas, profesor de la facultad de Derecho de la Universidad Nacional de Colombia e investigador del Centro de Estudios de Derecho, Justicia y Sociedad (Dejusticia) [2] publicó el libro "Normas de Papel: La cultura del incumplimiento de reglas" (García Villegas et al., 2009) donde se desarrolla el tema del incumplimiento en América Latina, sus orígenes y como se ve reflejado ese comportamiento en la sociedad actual. Como parte del desarrollo del tema en mención, uno de los estudios de caso se refiere específicamente al fraude académico comparando dos universidades colombianas (García Villegas et al., 2009). En el desarrollo de este caso, se realizó una investigación mediante una encuesta donde se tomaron como referencia las conductas que están tipificadas como fraude académico en el reglamento de una de las universidades que se estudiaron. Para el presente proyecto, se tomarán las clasificaciones realizadas en ese estudio (García Villegas et al., 2009) con respecto a los tipos de fraude que se pueden presentar en actividades académicas, junto con las clasificaciones de los motivos que se tienen tanto para cometer, como para no cometer fraude que se encuentran enunciadas en el instrumento de encuesta. Esto, con el fin de establecer las tipologías de lo que se entiende como fraude académico.

En el artículo de 2017 "Fraude académico en universitarios en Colombia: ¿Qué tan crónica es la enfermedad?" (Martínez & Ramírez, 2017) los investigadores de la universidad ICESI Enrique Ramírez y Lina Martínez, profundizaron en el estudio del fraude académico realizado por el grupo de trabajo del profesor García. De este estudio, se quiere resaltar los siguientes planteamientos:

---

[2] https://www.dejusticia.org/responsible/mauricio-garcia-villegas/

- En la medida que se presentan avances en tecnologías de la información, aumenta la tipología de conductas consideradas fraudulentas.
- Los métodos de evaluación no han cambiado a la par con las nuevas tecnologías.
- En el análisis de costo beneficio, las razones que tienen los estudiantes para cometer fraude son principalmente:
    - La percepción de tener una baja probabilidad de tener un castigo fuerte y ejemplarizante
    - La percepción que se tiene de una baja probabilidad de ser detectados por los profesores

(Martínez y Ramírez, 2017)

Estos planteamientos sustentan la justificación de este proyecto, pues existen brechas de formación en tecnologías de información entre la generación de estudiantes jóvenes, muchos de ellos al día con los más recientes adelantos tecnológicos, y la generación de docentes que los evalúan.

Son para estos últimos a quienes se les espera brindar una herramienta de análisis estadístico que les permita detectar comportamientos que podrían hacer referencia a fraude académico, y suministrar evidencias para que la IESoe determine la existencia del fraude, específicamente del tipo suplantación, para contribuir con el cumplimiento de los estándares de calidad de la educación colombiana

2. Objetivos

Objetivo General

Proveer a educadores con una herramienta de información analítica que les apoye al momento de revisar los resultados de evaluaciones en entornos virtuales y detectar anomalías en los comportamientos de estudio y evaluación de los estudiantes.

Objetivos Específicos

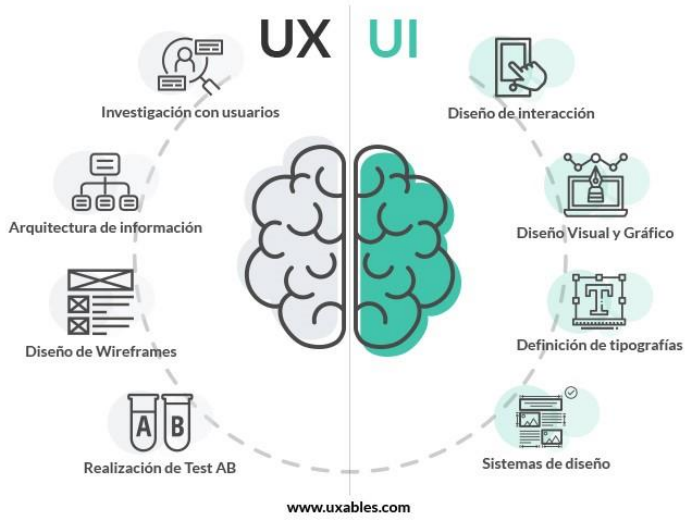
- Identificar la tipología existente en el sistema educativo de Colombia relacionada con la clasificación de los tipos de fraude académico, qué motiva a realizarlo, reconocimiento de sus consecuencias y su soporte normativo.
- Desarrollar Indicadores Claves de Proceso (KPIs) que faciliten la detección de comportamientos anómalos en un registro de información de actividades académicas.
- Evaluar la fortaleza que ofrecen los datos analizados para servir como evidencia indiscutible de comportamientos anómalos relacionados con el fraude académico, y para el caso del presente proyecto, específicamente de suplantación.
- Explorar la viabilidad de establecer una propuesta de análisis predictivo que permita anticiparse a comportamientos tendientes a realizar fraude académico.

3. Alcance del proyecto aplicado

Para el proyecto se desarrolló, de acuerdo con el Plan de Trabajo, dos (2) versiones de la herramienta para el análisis de datos, con objetivo de que el educador tenga evidencia sólida que le permita identificar comportamientos que podrían tratarse de fraude académico al interior de la IESoe y contribuir con la calidad del perfil de sus estudiantes. Estas versiones son: "versión 1" compuesto por un backend con cargue de reportes de Moodle y un frontend realizado en Power BI como primera aproximación a la interfaz de usuario y "versión 2" la cual mejora el concepto de storytelling en el frontend desplegado.

Para ello es importante que los datos-evidencia que arroje la herramienta sean fáciles de comprender. De ahí que el planteamiento en el Plan de Trabajo inicialmente fue el tratamiento de datos a nivel estadístico, para al final del proyecto concentrarse en su clara visualización y lectura, lo que se conoce como enfoque en la interfaz de usuario (UI) y la Experiencia de Usuario (UX) (ver Figura 1).

Figura 1 *Diferencias entre UI y UX*



Fuente: (Arias del Prado, 2020)

4. Cronograma

Tabla 1 *Cronograma*

A. Actividades iniciales	**20/may./2022**	**28/jun./2022**
U. Rosario	20/may./2022	01/jun./2022
Solicitud y aceptación Directora de Trabajo Diligenciar y compartir formatos para la aceptación y firma de directora: - Carta Compromiso - Acuerdo de Confidencialidad	20/may./2022	22/may./2022
Oficialización a Dirección de Maestría Inicial Generar carta formal firmada donde Estudiantes y Directora oficializan frente a la directiva su compromiso para el proyecto. Resultado (03/06/22) Correo de hoy del Director Daniel Diaz da el visto bueno para continuar con lo ya entregado (formatos de la directora firmados y propuesta ya cargada en E aulas)	27/may./2022	01/jun./2022
IESoe En este aparte se relacionan las gestiones realizadas con la IESoe para formalizar el proyecto.	03/jun./2022	28/jun./2022
Solicitud acceso información IESoe Se dirigió carta a la Dirección de Innovación y Virtualidad de la IESoe mostrando los formatos que serán presentados para la aprobación del contenido. Se adjuntan los formatos enviados por el profesor Carlos Franco para los acuerdos de confidencialidad y de presentación por parte de U. Rosario	03/jun./2022	03/jun./2022
Respuesta IESoe a Ethical Tracker Correo de respuesta de la IESoe: "Cordial saludo, profesor Daniel. Para revisar la petición requerimos: -Copia del proyecto de investigación con carta avalada por el docente tutor. -Carta de presentación de los estudiantes por parte de la maestría. -Documento detallado con la información a la que solicitan acceso y el uso que se le dará a la misma. Quedo atenta, Cordialmente, "	10/jun./2022	10/jun./2022
Respuesta Ethical Tracker a IESoe Respuesta “Buenas tardes y gracias por su respuesta. Con respecto a los documentos -Copia del proyecto de investigación con carta avalada por el docente tutor. R./ Estamos terminando el primer semestre de la maestría y en 2 semanas estaríamos presentando el anteproyecto. ¿Este documento sería el solicitado? -Carta de presentación de los estudiantes por parte de la maestría. R./ Ya contamos con este documento por parte de la universidad.	13/jun./2022	13/jun./2022

Solo necesitaríamos saber a quién estaría dirigida esa carta (área y/o persona) para solicitar la formalización de esta. -Documento detallado con la información a la que solicitan acceso y el uso que se le dará a la misma. R./ Lo vamos a construir de la manera más adecuada para claridad de las partes. Quedo atento a sus comentarios		
Respuesta 2 IESoe a Ethical Tracker Responde: "Buenas tardes profe, respondo punto a punto: 1. Sí profe, ese es el documento solicitado. 2. Debería en un principio estar dirigido a la Dirección. 3. Porfa, profe, este documento es clave. Quedo atenta a cualquier inquietud."	14/jun./2022	14/jun./2022
Reenvío información a nuevo contacto en la IESoe En la IESoe, la persona del cargo de contacto que hace seguimiento a la solicitud fue reemplazada. Se hace contacto personalmente y se reenvía la información previamente tratada para continuar con el proceso. Se solicita de nuevo confirmación de los documentos requeridos.	16/ago./2022	16/ago./2022
Respuesta 3 IESoe a Ethical Tracker Nuevo contacto confirma que documentos solicitados previamente son los requeridos. Se solicita un documento adicional explicando de manera detallada el manejo que se espera dar a los datos.	17/ago./2022	17/ago./2022
Envío documento a Director de Investigación U. Rosario para firma Se envía para firma a la Dirección de Investigación de la Universidad del Rosario el modelo de Carta de Presentación para presentar ante la IESoe.	23/sept./2022	23/sept./2022
Recibo de la Carta de Presentación firmada Recibo de la Carta de Presentación firmada por parte de la Dirección de Investigación de la Universidad del Rosario.	28/sept./2022	28/sept./2022
Envío documentos solicitados por IESoe Envío de los documentos solicitados por la IESoe.	06/oct./2022	06/oct./2022
Respuesta 4 IESoe a Ethical Tracker Se solicita un documento adicional no tan formal aclarando la "Información solicitada y manejo de datos" que solicita la IESoe para seguir con el proceso.	03/nov./2022	03/nov./2022
Envío documento adicional solicitado por IESoe Se hace envío del documento extra para solicitud de información.	03/nov./2022	03/nov./2022
Respuesta 5 IESoe a Ethical Tracker Al interior de la IESoe la solicitud fue enviada al Vicerrector de Servicios Digitales y al Director Jurídico de la Secretaria General de la institución e informan que no es viable la solicitud.	11/nov./2022	11/nov./2022
Pendiente: Envío de respuesta del Director Jurídico a Director de Innovación Reenviar respuesta del Director Jurídico al Director de Emprendimiento e Innovación para solicitarle mediación sobre el acceso y uso de los datos, remarcando que en los datos solicitados no se pide información de los estudiantes (nombres e identificación) y que este trabajo procura presentarle una herramienta de atención a la misma IESoe para apoyo a la detección de prácticas de suplantación.	12/dic./2022	14/dic./2022
B. Proyecto Ethical Tracker	**28/may./2022**	**01/abr./2023**

Semestre 1	**28/may./2022**	**24/jun./2022**
Doc 0 - Propuesta Proyecto Empresarial Entrega del formato "PROPUESTA PROYECTO EMPRESARIAL– MAESTRÍA EN BUSINESS ANALYTICS" Con la información inicial del tema seleccionado.	28/may./2022	28/may./2022
Doc 1 - Anteproyecto	01/jun./2022	24/jun./2022
Bibliografía de Contexto Revisión de la bibliografía pertinente para dar contexto al tema del proyecto.	01/jun./2022	14/jun./2022
Resumen Ejecutivo e Introducción **Anteproyecto** - Resumen Ejecutivo - Introducción	15/jun./2022	15/jun./2022
Objetivos Determinar - Objetivos Generales - Objetivos Específicos	16/jun./2022	16/jun./2022
Alcance del Proyecto y Fuentes de Información	17/jun./2022	17/jun./2022
Plan de Trabajo Elaboración del cronograma de actividades	18/jun./2022	18/jun./2022
Entrega de Anteproyecto Entrega del documento Anteproyecto.	21/jun./2022	21/jun./2022
Sustentación de avance Presentación del Anteproyecto.	25/jun./2022	25/jun./2022
Doc 1 - Análisis Estadístico (Anexo B) Información acerca del Análisis Estadístico inicial como Anexo Al documento del Anteproyecto.	01/jun./2022	20/jun./2022
Tratamiento Inicial de Información Depuración de la información inicial ejemplo de estudio. - Método de obtención - Anonimización de datos personales - Preprocesamiento	01/jun./2022	08/jun./2022
Análisis Inicial Análisis Estadístico Inicial sobre los datos disponibles para identificar comportamientos de estos.	09/jun./2022	15/jun./2022
Confirmar hipótesis sesgo Confirmar hipótesis de sesgo en la normal por valores Outliers (relacionados a fraude académico).	16/jun./2022	18/jun./2022
Patrones de comportamiento Identificación de patrones de comportamiento con apoyo de los datos.	19/jun./2022	20/jun./2022
Semestre 2	**06/ago./2022**	**09/dic./2022**
Sistemas de Información	06/ago./2022	20/ago./2022

Análisis y comprensión del sistema de información del cliente y su gobernanza.		
Analítica del Negocio - Identificación de las analíticas de negocio relacionadas con el tema del proyecto.	27/ago./2022	17/sept./2022
Doc 2 - Anteproyecto Actualizado Entrega de documento de Anteproyecto actualizado que incluye: - Sistemas de Información - Analítica del Negocio	24/sept./2022	24/sept./2022
Anteproyecto Actualizado (Presentación) Presentación del Anteproyecto Actualizado.	01/oct./2022	01/oct./2022
Analítica Predictiva Elaboración de una propuesta de viabilidad de análisis predictivo sobre la información del negocio trabajada hasta el momento.	08/oct./2022	29/oct./2022
Presentación de Avance Presentación de Avance que incluye: - Propuesta de la viabilidad de análisis predictivo en el negocio.	12/nov./2022	12/nov./2022
1a Versión herramienta para toma de decisiones Elaboración de propuesta al cliente de primera versión de la herramienta de análisis para Toma de decisiones.	19/nov./2022	03/dic./2022
Sustentación final semestre Presentación y sustentación de la 1a versión de la herramienta de análisis.	10/dic./2022	10/dic./2022
Semestre 3	**18/feb./2023**	**17/jun./2023**
Gestión de Productos Analíticos Compilación de los aportes que la asignatura de Gestión de Productos Analíticos brinda para la elaboración de proyectos de analítica y observar su pertinencia en el presente proyecto.	18/feb./2023	25/mar./2023
Análisis de Riesgos Compilación de los aportes que la asignatura Análisis de Riesgos brinda para la revisión de los riesgos y toma de decisiones en escenarios de incertidumbre.	18/feb./2023	10/mar./2023
Presentación de Avance Presentación del avance del proyecto según las indicaciones que sean dadas desde la maestría.	11/mar./2023	11/mar./2023
Trabajo de análisis para interfaz UX / UI Compilación de los aportes que la asignatura de Data Story Telling brinde para el óptimo manejo de la herramienta por parte del usuario final.	18/mar./2023	20/may./2023
Presentación de Avance Presentación del avance del proyecto según las indicaciones que sean dadas desde la maestría.	15/abr./2023	15/abr./2023
Desarrollo de la 2a versión de la herramienta Período para el desarrollo de la 2a versión de la herramienta para toma de decisiones donde se espera configurar los aspectos de experiencia e interfaz de usuario (UX /UI).	17/abr./2023	16/jun./2023
2a Versión herramienta para toma de decisiones Presentación de 2a versión de herramienta de análisis. Incluye: - Mejoras visuales y de comprensión (Data Story Telling si es viable)	17/jun./2023	17/jun./2023

Fuente: Elaboración propia (2022)

5. Descripción de la situación organizacional (Contexto)

Luego del cambio en las estrategias de estudio que trajo la pandemia en 2020, donde la virtualidad sirvió de herramienta clave para la continuidad del funcionamiento de gran parte de las instituciones de educación, la IESoe identificó la necesidad y demanda que tiene un amplio sector de su población tipo interesada en obtener un título universitario sin necesariamente incurrir en los altos costos de sostenimiento y movilidad que implica el desplazarse desde territorios apartados a los grandes centros urbanos donde se encuentran la mejor oferta de programas académicos. Para ese sector de población, estudiar de manera virtual desde los territorios se convirtió en una oportunidad de tener una mejora en sus expectativas de vida que no implica esfuerzos extraordinarios desde lo económico, situación favorable que se hizo evidente debido a la forzosa necesidad de hacer uso de los medios digitales para mantener una conexión con el mundo.

Sin embargo, así como se tiene ganancia en lo económico, este tipo de educación saca a relucir las brechas entre las formaciones académicas primaria y secundaria que tienen las zonas rurales frente a las zonas urbanas y que por años han sido tema de debate sobre la desigualdad existente en la educación (Semana, 2022). Esos estudiantes, que ahora tienen acceso desde lo remoto a una educación de mejor de calidad, deben hacer frente a una tal vez inesperada mayor exigencia para la que quizá no estaba preparados, con retos académicos superiores a los esperados, y cuya calidad de resultados finales terminan reflejándose en su bajo rendimiento académico.

Y es aquí, en este contexto, donde el estudiante de la IESoe se enfrenta al elegir sobre el qué camino tomar en su situación académica:

- **Buscar distintas estrategias académicas para mejorar el rendimiento académico**. Aquí el esfuerzo al interior de las IES se centra en brindarles oportunidades de espacios y tiempos para que los estudiantes tengan la posibilidad de encontrar un mecanismo de nivelación a los requeridos como mínimos en las asignaturas del programa que cursa el estudiante (sean presentados estos mínimos de manera explícita o no) . Esto se da a través de encuentros de tutorías, cursos de educación continuada y otros mecanismos que, aunque exijan del estudiante más tiempo dedicado a sus estudios, identifican de una manera más personalizada las deficiencias que se deban atender.

  Viene a ser decisión y trabajo del estudiante aprovechar todas esas opciones que sus instituciones les brindan.
- **Desertar.** Cuando se conversa en las aulas con estudiantes que están en una situación límite en sus estudios, y no ven salida a esa situación de bajo rendimiento académico, y sus actividades cotidianas no les permiten disponer de recursos, principalmente de tiempo, optan por aplazar sus estudios esperando una mejora de sus condiciones personales y laborales para retomarlos. Desafortunadamente, ese aplazamiento temporal, en muchas ocasiones pasa a ser permanente, dándose de ese modo una deserción no esperada.
- **Cometer fraude académico.** La situación límite a la que puede llegar un estudiante para salvar la nota de una asignatura, hace que se vea enfrentado a un dilema ético donde debe poner en un lado de la balanza el reconocer que su comprensión de los temas no ha sido el suficiente como para obtener una nota aprobatoria y resignarse a aceptar la reprobación; y por otro lado hacer lo que sea

necesario para no tener que generar más gastos de recursos (tiempo y dinero) y obtener el título que busca para consigo, obtener la tan esperada mejora de sus ingresos. Es en este instante donde el estudiante puede llegar a cometer fraude académico, y es sobre este punto donde se describirá la problemática del caso a continuación.

6. Descripción de la situación de caso

6.1. Problemática Principal

La virtualidad ha traído la oportunidad para que los estudiantes manejen la autodisciplina y el trabajo autónomo de su estudio. Desafortunadamente, si este paso se da de manera no esperada (por ejemplo, la pandemia) o sin previa contextualización de lo que es un estudio que no requiere supervisión, donde se hace un voto de confianza en el estudiante en su propia responsabilidad y que por lo tanto exige de él autocontrol y autodisciplina; los controles tradicionales existentes en los espacios de presencialidad para evitar el fraude académico se vuelven inútiles o simplemente desaparecen.

Y esto es lo que actualmente sucede en la IESoe, evidenciado en los registros de actividad de los estudiantes en las plataformas virtuales donde realizan las actividades académicas. La detección de estos comportamientos trajo consigo un llamado de atención a replantear los controles que son requeridos para evitar este tipo de comportamientos y buscar alternativas de solución al asunto.

6.2.Fraude Académico

Para comprender adecuadamente esta problemática, es necesario definir lo que se entiende como fraude académico.

Según el estudio realizado por los investigadores Enrique Ramírez y Lina Martínez de la universidad ICESI, el fraude académico no tiene una definición universalmente aceptada (Martínez & Ramírez, 2017) por lo que para este proyecto se toma la definición que estos autores proponen:

"(…) conjunto de comportamientos inapropiados o no permitidos en que incurre un estudiante, en relación con trabajos, exámenes, pruebas que se le asignan o requisitos que debe cumplir en ámbitos académicos." (Martínez & Ramírez, 2017, p.3)

A partir de esta definición, es necesario entonces diferenciar esos comportamientos para comprender por separado sus causas. Para ello, se parte de la propuesta que el grupo de trabajo del profesor García Villegas en su documento "Normas de Papel: la cultura del incumplimiento de reglas" hizo a través de un listado para una encuesta aplicada en universidades con los siguientes comportamientos considerados fraudulentos. Son estos comportamientos según García Villegas (2009, p.86):

1. Copiar las respuestas de un compañero en un examen.
2. Dejar que un compañero copie sus respuestas en un examen.
3. Copiar el trabajo de un compañero.
4. Prestar un trabajo que usted hizo para que lo copien.
5. Bajar un trabajo de Internet y presentarlo como propio.
6. Utilizar ideas de un autor sin citarlo.
7. Copiar o parafrasear apartes de otros trabajos, sin hacer la referencia correspondiente.
8. Presentar un certificado médico falso para justificar una inasistencia.
9. Utilizar herramientas que no están autorizadas en un examen (apuntes, calculadora).
10. Firmar una lista de asistencia a nombre de un compañero.
11. Incluir a alguien en un grupo sin que haya colaborado con el trabajo.
12. Aparecer como miembro de un grupo sin haber colaborado con el trabajo.
13. "Copiar y pegar" texto de Internet sin hacer la referencia correspondiente.

14. Presentar una prueba (examen o parcial) a nombre de un compañero.

Se debe tener en cuenta que para el año de realización de este estudio (2009) el impacto que tuvo el acceso de internet en el campo académico no era comparable al impacto que tiene hoy. La virtualidad de la educación en ese entonces no tenía el grado de desarrollo que tiene actualmente.

6.3.Plagio

La preocupación permanente por tener creaciones originales en los documentos académicos ha llevado a que los esfuerzos de las IES se enfoquen en resolver los temas relacionados con la protección de derechos de autor, especialmente los temas relacionados con el plagio, entendiendo "Plagiar" como "Copiar en lo sustancial obras ajenas, dándolas como propias." (RAE, 2022a). Herramientas con un importante grado de desarrollo como Turnitin[3], Plagium[4] (extensión para Google Docs) o Urkund[5], demuestran el énfasis que se da por combatir este tipo de fraude académico.

Pero es aquí donde el tema se vuelve confuso, al menos en la IESoe, pues allí se le denomina "plagio" a todo comportamiento fraudulento. Esto trae como consecuencia que, al querer realizar un procedimiento de sanción por fraude académico, se aplique un procedimiento que está concebido para casos de plagio. Y es aquí donde se genera un vacío del proceder frente a otros comportamientos fraudulentos, pues no todos los casos de fraude académico son efectivamente un plagio. Por lo que aquí resulta especialmente útil precisar que el plagio no es sinónimo de fraude académico sino sólo es una de sus modalidades.

---

[3] https://www.turnitin.com/es
[4] http://www.plagium.com/
[5] https://www.urkund.com/es/

Si se toman del listado propuesto por García Villegas, los comportamientos considerados fraudulentos y que se ajustan mejor a la definición de plagio serían los siguientes:

3. Copiar el trabajo de un compañero.

4. Prestar un trabajo que usted hizo para que lo copien.

5. Bajar un trabajo de Internet y presentarlo como propio.

6. Utilizar ideas de un autor sin citarlo.

7. Copiar o parafrasear apartes de otros trabajos, sin hacer la referencia correspondiente.

11. Incluir a alguien en un grupo sin que haya colaborado con el trabajo.

12. Aparecer como miembro de un grupo sin haber colaborado con el trabajo.

13. “Copiar y pegar” texto de Internet sin hacer la referencia correspondiente.

(García Villegas et al., 2009)

Este filtro, necesariamente deja entonces por fuera otro tipo de comportamientos que no encajan bien con la definición de plagio:

1. Copiar las respuestas de un compañero en un examen.

2. Dejar que un compañero copie sus respuestas en un examen.

8. Presentar un certificado médico falso para justificar una inasistencia.

9. Utilizar herramientas que no están autorizadas en un examen (apuntes, calculadora).

10. Firmar una lista de asistencia a nombre de un compañero.

14. Presentar una prueba (examen o parcial) a nombre de un compañero.

(García Villegas et al., 2009)

6.4.Mas Allá Del Plagio

Otros comportamientos fraudulentos distintos al plagio, observados y evidenciados con la experiencia docente adquirida en las plataformas virtuales de la IESoe, pero no incluidos en el listado de García Villegas, han sido los siguientes:

1. Obtener las respuestas de exámenes por mecanismos fraudulentos.
2. Realizar exámenes haciendo uso de las respuestas obtenidas por mecanismos fraudulentos.
3. Pagar a un tercero para la elaboración de un trabajo y presentarlo como de propia elaboración.

A estos comportamientos se le puede agregar el comportamiento número 14 del listado de García Villegas (2009):

14. Presentar una prueba (examen o parcial) a nombre de un compañero.

(p.86)

Acorde con la experiencia docente de uno de los autores, estos tres comportamientos, incluido el presentar una prueba a nombre de un compañero, son más frecuentes que los definidos como plagio en la IESoe. Y esto se da justamente por varios factores:

- **Centrarse en el plagio.** Se observa contraproducente para un efectivo control de estos comportamientos definirlos todos como plagio y no diferenciar los procedimientos para cada uno ellos.
- **La ausencia de controles efectivos de la IESoe hacia ese tipo de comportamientos**. Como consecuencia del punto anterior, no hay un proceder claro hacia ese tipo específico de comportamiento, y esto trae consigo que las acciones se limiten al alcance que los docentes por iniciativa

propia deseen realizar durante la evaluación, corriendo el riesgo de señalarlos equivocadamente como plagio y no saber cómo proceder ante ellos.

- **Desconocimiento del alcance de la tecnología LMS**. Las plataformas LMS que la IESoe facilita para sus colaboradores, y en especial los docentes, brinda múltiples herramientas para un óptimo desarrollo de las actividades de enseñanza, aprendizaje y seguimiento. Sin embargo, se evidencia un alto grado de desconocimiento de las herramientas de control por parte de los docentes.
- **Ausencia de herramientas para detección.** Debido al anterior desconocimiento de los LMS, los docentes carecen de medios que les permitan tener una evidencia clara y sólida para determinar cuándo se pueda presentar uno de estos comportamientos y tomar las acciones que sean necesarias.

6.5.Definición De Comportamientos

Para abordar la solución, primero hay que definir el problema. Los comportamientos mencionados como de reciente aparición, junto con el comportamiento número 14 del listado de García Villegas, se propone definirlos de la siguiente manera:

Hurto de información

1. Obtener las respuestas de exámenes por mecanismos fraudulentos.

Este comportamiento deberá ser revisado en detalle por el área jurídica de la IESoe, pues su definición es muy cercana a la del delito denominado “Acceso abusivo a un sistema informático” (Congreso de la República, 2000, art. 269A).

Uso de información hurtada

2. Realizar exámenes haciendo uso de las respuestas obtenidas por mecanismos fraudulentos.

Este comportamiento deberá ser revisado en detalle por el área jurídica de la IESoe, pues su definición es muy cercana a la del delito denominado “Receptación” (Congreso de la República, 2000, art. 447).

Falsedad Personal

3. Pagar a un tercero para la elaboración de un trabajo y presentarlo como de propia elaboración.
14. Presentar una prueba (examen o parcial) a nombre de un compañero. (García Villegas et al., 2009)

    Estos comportamientos se acercan a la definición de “Falsedad Personal” (Congreso de la República, 2000, art. 296) y correspondería al área jurídica de la IESoe estudiar sus efectos y verificar si deben ser incluidos en la normativa de la institución.

6.6.Herramienta Orientada A La Suplantación

A partir de los comportamientos antes mencionados, este trabajo propone una herramienta que facilita la detección de los denominados “Falsedad Personal” en entornos de enseñanza virtual. De acuerdo con la experiencia docente en estas plataformas, son estos comportamientos los que más se están registrando en las actividades de evaluación y para los cuales no existen herramientas de detección que faciliten el trabajo del docente y la toma de acciones frente a los mismos.

Para efectos del presente proyecto se propone ajustar el término “Falsedad Personal” y tratar este tipo de comportamientos como “Suplantación” por la propia definición que la lengua española tiene para este tipo de comportamientos:

> “Suplantar: Ocupar con malas artes el lugar de alguien, defraudándole el derecho, empleo o favor que disfrutaba.” (RAE, 2022b)

7. Descripción de las alternativas, estrategias o acciones que se toman en el análisis de la solución a la problemática

Tal como se describe en los capítulos que preceden, el entorno digital y las dinámicas académicas de la actualidad brindan un conjunto de herramientas al estudiante que pueden llegar a ser manejadas de manera no ética y es por ello que surge la necesidad de brindarle al docente y a la institución educativa herramientas que les permitan contrarrestar estas actividades; por lo que se propone la creación de una herramienta de identificación de comportamientos anómalos que alerte mediante una serie de KPI's sobre la posible existencia de una actividad no ética.

Para la construcción de los KPI's es primordial la identificación de un estado cuyas características sean "normales", es decir el escenario de tendencia promedio resultante de las dinámicas de la mayoría, el cual nos permitirá entonces identificar aquellos escenarios cuyas características sean "anormales", es decir que se alejan de forma estadísticamente importante de la tendencia central (outliers); esta dualidad permitirá establecer las bases comparativas para cotejar la información introducida al sistema y así exponer insights que permitan al usuario obtener un panorama del grupo de alumnos y sus acciones frente a las evaluaciones digitales.

La definición de un escenario normal y anormal está sujeta al comportamiento de las variables que lo integren, y estas a su vez surgen de la data con que se alimenta al sistema; así pues, se trabajarán las siguientes variables cuya descripción detallada se encuentran en el ANEXO B:

- **ID del estudiante:** Es el número que identifica al estudiante, el cual para efectos del frontend será anonimizado.

- **Tipo de Examen:** El reporte de Interacciones nos indica el tipo de evaluación realizada: Cuestionario, Tarea, Parcial, etc.; este insumo permite identificar las características de la evaluación.

- **Puntaje:** Es un valor numérico generado por el profesor de la asignatura, y cuantifica el nivel de aciertos del estudiante dentro de una evaluación.

- **Tiempo de duración de la evaluación por usuario (Minutos):** Es el tiempo que tarda un usuario realizando la evaluación y se obtiene a través de la diferencia de horas, minutos y segundos entre la acción "started" y la acción "submitted" de la columna "Descripción" del reporte de interacciones de usuario.

- **Indicador de Velocidad:** Esta variable surge de la división del puntaje obtenido del usuario (reporte de calificaciones) entre el tiempo de duración de la evaluación por usuario y sirve como herramienta para relacionar minutos vs calificación, teniendo de esta manera que a menor velocidad se implicaría menos puntaje y mayor tiempo de desarrollo y a mayor velocidad lo contrario.

- **Dirección IP: Latitud y Longitud (Distancia de IP vs IESoe Bogotá):** Para el uso de esta variable se desarrolló el código de Python mostrado en el ANEXO B, el cual permite tomar las direcciones IP de los reportes (colocándolas en un archivo .txt llamado codIP.txt) y conectar con una API en www.ipwho.is [6]quien

[6] http://www.ipwho.is/

expone data de geolocalización como región, ciudad, país, código postal, latitud, longitud, entre otras y finalmente traduce la información y la consolida en un archivo de txt llamado finalcodIP.txt.

Con el archivo generado en el proceso anterior, se desarrolló este indicador el cual permite medir los kilómetros que hay entre cada dirección IP por donde ingresó el usuario vs la localización de la IESoe en Bogotá, la idea es tener un punto de referencia estático que permita medir variaciones de localización entre sesiones, ya que los usuarios suelen conectarse mediante varias direcciones IP; para el desarrollo de este indicador tomamos la latitud y longitud de cada IP y la latitud y longitud de la IESoe en Bogotá y mediante la fórmula de Haversine (ver Figura 4 en el ANEXO B) se obtienen las distancias en kilómetros.

- **Número de interacciones en el sistema:** El reporte del Log de Usuario proporciona la cantidad de interacciones que ha tenido dentro de la plataforma (interacciones no solo de evaluaciones sino con la plataforma en sí).

- **Cantidad de IP’s manejados por el estudiante:** El reporte del Log de Usuario indica la dirección IP manejada por el usuario por lo que a partir de esta información se cuenta la cantidad de IP’s que maneja el usuario al momento de ingresar a las evaluaciones.

- **Indicador de repetición de IP’s en más de 1 estudiante:** El estudio de la cantidad de IP’s permite también identificar si uno (o varias) direcciones IP’s son usadas simultáneamente por más de 1 estudiante.

- **Indicador de realización de varias evaluaciones durante el mismo día de parte del estudiante:** Usando el reporte de calificaciones se obtiene la información de aquellos casos donde el estudiante activa y genera más de una evaluación en un mismo día.

- **Indicador de finalización de evaluación por parte del estudiante:** El Log de usuario presenta la información de si el usuario activó una evaluación y si la finalizó; esta información es importante para determinar el comportamiento de cada usuario frente a las evaluaciones.

- **Cantidad de visualizaciones que tuvo cada examen de parte del estudiante:** Cuando un usuario activa una evaluación el Log identifica hasta qué punto visualizó cada parte y este dato permite entender en qué punto para el usuario o hasta donde avanza.

- **Indicador de patrones para casos de varias evaluaciones un mismo día:** Para los casos de usuarios que realizan más de 1 evaluación el mismo día, se busca estudiar su interacción para identificar si existe un patrón en el tiempo de cada respuesta de manera para determinar si el comportamiento es atípico.

- **Identificador de velocidad atípica (Con Z-score):** Esta variable usa el Z-SCORE para estudiar si cada dato se puede considerar “atípico” o no.

Los KPI’s generados están expuestos en el software de B.I. (Power BI) para una interacción optima con el usuario. Para lograr el impacto buscado, a continuación, se presenta los análisis de los datos que son base para la solución propuesta:

- **Estudio de Análisis de Componentes Principales (ACP),** el cual permitió estudiar las relaciones entre las variables y principalmente generar una visualización que de manera clara expone la aglomeración de los usuarios alrededor de la tendencia global de sus variables (Escenario NORMAL) y así mismo expone a aquellos usuarios que se alejan de esta tendencia (Escenario ANORMAL) y que serían el foco de alerta a ser estudiada. (Ver ANEXO E)
- **Modelo Descriptivo,** con programación en Python se desarrolló un código que balancea la información mediante el uso de las librerías SMOTE y la técnica de UNIONES DE TOMEK para así alimentar un grupo de modelos de Machine Learning que finalmente arrojan las probabilidades de que un estudiante X esté incurriendo en una situación atípica. (Ver ANEXO E)

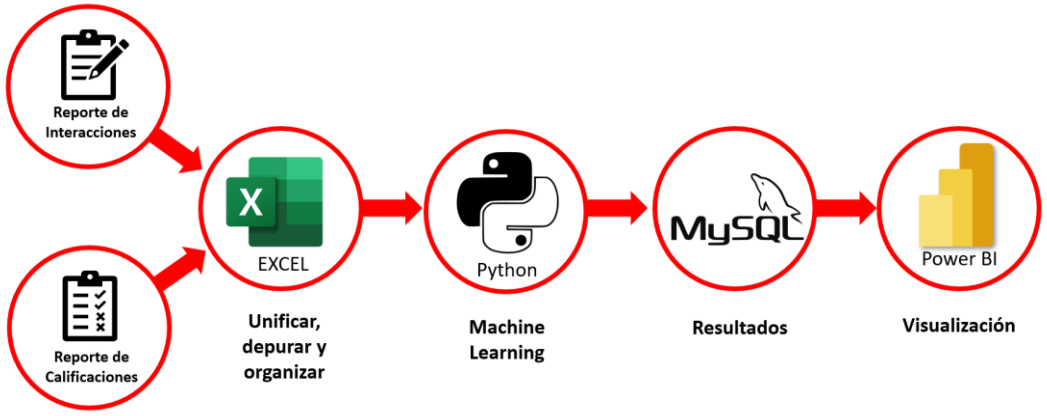
Las herramientas que permitieron los análisis descritos anteriormente son: programación en PYTHON mediante el uso de COLAB con librerías libres y el software R-STUDIO el cual también es de uso libre por lo que no se incurre en gasto alguno, sin embargo es importante recalcar la importancia de la organización previa de la data, para ello se hace uso de la herramienta EXCEL junto con el previo desarrollo de una serie de MACROS que garanticen el tratamiento estructurado de los reportes y así pues, como última salvedad cabe

aclarar la criticidad de introducir data sin errores, que sea: fiel, fidedigna, coherente y oportuna para así garantizar resultados de igual calidad.

8. Plan y recomendaciones de implementación y aplicación

La implementación planteada actualmente se estructura según el siguiente diagrama:

Figura 2 *Estructura del desarrollo del proyecto*



Fuente: Elaboración propia (2023)

1. Inicialmente se introducen los reportes que se extraen previamente de Moodle a un EXCEL que contiene una serie de MACROS que de manera automática organiza la información y genera los indicadores previamente descritos.
2. Este EXCEL alimentará el MODELO DESCRIPTIVO de Python para obtener la probabilidad de comportamiento atípico por alumno.
3. Adicionalmente el EXCEL alimentará un script en R-STUDIO con el cual se obtendrá el ACP que expondrá los grupos atípicos (escenario ANORMAL) y aquellos con comportamientos promedios (escenario NORMAL).
4. La información será cargada en una base de datos de MySQL para que sea de fácil consolidación y manejo.

5. Mediante el uso de un software de B.I. (Power BI) se generó un dashboard interactivo para que el usuario pueda navegar por cada uno de los indicadores y así hacer seguimiento a las alertas relacionadas.

Se buscará automatizar el proceso de manera que sea lo más práctico e intuitivo posible e igualmente se desarrolló un manual para el uso de la herramienta (Anexo E).

Finalmente, para medir la eficacia del modelo se realizará una serie de pruebas supervisadas al interior de la institución (esto debido a políticas internas del manejo de datos de la IESoe) con casos reales con el fin de hacer las mejoras que requiera la herramienta.

9. Descripción de las fuentes de información

El desarrollo de este proyecto busca construir una herramienta que basada en data histórica pueda exponer una serie de indicadores y visuales que sugieran el comportamiento ético o no ético de un grupo de usuarios dentro de un programa académico, para esto y en pro de sustentar los insights e hipótesis se toma la información proporcionada por la base de datos del software Moodle manejado en la IESoe; este software es una plataforma online que funciona como herramienta para los educadores y estudiantes, dándole a los primeros la posibilidad de publicar cursos enteros con actividades que son evaluadas de manera automática (entre otras cosas) y dándole a los segundos la posibilidad de interactuar en todo momento con las clases, evaluaciones y tareas, así como llevando históricos de sus progresos (entre otras cantidad importante de herramientas).

La capacidad del software para llevar históricos de los eventos de cada usuario que interactúe (estudiantes, profesores, administradores, etc.) le permite generar una serie de reportes con información de todo tipo basada en un modelo de datos estructurados, y son estos archivos los que inicialmente se toman como base para la construcción de esta fase del proyecto, por lo que los insights, análisis y desarrollos que se presentarán a lo largo de este documento estarán basados principalmente en los siguientes 2 reportes:

- **Reporte de calificaciones**: Este archivo de Excel proporciona principalmente la información sobre las evaluaciones realizadas por cada usuario para una materia en específico, incluyendo: fechas, notas y tiempos, dentro de su data más relevante.
- **Reporte de interacciones**: Este archivo de Excel funciona como una bitácora que principalmente registra cada movimiento realizado por un usuario, así como el

tipo de acción realizada y el receptor de esta acción, adicionando información importante como fechas, IP, origen de la conexión, entre otras.

La gestión de la información almacenada en estos reportes es un tema de alta sensibilidad, ya que presenta datos tanto públicos, como privados y semiprivados por lo que el manejo no sólo está amarrado a un tema ético sino también a un tema legal y de seguridad del uso de la data; a modo informativo, se presenta a continuación los datos manejados en cada reporte según su tipo:

Tabla 2 *Clasificación de datos en Reporte de Calificaciones*

INFORMACIÓN	**DESCRIPCIÓN**	**TIPO DE DATO**
Nombre de usuario	Indica el número de cedula del usuario	Público
Apellido(s)	Indica el/los apellido(s) del usuario	Público
Nombre	Indica el nombre del usuario	Público
Dirección de correo	Indica el e-mail del usuario	Público
Departamento	Indica la facultad relacionada a la materia que se está viendo	Público
Institución	Indica la región de la sede	Semiprivado
Estado	Indica si la evaluación fue finalizada o no	Semiprivado
Comenzado el	Indica la fecha y hora de inicio de la evaluación	Semiprivado
Finalizado	Indica la fecha y hora de finalización de la evaluación	Semiprivado
Tiempo requerido	Indica las horas, minutos y segundos que tomó la realización de la evaluación	Semiprivado
Calificación/5.00	Indica la calificación total obtenida	Semiprivado
P. 1 /1.00	Indica la calificación obtenida en la pregunta 1	Semiprivado
P. 2 /1.00	Indica la calificación obtenida en la pregunta 2	Semiprivado

P. 3 /1.00	Indica la calificación obtenida en la pregunta 3	Semiprivado
P. 4 /1.00	Indica la calificación obtenida en la pregunta 4	Semiprivado
P. 5 /1.00	Indica la calificación obtenida en la pregunta 5	Semiprivado

Fuente: Elaboración propia (2022)

Tabla 3 *Clasificación de datos en Reporte de Interacciones*

INFORMACIÓN	**DESCRIPCIÓN**	**TIPO DE DATO**
Hora	Indica fecha, hora, minuto y segundo donde comenzó la acción	Semiprivado
Nombre completo del usuario	Indica el nombre completo de la persona que realizó la acción	Público
Usuario afectado	Indica el nombre completo de la persona que recibió la acción	Público
Contexto del evento	Indica el nombre del componente que recibió la acción	Semiprivado
Componente	Indica el tipo de componente que recibió la acción	Semiprivado
Nombre evento	Indica la acción realizada	Semiprivado
Descripción	Indica la interacción realizada y el número asociado a esta	Semiprivado
Origen	Indica el tipo de portal que se usó para el ingreso a la plataforma	Privado
Dirección IP	Indica la dirección IP asociada a la conexión del usuario	Privado

Fuente: Elaboración propia (2022)

Para poder hacer uso de esta información de manera adecuada a nivel ético, legal y de seguridad se debe realizar un proceso de **Anonimización** (Ver ANEXO A).

10. Conclusiones

- Objetivo: Identificar la tipología existente en el sistema educativo de Colombia relacionada con la clasificación de los tipos de fraude académico, qué motiva a realizarlo, reconocimiento de sus consecuencias y su soporte normativo.

Podemos concluir que a la fecha dentro de la normatividad del país no existe una definición oficial y consensuada respecto a los tipos de fraude académico.

Los recientes casos de fraude académico que le han costado la anulación de títulos a diferentes personalidades del orden político en el país han motivado la toma de acción por parte de las IES puesto que su reputación como organismos que avalan las competencias académicas de sus estudiantes queda en entredicho al descubrirse casos cada vez más frecuentes tanto de plagio como de suplantación.

La principal consecuencia de obtener una titulación de manera fraudulenta es la afectación reputacional y de credibilidad de la IES, porque esto evidenciaría:

1. Falta de control y deficiencias en los procesos de evaluación de las competencias.
2. Falta de control en la selección de los docentes encargados de las asignaturas.
3. Falta de seguimiento a la realización de las actividades académicas.

De acuerdo con el alcance de la investigación el fraude académico en Colombia es un tema que principalmente se observa en las IES en sus reglamentos estudiantiles internos, documentos donde se menciona las faltas, su gravedad y las sanciones bajo los criterios que cada institución considere adecuados. La redacción sobre este tema en el caso de la IESoe al parecer fue realizada con base en las experiencias propias de sus años de funcionamiento y no como resultado de un

consenso del conjunto de IES. Para la IESoe todo lo que se puede entender como Fraude Académico tiene una sola denominación: Plagio. Para otras IES, se debe revisar en su respectivo reglamento estudiantil.

- Objetivo: Desarrollar Indicadores Claves de Proceso (KPIs) que faciliten la detección de comportamientos anómalos en un registro de información de actividades académicas.

Se identifican y analizan las variables relacionadas a los reportes con el uso de herramientas informáticas (Visual Studio Code y R-STUDIO) las cuales permiten identificar grupos de estudiantes con comportamientos atípicos y cuantificar las probabilidades relacionadas a estos; los KPI's manejados (información ampliada en el Anexo B) y plasmados en el dashboard de usuario son los siguientes:

- **Tiempo de duración de la evaluación por usuario:** Es el tiempo que tarda un usuario realizando la evaluación y se obtiene a través de la diferencia de horas, minutos y segundos entre la acción "started" y la acción "submitted" de la columna "Descripción" del reporte de interacciones de usuario.
- **Indicador de Velocidad:** Esta variable surge de la división del puntaje obtenido del usuario (reporte de calificaciones) entre el tiempo de duración de la evaluación por usuario y sirve como herramienta para relacionar minutos vs calificación, teniendo de esta manera que a menor velocidad se implicaría menos puntaje y mayor tiempo de desarrollo y a mayor velocidad lo contrario.

- **Dirección IP:** Para el uso de esta variable se desarrolló el código de Python mostrado en el ANEXO B, el cual permite tomar las direcciones IP de los reportes (colocándolas en un archivo .txt llamado codIP.txt) y conectar con una API en www.ipwho.is [7]quien expone data de geolocalización como región, ciudad, país, código postal, latitud, longitud, entre otras y finalmente traduce la información y la consolida en un archivo de txt llamado finalcodIP.txt.

- **Distancia (Km) de IP vs IESoe Bogotá:** Con el archivo generado en el proceso anterior se desarrolló este indicador, el cual permite medir los kilómetros que hay entre cada dirección IP por donde ingresó el usuario vs la localización de la IESoe en Bogotá, la idea es tener un punto de referencia estático que permita medir variaciones de localización entre sesiones, ya que los usuarios suelen conectarse mediante varias direcciones IP; para el desarrollo de este indicador se toma la latitud y longitud de cada IP y la latitud y longitud de la IESoe en Bogotá y mediante la fórmula de Haversine (ver Figura 4) se obtienen las distancias en kilómetros.

[7] http://www.ipwho.is/

- Objetivo: Evaluar la fortaleza que ofrecen los datos analizados para servir como evidencia indiscutible de comportamientos anómalos relacionados con el fraude académico, y para el caso del presente proyecto, específicamente de suplantación.

Los insights generados están soportados a nivel matemático y exponen evidencia de comportamientos anómalos que se sugieren sean tratados como alertas para una posterior investigación más detallada y a fondo de los grupos identificados; la herramienta sugiere y cuantifica la probabilidad de situaciones no éticas, luego debe ser tratado como un insight más no como una evidencia concreta.

Esta herramienta se apoya en 4 modelos de machine learning para construir las probabilidades con las que se generarán insights, y en búsqueda de exponer los resultados de mayor calidad se contrasta el nivel de calibración (AUC) de cada modelo y se selecciona el óptimo (Anexo E).

Entre las conclusiones más importantes surgidas del análisis de los datos se encuentran:

- Existe una clara anomalía en el comportamiento de las notas vs tiempo de evaluación que permite detectar de manera explícita los casos atípicos, como por ejemplo situaciones donde el estudiante resolvió una evaluación de 20 preguntas en menos de 3 minutos y obtuvo el máximo puntaje (5 puntos) cuando inicialmente se estipula 1 hora máximo para todo el proceso.
- Dentro de los grupos atípicos se evidencia como una constante, el cambio de localización de la conexión del estudiante al momento de presentar un examen, existiendo variaciones importantes como grupos de ciertas y

variadas zonas del país que al momento de ingresar a la evaluación aparecen registrados en conjunto en una sola localización (distinta a las iniciales).

- Se detecta una tendencia geolocalizada de comportamientos atípicos que permiten sectorizar zonas más propensas a esto, a modo de ejemplo de los resultados obtenidos a partir del paquete de información inicial, se detectaron como zonas con mayor propensión a comportamientos atípicos las ciudades de Santa Marta, Neiva y Bogotá.

- Objetivo: Explorar la viabilidad de establecer una propuesta de análisis predictivo que permita anticiparse a comportamientos tendientes a realizar fraude académico.

De acuerdo con lo estudiado en la elaboración de modelos predictivos, es necesario contar con datos históricos que permitan comprender mejor las secuencias de comportamientos pasados. Con esta información previa, si en el presente se empiezan a identificar el mismo tipo de secuencias iniciales de esos comportamientos, permitiría establecer parámetros de probabilidades sobre la decisión de cometer fraude académico.

Sin embargo, se identifica que la información qué más refleja la variabilidad en el comportamiento proviene de la geolocalización por lo cual una siguiente etapa para la construcción de un indicador efectivo a nivel de predicción debe estar alimentado con una información ampliada que incluya criterios socioeconómicos y

culturales en la localización como parte de la base, de manera que el modelo pueda ser más preciso y sugerir tendencias a futuro.

Referencias bibliográficas

Arias del Prado, J. (2020, enero 21). UX y UI. No se enfrentan, ¡Se complementan! *UXABLES | Blog*. http://www.uxables.com/diseno-ux-ui/ux-y-ui-no-se-enfrentan-se-complementan/

Congreso de la República. (2000, julio 24). Ley 599. *Diario Oficial 44097*. https://www.funcionpublica.gov.co/eva/gestornormativo/norma.php?i=6388

Díaz-López, D. (2022, junio 20). *Anonimización* [Power Point].

García Villegas, M., Henao, A., Mejía, J. F., & Ordoñez, C. (2009). 3. Fraude académico: Comparación entre dos universidades colombianas. En *Normas de papel: La cultura del incumplimiento de reglas* (p. 79 a 104). Siglo del Hombre.

Martínez, L., & Ramírez, E. (2017). Fraude académico en universitarios en Colombia: ¿Qué tan crónica es la enfermedad? *Educação e Pesquisa*, *44*(0). https://doi.org/10.1590/s1517-9702201706157079

Murgano, A. (2022). *Fórmula de Haversine* (Versión 1) [Python].

Patiño M., E., & Iguarán, A. (2018, junio 17). *Estudiantes en líos por fraude: ¿qué pasa con la ética de los jóvenes?* El Heraldo. https://www.elheraldo.co/barranquilla/estudiantes-en-lios-por-fraude-que-pasa-con-la-etica-de-los-jovenes-507738

RAE. (2022a). *Plagiar*. «Diccionario de la lengua española» - Edición del Tricentenario. https://dle.rae.es/plagiar

RAE. (2022b). *Suplantar*. «Diccionario de la lengua española» - Edición del Tricentenario. https://dle.rae.es/suplantar

Semana. (2022, mayo 19). *Estudio basado en resultados de pruebas Saber 11 muestra gran desigualdad en la educación*. https://www.semana.com/educacion/articulo/estudio-basado-en-resultados-de-pruebas-saber-11-muestra-gran-desigualdad-en-la-educacion/202236/

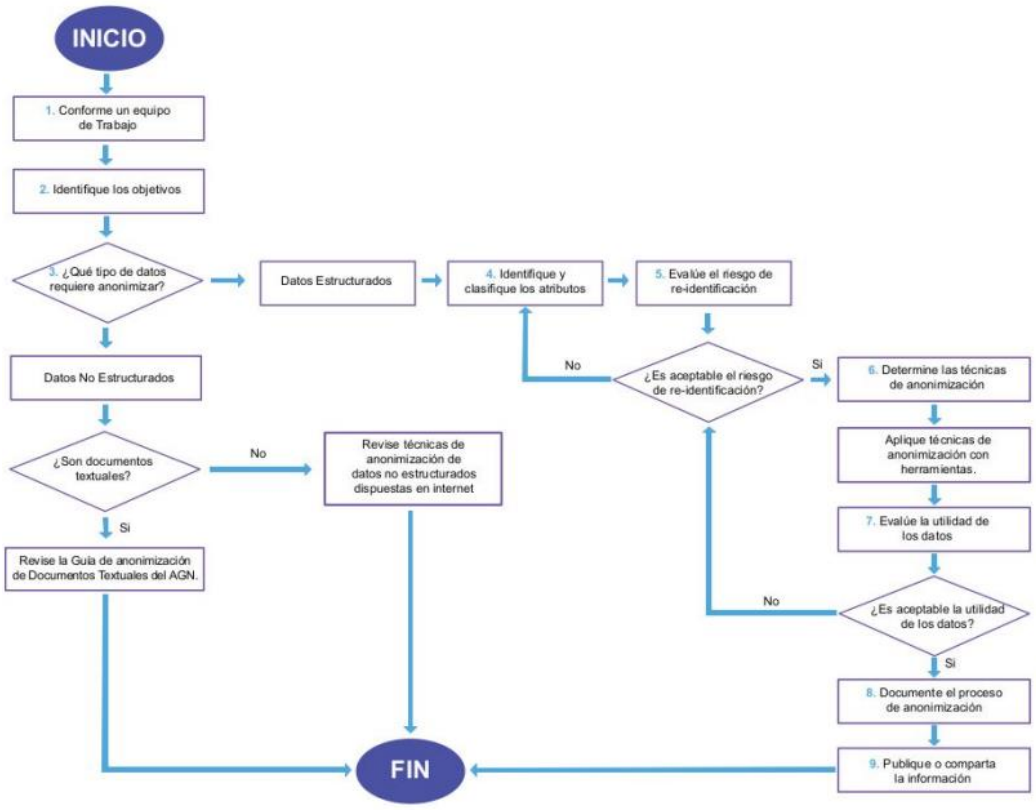
Anexos

Anexo A

Anonimización

La técnica de anonimización de datos es un proceso que busca transformar la información sensible mediante el uso de diversos métodos para llevarlos a un estado que sea inidentificable respecto a su versión original, de manera de proteger la data ante los riesgos de su divulgación amarrados a la legalidad de cada país; para realizar este proceso seguiremos la metodología indicada en el algoritmo del material académico de anonimización compartido por el profesor Daniel Díaz López de la Universidad del Rosario:

Figura 3 *Metodología de anonimización*



Fuente: (Díaz-López, 2022)

Atendiendo la metodología se siguieron los pasos indicados:

1. Conforme un equipo de trabajo

   Se conforma por los integrantes de este proyecto, Daniel Hernández y Ángel Murgano.

2. Identifique los objetivos

   Para esta fase trabajaremos con los siguientes datos objetivos que resultan ser DATOS ESTRUCTURADOS.

   a. Reporte de calificaciones: Nombre, Apellido(s), Tiempo requerido y Calificación/5.00.

   b. Reporte de interacciones: Hora, Nombre completo del usuario, Descripción y Dirección IP.

3. ¿Qué tipo de datos requiere anonimizar?

   Se tratarán DATOS ESTRUCTURADOS.

4. Identifique y clasifique los atributos:

Tabla 4 *Atributos de los datos*

INFORMACIÓN	**DESCRIPCIÓN**	**TIPO DE DATO**
Apellido(s)	Indica el/los apellido(s) del usuario	Público
Nombre	Indica el nombre del usuario	Público
Tiempo requerido	Indica las horas, minutos y segundos que tomó la realización de la evaluación	Semiprivado
Calificación/5.00	Indica la calificación total obtenida	Semiprivado
Hora	Indica fecha, hora, minuto y segundo donde comenzó la acción	Semiprivado
Nombre completo del usuario	Indica el nombre completo de la persona que realizó la acción	Público

Descripción	Indica la interacción realizada y el número asociado a esta	Semiprivado
Dirección IP	Indica la dirección IP asociada a la conexión del usuario	Privado

Fuente: Elaboración propia (2022)

5. Evalúe el riesgo de re-identificación

   Sí existe el riesgo de identificación, ya que claramente se asocia el nombre de los usuarios.

6. Determine las técnicas de anonimización

   Se usará la técnica de ENCRIPTACIÓN SIMÉTRICA para los siguientes casos:

   a. Se descarta el uso de nombre, apellido y nombre completo del usuario y por el contrario tomaremos de la Descripción el número de usuario asociado a la persona; a este número la aplicaremos ENCRIPTACIÓN SIMÉTRICA para anonimizarlo.
   b. A la dirección IP se aplica el mismo principio de encriptación y adicionalmente mientras no sea estrictamente necesario, no se mostrará el número resultante sino más bien la región, ciudad, código postal o información derivada de estos.

   6.1. Aplique técnicas de anonimización:

   Procederemos a encriptar la información llevando los números a letras mediante el uso de la siguiente clave:

Tabla 5 *Clave de anonimización*

Número	Letra
0	W
1	B
2	R
3	C
4	D
5	Q
6	E
7	X
8	A
9	Z

Fuente: Elaboración propia (2022)

La información quedará oculta de la siguiente manera:

Figura 4 *Información inicial*

Hora	Nombre completo del usuario	Usuario afectado	Contexto del evento	Componente	Nombre evento	Descripción	Origen	Dirección IP
[redacted]								
17/03/2022 08:37	ANGEL DAVID MURGANO CÁCERES	-	Curso: CREACIÓN DE EN	Sistema	Curso visto	The user with id '32445' viewed the section '6' of the course with id '54012'	web	201.234.176.252
[redacted]								
17/03/2022 08:36	DANIEL HERNANDEZ GOMEZ	-	Curso: CREACIÓN DE EN	Sistema	Curso visto	The user with id '32445' viewed the section '4' of the course with id '54012'	web	201.234.176.252

Fuente: Elaboración propia (2022)

Figura 5 *Información final*

Hora	Nombre completo del usuario	Usuario afectado	Contexto del evento	Componente	Nombre evento	Descripción	Origen	Dirección IP
[redacted]								
17/03/2022 08:37	DQRCC	-	Curso: CREACIÓN DE EN	Sistema	Curso visto	The user with id 'DQRCC' viewed the section '6' of the course with id 'QDW	web	RWB.RCD.BXE.RQR
[redacted]								
17/03/2022 08:36	CRDDQ	-	Curso: CREACIÓN DE EN	Sistema	Curso visto	The user with id 'CRDDQ' viewed the section '6' of the course with id 'QDW	web	RWB.RCD.BXE.RQR

Fuente: Elaboración propia (2022)

Anexo B

PRIMERA ETAPA (Semestre I) - ANÁLISIS DE DATOS: Variables Objetivo

Con los datos anonimizados se reduce el riesgo de identificación y se procede a conectar los reportes usando los nombres completos como llaves primarias y foráneas, para finalmente enlazar el número de usuario (encriptado) como la variable que determinará al individuo en cada acción y en cada resultado obtenido dentro de la data. Es importante recalcar que el objetivo de esta primera fase de análisis es el de poder construir una serie de indicadores que expongan un mapa del comportamiento del usuario frente a los escenarios de evaluación, y a partir de allí poder identificar patrones que nos permitan armar un medidor de conductas para clasificar a los individuos y más aún, para poder analizar posibles acciones éticas y no éticas.

Para lograr el desarrollo de este mapa de comportamiento debemos inicialmente enfocarnos en conseguir identificar lo que se consideraría como un comportamiento "normal" es decir un comportamiento sin ningún tipo de sesgo o dato atípico, con el cual podremos crear una regla para medir escenarios futuros e indicar con cierto grado de certeza si estos resultan contener situaciones típicas o situaciones atípicas, siendo esto nuestra brújula para el desarrollo posterior de la clasificación; para iniciar la construcción debemos primero enfocarnos en un grupo de variables que reflejen la conducta de los usuarios dentro de las evaluaciones en el tiempo, para así detectar cambios que pueden presentarse, es por ello que en esta primera etapa tomaremos el tiempo de duración de la evaluación por usuario, adicionalmente y con el fin de incluir la variable calificación crearemos un indicador llamado "velocidad" el cuál relacionará puntaje con tiempo de evaluación, también tomaremos la variable dirección IP para detectar cambios de localización en el tiempo y finalmente

crearemos un indicador llamado “Distancia (Km) de IP vs IESoe Bogotá” el cual relacionará distancias en kilómetros entre las IPs y la localización de la IESoe en Bogotá.

- **Tiempo de duración de la evaluación por usuario**: Es el tiempo que tarda un usuario realizando la evaluación y se obtiene a través de la diferencia de horas, minutos y segundos entre la acción “started” y la acción “submitted” de la columna “Descripción” del reporte de interacciones de usuario.
- **Indicador de Velocidad**: Esta variable surge de la división del puntaje obtenido del usuario (reporte de calificaciones) entre el tiempo de duración de la evaluación por usuario y nos sirve como herramienta para relacionar minutos vs calificación, teniendo de esta manera que a menor velocidad se implicaría menos puntaje y mayor tiempo de desarrollo y a mayor velocidad lo contrario.
- **Dirección IP**: Para el uso de esta variable hemos desarrollado el código de Python abajo mostrado, el cual nos permite tomar las direcciones IP de los reportes (colocándolas en un archivo .txt llamado codIP.txt) y conectar con una API en www.ipwho.is [8]quien nos expone data de geolocalización como región, ciudad, país, código postal, latitud, longitud, entre otras y finalmente traduce la información y la consolida en un archivo de txt llamado finalcodIP.txt.
- **DISTANCIA (Km) de IP VS IESoe BOGOTÁ:** Con el archivo generado en el proceso anterior desarrollamos este indicador el cual nos permite medir los kilómetros que hay entre cada dirección IP por donde ingreso el usuario vs la localización de la IESoe en Bogotá, la idea es tener un punto de referencia estático

[8] http://www.ipwho.is/

que nos permita medir variaciones de localización entre sesiones ya que los usuarios suelen conectarse mediante varias direcciones IP; para el desarrollo de este indicador tomamos la latitud y longitud de cada IP y la latitud y longitud de la IESoe en Bogotá y mediante la fórmula de Haversine (ver Figura 4) obtenemos las distancias en kilómetros.

Fórmula de Haversine:

$$Distancia = 2 * R * asin\sqrt{sin^2\left(\frac{\Delta lat}{2}\right) + \cos(lat1) * \cos(lat2) * sin^2\left(\frac{\Delta lon}{2}\right)}$$

Donde:
Lat = Latitud
Lon= Longitud
(lat1, lon1) = Latitud y longitud en el primer punto
(lat2, lon2) = Latitud y longitud en el segundo punto
DeltaLat = lat2-lat1
DeltaLon = lon2-lon1
R = Radio de la tierra: 6372.7954

Fórmula de Haversine en código de Python (Murgano, 2022)

```
import re
import json
from urllib.request import urlopen
import requests
a=[]
with open('codIP.txt','r') as archivo:
   for i in archivo:
     if i[-1]=='\n':
       a.append(i[0:len(i)-1])
     else:
       a.append(i)
```

```
for i in a:
  response = urlopen('http://ipwho.is/'+i)
  ipwhois = json.load(response)
  archivoescritura=open('finalcodIP.txt','a')

archivoescritura.write(ipwhois['ip']+','+ipwhois['city']+','+ipwhois['region']+','+ipwhois['co
untry']+','+ipwhois['postal']+','+str(ipwhois['latitude'])+','+str(ipwhois['longitude'])+'\n')
  archivoescritura.close()
```

SEGUNDA ETAPA (Semestre II) - ANÁLISIS DE DATOS: Variables Objetivo

En esta segunda etapa se adicionaron nuevas variables de manera de ampliar la fuente de insumos y poder generar nuevos insights que robustezcan los KPIs asociados; cabe aclarar que se mantiene la anonimización del ID de estudiante, se mantienen las variables de “Velocidad”, “Puntaje” y “Tiempo de duración de la evaluación por usuario (Minutos)” y la variable de “DISTANCIA (Km) de IP vs IESoe BOGOTÁ” se abre en “latitud y longitud” de manera de geolocalizar la posición y finalmente se adicionan las siguientes variables:

- **Tipo de Examen:** El reporte de Interacciones nos indica el tipo de evaluación realizada: Cuestionario, Tarea, Parcial, etc.; este insumo nos permite identificar las características de la evaluación.
- **Número de interacciones en el sistema:** El reporte del Log de Usuario nos proporciona la cantidad de interacciones que este ha tenido dentro de la plataforma (interacciones no solo de evaluaciones sino con la plataforma en sí).
- **Cantidad de IP’s manejados por el estudiante:** El reporte del Log de Usuario indica el IP manejado por el usuario por lo que a partir de esta información

contamos la cantidad de IP’s que maneja el usuario al momento de ingresar a las evaluaciones.

- **Indicador de repetición de IP’s en más de1 estudiante:** El estudio de la cantidad de IP’s nos permite también identificar si uno (o varios) IP’s son usados simultáneamente por más de 1 estudiante.
- **Indicador de realización de varias evaluaciones durante el mismo día de parte del estudiante:** Usando el reporte de calificaciones obtenemos la información de aquellos casos donde el estudiante activa y genera más de 1 evaluación en un mismo día.
- **Indicador de finalización de evaluación por parte del estudiante:** El Log de usuario nos presenta la información de si el usuario activo una evaluación y si la finalizo; esta información es importante para determinar el comportamiento de cada usuario frente a las evaluaciones.
- **Cantidad de visualizaciones que tuvo cada examen de parte del estudiante:** Cuando un usuario activa una evaluación el Log identifica hasta qué punto este visualizo cada parte y este dato nos permite entender en qué punto para el usuario o más bien hasta donde avanzan.
- **Indicador de patrones para casos de varias evaluaciones un mismo día:** Para los casos de usuarios que realizan más de 1 evaluación el mismo día, se busca de estudiar su interacción para identificar si existe un patrón en el tiempo de cada respuesta de manera que determinar si el comportamiento es atípico.
- **Identificador de velocidad atípica (Con Z-score):** Esta variable usa el Z-SCORE para estudiar si cada dato se puede considerar “atípico” o no.

Anexo C

ANÁLISIS DE DATOS: Detección de datos atípicos

Cómo se indicó anteriormente, nuestro objetivo es el de construir el comportamiento del usuario frente a las evaluaciones, por lo que una vez determinadas las variables a usar debemos enfocarnos en identificar los casos atípicos y así con ellos podamos armar un mapa de la situación normal que usaremos como base comparativa posteriormente para la detección de situaciones anormales, en este sentido tomamos inicialmente la variable VELOCIDAD para el estudio ya que esta relaciona puntaje con tiempo y nos permite exponer situaciones de minutos vs calificaciones que sean anómalas.

Para definir que un dato es anómalo frente a un conjunto usamos la fórmula de Edward Altman llamada Z SCORE o PUNTAJE Z (ver Figura 5), la cual genera un puntaje de la data usando la desviación estándar y el promedio y nos indica que de poseer un valor absoluto por encima de 3 estamos frente a un caso atípico (y viceversa).

Formula Z SCORE

Formula de la Media Poblacional:

$$\mu = \frac{\sum x}{n}$$

Fórmula para desviación estándar poblacional:

$$\sigma = \sqrt{\frac{\sum (x - \mu)^2}{n}}$$

Formula de Z-Score

$$z\ score = \frac{(x - \mu)}{\sigma}$$

Finalmente obtenemos de resultado que los siguientes usuarios tuvieron un PUNTAJE Z superior a 3 por lo que los consideramos datos atípicos e incluso observando la data se concluyen casos claramente anormales con puntajes máximos (5 puntos) en tiempos incluso menores a 1 minuto, es decir usuarios que respondieron perfectamente evaluaciones de 20 preguntas en segundos.

Tabla 6 *Puntaje Z*

USUARIO	Z SCORE	VELOCIDAD
RWEBQ	3,83	3,50
RAWRZ	4,59	3,50
DZXZQ	3,57	3,50
ABEE	4,59	5,00
BCZRE	4,59	5,00
XCEA	4,59	5,00
EXCC	4,59	5,00
CCDQE	4,59	3,50
BRCWC	4,59	5,00
XXDB	4,59	5,00
DWADR	4,59	3,50
CCCWR	4,33	3,50
QQRX	3,83	4,25
EDAQQ	4,08	3,50
RBEBR	4,59	3,50
QAWQ	3,07	3,50
BRWDQ	4,59	5,00
ZDCE	3,83	5,00
BEXE	4,59	5,00
RDZWB	4,59	3,50
QDQWD	4,59	3,50
BZBCX	4,59	3,50
XBRB	4,59	5,00
XXQE	4,59	5,00
BXQEQ	4,59	5,00
BEEAR	4,59	5,00
RCRAA	4,59	3,50

BZAXX	4,59	3,50
RDRZD	4,59	3,50
BDEQW	4,59	5,00

Fuente: Elaboración propia (2022)

Con la identificación de estos usuarios con velocidades atípicas procedemos a construir un mapa para cada variable con su situación normal (retirando estos usuarios) y con su situación anormal (manteniendo los usuarios atípicos) para así comparar los casos de la data actual.

Anexo D

PRIMERA ETAPA (Semestre I) - ANÁLISIS DE DATOS: Tiempo de duración en la evaluación por usuario

Con la información de tiempos de duración por usuario para los parciales 1, 2 y 3 se procede a generar un estudio estadístico basándonos en su comportamiento bajo una distribución normal, generando las curvas abajo mostradas y los siguientes indicadores:

- Coeficiente de asimetría: Es un coeficiente que indica el grado de desface/asimetría de una curva, teniendo la siguiente indicación:
    - Coef<0: curva asimétrica a la izquierda.
    - Coef=0: curva simétrica.
    - Coef>0: curva asimétrica a la derecha.
- Sesgo: También conocido como el coeficiente de asimetría de Pearson y genera un coeficiente similar al anterior con las mismas características de medición respecto al cero.
- Curtosis: Es un indicador del grado de concentración de los datos respecto a la zona central de la curva y se divide en:

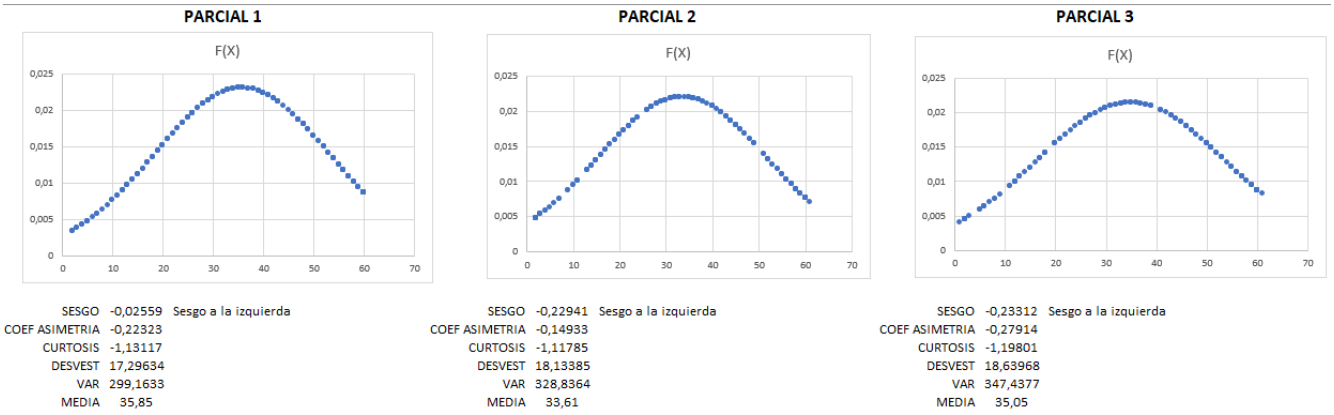
- Leptocúrtica: Donde existe gran concentración de los datos en torno a la media: (indicador > 3)

- Mesocúrtica: Donde existe concentración normal de los datos en torno a la media: (indicador = 3)

- Platicúrtica: Donde existe baja concentración de los datos en torno a la media: (indicador < 3)

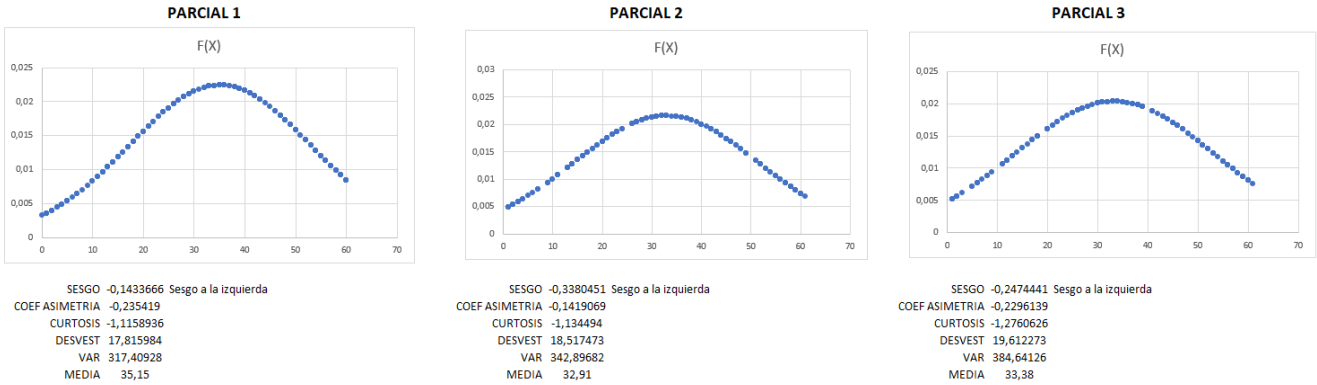
- Desviación Estándar: Es la raíz cuadrada de la varianza y representa la dispersión de los datos respecto a la media.
- Varianza: Es la medida de dispersión de los datos respecto a la Media.
- Media: Es el promedio de los datos y dentro de la curva representa el caso de mayor frecuencia.

Figura 6 *Escenario sin usuarios atípicos*



Fuente: Elaboración propia (2022)

Figura 7 *Escenario con usuarios atípicos*



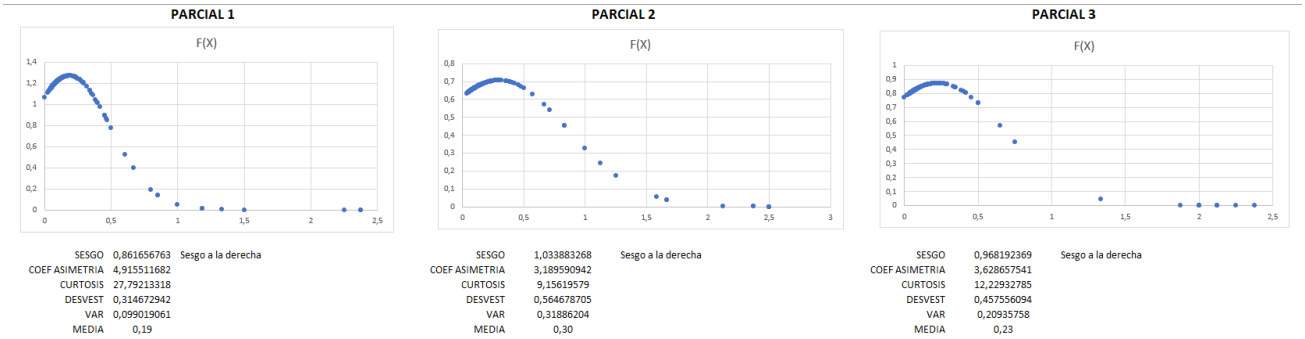
Fuente: Elaboración propia (2022)

Basado en las gráficas y métrica anterior podemos inferir que en los casos con usuarios atípicos tenemos una situación de sesgo de la curva hacía la izquierda mucho más pronunciada y más platicúrtica (más plana), adicional las medias son menores lo que sugiere finalmente es que los tiempos de desarrollo de las evaluaciones tienen a ser menor con la inclusión de los casos anormales, en otras palabras la anormalidad se presenta a tiempos más bajos sin embargo, es menester estudiar cómo son estos tiempos vs las calificaciones para determinar si son casos de baja nota y bajo tiempo o por el contrario alta nota y bajo tiempo (revisión de la velocidad a continuación).

ANÁLISIS DE DATOS: Velocidad

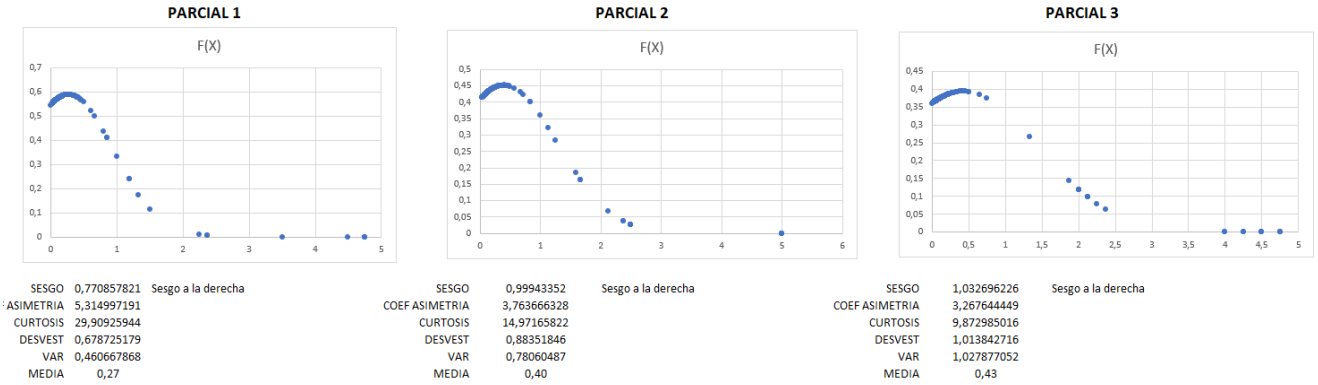
Con la información de VELOCIDAD por usuario para los parciales 1, 2 y 3 se procede a generar un estudio estadístico basándonos en su comportamiento bajo una distribución normal, generando las siguientes curvas:

Figura 8 *Escenario sin usuarios atípicos (velocidad)*



Fuente: Elaboración propia (2022)

Figura 9 *Escenario con usuarios atípicos (velocidad)*



Fuente: Elaboración propia (2022)

Analizando las curvas anteriores se identifica que en todos los casos existe un sesgo a la derecha y una curtosis del tipo platicúrtica que implica mayor aglomeración de datos en

torno a la media, esto nos habla de un comportamiento en donde suele ocurrir que el usuario obtiene notas altas a mayor tiempo de evaluación; adicionalmente podemos detectar que en el escenario normal los casos más alejados de la media manejan velocidades por debajo de 2.5 sin embargo, en el escenario atípico tenemos velocidades de 4 en adelante lo que significa usuarios que desarrollaron todo el parcial en segundos y sacaron notas muy altas (las probabilidades de esta situación serán estudiadas en fases futuras, sin embargo efectivamente representa un caso atípico).

SEGUNDA ETAPA (Semestre II) - ANÁLISIS DE DATOS

El análisis de los nuevos datos adicionados es realizado en el ANEXO E usando la técnica de ACP.

Anexo E

PRIMERA ETAPA (Semestre 1) - DESARROLLO DEL ANÁLISIS DE VARIABLES:

Se busca generar un análisis a fondo de las variables para identificar las tendencias de estas y las correlaciones entre ellas, para esto el estudio se enfoca en un análisis factorial el cual es una herramienta que permite cotejar N variables y estudiar gráficamente su comportamiento incremental y niveles de relación. Se ejecutó la herramienta usando código en R, y se obtuvieron los siguientes insights:

- Realizando el análisis de las variables se identifica que “KPI Velocidad de Notas es la que más contribuye al análisis con 53.24% dentro del espectro de dimensiones (Dim.1, 2, 3 y 4).

Tabla 7 *Análisis de variables por contribución*

	Dim.1	Dim.2	Dim.3	Dim.4
PUNTAJE	13.85	54.51	1.91	29.73
MIN EVALUACION	19.57	45.12	4.92	30.40
Distancia (Km) desde IP vs Bogota	13.58	0.13	86.18	0.11
KPI Velocidad de Notas	53.00	0.24	6.99	39.76

Fuente: Elaboración propia (2022)

- Se desarrolla un análisis para medir la calidad de representación de cada variable dentro del modelo para su espectro de dimensiones (Dim.1, 2, 3 y 4).

Tabla 8 *Análisis de variables por calidad de representación*

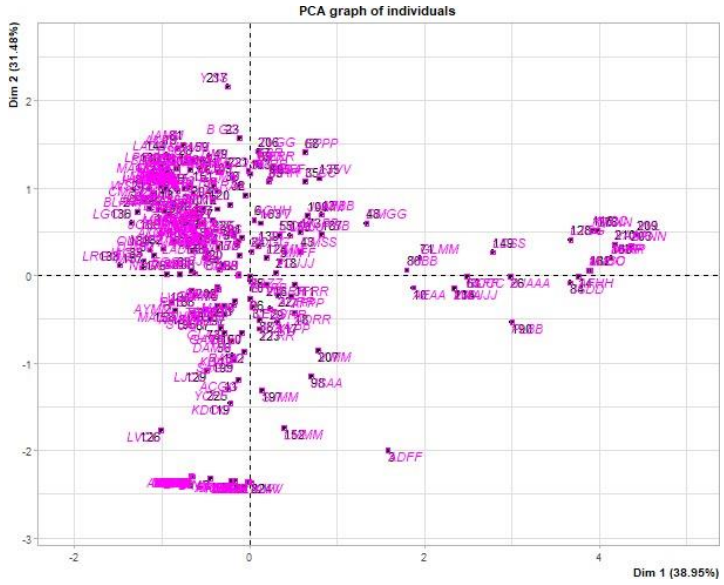
	Dim.1	Dim.2	Dim.3	Dim.4
PUNTAJE	0.22	0.69	0.02	0.08
MIN EVALUACION	0.30	0.57	0.04	0.08
Distancia (Km) desde IP vs Bogota	0.21	0.00	0.79	0.00
KPI Velocidad de Notas	0.83	0.00	0.06	0.11

Fuente: Elaboración propia (2022)

  - La variable que tiene menor calidad de representación es la Distancia dado que su Cos^2 es 0,21.
  - La variable que tiene mayor calidad de representación es la Velocidad dado que su Cos^2 es 0,83.

- Generamos un análisis ACP donde graficamos el comportamiento en 2 dimensiones frente a las variables usadas, donde podemos observar que existen estudiantes en los cuadrantes del lado derecho que se alejan de manera importante del promedio:

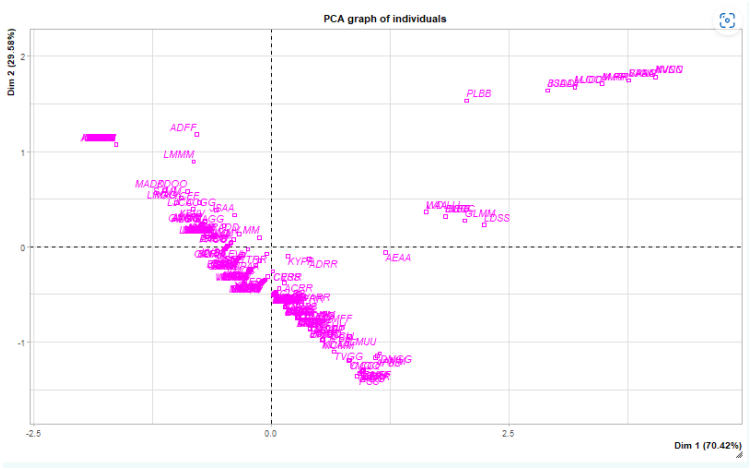
Figura 10 *Análisis ACP inicial*



Fuente: Elaboración propia (2022)

- Ampliando de manera más detallada el análisis del punto anterior se ajustó la gráfica para que solo presentara la data de puntaje y de velocidad de manera que se puede representar más claramente un grupo atípico (Cuadrante derecho superior) con velocidades muy altas y puntajes altos respecto al promedio.

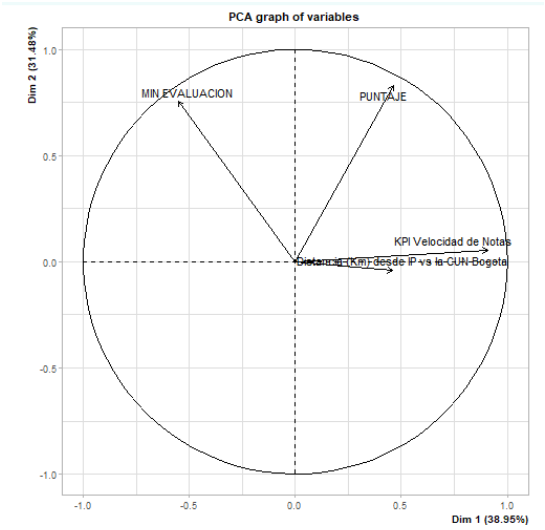
Figura 11 *Análisis ACP ajustado*



Fuente: Elaboración propia (2022)

- Estas variables que se alejan del promedio presentan las siguientes características según el análisis del ACP de la gráfica de variables:

Figura 12 *Análisis ACP (gráfica de variables)*



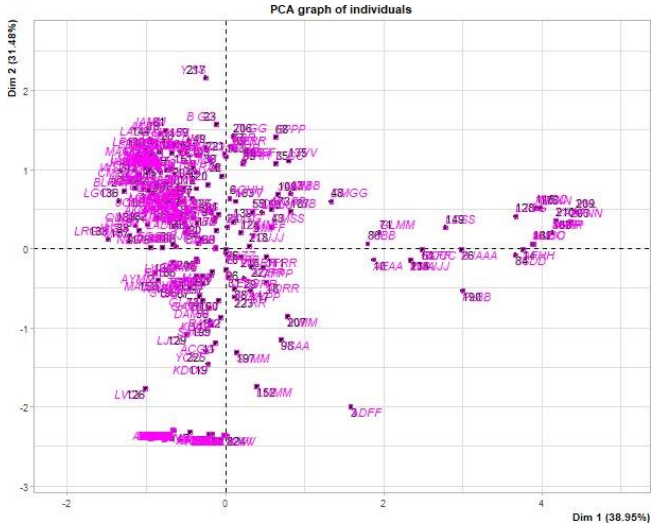
Fuente: Elaboración propia (2022)

- Entre más al extremo derecho manejan más velocidad de notas, lo que quiere

decir que obtienen mayor nota en menos tiempo y revisando la data de relacionada a estos casos se tiene usuarios que sacan notas cercanas al máximo (entre 4 y 5 puntos) en tiempos tan cortos como 1 minuto o menos, lo que conlleva a tener una alerta sobre estos usuarios.

- Dado que el ángulo entre el KPI de velocidad y el de KM es muy bajo (menor a 45°) se infiere una relación estrecha entre estas dos variables lo que en conclusión implica que los alumnos con mayores velocidades se encuentran más alejados de la IESoe.
- La relación entre MIN Evaluación y velocidad es inversa (>90%) lo que implica que en la data se presentan casos extremos que relacionan notas muy altas a velocidades muy bajas, afectando de esta manera la correlación.
- El cuadrante derecho inferior indica grupo de usuarios quienes tuvieron una alta velocidad, pero en su mayoría bajo puntaje; esto también va de la mano con la relación entre velocidad y puntaje la cual presenta ser positiva (<90°) según la gráfica, pero con menor correlación que con la indicada hacia los KM.
- El cuadrante superior izquierdo concentra a la mayor parte de los estudiantes e indica los casos de altos tiempos de evaluación con relaciones de puntajes divididas en un aparente 50%-50%.

Figura 13 *Análisis ACP*

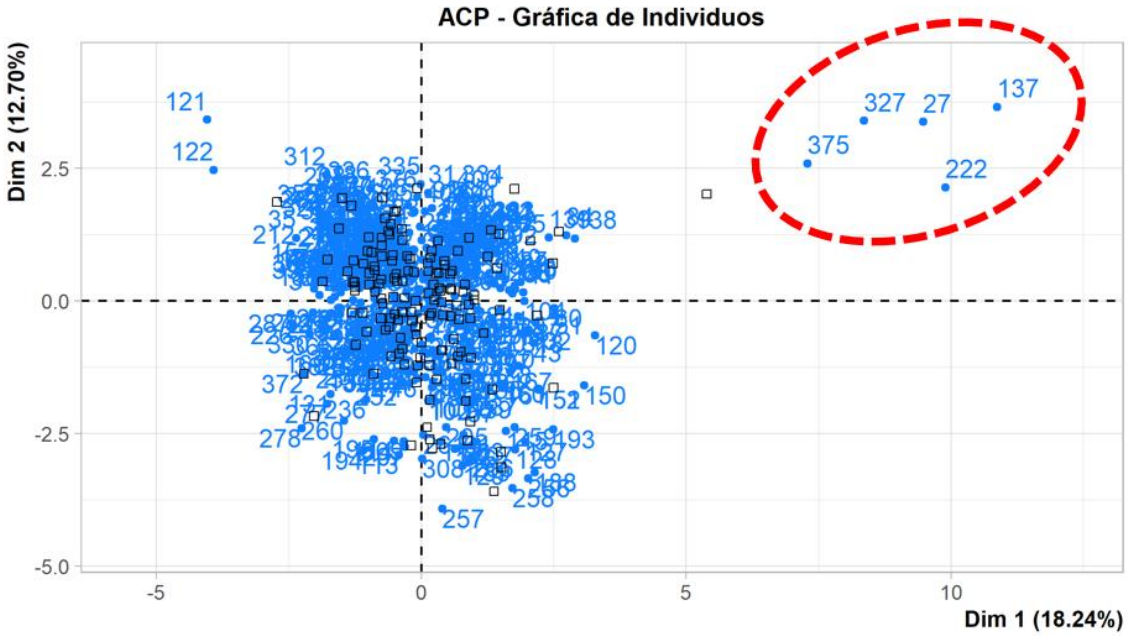


Fuente: Elaboración propia (2022)

SEGUNDA ETAPA (Semestre II) - DESARROLLO DEL ANÁLISIS DE VARIABLES:

Con las nuevas variables adicionadas en esta segunda etapa se realizó el análisis ACP (Análisis de Componentes Principales) el cual nos permitió exponer los grupos cuyos comportamientos son atípicos, dando de esta manera una alerta sobre el grupo de usuarios que deben ser estudiados más a fondo; como etapa final, se incluyeron estas visualizaciones dentro del BI para poder exponer de manera clara los grupos de interés.

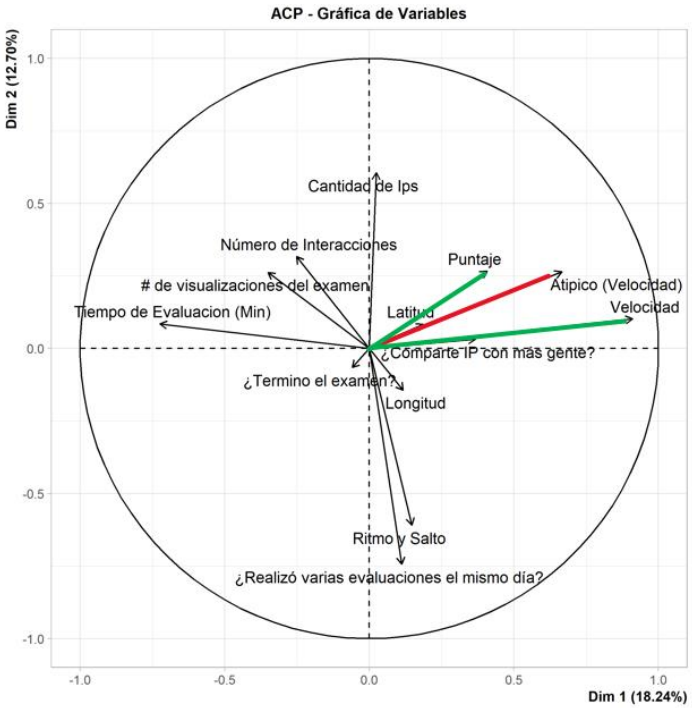
Figura 14 *Atípicos en análisis ACP*



Fuente: Elaboración propia (2022)

Adicionalmente se realizó un análisis factorial para las variables del cual se obtuvieron los siguientes insights respecto a la correlación vs la variable “Atípico”:

Figura 15 *Atípicos en análisis factorial*



Fuente: Elaboración propia (2022)

- La velocidad y puntaje tienen estrecha relación entre sí, lo cual es coherente dado que el indicadore Velocidad surge del uso del puntaje.
- La variable “Tiempo de Evaluación (Min)” posee una estrecha relación inversa, es decir a menor velocidad tiene a ser más “Atípico”.
- Variables como “latitud y longitud” poseen una fuerte relación con los casos atípicos.
- Si se realizaron varias evaluaciones el mismo día sugiere no tener una relación fuerte con la variable “Atípico”.

DESARROLLO DEL MODELO PREDICTIVO:

Con la finalidad de generar una herramienta que permita estudiar el comportamiento de los estudiantes basado en las variables anteriormente mostradas, nos apoyamos en la ciencia del Machine Learning la cual usando una serie de modelos estadísticos y alimentándose de la data histórica nos permite generar escenarios probabilísticos para el análisis predictivo. El foco principal de esta herramienta será la de indicar un porcentaje de probabilidad de que un usuario X esté teniendo un comportamiento atípico, para esto el modelo de Machine Learning será alimentado con data histórica que presenta casos normales y casos atípicos para "entrenar" un modelo matemático que intrínsicamente identifique escenarios fuera del comportamiento común.

Existe una gran variedad de modelos de Machine Learning, cada uno basado en una teoría matemática distinta, pero con el mismo foco de analizar los comportamientos históricos; en este sentido para el proyecto se hizo uso de los siguientes modelos:

- Árbol de Clasificación.
- Gradient Boosting.
- Random Forest.
- Máquina de Soporte Vectorial.

El componente diferenciador y el factor determinante para tomar un modelo u otro a medida que la base de información crezca o cambie será el AUC o también llamado “área bajo la curva”, esta es una medida que indica el nivel de calidad de un modelo, qué tan calibrado está y que tan asertivo es para predecir la data de prueba; el AUC es una medida que va del 0.5 al 1 y que para efectos de nuestra herramienta, tomaremos como brújula para decidir el modelo a manejar; así pues una parte importante de la herramienta será la de cotejar internamente los resultados y ofrecer al usuario la mejor y más acertada opción.

NOTA: Dentro del desarrollo se busca el poder obtener un AUC alto sin embargo cabe aclarar que a nivel estadístico un modelo óptimo debe manejarse con AUC de entre 0.7 y 0.9 ya que un AUC de 1 no es coherente por referirse a un resultado probabilístico perfecto el cual no es realista.

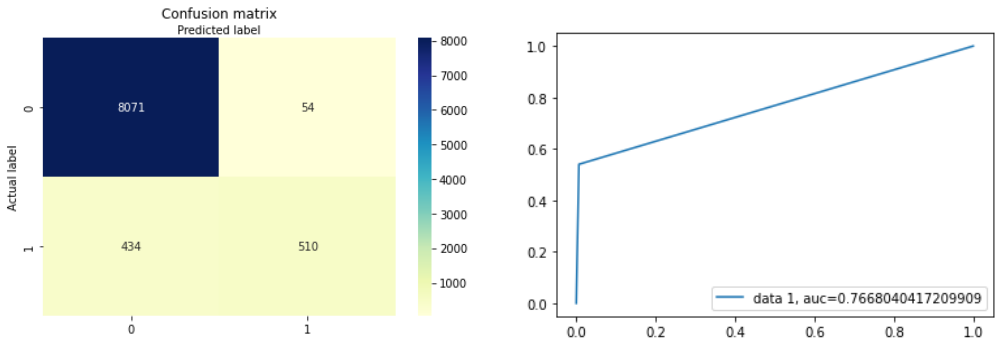
Actualmente la herramienta fue generada en código Python y se desarrolló una interfaz de usuario que permite un manejo óptimo; actualmente, el código se divide en 4 etapas:

1. **Cargue de la información:** En esta etapa se cargan las librerías a usar para el análisis descriptivo y se realiza el cargue de la base de datos proporcionada por la IESOE.
2. **Análisis descriptivo de la data:** Se realiza una revisión general de la base de datos para comprender su comportamiento y ajustar data vacía o nula.

3. **Preparación de la data:** Se define la variable objetivo y las variables independientes, adicionalmente se configura los parámetros de entrenamiento del modelo. Adicionalmente se usa la librería de PYTHON llamada SMOTE y la técnica de UNIONES DE TOMEK para poder realizar un balance de la data de manera que el modelo pueda tener un entrenamiento más claro respecto al grupo minoritario que vienen siendo los estudiantes con comportamiento de velocidad atípica.
4. **Modelación:** Se entrena cada modelo con la base de datos histórica, se obtiene la matriz de confusión de cada uno y se estudia el AUC generado; para efectos informativos, compartimos un ejemplo desarrollado con la data de estudiantes de un semestre de uno de los cursos de la IESoe:

- **Modelo de Árbol de Clasificación:**

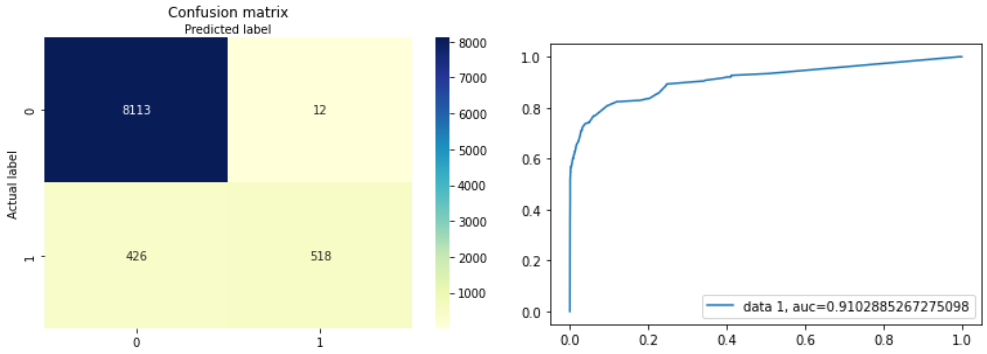
Figura 16 *AUC para Modelo de Árbol de Clasificación*



Fuente: Elaboración propia (2022)

- **Modelo de Gradient Boosting:**

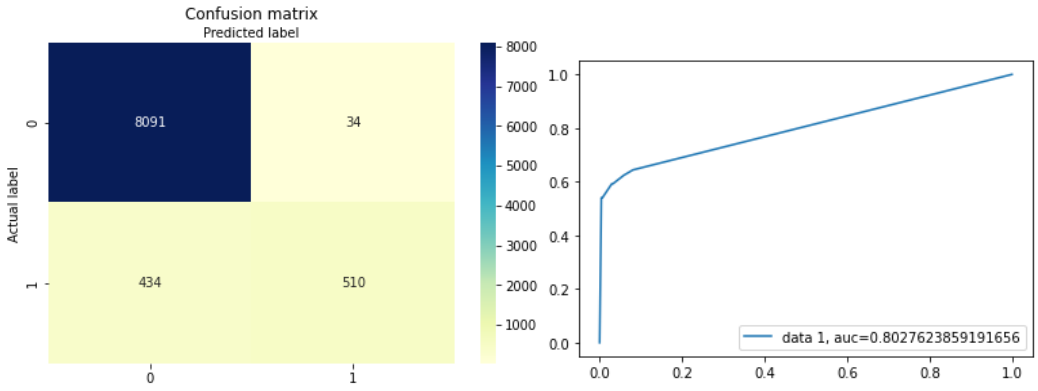
Figura 17 *AUC para Modelo de Gradient Boosting*



Fuente: Elaboración propia (2022)

- **Modelo de Random Forest:**

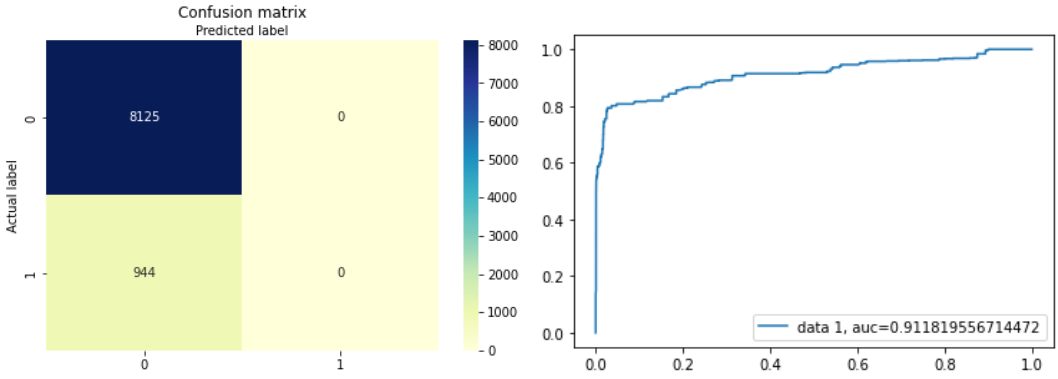
Figura 18 *AUC para Modelo de Random Forest*



Fuente: Elaboración propia (2022)

- **Modelo de Máquina de Soporte Vectorial:**

Figura 19 *AUC para Modelo de Máquina de Soporte Vectorial*



Fuente: Elaboración propia (2022)

5. **Predicción:** Luego de entrenar los modelos y generar los indicadores, la herramienta toma la opción con el AUC más alto, indica el mejor modelo y comparte la predicción; para efectos del ejemplo anterior tenemos:

```
El modelo más optimo es la Maquina de Soporte Vectorial
La probabilidad que el alumno X sea atipico es de 34.5 %
```

TERCERA ETAPA (Semestre III) - DESARROLLO DE LA APLICACIÓN

Ya habiendo identificado las variables críticas que alimentaran el modelo y ya habiendo desarrollado el código en Python que permitirá mediante el uso de machine learning arrojar un porcentaje de probabilidad frente a los comportamientos atípicos, se desarrollará una aplicación que permitirá facilitarle al usuario la interacción con la herramienta y le expondrá una serie de KPI's basados en la data introducida.

Esta etapa se dividirá en los siguientes puntos:

1. **Base de datos:** Se detalla la estructura en el servidor donde es almacenada la información trabajada en el backend mediante lenguaje SQL.
2. **Backend:** Se explica la estructura en Python que interactúa con la información y genera el output que se carga en el servidor en internet.
3. **Frontend:** Es la interfaz gráfica mediante la cual el usuario interactúa y carga los reportes que serán usados como insumo para la herramienta.
4. **Dashboard:** Es una interfaz en Power BI donde se muestran las analíticas de los KPI's generados por el modelo y almacenado en la base de datos.
5. **Manual:** Se indica los pasos para el manejo de la herramienta.

BASE DE DATOS:

La base de datos se encuentra en un servidor privado en internet con el cual interactuamos mediante una conexión de Python-SQL y unas credenciales para garantizar la seguridad; la base de datos contiene las siguientes tablas:

- **DATAUSUARIOS:** En esta tabla se encuentra la información histórica de cada indicador calculado para cada evaluación realizada por cada estudiante; esta tabla va creciendo con cada ejecución y la variable atípica va siendo recalculada en cada ejecución de manera que el modelo se va calibrando, permitiendo mejorar la precisión por usuario en el tiempo, así como permitiendo ajustarse a la evolución del comportamiento en el tiempo. Es importante recalcar que esta tabla almacena el ID y IP anonimizado de cada usuario.

- **INFO_IP:** Esta tabla recopila la latitud y longitud de cada IP (IP anonimizado); esta tabla va creciendo con cada ejecución y permite optimizar el uso de los GET-POST ya que el rastreo comienza primero cotejando esta base histórica.

- **PROB_USUARIOS:** Esta tabla contiene la probabilidad de comportamiento atípico por usuario (ID anonimizado), el nombre del modelo con mejor precisión y su AUC; esta tabla no crece con cada ejecución sino más bien muestra únicamente los usuarios del reporte ingresado en la consulta puntual y en la siguiente ejecución la data contenida es borrada y sustituida por la data de la consulta de esa nueva ejecución.

- **PROB_USUARIOS_HIST:** Esta tabla contiene la probabilidad de comportamiento atípico por usuario (ID anonimizado), el nombre del modelo con mejor precisión y su AUC; al contrario de la anterior tabla, esta tabla va creciendo con cada ejecución y la idea es que permita mantener un histórico de la probabilidad de comportamiento atípico por cada estudiante que permita analizar los cambios en el tiempo.

- **TEMP_ENSAMBLE:** Esta tabla mantiene de manera temporal la misma información de la tabla DATAUSUARIOS para poder ser usada dentro de una parte del proceso de backend; en cada ejecución la información contenida es eliminada y cargada con la data histórica y la nueva data.

BACKEND:

El Código de esta herramienta está creado en Python y maneja las librerías abajo mostradas (al final de este punto); el código construye la herramienta y todas sus funcionalidades, permitiendo principalmente el cruce de la data, su ajuste, anonimización, construcción de indicadores, calibración de modelos, entre otros.

A continuación, se indica acción tras acción realizada por el código para el procesamiento de la data:

1. Al cargar el reporte de notas, el código identifica si se cargó la data correcta o no mediante la revisión de los encabezados del reporte y luego arroja un mensaje indicando si este es correcto (cargue exitoso) o no (reporte incorrecto).
2. Al cargar el reporte de log, el código identifica si se cargó la data correcta o no mediante la revisión de los encabezados del reporte y luego arroja un mensaje indicando si este es correcto (cargue exitoso) o no (reporte incorrecto).
3. Al iniciar la ejecución mediante la activación del botón “Analizar archivos” el código verifica si se cargaron correctamente los reportes o si falta uno o si faltan los dos y de existir un problema arroja un mensaje solicitando al usuario el correcto cargue y no continua con la ejecución, caso contrario continua los siguientes pasos.
4. Para cada reporte se crea una llave con los datos de nombres+apellidos+fecha&hora de inicio de evaluación.
5. Con el uso de las llaves previamente generadas se combinan los dos archivos para obtener uno solo con toda la información.
6. Se ajusta la data de “tiempo de evaluación” a minutos.

7. Se genera la variable "Interacciones" mediante el conteo de las llaves; con esto se identifica el número de acciones realizadas por cada usuario.

8. Se filtran todos los eventos diferentes a aquellos relacionados con el inicio de cada evaluación y luego se eliminan.

9. Para cada usuario, se genera la variable "Velocidad" dividiendo puntaje entre tiempo.

10. Para cada usuario, se genera la variable binaria "MuchosExamenes" mediante el conteo de la llave, indicando 0 si no hay llaves repetidas y 1 si las hay; la existencia de una llave repetida implica que un estudiante realizó más de una evaluación el mismo día.

11. Se genera la variable binaria "CantidadInteraccionesExamen" mediante el conteo de todos aquellos eventos (columna “evento”) que comienzan con la palabra "visto", de manera que se cuantifica el número de interacciones que tuvo el estudiante con la evaluación.

12. Se crea una tabla con todas las variables incluyendo las anteriormente generadas.

13. La cedula (ID) y la IP de los usuarios se anonimizan.

14. Se eliminan las columnas que hacen referencias a nombres y apellidos (incluyendo la llave)

15. Se importa la data que se encuentra en la tabla DATAUSUARIOS y se consolida con la tabla indicada en el punto 12.

16. Se eliminan duplicados para evitar que se repita la información.

17. Se genera la variable "CantidadIPs" mediante el conteo de las repeticiones de cada IP por estudiante; esto nos permite cuantificar cuántas IP’s usa el estudiante para conectarse.

18. Se genera la variable binaria "MuchosIP" mediante la construcción de una tabla temporal con IP y Cedula (ID) en la cual se eliminan los duplicados y se cuentan las IPs, asignando 1 si se repite más de una vez y 0 en caso contrario; esto nos permite indicar si existe una misma IP usada por más de un estudiante.

19. La tabla resultante con todas las variables es subida a la base de datos, específicamente a la tabla "TEMP_ENSAMBLE" cuyo contenido es previamente vaciado para asegurar que solo se cargue la nueva información junto con la data histórica (punto 15).

20. Se realiza la importación de la data contenida en la tabla "INFO_IP"

21. Se realiza el análisis de geolocalización: para ello se activa un proceso por etapas donde inicialmente se toma cada IP y se busca la "latitud" y la "longitud" dentro la tabla “INFO_IP” y de no encontrarse allí se procede a des-anonimizar cada IP para iniciar un proceso de consultas del tipo Get-Post con diferentes páginas en internet, iniciando con www.elhacker.net/geolocalización.html, y si la información no es conseguida allí entonces se procede a realizar la búsqueda en la página es.geoipview.com, y si persiste el problema entonces se busca en la página ipwho.is y finalmente de no conseguirse la información se coloca "latitud" y "longitud" 0 (cero), lo cual permite que en el dashboard final sea retirada la data que no esté geolocalizada mediante el filtraje y luego en una próxima ejecución se buscará nuevamente los casos 0 (cero).

22. Cada IP nuevo obtenido por fuera de la tabla "INFO_IP" es incluido en esta de manera anonimizada para tenerlo disponible en una próxima ejecución.

23. Se toma la data de la tabla "TEMP_ENSAMBLE" y se divide en 2 tablas: una con la data Histórica y otra con la Nueva data adicionada.

24. Se toma la tabla de data Histórica y se crean dos tablas adicionales: una con la variable objetivo (“Atípico”) y otra con las demás variables.

25. Con la tabla de data Histórica se realiza un proceso de balanceo mediante la librería SMOTE que estudia la información y genera muestras artificiales para poder balancear la cantidad de data atípica y típica y así disminuir la probabilidad de generar Overfitting.

26. Luego se realiza una calibración de la data anterior con la técnica de UNIONES DE TOMEK mediante la librería SMOTETomek, de manera que se genera una brecha entre la data típica y la atípica para así acentuar las diferencias.

27. Con la data resultante entrenamos 4 modelos: Árbol de Clasificación, Gradient Boosting, Random Forest y Maquina de Soporte Vectorial; usamos estos modelos entrenados para calcular la probabilidad de comportamiento atípico para la Nueva data.

28. Al entrenar cada modelo se obtiene el AUC (Area Under the Curve) de estos, el cual nos indica la calidad de la predicción, luego tomamos el caso cuyo AUC sea más cercano a 0.9.

27. Con los resultados construimos una nueva tabla que contiene el ID de usuario anonimizado, la probabilidad de ser atípico, el AUC y el nombre del modelo con mejor AUC; esta nueva tabla se carga en la base de datos en 2 sitios: PROB_USUARIOS (eliminando previamente el contenido que allí se encuentre) y en la tabla PROB_USUARIOS_HIST.

28. Se ajusta la variable binaria "Atípico", llevando a 1 a aquellas probabilidades mayores o iguales a 0.5 y llevando a 0 los casos contrarios.

29. Se toma la tabla final que contiene la data Histórica y la Nueva data (con el valor "Atípico" calculado) y se sube a la tabla DATAUSUARIOS (eliminando previamente el contenido) de manera que pueda ser usada en su totalidad como data Histórica en las próximas ejecuciones.

30. Se le muestra un mensaje al usuario indicando que el proceso ha finalizado exitosa e inmediatamente se apertura una pestaña en el navegador favorito con el Dashboard en Power BI, el cual se alimenta de la información de todas las tablas con que trabajamos y que luego de la ejecución presentan la información que acabamos de ingresar.

librerías de Python usadas: requests, pandas, numpy, seaborn, sklearn, pylab, imblearn, collections, matplotlib, re, json, urllib, bs4, beautifulsoup, tkinter, datetime, ntpath, os, webbrowser, random, pyodbc y sql.connector.

FRONTEND:

Con la ayuda de la librería TKINTER generamos la siguiente interfaz gráfica con la cual interactúa el usuario:

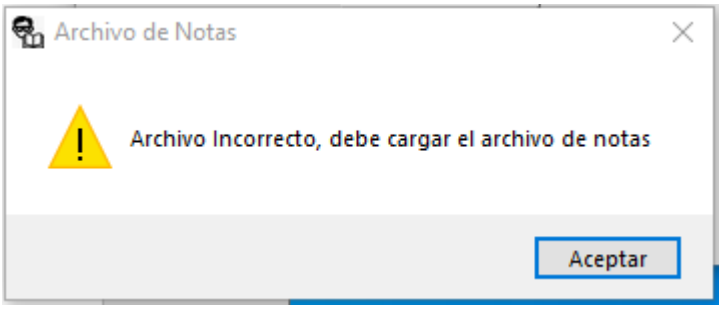
Figura 20 *Interfaz gráfica (FrontEnd)*

Fuente: Elaboración propia (2023)

Esta interfaz maneja los siguientes botones:

- **Botón Leer Archivo de Notas:** Al accionar este botón se desplegará una ventana donde el usuario navegará entre directorios para poder elegir el reporte de Notas a usar; en caso de elegir un archivo incorrecto se desplegará el siguiente mensaje:

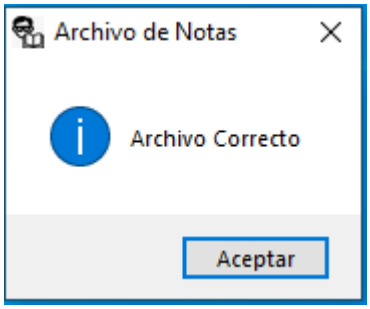
Figura 21 *Mensaje de carga incorrecta*



Fuente: Elaboración propia (2023)

Y en caso de elegir el archivo correcto se indicará el siguiente mensaje:

Figura 22 *Mensaje de carga correcta*



Fuente: Elaboración propia (2023)

- **Botón Leer Archivo de Logs:** Al accionar este botón se desplegará una ventana donde el usuario navegará entre directorios para poder elegir el reporte de Logs a usar y al igual que el botón anterior se desplegarán diferentes mensajes dependiendo de si se carga o no el archivo correcto.

- **Botón Analizar Archivos:** Al accionar este botón se activará el proceso de análisis de la información, y de haber sido cargado correctamente los anteriores

reportes entonces al finalizar el proceso de análisis se mostrará un mensaje indicando que el proceso fue satisfactorio, y en caso contrario se mostrará un mensaje indicando si falta cargar alguno de los dos reportes o ambos inclusive.

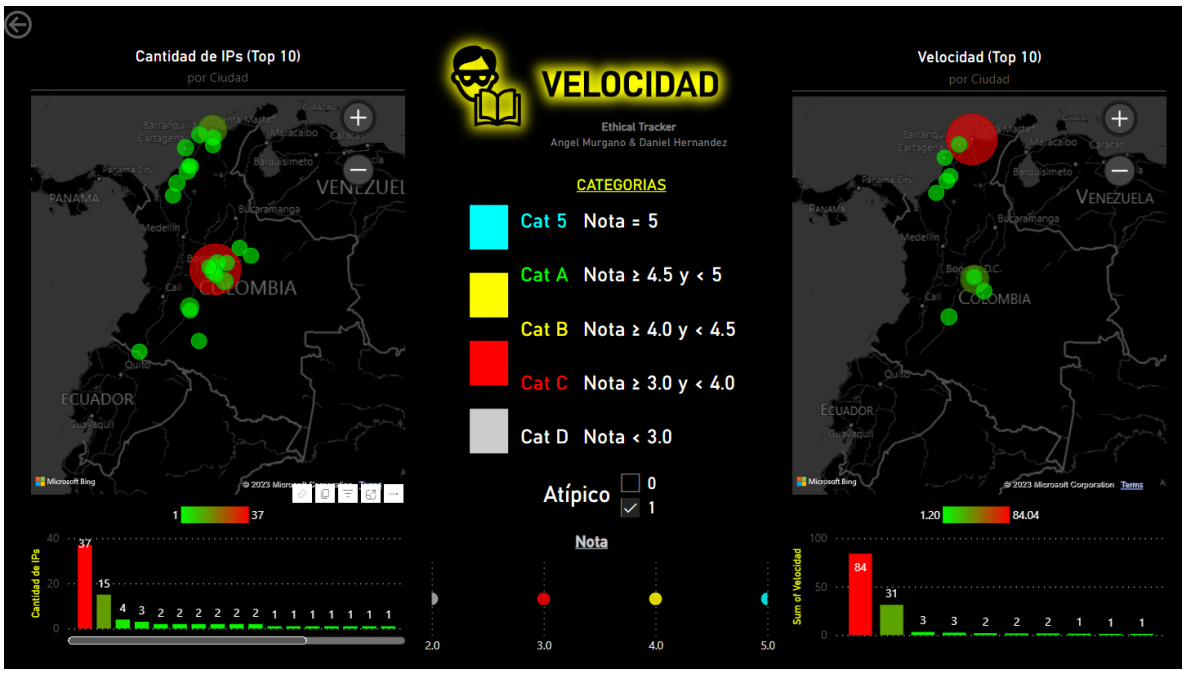
DASHBOARD:

Apoyados en la data disponible en la base de datos del servidor (la cual se crea luego de todo el proceso del backend) se alimenta un cuadro de control generado en POWER BI el cual permite al usuario interactuar con una serie de elementos gráficos que sintetizan la información y exponen de manera sencilla e intuitiva los KPI's que serán nuestro recurso para entender el comportamiento de los estudiantes frente a las evaluaciones.

A continuación, se presenta una descripción de las páginas contenidas en el dashboard (el manejo de este se presenta en la sección "MANUAL"):

- **Página de "VELOCIDAD":** En esta sección se presenta información de la geolocalización de los estudiantes al momento de una evaluación (mapa izquierdo) acentuando la zona que reúne a mayor cantidad de individuos y adicionalmente se presenta de manera geolocalizada el KPI llamado "Velocidad" el cual cuantifica el comportamiento de los estudiantes respecto a sus notas vs los tiempos de desarrollo de la evaluación y de esta manera se presenta en el mapa las zonas donde este indicador muestra datos atípicos (velocidades altas que implican nota alta en muy poco tiempo).

Figura 23 *Dashboard para análisis de Velocidad*

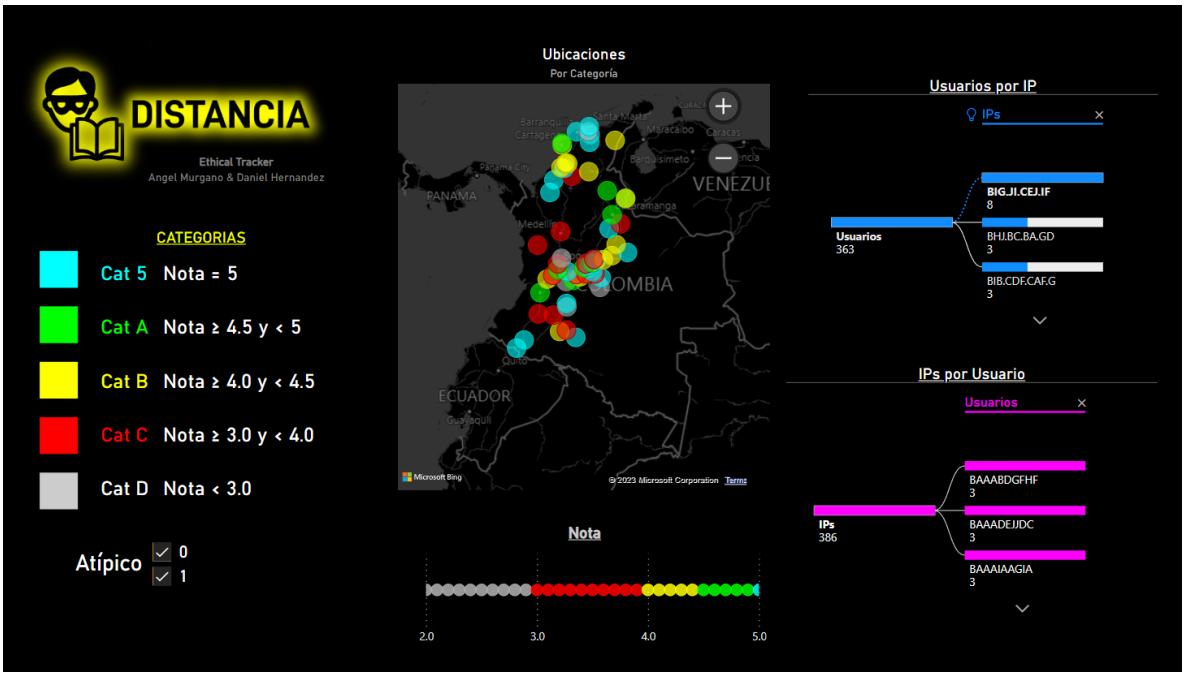


Fuente: Elaboración propia (2023)

- **Página de "DISTANCIA":** En esta página se expone la información de la geolocalización de los estudiantes según una serie de categorías relacionadas al rango de puntaje obtenido (mapa central y cuadro izquierdo) y adicionalmente se presenta en la parte superior izquierda la cantidad de usuarios por IP, este es un KPI que expone una lista con aquellas IP's donde al momento de la evaluación están entrando más de un usuario; finalmente en la parte inferior izquierda se presenta el KPI de IP's por usuario donde se expone una lista de el número de IP's que cada usuario usa en total al momento de interactuar con la plataforma universitaria (no solo en momentos de evaluación).

Cada una de las páginas indicadas del dashboard vienen con una funcionalidad para poder mostrar los KPI's de todos los estudiantes o segmentándolos entre Atípicos y No Atípicos, de manera de poder contrastar la información y enfocarse en aquellos casos que sugieran una alerta para su revisión más detallada.

Figura 24 *Dashboard para análisis de Distancia*



Fuente: Elaboración propia (2023)

MANUAL:

USO DEL PROGRAMA:

El programa es alimentado por 2 reportes de Excel generados por la plataforma MOODLE la cual es el entorno virtual donde los alumnos interactúan tanto para las clases como para las evaluaciones; los reportes generados vienen en un formato previamente

establecido por la institución y contienen información sobre las interacciones (archivo de Logs) y las notas de las evaluaciones (archivo de Notas).

Paso 1. Cargar Archivo de Notas: Se hace clic en el botón “Leer Archivo de Notas” y se navega en la ventana emergente para elegir el reporte a cargar.

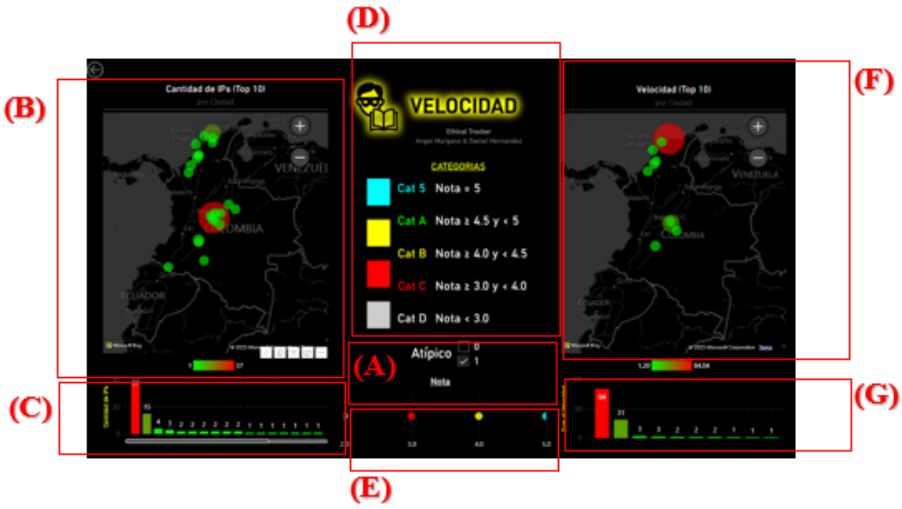
Paso 2. Cargar Archivo de Logs: Se hace clic en el botón “Leer Archivo de Logs” y se navega en la ventana emergente para elegir el reporte a cargar.

Paso 3. Tocar el botón Analizar archivos: Se hace clic en el botón “Analizar Archivos” y se espera que se genere el mensaje de proceso exitoso.

USO DEL DASHBOARD:

Página de Velocidad:

Figura 25 *Áreas de análisis en dashboard de Velocidad*



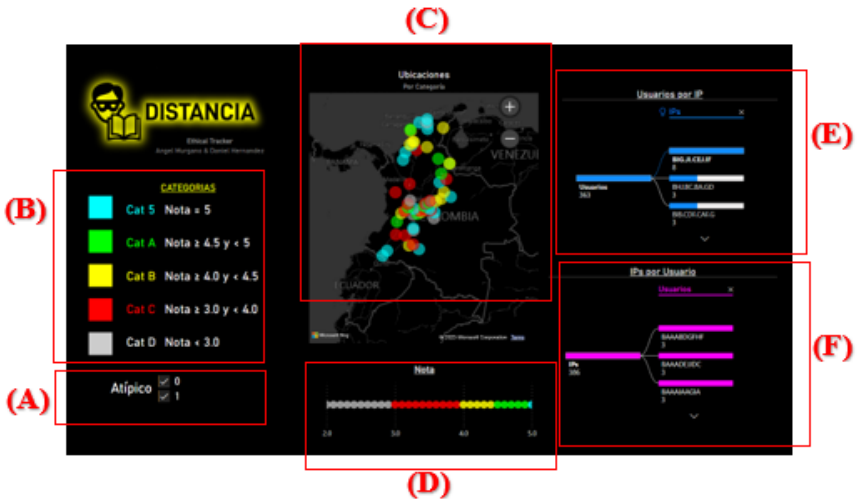
Fuente: Elaboración propia (2023)

Al seleccionar algún punto dentro de las siguientes opciones todas las gráficas se ajustarán para mostrar la data relacionada a dicho punto; para volver al estado inicial solo se deberá tocar nuevamente el punto previamente seleccionado.

A. Con este cuadro de control podemos activar y desactivar la visualización de toda la data en general, la data atípica y la data no atípica, de manera que todas las demás gráficas mostraran la información según esta configuración.
B. Aquí podemos geolocalizar las zonas donde se conectan menos o más estudiantes al momento de una evaluación.
C. Aquí podemos identificar el número general de conexiones por zona.
D. Se presenta un mapa de categorías según el rango de notas.
E. Se indica cómo se distribuye en general las notas.
F. Se presenta un mapa donde podemos ver cómo es la distribución de la variable “VELOCIDAD” según la zona y donde se encuentra aquellas zonas con mayor incidencia de velocidades atípicas.
G. Se indica cómo se distribuye en general la variable “VELOCIDAD”.

Página de Distancia:

Figura 26 *Áreas de análisis en dashboard de Distancia*



Fuente: Elaboración propia (2023)

Al seleccionar algún punto dentro de las siguientes opciones todas las gráficas se ajustarán para mostrar la data relacionada a dicho punto; para volver al estado inicial solo se deberá tocar nuevamente el punto previamente seleccionado.

A. Con este cuadro de control podemos activar y desactivar la visualización de toda la data en general, la data atípica y la data no atípica, de manera que todas las demás gráficas mostraran la información según esta configuración.
B. Se presenta un mapa de categorías según el rango de notas.
C. Se presenta un mapa donde podemos ver cómo es la distribución de las categorías de Notas según la zona.
D. Se indica cómo se distribuye en general las notas.

E. Se presenta las IP’s que poseen más de 1 usuario al momento de realizar una evaluación.

F. Se presenta la cantidad de IP’s manejadas por los estudiantes en todas sus interacciones.