MÁSTER EN INTELIGENCIA DE NEGOCIOS Y BIG DATA
2018/2019

Proyecto Fin de Máster

Análisis de Clientes para Casinos

AUTORES
Camilo Rafael Frias López
Darlin Hidekel Villa López
Jhairo Núñez García
Luis Miguel La Paz De La Cruz
Walewsky Terrero Turbi

TUTORA
Patricia Benito

I. ÍNDICE GENERAL

II. ÍNDICE DE ILUSTRACIONES

III. ÍNDICE DE TABLAS

IV. ÍNDICE DE ANEXOS

1. INTRODUCCIÓN

El siguiente trabajo de investigación es un razonamiento documentado, el cual pretende aportar para engrandecer, en la medida de lo posible, el conocimiento en pro de la inteligencia de negocios y el manejo de los datos.

Partiendo de la necesidad que poseen las empresas hoy en día de aprovechar las informaciones acumuladas como consecuencia de sus transacciones y de incrementar los beneficios económicos mediante la implementación de nuevas tecnologías, hemos decido realizar este proyecto con el objetivo de exhibir los beneficios de la implementación de un sistema Business Intelligence (BI) y Big Data (BD) dentro de una organización.

Corroborando con esta necesidad y realizando un diagnóstico de la empresa INMABUIN, fueron identificados puntos de mejora en sus procesos, manejo de datos e implementación de nuevas tecnologías, así como la oportunidad de poder contribuir con la generación de nuevos ingresos a través de nuevas técnicas de selección de clientes rentables. Esto ha contribuido a la creación de un caso de negocio robusto.

El desarrollo de este proyecto nos permitió poder aplicar los conocimientos que hemos adquirido en el máster de BI y BD, y sumando esta experiencia, contribuir al crecimiento profesional de sus integrantes.

2. PRESENTACIÓN DEL PROBLEMA

El mundo del entretenimiento está conformado por diversos sectores, uno de estos es el sector de juego y ocio, al cual pertenece el negocio de los casinos. Por mucho tiempo el ser humano ha buscado la forma de descansar, o mejor dicho de cambiar de actividad, hacia actividades más divertidas y placenteras. Dentro de estas están los juegos de azar. Los juegos de azar se remontan a tiempos inmemoriales, en la antigua China hace más de 4,000 años. Un juego de casino es considerado como tal cuando se hacen apuestas económicas respecto al resultado u opción diferente respecto a una actividad.

Hoy en día contamos con cientos de compañías que se dedican a proveer de espacios y actividades de entretenimiento. En este sentido, los casinos son de los más populares, más concurridos y con mayor movimiento económico. A pesar del tiempo que tienen existiendo y los avances tecnológicos que ha alcanzado la humanidad, la forma de operar de los casinos se mantiene casi intacta a como fue en sus orígenes.

La empresa INMABUIN es una compañía de software dominicana dedicada al diseño, desarrollo e implementación de programas y sistemas de gestión de información e inteligencia de negocios. Esta empresa cuenta con varios softwares instalados en varios casinos de República Dominicana y España. Y a pesar de que dichas empresas (los casinos clientes de INMABUIN) cuentan con una infraestructura de sistemas para llevar a cabo todas sus transacciones, tanto las de tipo administrativo y financiera, así como las de registro de apuestas, todavía quedan muchas deficiencias a nivel de datos en lo que respecta al registro y análisis de las apuestas.

Uno de los principales problemas radica en el hecho de que, en las mesas de juegos manuales, el registro de las apuestas resulta difícil y tedioso de llevar a cabo de forma detallada. Este registro detallado es importante debido a que a partir de este es que se podrán realizar los análisis de la rentabilidad del cliente y de los posibles fraudes en tiempo real. Este registro se lleva a cabo de forma manual, por parte de un jefe de Pit (un Pit es un grupo de mesas colocadas alrededor de un mismo centro) colocado en el

centro de cada Pit, de forma tal que registre las jugadas de los clientes de las mesas de dicho Pit. Este registro se lleva a cabo de forma arbitraria, registrando solo las jugadas que considera más relevantes de los clientes que considera más importantes.

Una de las razones por las cuales el registro de apuestas se lleva a cabo de la forma explicada en el párrafo anterior, es debido a que tener una persona por mesa registrando cada jugada haría que el espacio se viera muy cargado de empleados, lo cual no es muy estético, y en este tipo de negocios la estética es primordial. Además, para poder hacerlo de esa forma el empleado debería de conocer a cada cliente, y aunque hay muchos clientes que son conocidos, los no conocidos serían difícil de identificar luego que están en las mesas. A ningún cliente le gusta que le pregunten su nombre o número luego que está sentado jugando en una mesa, esto se puede hacer solo al entrar al casino.

El hecho de registrar únicamente una pequeña porción de las jugadas (alrededor del 20%) hace que los datos a analizar no sean precisos. Esto trae como repercusión que algunos clientes que no son rentables para el casino sean considerados como tal, debido a la falta de datos que ayude a comprobar dicha condición. En este sentido, muchos clientes no rentables reciben cortesías (las cortesías son regalos hechos a los clientes, ya sea: comidas, bebidas, hospedaje, transporte, entre otros) por parte del casino, sin ser merecedores de ellas, generando únicamente pérdidas al negocio. Al mismo tiempo, clientes nuevos que están dejando mucha ganancia al casino, no reciben las cortesías necesarias que le motiven a seguir jugando y a volver nuevamente al casino.

Por otro lado, están los casos de fraudes, los cuales son actualmente detectados por medio a la experiencia del personal de seguridad que vigila los clientes y las jugadas registradas y analizadas de forma manual. Para ello se toman las entradas realizadas por los jefes de Pit para ser analizados por un personal especializado en ello. De lo que dicho analista concluya, el personal de seguridad se mantiene atento a dichos clientes y empleados. Este proceso es muy lento y tedioso, ya que posiblemente cuando descubran un fraude es porque ya el personal, tanto cliente como empleado, ha engañado al casino por un buen tiempo.

Por último, está el hecho de tomar decisiones respecto a la distribución de las mesas, cuales mesas abrir en temporadas específicas, así como los días y horas que resulta más conveniente ofrecer determinados tipos de cortesías generales. Hasta este momento todas esas decisiones son tomadas por los gerentes de turno (personal que dirige las salas por turnos específicos) de forma arbitraria, sin ningún análisis de base. De igual forma es necesario analizar las temporadas, días y horas en que se vuelve necesario aplicar acciones comerciales con el objetivo de aumentar la visita de clientes.

A partir de estas problemáticas se entiende que se deben suplir dos necesidades que actualmente presentan estos casinos: el registro fidedigno de todas las transacciones de apuestas de todos los clientes, y el análisis de los datos recabados con el objetivo de evitar fraudes y premiar con cortesías a los clientes con el objetivo de aumentar la rentabilidad de estos, así como aplicar estrategias de marketing en el momento oportuno. La incorporación de soluciones a los problemas planteados implicaría que INMABUIN, como empresa que suple servicios de datos, mejore su posicionamiento en el mercado con soluciones innovadoras, incrementando al mismo tiempo, la posibilidad de captación de nuevos clientes (casinos).

3. DEFINICIÓN DE LAS SOLUCIONES

Es por todo lo anterior que se plantean las siguientes soluciones a los problemas propuestos:

- **Automatización del registro de las transacciones de las jugadas de los clientes.** Para ello se propone la creación de una infraestructura electrónica como estrategia de captación basada en

antenas RFID en las mesas, capaz de recolectar toda la información automatizada (ID de los clientes y apuestas en las mesas) sin necesidad de la intervención de ningún personal para su registro en las bases de datos, solucionando el inconveniente de que el cliente no quiere dar sus datos y reduciendo la dependencia de personal a la hora de los registros, evitando la captación de datos aproximados, siendo estos sustituidos por datos exactos tomados desde antenas que realizarán el registro automático de datos.

● **Modelo predictivo del comportamiento de los clientes.** Se creará un modelo predictivo en Python en base a los datos históricos existentes en la base de datos de MS-SQL para poder clasificar los clientes y poder predecir en tiempo real el comportamiento de estos. De esta forma el sistema podrá proponer cortesías específicas y acciones a hacer a favor de determinados clientes. Además, tendrá la opción de preguntar al sistema si considera que un cliente específico "merece" o le corresponde una cortesía dada.

● **Visualización del comportamiento de jugadas por temporada.** Se crearán gráficas orientadas al análisis de las jugadas por temporadas (trimestres, meses, semanas, días de la semana, horas) de forma tal que el sistema pueda proponer el momento preciso para que se hagan campañas de marketing, con el objetivo de incrementar los beneficios.

● **Sistema de alarma basado en los movimientos de las jugadas de los clientes.** Se creará un modelo que determine si el movimiento de un cliente es característico de posible fraude, de forma tal que envíe alertas a personas específicas para que se mantengan atentos y tomen acciones recomendadas.

● **Creación de dos cuadros de mando integral, uno a nivel gerencial y otro a nivel presidencial.** Se creará un cuadro de mando que le dé los pormenores de lo que está ocurriendo en el casino a los gerentes de turno, de forma tal que puedan tomar decisiones en el momento basados en datos reales. Además, se creará un cuadro de mando para la presidencia de la compañía, donde podrán ver lo que ocurre en el momento, así como lo ocurrido en sesiones anteriores, y comparativas entre sesiones y temporadas.

<table>
<tr><td>Institución</td><td>Tecnología</td><td rowspan="4">Imposibilidad de gestionar apropiadamente la información de las apuestas en los juegos de mesas, de forma tal que sea útil para tomar mejores decisiones durante el tiempo de juego de los clientes.</td></tr>
<tr><td>♣ Desconocimiento de las soluciones electrónicas.
♣ Desconocimiento de las técnicas de BI y BD.</td><td>♣ Falta de estructura electrónica.
♣ Falta de herramienta de visualización.</td></tr>
<tr><td>♣ Aman sus juegos tradicionales.
♣ Movimientos azarosos respecto a las jugadas.</td><td>♣ Se registran pocas transacciones de las jugadas.
♣ Los datos no son suficientes para el análisis eficiente.</td></tr>
<tr><td>Usuarios</td><td>Información</td></tr>
</table>

Tabla 1 - DIAGRAMA ISHIKAWA: Problema-Causa (A nivel de Casino)

4. RESULTADOS ESPERADOS

Tomando en cuenta las problemáticas presentadas por los clientes de INMABUIN, y las soluciones propuestas, se esperan obtener los siguientes resultados, luego de la implementación de dichas soluciones:

- Al implementar un sistema IoT que registre de forma automática los movimientos de las jugadas de los clientes, el casino dispondrá de los datos suficientes para hacer los análisis necesarios para mejorar la productividad de las salas de mesas.

- El uso del modelo predictivo del comportamiento de los clientes provocará que estos incrementen su rentabilidad para el casino.

- El uso de gráficas que analicen el comportamiento de jugadas por temporada provocará un incremento en las ganancias comparativamente a las temporadas previas.

- El uso del sistema de alarma basado en los movimientos de las jugadas de los clientes disminuirá las jugadas fraudulentas.

- El uso del cuadro de mando gerencial hará más eficiente la toma de decisiones de los gerentes de turno, de forma tal que el resultado genere menos pérdidas y propicie mayores ganancias.

- El uso del cuadro de mando presidencial hará más eficiente la toma de decisiones de los gerentes de turno, de forma tal que el resultado genere menos pérdidas y propicie mayores ganancias.

5. INVESTIGACIÓN, TOMA DE DATOS Y VALIDACIÓN

Es importante a la hora confeccionar, o bien desarrollar un modelo de negocios, que el mismo esté basado en función de información documentada que permita tomar las decisiones más adecuadas a priori, y así aumentar la objetividad durante la implementación del proyecto, a la par que disminuya los riesgos propios de la construcción del negocio en base a percepciones y/o concepciones subjetivas.

En tal sentido, el equipo de investigación se dispone a profundizar sobre los pasos tomados para corroborar, o en su defecto demostrar, que las asunciones previas no se corresponden con la realidad del mercado objeto de estudio. Y es por tal razón que en este marco que se desarrolla a continuación se siembran las bases para la fundamentación de la investigación.

Debemos recordar que en la definición del problema a abordar se planearon unas hipótesis o conjeturas, las cuales se deben analizar a fines de confirmar la veracidad del planteamiento. Recordemos que el alcance, o bien el objeto de investigación, es determinar para los clientes de INMABUIN, cómo medir la rentabilidad de los visitantes a establecimientos de ocio y juegos de azar. Esto para acrecentar la rentabilidad en general de dichos casinos y que los mismo actúen de forma más eficiente, de cara a la gestión por cliente sobre qué cortesías merecen dichos clientes, sin la necesidad de que la decisión sea tomada por una persona en base a su experiencia particular.

En lo que respecta a los registros de datos a utilizar, los mismos se han obtenido vía una de las entidades de ocio y recreación que funge como cliente de la empresa INMABUIN. Donde vale la pena destacar que se han cuidado los aspectos previstos en las Ley de Seguridad de la Información, al artículo 5 (Deber de confidencialidad) de la ley Orgánica 3/2018 de España, a través de la cual se debe respetar y proteger los datos de las personas con miras a garantizar los derechos digitales de dignidad, intimidad y privacidad

de los individuos; y la Ley 172-13 de la Constitución Dominicana, donde se establece que el tratamiento de los datos de personales y otros no es permitido sin el consentimiento del titular.

Delimitamos el alcance de nuestros esfuerzos a las mesas de apuestas, en las cuales se generan la mayor proporción de las ganancias de los casinos, en contraposición de las máquinas electrónicas. Por lo que apoyar el proceso de determinar los perfiles y las rentabilidades de los tipos de clientes que se encuentran apostando en los centros lúdicos permitirá tomar decisiones que eficienticen la gestión de las cortesías para maximizar la función de beneficio por cliente, lo que se traduciría como beneficios mayores para los centros de ocio y apuestas.

En lo que respecta a esta indagación según (Hernández Sampieri, Fernández Collado, & Baptista Lucio, 2014) "hay dos enfoques primordiales de la exploración científica: el primero sería el enfoque cuantitativo y el enfoque cualitativo".

En tal sentido, para los fines de la investigación que nos compete se eligió el enfoque cuantitativo, ya que, utilizaremos tecnologías de la información propias de BI y BD para obtener estadísticas sobre el desempeño de la rentabilidad por cliente. Todo esto con el fin de determinar modelos de comportamiento, tipos de cliente, más información sobre el desempeño de las mesas de juego, etc.

La técnica utilizada para los fines de este trabajo de investigación es el método analítico. El cual radica en la separación de las partes que componen un todo, con la intención de observar las causas y los efectos. Este estudio no es más que la observación y análisis de un suceso en particular. Dicha técnica se utilizó para dar tratamiento sistemático a la data al procesar, analizar y registrar los resultados obtenidos.

Las fuentes que se utilizaron fueron primarias y secundarias, las primeras basadas en fuentes bibliográficas y documentales, tales como, libros, páginas web, bases de datos, entre otras. Por otro lado, el estudio de campo se realizó a través de entrevistas con la finalidad de corroborar la implementación del prototipo y sus resultados esperados, de cara al tópico que hemos estudiado durante el proceso de la investigación.

Las técnicas de investigación usadas han sido:

- Consulta a la documentación referente al tema.
- Reunión y consulta con expertos en la materia.
- Análisis de datos históricos por bases de datos de casinos españoles.
- Realización de análisis estadísticos

Por lo que a continuación, se listan las hipótesis plantadas y los cuestionamientos en base a los cuales analizamos la realidad tanteada en esta investigación, la metodología con criterio científico utilizada para determinar la veracidad de estas y la exploración de la base de datos obtenida por parte de uno de nuestros clientes para la validación y desarrollo de los procesos de análisis.

#	Hipótesis	#	Pregunta
1	Clientes: Casinos no poseen la capacidad de segmentar sus clientes.	1.1	¿Se conoce la cantidad de clientes que visitan los casinos?
		1.2	¿Es posible segmentar el tipo de clientes en base a su comportamiento dentro de los casinos?
		1.3	Utilizando una escala del 1 al 5, donde 1 representa poco de acuerdo y 5 muy de acuerdo, ¿Es efectivo el proceso de identificación de clientes?

		1.4	¿Cómo es el método de identificación de clientes?
		1.5	¿Cómo se anotan las jugadas, ganancias y/o pérdidas de los clientes de los casinos?
		1.6	¿Cuáles datos crees que serían necesarios para poder segmentar los clientes?
2	Apuestas: No se conocen los patrones de apuestas de los clientes que asisten a los casinos.	2.1	¿Es posible determinar el patrón de apuestas por cliente con la metodología actual?
		2.2	¿Cómo se realiza actualmente?
		2.3	Utilizando una escala del 1 al 5, donde 1 representa poco de acuerdo y 5 muy de acuerdo, ¿Es satisfactorio el método actual de identificación de patrones?
		2.4	¿Qué se puede mejorar para determinar mejor los patrones de juego de los clientes?
3	Campañas: Los casinos necesitan realizar campañas para atraer más clientes	3.1	¿Se realizan campañas para atraer clientes de forma periódica?
		3.2	¿Qué criterios se usan determinar cuándo hacer campañas?
		3.3	Utilizando una escala del 1 al 5, donde 1 representa poco de acuerdo y 5 muy de acuerdo, ¿Son efectivas las campañas realizadas por los casinos?
4	Customer journey: No se puede conocer la experiencia del cliente dentro del casino.	4.1	¿Es posible determinar el nivel de experiencia del cliente?
		4.2	¿Crees que sería útil para la mejora continua , tanto de empleados como del negocio, implementar alguna técnica de encuestas al cliente para saber su experiencia en el casino
5	Datos: En las mesas del casino, los datos se registran de forma manual.	5.1	¿Es cierto que los datos se registran de forma manual? En caso de ser Sí ¿Cuáles datos se registran?
		5.2	Utilizando una escala del 1 al 5, donde 1 representa poco de acuerdo y 5 muy de acuerdo, ¿Es fácil registrar los datos de los clientes del casino?
		5.3	¿Conocen alguna alternativa que facilite este proceso?
6	Rentabilidad: No se conoce la rentabilidad por segmento de cliente que juega en el casino.	6.1	¿Se conoce el ROI por mesa? En caso de ser afirmativo ¿Cómo se calcula?
		6.2	¿Se conoce el ROI por cliente? En caso de ser afirmativo ¿Cómo se calcula?
		6.3	¿Cuánto tiempo de análisis conlleva analizar la rentabilidad por cliente y por mesa?
7	Seguridad: Existe la necesidad de un sistema de detección de operaciones fraudulentas en el casino.	7.1	¿Poseen algún sistema de visualización de fraudes, en base al comportamiento sospechoso de juego de los clientes?
		7.2	En caso de ser cierto ¿Cómo funciona dicho sistema?
		7.3	¿El sistema de seguridad posee dashboard con la capacidad de generar alertas ante comportamientos sospechosos?

		7.4	Utilizando una escala del 1 al 5, donde 1 representa poco de acuerdo y 5 muy de acuerdo, ¿Considera fácil de usar dicha herramienta de seguridad?
		7.5	En caso de ser falso ¿Cuáles datos crees que se pueden tomar en cuenta para pronosticar algún fraude?
8	Al implementar un sistema IoT que registre de forma automática los movimientos de las jugadas de los clientes, el casino dispondrá de los datos suficientes para hacer los análisis necesarios para mejorar la productividad de las salas de mesas	8.1	¿Cómo se mide actualmente el comportamiento de jugadas en las mesas?
		8.2	Utilizando una escala del 1 al 5, donde 1 representa poco de acuerdo y 5 muy de acuerdo, ¿Considera la metodología actual como eficiente?
		8.3	¿Existe alguna tecnología implementada por otro casino capaz de medir de forma automática las jugadas de los clientes?
		8.4	Utilizando una escala del 1 al 5, donde 1 representa poco de acuerdo y 5 muy de acuerdo. ¿Estaría de acuerdo con implementar una tecnología automatizada para facilitar el proceso de captura de datos?
9	El uso del modelo predictivo del comportamiento de los clientes provocará que estos incrementen su rentabilidad para el casino	9.1	¿Han escuchado acerca de los modelos predictivos?
		9.2	Utilizando una escala del 1 al 5, donde 1 representa poco de acuerdo y 5 muy de acuerdo ¿Considera que se puede mejorar la rentabilidad por medio a modelos predictivos?
10	El uso del sistema de alarma basado en los movimientos de las jugadas de los clientes predecirá las jugadas fraudulentas en el casino de forma más certera.	10.1	¿El casino posee actualmente un cuadro de mando con información sobre jugadas sospechosas?
		10.2	¿Cómo se mide actualmente?
		10.3	Utilizando una escala del 1 al 5, donde 1 representa poco de acuerdo y 5 muy de acuerdo, ¿Considera la metodología actual como eficiente?
11	El uso del cuadro de mando gerencial hará más eficiente la toma de decisiones de los gerentes de turno, de forma tal que el resultado genere menos pérdidas y propicie mayores ganancias	11.1	¿El casino tiene actualmente un cuadro de mando con información la capacidad de medir la rentabilidad por cliente?
		11.2	¿Cómo miden actualmente?
		11.3	¿Obtienen la información en tiempo real?
		11.4	Utilizando una escala del 1 al 5, donde 1 representa poco de acuerdo y 5 muy de acuerdo, ¿La metodología actual permite tomar decisiones a tiempo?
12	Clientes: no desean compartir su información luego que se encuentran en la mesa de juego.	12.1	¿Qué proporción de los clientes se registran en el casino?
		12.2	¿Posee el casino un sistema de registro de los nuevos clientes?
		12.3	¿Cómo funciona el sistema de registro de los clientes?
		12.4	¿Qué entiende usted se podría mejorar para incrementar la cuota de clientes que se inscriben en el sistema de fidelidad del casino?

		12.5	Para los clientes: ¿Tendría usted como cliente inconveniente en registrarse en el casino, y a cambio, recibir ofertas, cortesías y participar en concursos?
13	Las Cortesías no se asignan correctamente a los clientes rentables dentro del casino.	13.1	¿Qué tanto ayuda la asignación de cortesías a los clientes para aumentar la rentabilidad del negocio?
		13.2	¿Cuáles puntos se toman en cuenta para conceder cortesías a cada cliente?
		13.3	¿Se lleva un control o registro de estas cortesías?
		13.4	De ser así ¿Cuáles datos se almacena sobre estas?

Tabla 2 - Cuestionario de entrevistas para validación de Hipótesis

6. EXPERIENCIAS OBTENIDAS, VALIDACIÓN DE LOS HECHOS

Luego de culminar el proceso de entrevistas y escrutinio de la información colectada al visitar un centro de entretenimiento lúdico o casino en Santo Domingo, República Dominicana, el equipo investigador ha llegado a las siguientes revelaciones:

Los casinos no cuentan con una metodología, o bien una herramienta tecnológica que les ayude a saber con certeza la cantidad de clientes que visitan dichos establecimientos, dado que los clientes pueden entrar una y otra vez al casino, y por ello el conteo no es efectivo. Algunos implementan conteo manual, o en su defecto, conteo de entradas con contadores de puertas giratorias unidireccionales.

Otro hecho a considerar es que, pese a que los clientes de los casinos deben suministrar su identificación oficial (cédula, DNI, etc.) para acceder al establecimiento, o comprar fichas con fines de juego, solo una minoría de dichos clientes accede a ser vinculado de forma directa con el casino, esto responde, no a temor de estar registrado en este, sino más bien, al temor de que alguien externo pueda relacionarlo con este mundo del entretenimiento.

Este hecho dificulta algunas tareas esenciales para una buena gestión del establecimiento, como lo es el hecho de poder registrar los patrones de juego y el comportamiento de los clientes durante su estadía en el casino de una forma precisa y completamente confiable; este proceso se realiza por medio de la observación, apoyado en la experiencia o conocimiento de los clientes que posea la persona responsable de llevar a cabo dichos registros.

Continuando en el mismo tenor, cabe destacar que la información colectada mediante este proceso de observación no registra todos los datos requeridos, ya que, la persona responsable, únicamente se limita a registrar de forma manual, mediante anotación en papel, las compras y ganancias de los clientes más relevantes, es decir, aquellos con un hábito de juego frecuente, o aquellos que se encuentren jugando cifras significativas.

Esta práctica de registro manual es una actividad que da lugar al error humano, no solo por el hecho de que puede incluso omitir jugadas importantes, sino también porque al escribir los datos tiene la posibilidad de anotar valores erróneos, a esto le sumamos el bajo nivel de eficiencia en el proceso.

Otra de las tareas esenciales que no es posible realizar con un alto grado de eficacia, o pudiéramos decir, de forma más precisa y correcta, es la asignación u ofrecimiento de cortesías a los clientes. Las cortesías, como su nombre indica, son gestos que realiza el casino a sus clientes al brindarles bebidas, comidas y hasta alojamiento, con el propósito de que cada cliente se sienta con la mejor experiencia posible dentro de las áreas de juego.

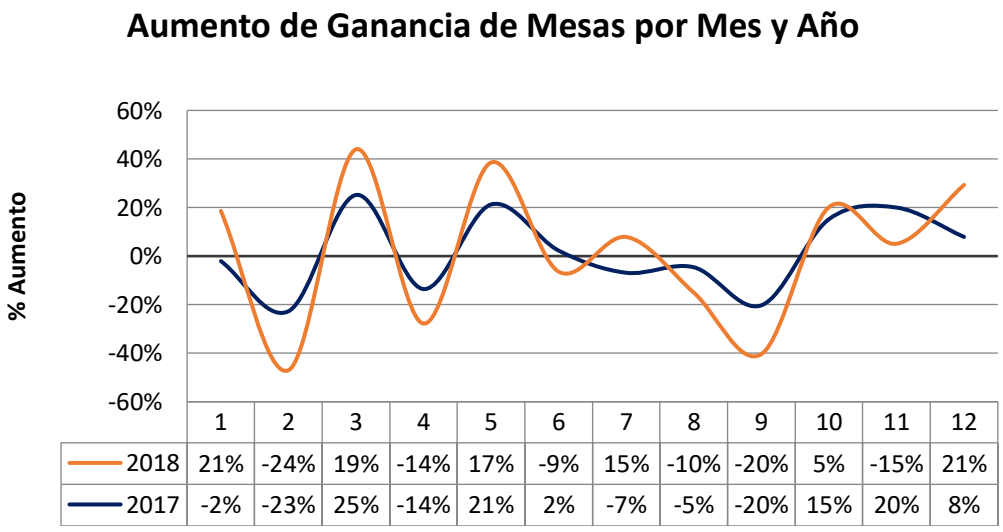
El problema en torno a este último punto radica en que, debido a que no es posible generar o mantener un registro del comportamiento de los clientes, y el hecho de que los casinos no sean capaces de identificar de forma correcta cada cliente, no se puede conocer de forma certera el nivel de rentabilidad de estos, lo cual lleva a los establecimientos a la asignación no correcta de dichos gestos brindados por el casino.

Otra de las debilidades presentes actualmente alrededor del proceso de registro en las mesas de juego es que, dado que es un proceso manual, los ejecutivos no tienen acceso a la información en el tiempo propicio. Para obtener un análisis de los datos, y a su vez, tomar decisiones en base a la información generada a partir de estos datos, se requiere un tiempo elevado.

Un punto para resaltar es la necesidad de contar con una herramienta que ayude a los casinos a evaluar de forma constante patrones conocidos de fraude; actualmente, todo el análisis de actividades sospechosas de fraude es realizada por personal entrenado en el comportamiento humano, y mediante la observación de la conducta de las personas. Estos profesionales son capaces de detectar actividades sospechosas, pero con la desventaja de que sus conclusiones toman demasiado tiempo.

Sin embargo, en la actualidad no es posible dar seguimiento, mucho menos determinar de forma automática, patrones de juego que resulten en actividades sospechosas que pueden ser tratadas de forma proactiva.

Por último, se pudo constatar que existen en el año temporadas en las cuales es recurrente la disminución y aumento de las ganancias, por lo que se hace necesario que se esté alerta a estos cambios, de forma tal que el casino pueda aplicar estrategias encaminadas al aumento de estas.



	1	2	3	4	5	6	7	8	9	10	11	12
2018	21%	-24%	19%	-14%	17%	-9%	15%	-10%	-20%	5%	-15%	21%
2017	-2%	-23%	25%	-14%	21%	2%	-7%	-5%	-20%	15%	20%	8%

Ilustración 1- Evolución de las ganancias de uno de los casinos evaluados

7. ANÁLISIS Y DIAGNÓSTICO/PLAN ESTRATÉGICO-ACCIÓN

7.1 Análisis y Diagnóstico

Llegados a este punto, es necesario ver y examinar la información que tenemos, de modo que podamos analizar los aspectos de nuestra propuesta y cómo la misma puede ser un negocio viable.

Análisis DAFO

El análisis de debilidades, amenazas, fortalezas y oportunidades nos ayudará a conocer la situación del proyecto, analizando sus características internas (debilidades y fortalezas) y la situación externa (amenazas y oportunidades) de modo que tengamos un mejor entendimiento de este.

Ilustración 2 - Análisis DAFO

Debilidades

El equipo cuenta con experiencia limitada en el desarrollo de soluciones IoT, lo cual supone un eslabón débil en toda la cadena que compone el proyecto, pese a que el equipo tiene experiencia técnica, las soluciones IoT no son su fuerte.

Por otro lado, es la primera vez que el equipo trabaja en una solución para el sector casino la cual integra un conjunto de elementos que son nuevo para dicho sector.

La última debilidad identificada se encuentra en la resistencia de los clientes en proveer sus datos personales o la recolección de datos de sus transacciones, apoyado en la Ley de Seguridad de la Información, al artículo 5 (Deber de confidencialidad) de la Ley Orgánica 3/2018 de España, a través de la cual se debe respetar y proteger los datos de las personas con miras a garantizar los derechos digitales de dignidad, intimidad y privacidad de los individuos; y la Ley 172-13 de la Constitución Dominicana,

donde se establece que el tratamiento de los datos de personales y otros no es permitido sin el consentimiento del titular. Partiendo de esto, la cantidad de datos recolectados depende considerablemente del grado de aceptación del cliente a estas nuevas metodologías

Amenazas

Debido a que el sector al cual dirigimos nuestro proyecto es el sector casino, tenemos retos importantes a considerar, los cuales se convierten en amenazas, no solo en la implantación del proyecto, sino en el uso cotidiano del mismo.

Lo primero que vemos como una posible amenaza es que los altos ejecutivos de los casinos tengan alguna resistencia al cambio y que por tanto no sean capaces de adoptar el proyecto. Pese a que la solución es una necesidad latente, el sector es muy cuidadoso en las medidas o cambios que implementa, ya que todo el negocio se trata de que los clientes se sientan bien y apuesten.

Esto nos lleva a la segunda y principal amenaza, puede ser que a los clientes no les guste el hecho de que se capturen los datos de su comportamiento mientras juegan, lo cual puede poner en riesgo la fidelidad de los clientes con el casino.

Dentro de otras posibles amenazas está el hecho de que la inversión para la transformación de la empresa (casino) pueda resultarles elevada al momento de la ejecución inicial del proyecto.

Pese a que no encontramos una empresa que brinde los servicios de la solución que estamos proponiendo, consideramos como amenaza la posible aparición de un competidor.

Fortalezas

Nuestra principal fortaleza es que contamos con una solución innovadora. Hasta donde pudimos verificar, no existe competencia en el mercado que brinde los servicios de la solución propuesta.

Además, contamos con la alta reputación en el mercado de INMABUIN, quien tiene como clientes una amplia cartera de casinos, de entre los cuales podemos obtener nuestros “early adopters”. A la vez contamos con un equipo multidisciplinar capaz de llevar todo el proyecto desde su planteamiento hasta su implementación.

Pese a que la implementación total del proyecto es elevada, contamos con un producto mínimo viable de bajo costo y una estructura de costos reducida.

Oportunidades

Con nuestra propuesta los casinos tendrán oportunidades valiosas que les posicionará por encima de aquellos que no la posean. Podrán, de una manera más eficiente, segmentar los clientes de acuerdo con su comportamiento y rentabilidad, permitiendo con esto una gestión de clientes más precisa

Con esto será posible establecerse por encima de la competencia en dos aspectos importantes: la gestión de cortesías y el análisis para la detección de posibles acciones fraudulentas.

7.2 Definición Modelo de Negocio

A continuación, queremos analizar la estructura o componentes de la solución propuesta, resaltando en ello a quien está dirigida y cuáles elementos intervienen en la misma. Gracias al canvas del modelo de negocio podemos verlo en un solo vistazo en la siguiente gráfica.

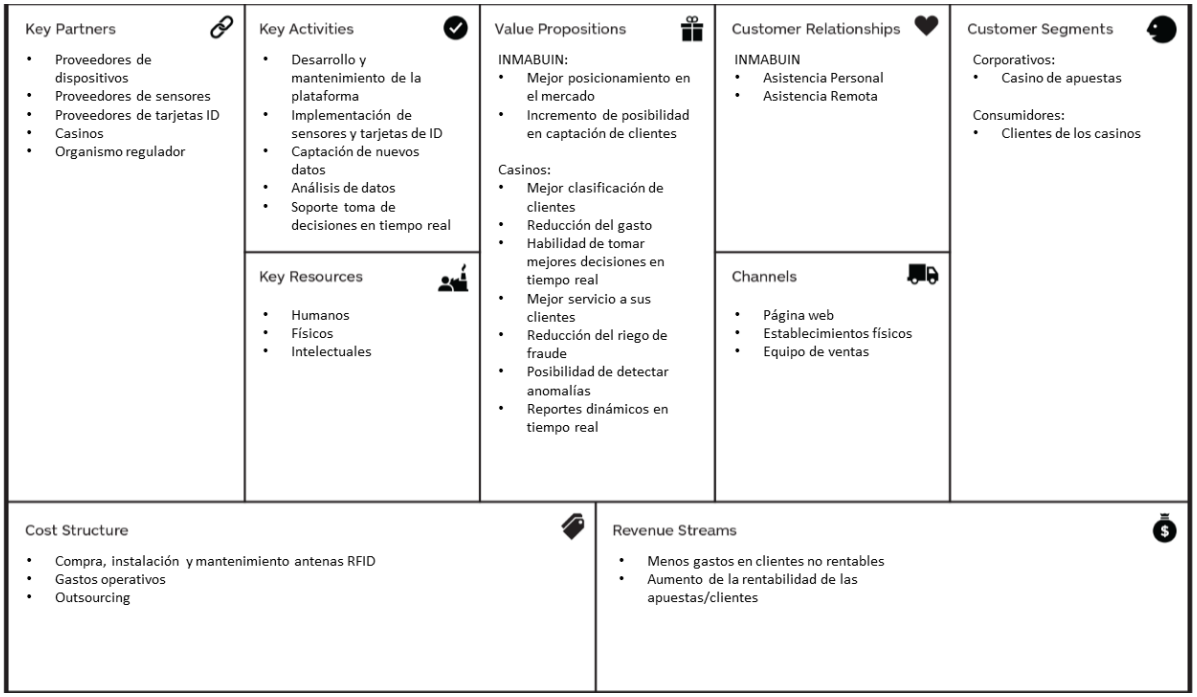


Ilustración 3 - Modelo de Negocio

Propuesta de Valor

Nuestra propuesta de valor es la solución a necesidades latentes en el sector de casino de apuestas, con el fin de lograr mejora en el servicio, gestión de riesgos, gestión de gastos y acceso a mejor información.

Dicha propuesta podemos verla en dos vertientes, aquella que afecta o añade valor a la empresa INMABUIN (empresa intermediaria a quien se le está haciendo la propuesta) y aquella que añade valor al objeto de la solución, los casinos clientes de INMABUIN. Por tanto, el valor añadido podemos verlo de la siguiente manera:

INMABUIN

- Mejor posicionamiento en el mercado.
- Incremento de posibilidad en captación de clientes.

Casinos

- Mejor clasificación de clientes.
- Reducción de gastos.
- Habilidad de tomar mejores decisiones en tiempo real.

- Mejor servicio a sus clientes.
- Reducción del riesgo de fraude.
- Posibilidad de detectar anomalías.
- Reportes dinámicos en tiempo real.

Actividades Claves

Existe un conjunto de actividades clave para el desarrollo y éxito de la solución planteada. Podemos ver las mismas a continuación:

- **Compra e implementación de sensores y tarjetas para el registro de clientes y sus movimientos.**

Nuestra solución plantea un seguimiento más minucioso a las jugadas que realizan los clientes, de ahí la necesidad de implementar dispositivos IoT que nos ayuden a mantener este seguimiento.

- **Captación de nuevos datos en la plataforma existente.**

Una vez los dispositivos IoT estén en función, estos se encargarán de reunir la información necesaria para luego transmitirla como último destino a la aplicación existente, pudiendo la información ser utilizada para los múltiples fines que se presentan en la propuesta.

- **Recogida y análisis de datos.**

Una vez los datos se encuentran en las bases de datos actuales, se procede al tratamiento de estos con los fines de clasificar los clientes, y dada estas clasificaciones darle el tratamiento adecuado.

- **Soporte a toma de decisiones en tiempo real.**

Los ejecutivos de los casinos podrán tener acceso en tiempo real a un conjunto de informes que en la actualidad no les es posible consultar con rapidez, debido al proceso actual de recogida y captura de datos.

- **Desarrollo, mantenimiento y actualización de las plataformas.**

Es necesario tener en cuenta que no solo son dispositivos, sino que nuevas aplicaciones intermedias se estarán utilizando para llevar la solución a la realidad, por lo que dentro de las actividades necesarias tenemos las del desarrollo y mantenimiento de estas.

Con el desarrollo de estas actividades buscamos implementar y mantener la solución de forma que brinde y añada valor al proceso y funcionamiento actual de los casinos en sus mesas de juegos manuales.

Segmento Clientes

La propuesta presentada va dirigida a la empresa INMABUIN, siendo esta solución una propuesta de la mencionada empresa para sus clientes, casinos de apuestas.

Dicho esto, podemos decir que, por transición, el segmento de clientes a los cuales la propuesta está dirigida son los casinos de apuestas, aun existiendo la empresa INMABUIN como intermediaria y como único cliente a quien va dirigida la propuesta.

Relación con los Clientes

Dentro de las relaciones de clientes, para el caso de INMABUIN, la propuesta presenta un esquema de atención personalizada e individual para cada cliente. Adicionalmente, INMABUIN cuenta con una estructura robusta que le permite brindar asistencia a sus clientes de forma remota.

El objetivo principal en cuanto a la relación con los clientes es la de alcanzar la fidelización de estos, permitiéndoles saber en todo tiempo que pueden contar con INMABUIN en el momento que lo necesiten.

Socios Claves

Como se sabe bien, ningún proyecto puede caminar solo, por tal razón es necesario que contemos con un conjunto de asociados importantes para el desarrollo de este, y a la vez para su continuidad.

Dentro de estos asociados claves queremos mencionar los siguientes:

- Proveedores de dispositivos
- Proveedor de lectores RFID
- Proveedores de tarjetas RFID
- Proveedores de fichas con etiquetas RFID
- Organismo regulador
- Casinos

Cada uno de estos juega un papel importante en el proceso de la solución, por lo que podemos decir que sin ellos no es posible llevar a cabo el proyecto.

Canales de Distribución

Pese a que la propuesta es para la empresa INMABUIN, estaríamos con ellos en el proceso de implementar, no solo a sus clientes actuales, sino también durante el proceso de captura de nuevos clientes interesados en la nueva solución que INMABUIN estará brindando.

Por tal razón, los canales de distribución de la solución serán llamadas por medio de centros de contacto, anuncios publicitarios a través de la página web de INMABUIN y equipos de venta que estarán visitando casinos para ofertarles la propuesta.

Recursos Claves

En esta solución el principal recurso clave son los datos, ya que es la materia prima del proyecto. Sin embargo, tenemos una serie de recursos que debemos puntualizar, ya que sin ellos no es posible la ejecución de lo planteado. Podemos separarlos de la siguiente manera.

- **Recursos Humanos**

El factor humano siempre viene a ser el recurso más importante, y por eso lo mencionamos de primero. Entre esos recursos humanos necesarios tenemos analistas de datos, ejecutivos de negocios, personal de TI (estos han de tener conocimientos en soluciones de IoT), así como personal de ventas.

- **Recursos Tecnológicos**

Estos recursos son la base del proyecto y son los que nos permitirán obtener nuestra materia prima, los datos. Entre los recursos tenemos: antenas RFID (estas deben ser construidas para estos fines), tarjetas RFID, fichas con RFID integrados, redes de comunicaciones, microordenadores para control de lectores de las tarjetas) y pantallas de visualización para los dashboards.

Estructura de Costes

Dada la naturaleza de la propuesta, nuestra estructura de costes la presentaremos en dos fases, la primera fase nos dará el coste inicial de implementación en una mesa de juego, mientras que la segunda nos dará el coste de expansión de la solución a otras mesas de juego.

Fase 1 - Desarrollo Inicial e implementación en una mesa de 8 jugadores

Concepto	Importe	Observaciones
Personal	27.000,00 €	Incluye personal de instalación y configuración, analistas de datos y técnicos de comunicación.
Dispositivos IoT de Lectura	4.410,00 €	
Lector RFID Área de Caja	450,00 €	
Software	9.000,00 €	Esto incluye el desarrollo de las tres piezas de software que serían necesarias.
Microordenador Controlador IoT	99,00 €	
Fichas con RFID (lote de 100)	170,00 €	
Tarjetas RFID (lote de 10)	12,00 €	
Total	41.141,00 €	

Tabla 3 - Costes de implementación primera fase

Fase 2 - Instalación de mesas de juego adicionales

Concepto	Importe	Observaciones
Personal	3.000,00 €	Incluye personal de instalación y configuración, analistas de datos y técnicos de comunicación.
Dispositivos IoT de Lectura	4.410,00 €	
Lector RFID Área de Caja	450,00 €	

Microordenador Controlador IoT	99,00 €	
Total	7.959,00 €	

Tabla 4 - Costes de implementación segunda fase

En cuanto a las fichas y las tarjetas, hemos incluido el costo de dos lotes solamente durante la primera fase, la cantidad de lotes ha de ser determinada por el casino a través de la empresa INMABUIN.

Fuentes de Ingresos

Debido a la naturaleza de la propuesta, la cual está dirigida a una empresa específica, pero que a su vez es para un sector cliente de dicha empresa, nos gustaría presentar las fuentes de ingresos desde dos perspectivas distintas.

Desde la perspectiva de INMABUIN los ingresos se generarán a partir de las implementaciones de la solución, de la captación de nuevos clientes debido a la solución y por el cobro de servicios de mantenimiento de esta.

Pero en cuanto a la segunda perspectiva, esta es, para los casinos, la fuente de ingresos será la reducción y optimización del gasto en cortesías y atención al cliente, lo que se traduce en una mayor rentabilidad para el mismo. Además, Se reducen gastos por fraude, y aumentan los ingresos al aumentar las apuestas.

7.3 Plan de acción

7.3.1 Definición del alcance del proyecto: objetivos y métricas

Este proyecto persigue una serie de objetivos concretos y precisos, los cuales definen el alcance de la propuesta.

Objetivo General

Ser capaz de recolectar información más precisa sobre las jugadas de los clientes en las distintas mesas de juego manuales de forma que se pueda utilizar esta información para el análisis de perfiles de cliente en la sectorización de estos.

Objetivos Específicos

- Registrar todos los movimientos de fichas de los clientes.
- Segmentar los clientes de acuerdo con su comportamiento.
- Identificar el nivel de cortesía que merece un cliente tomando como base su rentabilidad para el casino.
- Reconocer en tiempo real cuando el movimiento de un cliente presenta la posibilidad de ser fraudulento.
- Proveer de información real y actualizada respecto a las jugadas de los clientes de forma tal que los gerentes de turno puedan tomar mejores decisiones.
- Proveer de información comparativa respecto a las jugadas en general a la presidencia del casino, de forma tal que les sea más fácil implementar cambios estructurales.

Métricas

- El porcentaje de movimientos por registrar no supera el 5%
- Mientras mayor es el porcentaje de rentabilidad del cliente, mayor es el costo de la cortesía ofrecida al cliente.
- El porcentaje de transacciones fraudulentas disminuye cada mes.
- Las quejas de los clientes disminuyen cada mes.
- La rentabilidad del casino aumente cada mes.

7.3.2 Análisis de actividades: modelo lógico - arquitectura técnica

Para el desarrollo exitoso de este proyecto se hace necesario especificar cuáles son las actividades claves dentro del mismo que nos llevarán desde el punto cero hasta la meta.

Pero para que podamos tener un mejor entendimiento de dichas actividades les invitamos a ver el siguiente diagrama de arquitectura de la solución.

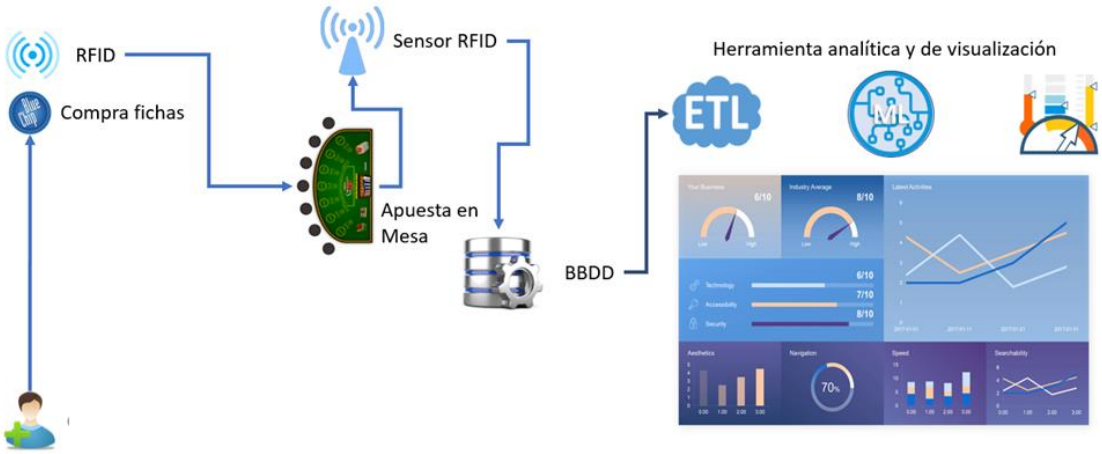


Ilustración 4 - Diagrama Arquitectura Solución

Visto el diagrama anterior podemos tener una idea más clara de las partes que componen el proyecto, y, por tanto, podemos tener una mejor estimación de las actividades necesarias para el desarrollo de este.

Pero tomando en consideración cada parte según la vemos en el diagrama, nos damos cuenta de que es imperativo separar el conjunto de actividades necesarias en una serie de fases que nos permitan ir alcanzando el desarrollo e implementación de distintos componentes o partes de la solución de forma separada, y que a la vez nos permita paralelizar la ejecución.

Hemos separado todo el proceso en 7 fases que nos permitirán un mejor aprovechamiento del tiempo, dado que algunas de ellas se pueden paralelizar, permitiendo esto una agilización en el proceso de implementación. A continuación, cada de ellas.

Fase 0: Planificación y Adquisición

En esta fase preliminar queremos realizar los preparativos iniciales previo al inicio de las actividades de despliegue de la solución. Con esto perseguimos que, una vez iniciada la implementación, todos los requerimientos necesarios para la consecución del proyecto se hayan cumplido.

Por tanto, será en esta fase en la cual se ha de comprar cada una de las partes necesarias, tanto de software como de hardware, para la implementación; además, será aquí también donde se ajustarán las fechas en caso de ser necesario.

Actividades a realizar:

- Compra de lectores RFID.
- Compra de Micro ordenares.
- Compra de tarjetas RFID.
- Compra de fichas RFID.
- Compra de licencias de software.
- Reevaluación de fechas de inicio y entrega.

Fase 1: Estudio de los Datos

Esta primera fase tiene como fin generar un modelo de datos el cual podamos utilizar para los objetivos últimos de esta propuesta: utilizar modelos predictivos que nos permitan segmentar los clientes en múltiples frentes.

Para ello es necesario la realización de un estudio de cómo el casino genera, procesa y almacena los datos para cada uno de los servicios que brinda, y que luego nos servirá para la clasificación de los clientes. Así mismo, es en esta fase donde se ha de definir cómo los nuevos datos generados por la solución se integrarán a los datos actuales, y cómo esto afectaría a los sistemas presentes en el casino.

Una vez obtenido un mapa de los datos actuales en conjunción con los que se han de agregar podremos generar un modelo de datos que nos permitirá más adelante usarlo para la construcción de nuestros modelos predictivos.

Actividades a realizar:

- Generar modelo de datos actual.
- Identificar las fuentes de los datos.
- Integrar nuevos datos en el modelo de datos existente.
- Validar nuevo modelo de datos.

Fase 2: Compra de Fichas (cliente)

Por compra de fichas nos referimos al área del casino donde el cliente, al llegar al establecimiento, procede a comprar las fichas que ha de utilizar en las mesas de juego.

En esta fase perseguimos dejar operativa esta área del negocio, mediante la instalación de los equipos necesarios para la codificación de las fichas, los equipos de comunicación y el software que permitirá la asignación de las fichas a un determinado cliente. Este componente del proyecto es bien importante ya que es el punto de entrada a la propuesta de esta solución.

Actividades a realizar:

- Desarrollo de componente de software para área de caja.
- Integración con sistemas actuales.
- Instalación y configuración de equipo RFID.

Fase 3: Mesas de Juego Inteligentes

Aquí procederemos al ensamblaje de las mesas de juego con capacidad RFID, lo cual es el componente IoT de la propuesta. Esto consiste en una serie de dispositivos que permitirán dar seguimiento al comportamiento de juego en cada una de las mesas.

Cada dispositivo lee, registra y envía los datos a una unidad central dentro de la misma mesa, por lo que cada mesa se comporta como un solo dispositivo en su totalidad, luego, dicha unidad central se encarga de enviar los datos a los servidores del casino para que la misma sea persistida en las bases de datos.

A la vez se han de instalar las conexiones de redes que permitan a las mesas poder comunicarse con los servidores.

Al final de esta fase, tendremos mesas inteligentes capaces de recolectar información por si mismas, guardarla localmente y enviarla a servidores de almacenamiento de datos.

Actividades a realizar:

- Desarrollo de componente de software para la mesa de juego.
- Desarrollo de componente de software para recibir datos desde las mesas.
- Instalación y configuración de lectores RFID en la mesa.
- Instalación de microordenador (incluida comunicación de red).
- Instalación de nuevos componentes de software desarrollados.
- Integración del software de mesa con el software de recibimiento de datos.

Fase 4: Data Warehouse y ETL

Sera necesario la creación, de no existir, de un data warehouse que nos permita guardar la información en el modelo correcto para el análisis de datos, tanto por parte de los modelos predictivos como para la visualización de estos. Por tanto, es el objetivo de esta fase el diseño e implementación de dicho almacén de datos. Si el casino ya posee uno, entonces debemos agregar nuestro modelo de datos al mismo.

Además, los procesos de extracción, transformación y carga de datos (ETL) deben quedar diseñados e implementados con los procesos de limpieza y calidad de datos que sean requeridos.

Es necesario puntualizar que, aunque nuestro cliente principal (directo) es la empresa INMABUIN, los procesos expuestos anteriormente, han de ser realizados en el casino al cual INMABUIN preste servicios.

Actividades a realizar:

- Diseño del almacén de datos.
- Diseño de procesos ETL.
- Construcción del almacén de datos.
- Implementación de procesos de ETL.

Fase 5: Análisis de Datos mediante Modelos Predictivos (ML)

Una vez los compontes que forman la base de la solución están en lugar, lo cual hemos logrado con la ejecución de las fases previas, podemos iniciar a sacar el máximo provecho a los datos, por lo cual iniciamos con la penúltima fase, siendo esta parte el objetivo de toda la solución.

Aquí el propósito es el diseño y construcción de los modelos predictivos que forman parte de esta propuesta. Luego de construidos dichos modelos, procederemos a entrenarlos con los datos actuales. Una vez entrenado cada modelo, procedemos a la implantación de estos para que sus predicciones puedan ser aprovechadas por cada una de las aplicaciones que operan en el negocio.

Actividades a realizar:

- Evaluación de algoritmos predictivos.
- Construcción de los modelos predictivos.
- Evaluación de los modelos predictivos.
- Implantación de los modelos predictivos.

Fase 6: Visualización de Datos

Llegados a esta fase hemos logrado casi todo lo necesario para el éxito del proyecto, pero aun nos falta una parte importante del mismo, cómo visualizaremos los datos que hemos limpiado y preparado para análisis.

El propósito en esta última etapa de implementación del proyecto es generar los reportes y cuadros de mando apropiados para cada una de las partes interesadas en la cadena ejecutiva del casino, lo cual les permitirá tener acceso a información de calidad en tiempo real.

Actividades a realizar:

- Diseño de reportes.
- Diseño de controles de mando.
- Construcción de reportes.
- Construcción de cuadros de mando.
- Implementación de reportes y cuadros de mando.

Cuando vemos cada una de las fases, podemos darnos cuenta de lo siguiente: la fase 1 debe realizarse primero de forma secuencial, ya que de los resultados de esta es que podremos avanzar en las demás; pero las demás pueden realizarse de forma paralela, habilitando con esto una optimización en el tiempo de implementación.

Concluido el despliegue de la solución, es necesario iniciar con el proceso posventa.

Proceso Posventa

Formación a usuarios y técnicos. Se realizarán jornadas de capacitación a los usuarios que harán uso del sistema y a los técnicos que estarán como responsables, tanto en el casino como en INMABUIN. Además, se suplirán manuales para usuario y manuales técnicos.

Soporte. Como toda herramienta viene con sus dudas y también con problemas, generalmente relacionados al desconocimiento de la nueva herramienta, es compromiso por parte de nosotros brindar servicio de soporte durante los próximos 6 meses a la implementación, con el fin de aclarar toda duda o para la resolución de los inconvenientes que puedan surgir.

7.3.3 Análisis de recursos: talento humano y recursos físicos

Dentro del análisis de los recursos requeridos para la implementación de este proyecto, debemos considerar dos tipos de recursos que serán las piezas fundamentales de nuestra propuesta: Los recursos humanos y recursos físicos.

En el análisis de los recursos humanos, es necesario considerar los siguientes factores:

Estructura Organizativa:

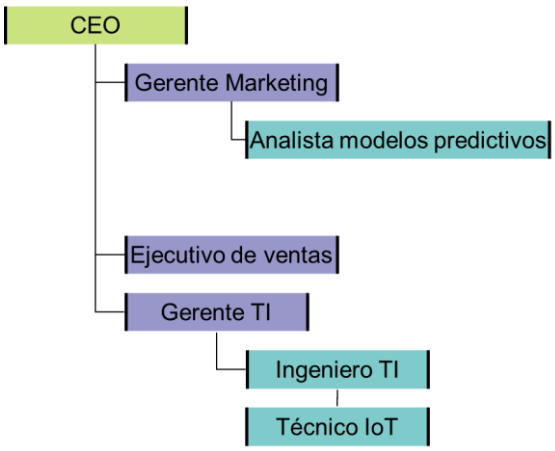


Ilustración 5 - Estructura Organizativa

Denominación del recurso humano y tareas asignadas:

Para cada usuario requerido en la implementación y sostenibilidad del proyecto, se han definido funciones y responsabilidades que justifiquen la asignación de estos recursos al proyecto. Estas funciones se pueden visualizar en el siguiente diagrama talento – tarea:

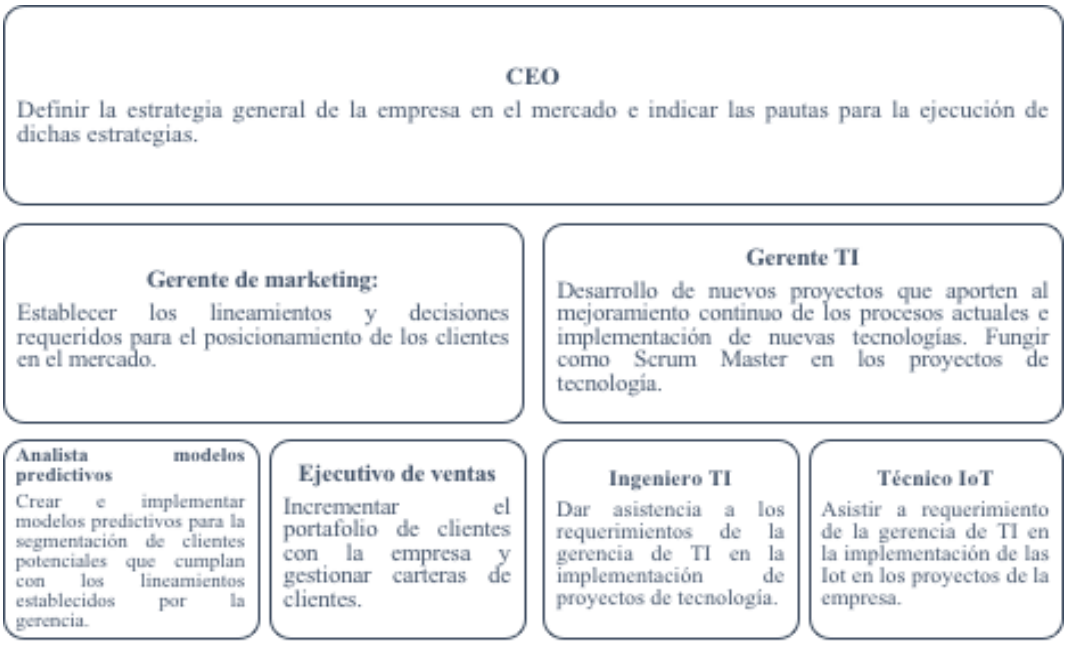


Ilustración 6 - Asignación de tareas del personal

7.3.4 Gestión del tiempo (cronograma)

Para representar la gestión del tiempo sobre cómo se realizarán las diferentes actividades expuestas anteriormente, utilizaremos como herramienta el diagrama de Gantt, el cual nos permitirá visualizar de forma clara y concisa el alcance temporal del proyecto.

Independientemente de que tenemos una serie de actividades que deben realizarse en forma secuencial, la mayor cantidad de estas pueden ir ejecutándose de forma paralela, lo cual nos brinda una mejor gestión del tiempo, permitiendo esto la realización del proyecto en un menor número de días.

Debido al gran número de actividades, primero presentaremos un diagrama con todas las fases, pero sin detallar las actividades contenidas dentro de ellas. Esto permitirá ver el proyecto como un todo a través del tiempo; luego, iremos presentando diferentes diagramas por cada fase, de modo que podamos ver con precisión cómo ha de ser la ejecución del proyecto.

A continuación, una visión general de principio a fin del proyecto.

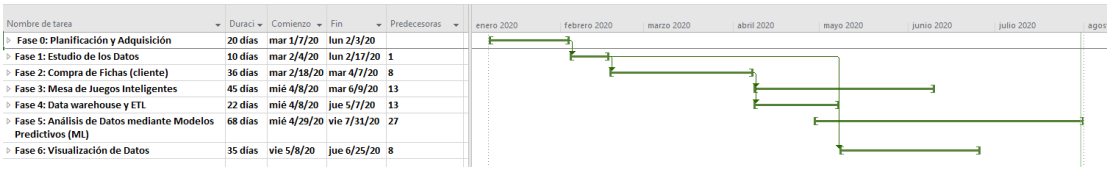


Ilustración 7 - Cronograma de gestión de tiempo proyecto general

Como podemos ver, el proyecto tiene una estimación de 7 meses calendario, iniciando el lunes 07 de enero de 2020 y finalizando el 31 de julio de 2020. Esta estimación no toma en consideración fines de semanas como días laborables; sin embargo, no están marcados los días festivos, por lo que podríamos esperar mínimas variaciones debido a estos, pero que no deben impactar significativamente en el desarrollo del proyecto. Tomando en cuenta esta última puntualización, nos queda que el proyecto tiene una estimación de 149 días laborables.

También podemos visualizar en el diagrama anterior el nivel de paralelización que podremos tener en la implantación de la solución.

Fase 0: Planificación y Adquisición

Esta fase podrá realizarse en un total de 20 días laborables, desde el 07/01/2020 hasta el 03/02/2020.

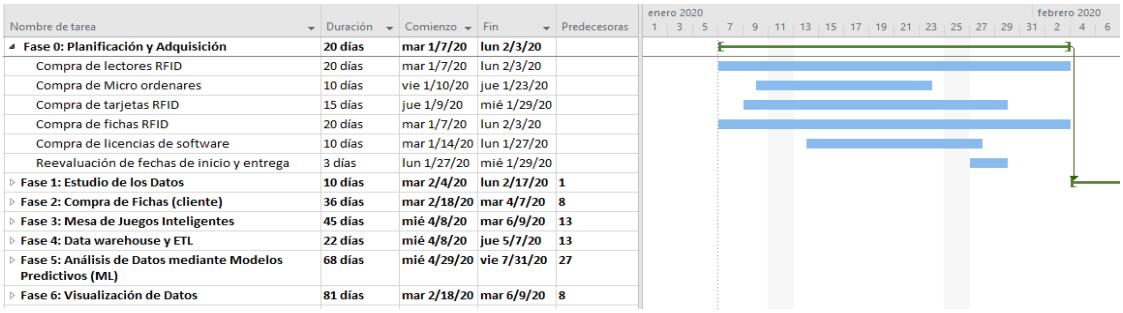


Ilustración 8 - Cronograma de gestión de tiempo Fase 0

Fase 1: Estudio de los Datos

Esta fase podrá realizarse en un total de 10 días laborables, desde el 04/02/2020 hasta el 17/02/2020.

Nombre de tarea	Duración	Comienzo	Fin	Predecesoras
Fase 0: Planificación y Adquisición	**20 días**	**mar 1/7/20**	**lun 2/3/20**	
Fase 1: Estudio de los Datos	**10 días**	**mar 2/4/20**	**lun 2/17/20**	**1**
Generar modelo de datos actual	3 días	mar 2/4/20	jue 2/6/20	1
Identificar las fuentes de los datos	3 días	vie 2/7/20	mar 2/11/20	9
Integrar nuevos datos en el modelo de datos ex	2 días	mié 2/12/20	jue 2/13/20	10
Validar nuevo modelo de datos	2 días	vie 2/14/20	lun 2/17/20	11
Fase 2: Compra de Fichas (cliente)	**36 días**	**mar 2/18/20**	**mar 4/7/20**	**8**
Fase 3: Mesa de Juegos Inteligentes	**45 días**	**mié 4/8/20**	**mar 6/9/20**	**13**
Fase 4: Data warehouse y ETL	**22 días**	**mié 4/8/20**	**jue 5/7/20**	**13**
Fase 5: Análisis de Datos mediante Modelos Predictivos (ML)	**68 días**	**mié 4/29/20**	**vie 7/31/20**	**27**
Fase 6: Visualización de Datos	**81 días**	**mar 2/18/20**	**mar 6/9/20**	**8**

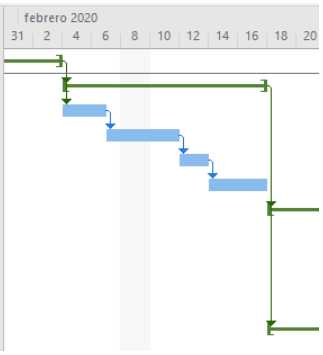


Ilustración 9 - Cronograma de gestión de tiempo Fase 1

Fase 2: Compra de Fichas (cliente)

Esta fase podrá realizarse en un total de 36 días laborables, desde el 18/02/2020 hasta el 07/04/2020.

Nombre de tarea	Duración	Comienzo	Fin	Predecesoras
Fase 0: Planificación y Adquisición	**20 días**	**mar 1/7/20**	**lun 2/3/20**	
Fase 1: Estudio de los Datos	**10 días**	**mar 2/4/20**	**lun 2/17/20**	**1**
Fase 2: Compra de Fichas (cliente)	**36 días**	**mar 2/18/20**	**mar 4/7/20**	**8**
Desarrollo de componente de software para área de caja	30 días	mar 2/18/20	lun 3/30/20	8
Integración con sistemas actuales	5 días	mar 3/31/20	lun 4/6/20	14
Instalación y configuración de equipo RFID	1 día	mar 4/7/20	mar 4/7/20	15
Fase 3: Mesa de Juegos Inteligentes	**45 días**	**mié 4/8/20**	**mar 6/9/20**	**13**
Fase 4: Data warehouse y ETL	**22 días**	**mié 4/8/20**	**jue 5/7/20**	**13**
Fase 5: Análisis de Datos mediante Modelos Predictivos (ML)	**68 días**	**mié 4/29/20**	**vie 7/31/20**	**27**
Fase 6: Visualización de Datos	**81 días**	**mar 2/18/20**	**mar 6/9/20**	**8**

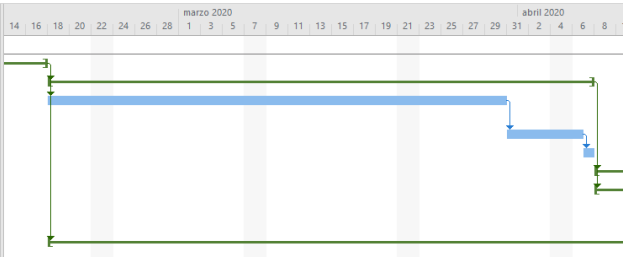


Ilustración 10 - Cronograma de gestión de tiempo Fase 2

Fase 3: Mesa de Juegos Inteligentes

Esta fase podrá realizarse en un total de 45 días laborables, desde el 08/04/2020 hasta el 09/06/2020.

Nombre de tarea	Duración	Comienzo	Fin	Predecesoras
Fase 0: Planificación y Adquisición	**20 días**	**mar 1/7/20**	**lun 2/3/20**	
Fase 1: Estudio de los Datos	**10 días**	**mar 2/4/20**	**lun 2/17/20**	**1**
Fase 2: Compra de Fichas (cliente)	**36 días**	**mar 2/18/20**	**mar 4/7/20**	**8**
Fase 3: Mesa de Juegos Inteligentes	**45 días**	**mié 4/8/20**	**mar 6/9/20**	**13**
Desarrollo de componente de software para la mesa de juego	30 días	mié 4/8/20	mar 5/19/20	13
Desarrollo de componente de software para recibir datos desde las mesas	15 días	mié 5/20/20	mar 6/9/20	18
Instalación y configuración de lectores RFID en la mesa	3 días	mié 4/8/20	vie 4/10/20	
Instalación de microordenador (incluida comunicación de red)	2 días	lun 4/13/20	mar 4/14/20	20
Instalación de nuevos componentes de software desarrollados	2 días	mié 4/15/20	jue 4/16/20	21
Integración sistema de software de mesa con software de recibimiento de datos	1 día	vie 4/17/20	vie 4/17/20	22
Fase 4: Data warehouse y ETL	**22 días**	**mié 4/8/20**	**jue 5/7/20**	**13**
Fase 5: Análisis de Datos mediante Modelos Predictivos (ML)	**68 días**	**mié 4/29/20**	**vie 7/31/20**	**27**
Fase 6: Visualización de Datos	**81 días**	**mar 2/18/20**	**mar 6/9/20**	**8**

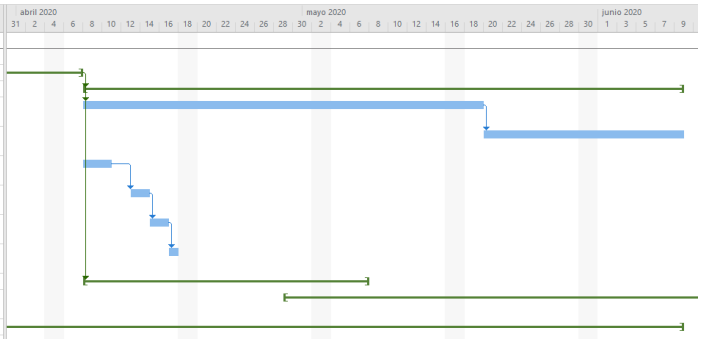
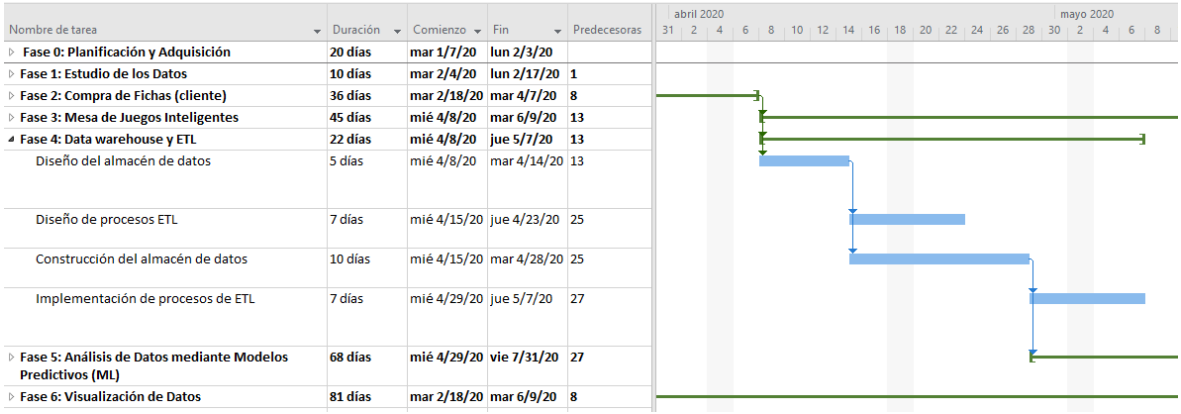


Ilustración 11 - - Cronograma de gestión de tiempo fase 3

Fase 4: Data warehouse y ETL

Esta fase podrá realizarse en un total de 22 días laborables, desde el 08/04/2020 hasta el 05/07/2020.

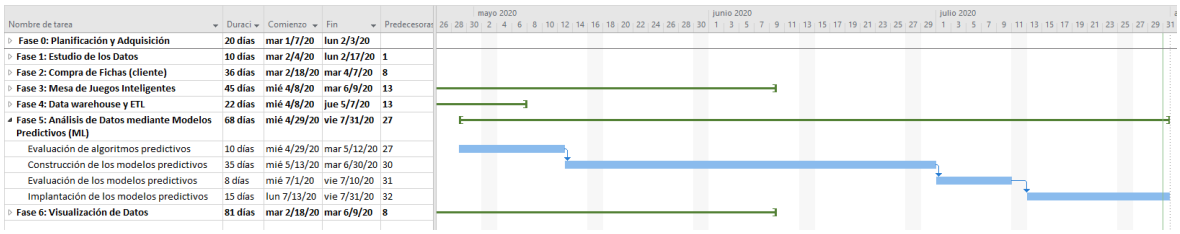


Nombre de tarea	Duración	Comienzo	Fin	Predecesoras
Fase 0: Planificación y Adquisición	**20 días**	**mar 1/7/20**	**lun 2/3/20**	
Fase 1: Estudio de los Datos	**10 días**	**mar 2/4/20**	**lun 2/17/20**	**1**
Fase 2: Compra de Fichas (cliente)	**36 días**	**mar 2/18/20**	**mar 4/7/20**	**8**
Fase 3: Mesa de Juegos Inteligentes	**45 días**	**mié 4/8/20**	**mar 6/9/20**	**13**
Fase 4: Data warehouse y ETL	**22 días**	**mié 4/8/20**	**jue 5/7/20**	**13**
Diseño del almacén de datos	5 días	mié 4/8/20	mar 4/14/20	13
Diseño de procesos ETL	7 días	mié 4/15/20	jue 4/23/20	25
Construcción del almacén de datos	10 días	mié 4/15/20	mar 4/28/20	25
Implementación de procesos de ETL	7 días	mié 4/29/20	jue 5/7/20	27
Fase 5: Análisis de Datos mediante Modelos Predictivos (ML)	**68 días**	**mié 4/29/20**	**vie 7/31/20**	**27**
Fase 6: Visualización de Datos	**81 días**	**mar 2/18/20**	**mar 6/9/20**	**8**

Ilustración 12 - Cronograma de gestión de tiempo fase 4

Fase 5: Análisis de Datos mediante Modelos Predictivos (ML)

Esta fase podrá realizarse en un total de 68 días laborables, desde el 29/04/2020 hasta el 31/07/2020.

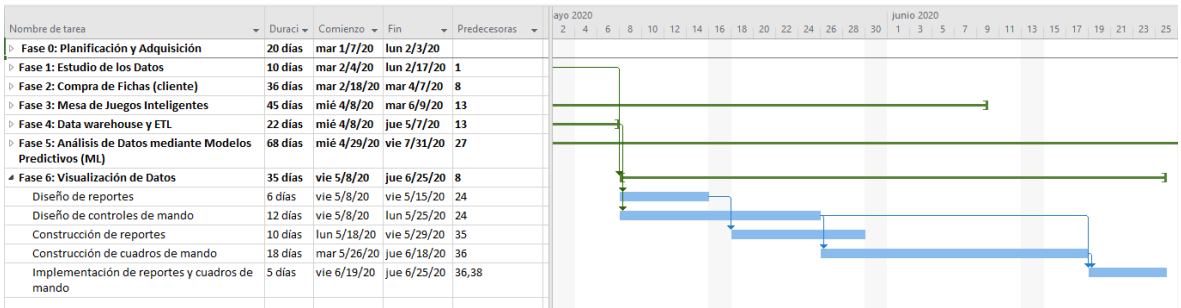


Nombre de tarea	Duraci	Comienzo	Fin	Predecesoras
Fase 0: Planificación y Adquisición	**20 días**	**mar 1/7/20**	**lun 2/3/20**	
Fase 1: Estudio de los Datos	**10 días**	**mar 2/4/20**	**lun 2/17/20**	**1**
Fase 2: Compra de Fichas (cliente)	**36 días**	**mar 2/18/20**	**mar 4/7/20**	**8**
Fase 3: Mesa de Juegos Inteligentes	**45 días**	**mié 4/8/20**	**mar 6/9/20**	**13**
Fase 4: Data warehouse y ETL	**22 días**	**mié 4/8/20**	**jue 5/7/20**	**13**
Fase 5: Análisis de Datos mediante Modelos Predictivos (ML)	**68 días**	**mié 4/29/20**	**vie 7/31/20**	**27**
Evaluación de algoritmos predictivos	10 días	mié 4/29/20	mar 5/12/20	27
Construcción de los modelos predictivos	35 días	mié 5/13/20	mar 6/30/20	30
Evaluación de los modelos predictivos	8 días	mié 7/1/20	vie 7/10/20	31
Implantación de los modelos predictivos	15 días	lun 7/13/20	vie 7/31/20	32
Fase 6: Visualización de Datos	**81 días**	**mar 2/18/20**	**mar 6/9/20**	**8**

Ilustración 13 - Cronograma de gestión de tiempo fase 5

Fase 6: Visualización de Datos

Esta fase podrá realizarse en un total de 35 días laborables, desde el 08/05/2020 hasta el 25/06/2020.



Nombre de tarea	Duraci	Comienzo	Fin	Predecesoras
Fase 0: Planificación y Adquisición	**20 días**	**mar 1/7/20**	**lun 2/3/20**	
Fase 1: Estudio de los Datos	**10 días**	**mar 2/4/20**	**lun 2/17/20**	**1**
Fase 2: Compra de Fichas (cliente)	**36 días**	**mar 2/18/20**	**mar 4/7/20**	**8**
Fase 3: Mesa de Juegos Inteligentes	**45 días**	**mié 4/8/20**	**mar 6/9/20**	**13**
Fase 4: Data warehouse y ETL	**22 días**	**mié 4/8/20**	**jue 5/7/20**	**13**
Fase 5: Análisis de Datos mediante Modelos Predictivos (ML)	**68 días**	**mié 4/29/20**	**vie 7/31/20**	**27**
Fase 6: Visualización de Datos	**35 días**	**vie 5/8/20**	**jue 6/25/20**	**8**
Diseño de reportes	6 días	vie 5/8/20	vie 5/15/20	24
Diseño de controles de mando	12 días	vie 5/8/20	lun 5/25/20	24
Construcción de reportes	10 días	lun 5/18/20	vie 5/29/20	35
Construcción de cuadros de mando	18 días	mar 5/26/20	jue 6/18/20	36
Implementación de reportes y cuadros de mando	5 días	vie 6/19/20	jue 6/25/20	36,38

Ilustración 14 - Cronograma de gestión de tiempo fase 6

Como mencionamos al inicio, los días festivos no fueron tomados en consideración, por lo que representarán una variación mínima que no debería impactar significativamente el proyecto.

8. FUENTES DE INFORMACIÓN Y TAXONOMÍAS

Para que un proyecto de BD sea exitoso se debe contar con un sin número de datos históricos, ya sea estructurados o no, lo que nos va a permitir analizar mejor la información y crear modelos predictivos que apoyarán a las empresas en sus tareas del día a día y a que estas logren un mejor manejo de toma de decisiones.

Nuestro cliente INMABUIN nos proporcionó las fuentes de datos de dos de sus casinos, uno ubicado en República Dominicana, y otro en España, tomando en cuenta que se nos pidió cambiar nombres, eliminar DNI, números de cédula y pasaportes, para mantener el anonimato y resguardar la identidad de sus clientes.

Estas fuentes de datos provienen de datos estructurados y relacionales bajo el gestor de base de datos SQL Server de Microsoft, donde se encuentran registradas las transacciones diarias de los casinos provenientes de sus sistemas transaccionales con datos históricos desde el año 2014 al 2018.

La estructura inicial consta de 4 bases de datos, las cuales antes de empezar a trabajar con los modelos fueron analizadas por el equipo y se eliminaron algunas tablas que no se necesitarían.

- **BDCortesia**: los registros principales que esta base de datos almacena son datos de las cortesías ofrecidas a los clientes y el inventario de productos con que cuentan los casinos para ofrecer de cortesía. En primera instancia, la base de datos contenía 23 tablas, y después de realizado el análisis de datos se procede a eliminar unas cuantas tablas que almacenaban datos de módulos de sistemas, pantallas e idiomas, las cuales no afectan ni influyen en nuestro modelo de datos.

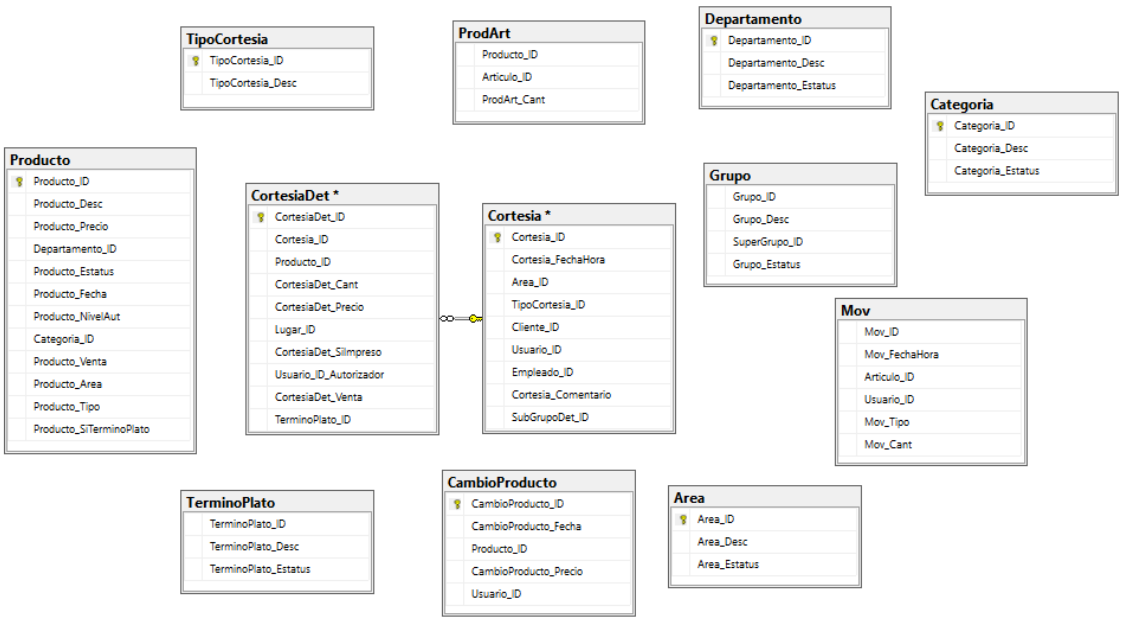


Ilustración 15 - Diagrama BDCortesia

- **BDCliente:** Esta es la base de datos más importante y la principal fuente de datos de nuestros modelos y cuadros de mando. Aquí se encuentran registrados los datos de los clientes, sus visitas, sus apuestas aproximadas y las mesas en las cuales ha jugado, recordando que aquí solo se almacenan datos estimados y no de todas las visitas realizadas en los casinos.

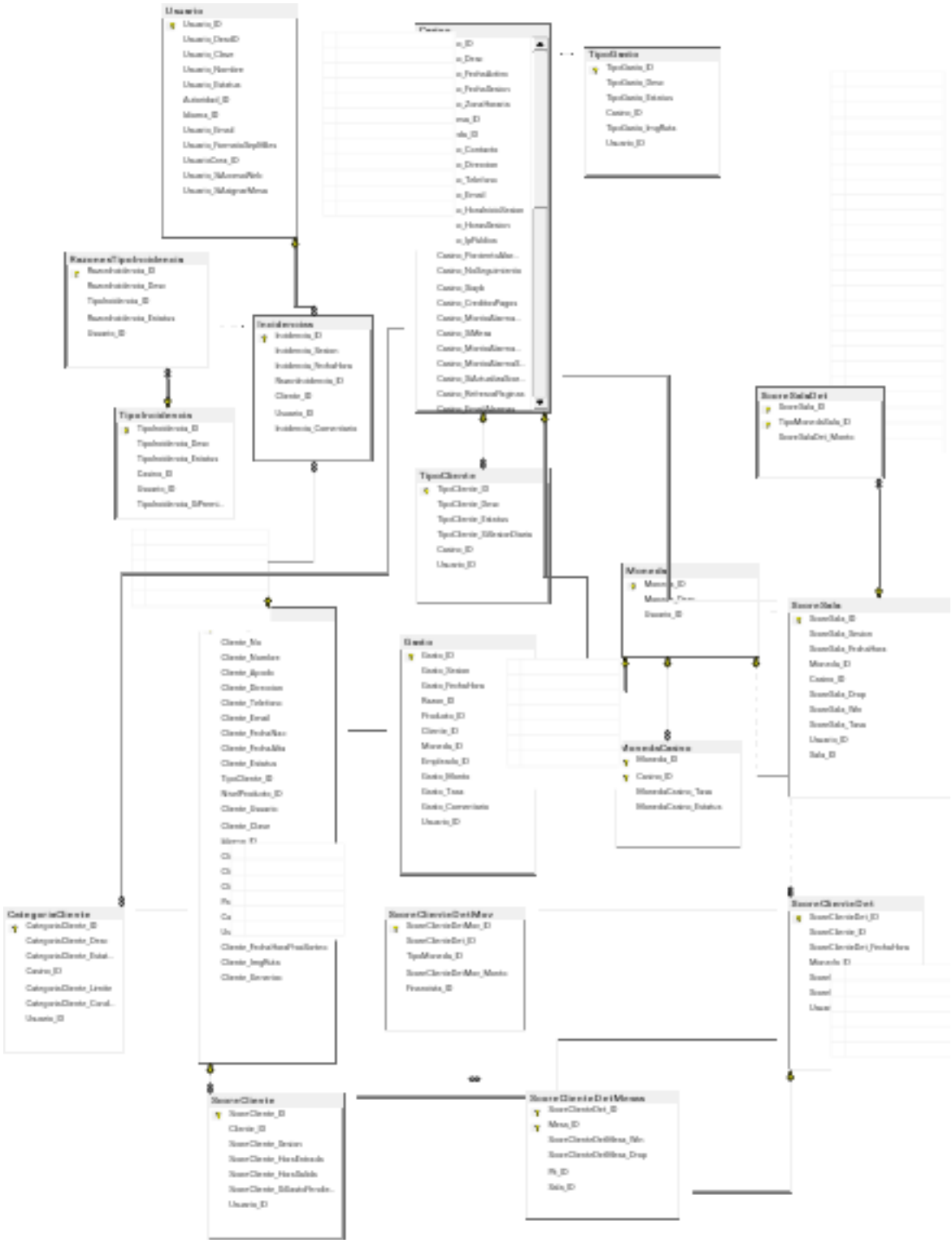


Ilustración 16 - Diagrama BDCliente

- **BDFidelizacion:** Esta base de datos almacena los clientes que pertenecen al programa de fidelización. Estos clientes están formalmente registrados y reciben en sus correos ofertas y promociones del casino, con cada visita acumulan puntos que pueden cambiar por fichas. En esta

base de datos se registra cada visita de estos clientes, lo que nos ayudará en los modelos predictivos para comparar el comportamiento de los clientes fidelizados con los que no lo están

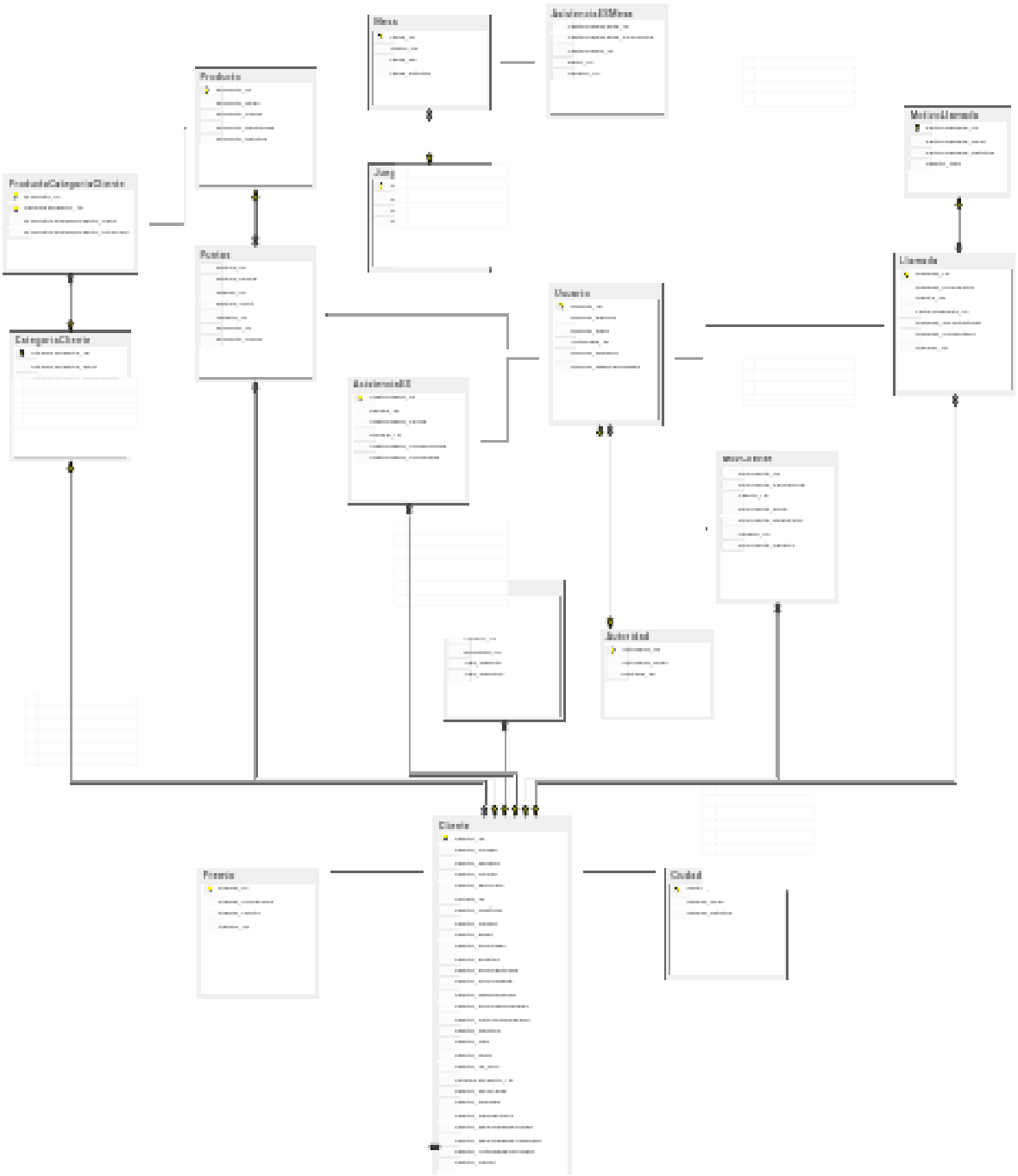


Ilustración 17 - Diagrama BDFidelización

- **BDMarketing:** Aquí se almacenan los datos relacionados a las campañas de marketing realizadas en los casinos. Podemos encontrar datos sobre los costos de estas campañas, los premios ofrecidos, fechas y horas en las cuales se realizaron las campañas de marketing, y además el cliente que ganó en las promociones y eventos realizados.

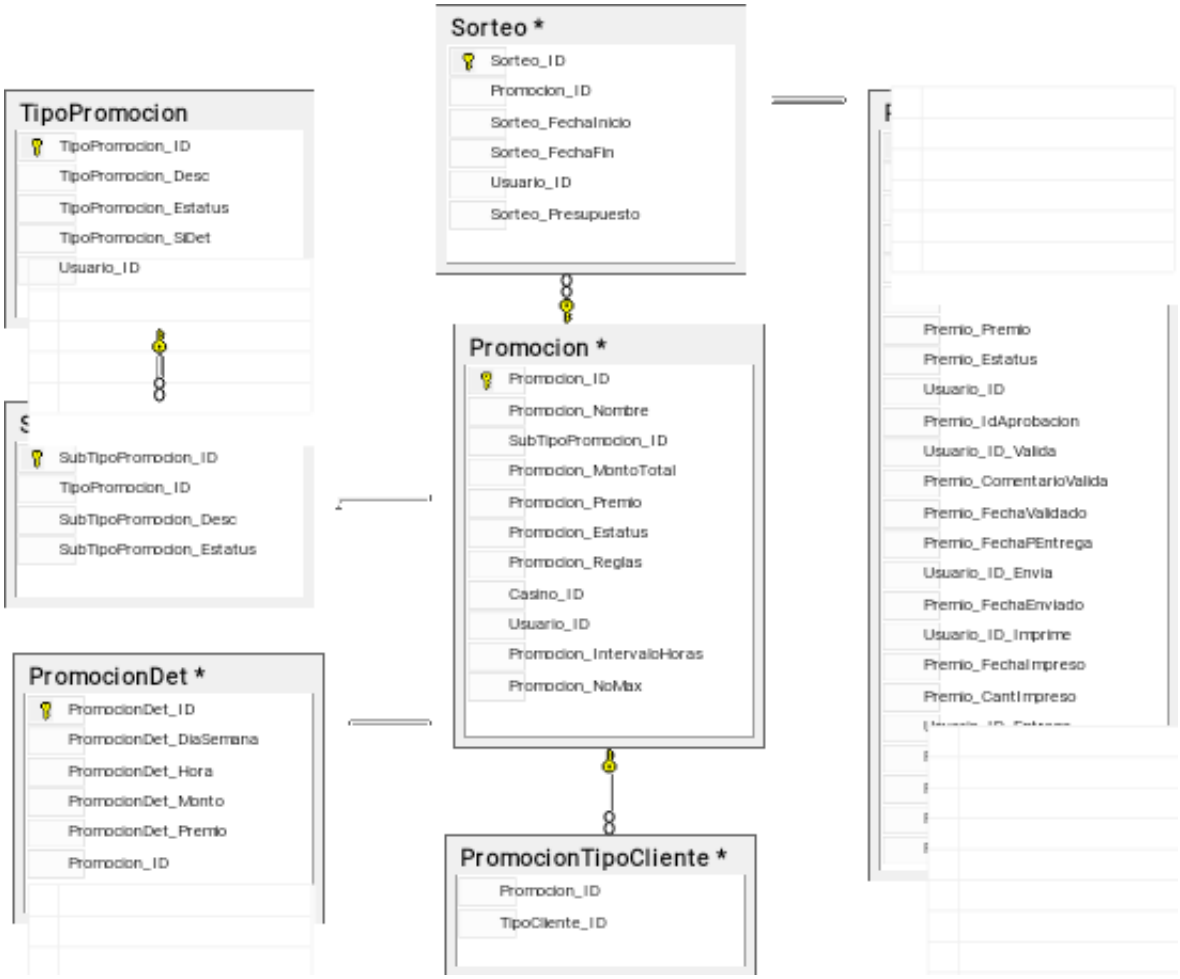


Ilustración 18 - Diagrama BDMarketing

Se realizaron unas cuantas vistas con consulta de datos relacionadas con los clientes para así alimentar los cuadros de mandos y los modelos en Python.

USO	Instrucción SQL de la Vista
Modelo para Segmentación de Clientes	SELECT Cliente.Cliente_ID, Cliente.Cliente_Nombre, Cliente.Cliente_FechaNac, Cliente.Cliente_FechaAlta,cliente.Cliente_Genero , TipoCliente.TipoCliente_Desc, pais.Pais_Desc, CategoriaCliente.CategoriaCliente_Desc, ScoreCliente.ScoreCliente_ID,ScoreCliente.ScoreCliente_HoraEntrada, ScoreCliente.ScoreCliente_HoraSalida, Usuario.Usuario_Nombre, ScoreClienteDet.ScoreClienteDet_ID, ScoreClienteDet.ScoreClienteDet_FechaHora, Moneda.Moneda_Desc, ScoreClienteDetMesas.Mesa_ID, ScoreClienteDetMesas.ScoreClienteDetMesa_Drop,ScoreClienteDetMesas.ScoreClienteDetMesa_Win, TipoMoneda.TipoMoneda_Desc, ScoreClienteDetMov.ScoreClienteDetMov_Monto from Cliente left join Pais on cliente.pais_ID = pais.pais_id LEFT join TipoCliente on Cliente.TipoCliente_ID =TipoCliente.TipoCliente_ID

		left join CategoriaCliente on CategoriaCliente.CategoriaCliente_ID = Cliente.CategoriaCliente_ID right join ScoreCliente on ScoreCliente.Cliente_ID = Cliente.Cliente_ID left join Usuario on ScoreCliente.Usuario_ID = Usuario.Usuario_ID right join ScoreClienteDet on ScoreCliente.ScoreCliente_ID = ScoreClienteDet.ScoreCliente_ID left join Moneda on ScoreClienteDet.Moneda_ID = Moneda.Moneda_ID right join ScoreClienteDetMesas on ScoreClienteDetMesas.ScoreClienteDet_ID =ScoreClienteDet.ScoreClienteDet_ID left join ScoreClienteDetMov on ScoreClienteDetMov.ScoreClienteDet_ID = ScoreClienteDet.ScoreClienteDet_ID left join TipoMoneda on ScoreClienteDetMov.TipoMoneda_ID = TipoMoneda.TipoMoneda_ID
Modelo para detección de posible fraude.		Select Cliente_No, Cliente_Genero, Pais_Desc, CategoriaCliente_ID, CategoriaCliente_Desc, TipoCliente_ID, TipoCliente_Desc, ScoreCliente_Sesion, ScoreCliente_HoraEntrada, ScoreCliente_HoraSalida, ScoreClienteDet_FechaHora, ScoreClienteDetMov_Monto, TipoMoneda_Desc, Usuario_DescID, CASE WHEN --R_Fraude >= 800 and (ScoreClienteDetMov_Monto>10000 and TipoCliente_Desc='GENERAL' and (datepart(hh,ScoreClienteDet_FechaHora)<=10 or datepart(hh,ScoreClienteDet_FechaHora)>=20)) THEN 1 ELSE 0 END AS Fraude from (SELECT dbo.Cliente.Cliente_No, dbo.Cliente.Cliente_Genero, dbo.Pais.Pais_Desc, dbo.Cliente.CategoriaCliente_ID, dbo.CategoriaCliente.CategoriaCliente_Desc, dbo.Cliente.TipoCliente_ID, dbo.TipoCliente.TipoCliente_Desc, dbo.ScoreCliente.ScoreCliente_Sesion, dbo.ScoreCliente.ScoreCliente_HoraEntrada, dbo.ScoreCliente.ScoreCliente_HoraSalida, dbo.ScoreClienteDet.ScoreClienteDet_FechaHora, dbo.ScoreClienteDetMov.ScoreClienteDetMov_Monto, dbo.TipoMoneda.TipoMoneda_Desc, dbo.Usuario.Usuario_DescID, RAND(CHECKSUM(NEWID())) * 1000 AS R_Fraude FROM dbo.Cliente INNER JOIN dbo.ScoreCliente ON dbo.Cliente.Cliente_ID = dbo.ScoreCliente.Cliente_ID INNER JOIN .ScoreClienteDet ON dbo.ScoreCliente.ScoreCliente_ID = dbo.ScoreClienteDet.ScoreCliente_ID INNER JOIN dbo.ScoreClienteDetMov ON dbo.ScoreClienteDet.ScoreClienteDet_ID = dbo.ScoreClienteDetMov.ScoreClienteDet_ID INNER JOIN dbo.TipoMoneda ON dbo.ScoreClienteDetMov.TipoMoneda_ID = dbo.TipoMoneda.TipoMoneda_ID INNER JOIN dbo.Pais ON dbo.Cliente.Pais_ID = dbo.Pais.Pais_ID INNER JOIN dbo.CategoriaCliente ON dbo.Cliente.CategoriaCliente_ID = dbo.CategoriaCliente.CategoriaCliente_ID INNER JOIN dbo.TipoCliente ON dbo.Cliente.TipoCliente_ID = dbo.TipoCliente.TipoCliente_ID INNER JOIN dbo.Usuario ON dbo.ScoreClienteDet.Usuario_ID = dbo.Usuario.Usuario_ID WHERE (dbo.ScoreClienteDetMov.ScoreClienteDetMov_Monto <> 0) AND (dbo.Cliente.Cliente_Genero IN ('M', 'F')) AND (dbo.Pais.Pais_Desc <> '') AND (NOT (dbo.ScoreCliente.ScoreCliente_HoraSalida IS NULL))) Datos ORDER BY ScoreClienteDet_FechaHora

Tabla 5 - Vistas SQL

9. FICHAS TÉCNICAS DE LAS TECNOLOGÍAS DESARROLLADAS

9.1 Tecnología IoT

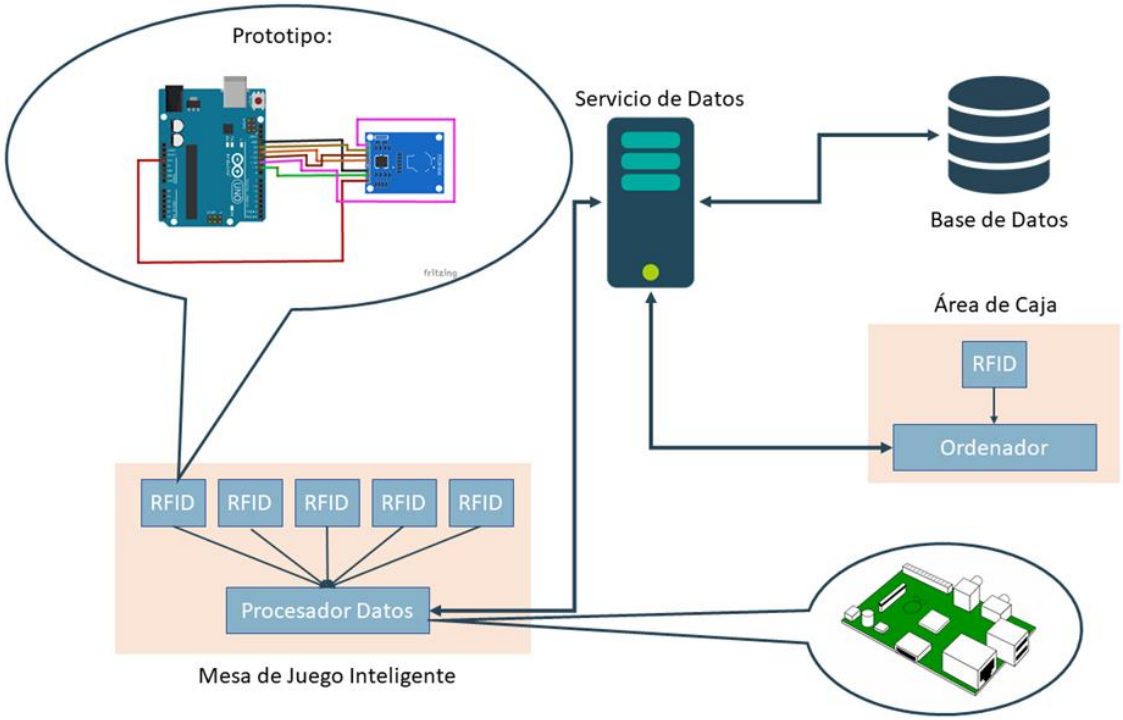


Ilustración 19 - Diagrama IoT

En el diagrama anterior podemos visualizar todos los componentes necesarios para que la solución de IoT funcione.

Componentes IOT

Mesa de juego inteligente. Este es el centro de la solución, ya que es la responsable, de forma autónoma, de registrar los datos de los clientes mientras éstos juegan.

Esta posee dos componentes esenciales: los lectores RFID y el procesador de datos. Los lectores RFID se encargarán de leer las fichas que están en juego, y enviar esta información al procesador de datos, este a su vez se encargará de identificar cada ficha y relacionarla con el jugador correspondiente o con la casa (es decir, que ya no pertenece a un jugador, sino al casino), para luego enviar esa información al servidor de datos.

En nuestro prototipo, el lector RFID está formado por una placa Arduino Uno, un microcontrolador programable, y por un lector RFID MFRC522. Este prototipo tiene limitaciones de alcance, pero muestra la solución en funcionamiento. El otro componente de la mesa de juego inteligente, el procesador de datos, es un microordenador RaspberryPi, el cual contiene el sistema operativo Windows IoT y una aplicación que se encarga de procesar la información proveniente de los lectores RFID.

Servicio de Datos. El servicio de datos es una aplicación desplegada en un servidor que brinda una serie interfaces de conexión. Esta recibe los datos enviados por el procesador de datos en la mesa inteligente y del ordenador (u ordenadores) en el área de caja, y se encarga a su vez, de mantener sincronizada

toda la información de las fichas que entran y salen de las áreas de juego, manejando además la persistencia de los datos en la base de datos.

Área de Caja. En el área de caja tendremos un ordenador común, el cual tendrá conectado un lector/escritor de RFID, este permitirá registrar las fichas que se han de jugar, así como aquellas que ya el cliente está intercambiando por dinero.

La aplicación instalada en este ordenador se comunica con el servicio de datos para hacerle saber a cuáles clientes se asignaron nuevas fichas, o cuales fichas ya están de regreso al casino sin clientes asignados debido a un intercambio de estas.

Limitaciones y Alcance del Prototipo IoT

Como mencionamos previamente, el prototipo desarrollado tiene sus limitaciones. La capacidad de lectura de los lectores RFID es de dos fichas simultaneas, a una distancia de 8 centímetros. Así mismo, la cantidad de lectores controlados por el procesador de datos (RaspberryPi) es de 4 dispositivos (Arduino UNO).

El alcance del prototipo se extiende solo al manejo de 3 clientes preestablecidos, los cuales han de estar en una mesa inteligente con sus fichas previamente asignadas. Podremos ver el funcionamiento de cómo las fichas se mueven dentro de la mesa hasta la persistencia de esta información en la base de datos.

9.2 Modelo de predicción de fraude

Debido a que los casinos son negocios que constantemente se enfrentan a situaciones fraudulentas, tanto de parte de sus clientes como de sus mismos empleados, se hace necesario, a partir de los datos disponibles, crear un sistema de alarmas que ayude a que el personal de seguridad y a la gerencia del casino estén en mayor alerta frente a aquellas jugadas que puedan representar un movimiento fraudulento.

Es para esto que se creará un modelo de analítica de operaciones, orientado a la detección de posible fraude frente a un movimiento o jugada llevado a cabo por un cliente. Para la creación del modelo se llevarán a cabo cuatro pasos: selección de las variables y reducción dimensional, aplicación de la técnica de PCA, creación de clústeres (clustering), y por último el análisis de la matriz de confusión.

Luego de creado el modelo para cada casino, se procederá a anexarlo al sistema de toma de datos, de forma tal que cada transacción realizada sea analizada por dicho modelo. Aquellas transacciones que de acuerdo con modelo resulten con una alta probabilidad de fraudulencia, se procederá a enviar un correo de alerta a las personas indicadas por el casino para el turno en cuestión.

Selección de variables y reducción dimensional

Para la muestra del prototipo, se seleccionaron los datos correspondientes a un casino de la empresa INMABIUIN, a los cuales se les asignó el atributo de Fraude a aquellas transacciones que, de acuerdo con los criterios del casino, son transacciones propensas a fraude. Luego, en cada casino se iniciará un proceso de asignación de transacciones fraudulentas para qué, a partir del tiempo puedan contar con una base de casos factible para la generación del modelo.

En ese caso se tomaron las siguientes variables: género del cliente, país del cliente, tipo de cliente, categoría de cliente, tipo de moneda, empleado en la mesa y monto de la transacción.

Aplicación de la técnica de PCA

Luego que se seleccionaron las variables y se redujeron dimensionalmente, se procedió a aplicar la técnica de PCA, con el objetivo de identificar cuales son los valores de las variables más significativas respecto al posible Fraude. En este caso se seleccionaron 6, debido a que eran las que tenían mayor influencua: tipo de moneda TICKET_ELECT y los empleados S. García, C. Velasco, J De Peddro, S. Aspiunza y el del Pit2.

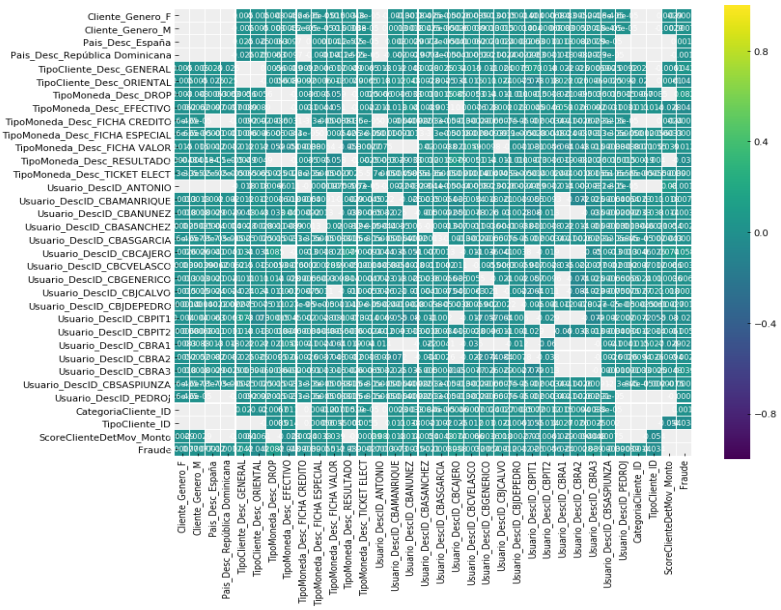


Ilustración 20 - Variables PCA

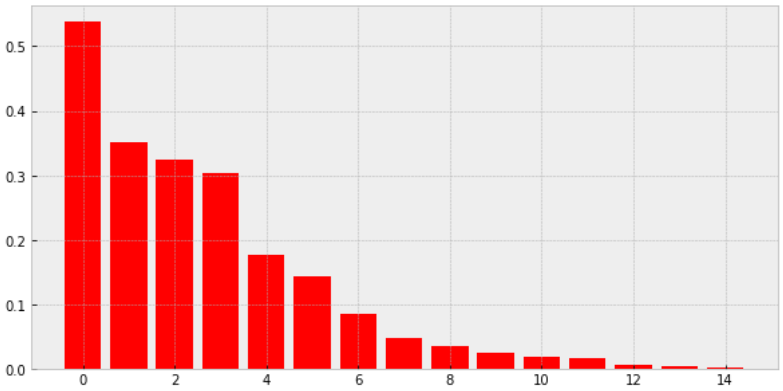


Ilustración 21 - Técnica PCA

En este sentido se puede observar que dichas dimensiones tienen baja correlación uno con otra, de forma tal que son apropiadas para determinar si una transacción puede ser o no fraudulenta.

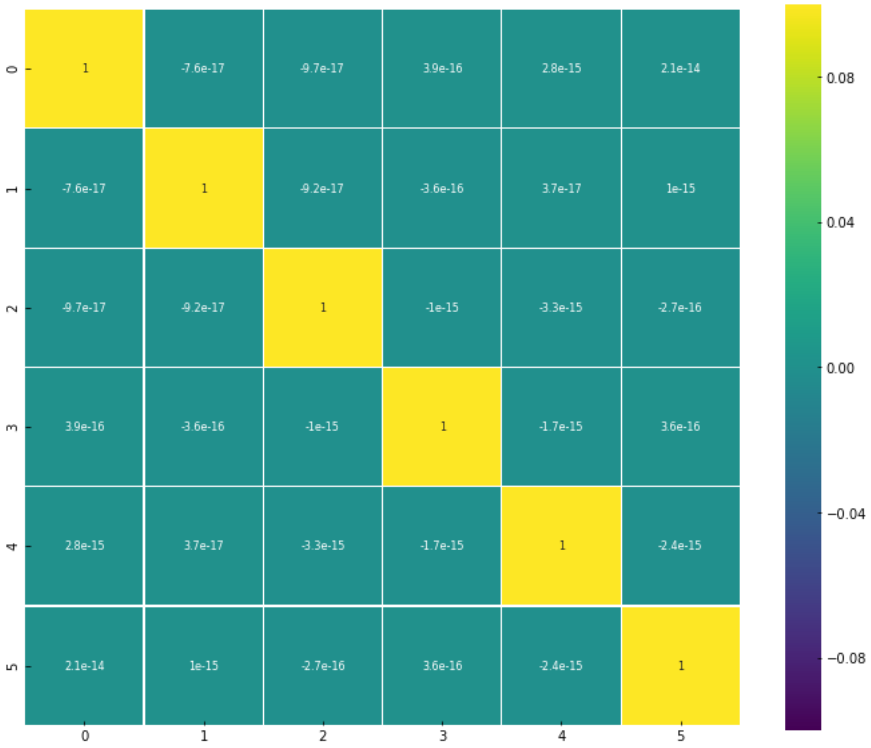


Ilustración 22 - Correlación de variables

Clustering

Luego de tener las dimensiones a utilizar, se procedió a determinar la cantidad de clústeres necesarios para agrupar las nuevas transacciones. En este caso, a través del uso de método del codo, se determinó que 4 clústeres sería una cantidad apropiada para este modelo.

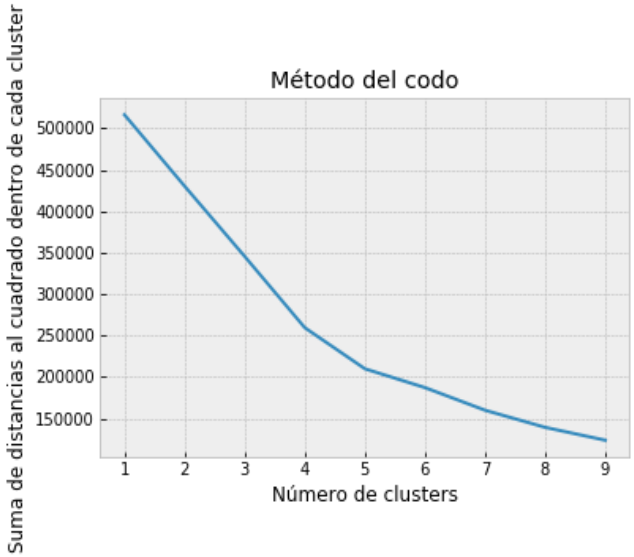


Ilustración 23 - Método del codo

Luego se procedió a identificar los valores distribuidos para cada modelo y elaborar una gráfica para cada uno. De estas gráficas, se puede identificar que tanto el clúster 1 como el 3 están libre de casos fraudulentos, por lo que todos los casos que caigan dentro de dichos clústeres son estimados como libres de fraude.

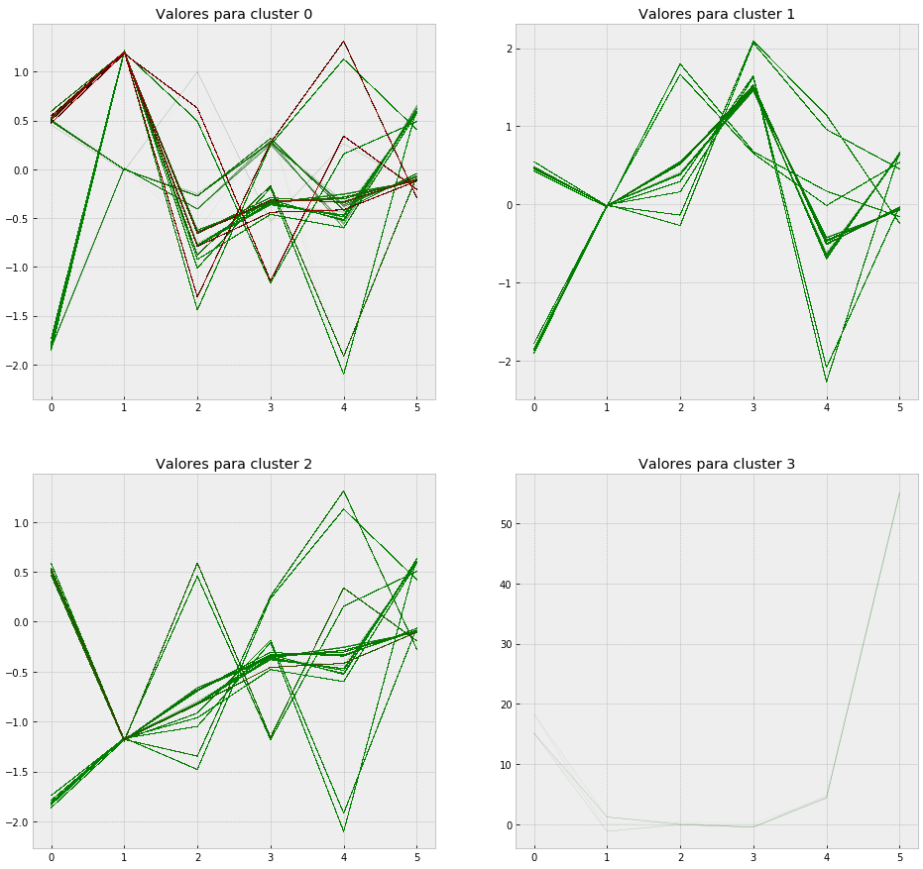
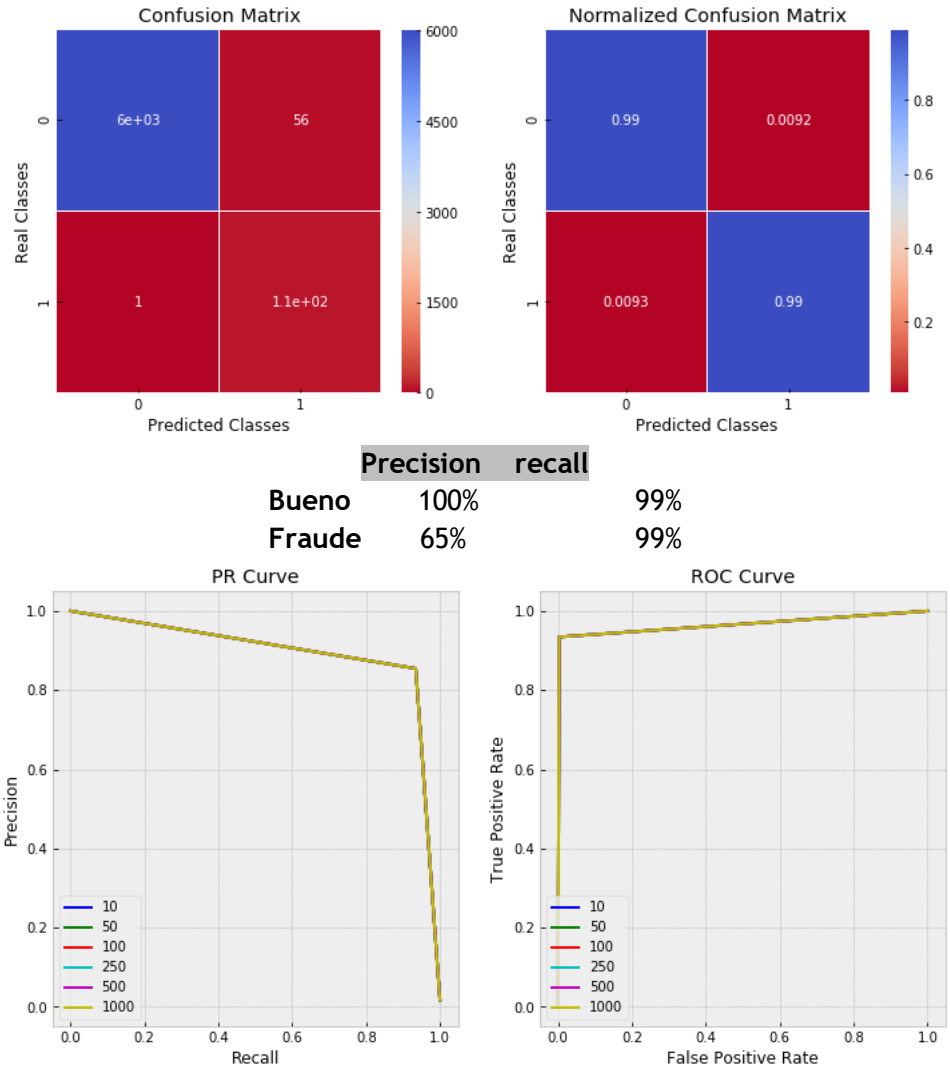


Ilustración 24 - Clúster fraude

Matriz de confusión

De los clústeres en los cuales aparecen transacciones fraudulentas, clúster 0 y clúster 2, se procedió a analizar su capacidad predictiva, tanto a través de la normalización de la matriz de confusión, como de la elaboración de una gráfica PR y una ROC.

Cluster 0

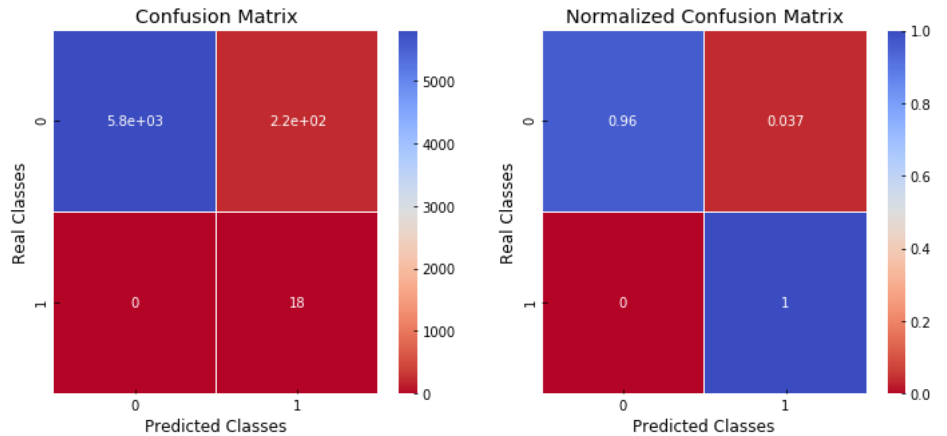


	Precision	recall
Bueno	100%	99%
Fraude	65%	99%

Ilustración 25 - Matriz de confusión clúster 0

Como se puede observar, el clúster 0 es muy preciso al identificar las transacciones buenas, pero para identificar las fraudulentas un poco menos, ya que toma algunos que no son fraudulentos como tales. Sin embargo, en ambos casos es bueno para identificar las que corresponden a cada caso.

Cluster 2



	Precision	Recall
Bueno	100%	96%
Fraude	8%	100%

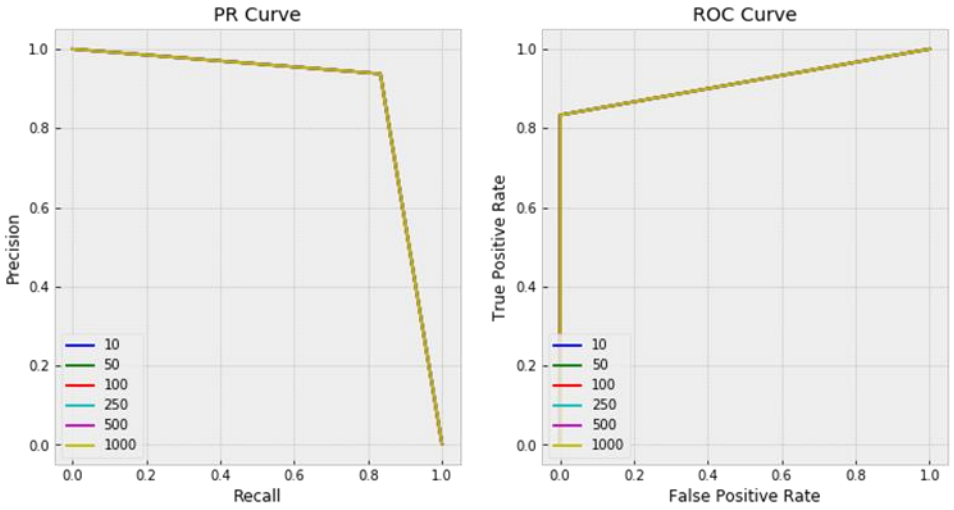


Ilustración 26 - Matriz de confusión clúster 2

El clúster 2, por su parte, es muy preciso al identificar las transacciones buenas, pero para identificar las fraudulentas es muy ineficiente. Sin embargo, en ambos casos es bueno para identificar las que corresponden a cada caso, aunque tenga mayor habilidad en el caso de los fraudulentos.

9.3 Modelo de segmentación de clientes

Unos de los aportes que se le dará a los casinos de INMABUIN, es la posibilidad de poder segmentar sus clientes y conocer las características de sus jugadas para poder así asignarle cortesías merecidas a cada cliente y poder segmentar su mercado.

Según Philip Kotler "la segmentación del mercado es la subdivisión del mercado en el sub-conjunto homogéneo de clientes, en cualquier subconjunto cabe la posibilidad de ser seleccionadas como objetivo de marketing con el que se alcanzó con la mezcla de marketing distinta". Es por esto que es tan importante segmentar el mercado ya que se pueden realizar campañas de marketing adecuadas según cada segmento, aumentar así la rentabilidad del negocio y captar mayor número de clientes.

Para nuestro modelo de segmentación se utilizaron datos existen de las bases de datos suministradas por nuestro cliente, teniendo en cuenta que estos datos no son datos exactos y que solo se capturan en ellos aproximadamente el 20% de las jugadas y clientes visitados. Se entrenaron los modelos con ellos para luego que se implante el IoT y las mesas inteligentes, donde sí se tendrán datos exactos y mejor registrados, entonces volver a entrenar los modelos y tener así un modelo más exacto.

Con miras al futuro cuando las bases de datos de las compañías estén más adecuadas a la realidad de esta, se pretende volver a implementar la segmentación para la reducción de falsos positivos en la matriz de confusión.

Obtención y limpieza de datos

Inicialmente se agregó un archivo csv con un total de 688.977 registros históricos y 21 columnas de las cuales se eliminaron 2 que no serían necesarias, luego se agregó una columna adicional donde se marcaba

si el cliente ganó o no, y finalmente se convirtieron las variables categóricas en numéricas para que pudieran ser entrenadas por el modelo. Al final se trabajó con una data frame de 688.977 x 28

Entrenamiento de modelo K MEANS

Con los datos ya tratados se procede a realizar el método del codo para la elección óptima de clúster a utilizar. Según la gráfica que se presenta a continuación podemos utilizar de 5 a 10 clúster. De estos se escogieron 5 clúster para trabajar con el modelo k means en Python.

Ilustración 27 - Método del codo segmentación clientes

Se obtuvieron un total de 5 clúster donde pudimos obtener 0-4 etiquetas, las cuales se clasificaron según su comportamiento y se agregaran a los cuadros de mando para que se puedan identificar clientes rentables. Estas etiquetas se representan en la siguiente gráfica:

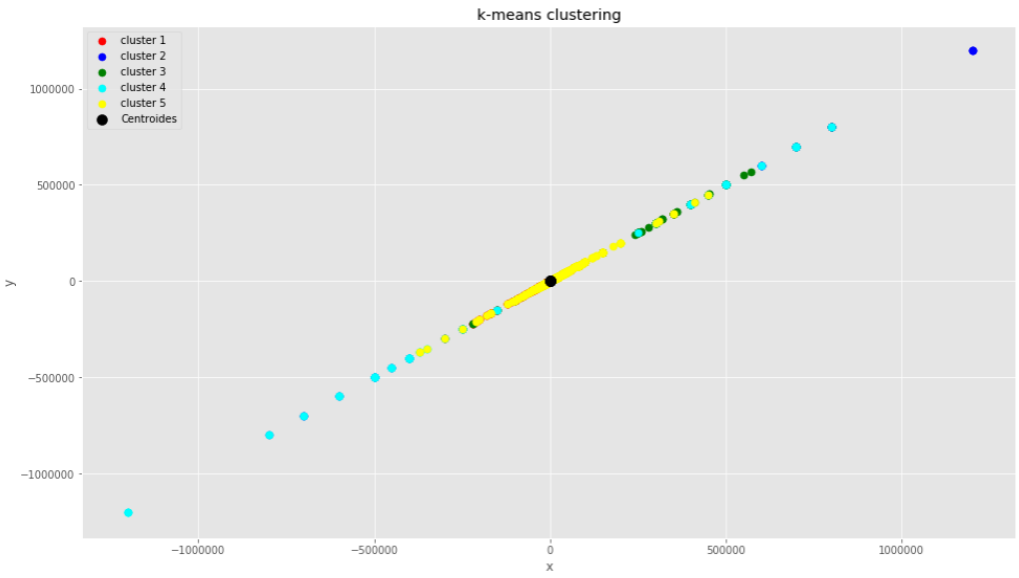


Ilustración 28 - Clientes segmentados

9.4 Cuadros de Mando

En este apartado, describimos las características de la tecnología para la visualización de datos y razones para elegir una herramienta para la confección del cuadro de mando integral, en base al cual se apoya la toma de decisiones como parte de la propuesta de valor para la mejora de recolección de datos en las mesas de casinos.

Estos cuadros de mando son una herramienta, que, gracias a la inteligencia de negocios, nos permite crear visualizaciones de datos, basadas en fuentes existentes de información histórica de la compañía, las cuales son transformadas en un modelado de información con formatos visuales que permiten a la gerencia un mejor análisis de la información y un apoyo a la toma de decisiones.

Al momento de seleccionar el sistema que sirva de lienzo para elaborar el tablero de control, se deben considerar varios factores, y es por eso por lo que hemos recurrido al Cuadro Mágico Gartner (el cual se publica periódicamente, por parte de un equipo de consultores de gran renombre expertos en análisis de mercados de nuevas tecnologías, donde se hace un ranking de los sistemas de inteligencia de negocios). También hay que considerar que el mercado de herramientas de inteligencia de negocios se actualiza mensualmente y por tal motivo hay varios sistemas que tienen mejoras frecuentemente.

A continuación, se muestra el ranking de las herramientas de visualización de datos:

Ilustración 29 - Cuadro Mágico de Gartner

Tal y como se puede observar, el cuadro no es más que un plano, compuesto por cuatro cuadrantes que describen el tipo de plataforma de inteligencia de negocios, identificados por el nombre de la empresa que los confeccionan, donde:

- Los líderes, son aplicaciones de BI (inteligencia de negocios) de empresas con alta habilidad para ejecutarse (con alta experiencia de usuario y fácil de usar) y a la par también permite una visión más holística de la información que se extrae de la data (completitud de visión). Se puede decir

que, ampliamente aceptados por el mercado, a la vez que están preparados para las necesidades del futuro.
- Los aspirantes, son sistemas de BI con alta habilidad para ejecutarse, pero su primordial falta es que no poseen una visión innovadora.
- Los jugadores de nichos son empresas con sistemas de BI con baja habilidad para ejecutarse, que también carecen de una visión innovadora (baja completitud de visión). Se puede decir que estos son creados para realizar un análisis en particular, dejando de lado un amplio espectro de herramientas de análisis y visualización que comprende el BI hoy en día.
- Los visionarios, al igual que los líderes, si poseen alta competitividad de visión, pero dado que generalmente son empresas disruptivas y de nueva entrada en el mercado, no tienen alta habilidad para ejecutarse.

Con respecto a los ejes del cuadrante mágico también podemos observar que en el eje horizontal representa la capacidad de visión de la plataforma de BI, mientras que el eje vertical a qué tan operable es dicho sistema. Por estas razones seleccionamos una herramienta perteneciente al cuadrante de los líderes. Y este sistema no es otro que Power BI Desktop de Microsoft.

La propia empresa define esta plataforma como un servicio de análisis empresarial que ofrece información relevante para el negocio, permitiendo tomar decisiones rápidas e informadas. Dentro de sus beneficios se listan:

- Capacidad de transformar los datos en imágenes impresionantes y compartirlos con colegas en cualquier dispositivo.
- Examinar y analizar visualmente los registros, en ficheros locales y/o en la nube, todo con la ventaja de ser representado en una sola vista.
- Colaborar y compartir paneles personalizados e informes interactivos.
- Escalar a través de la empresa con gobernanza y seguridad agregadas.
- Permitir la configuración del informe y sus visualizaciones dependiendo del equipo (tablets, ordenadores y celulares).

Datos técnicos relevantes herramienta Power BI Desktop versión Pro

Tipo de especificación	Detalle
Navegador Web	Internet Explorer 10 o mayor.
Sistema operativo	Windows 10, Windows 7, Windows 8, Windows 8.1, Windows Server 2008 R2, Windows Server 2012, Windows Server 2012 R2.
Tipo de sistema	32-bit (x86) y 64-bit (x64)
Tamaño del archivo	262.8 MB
Ventajas adicionales de la versión Pro	• Autoservicio y BI moderna en la nube. • Colaboración, publicación, intercambio y análisis ad hoc. • Completamente administrado por Microsoft.

Tabla 6 - Tabla especificaciones Power BI Desktop Pro

A simple vista se puede apreciar que las especificaciones para la instalación de la herramienta no son muy demandantes para los ordenadores personales actuales, aun así, presenta la ventaja de consultas a bases de datos con cantidades de información consideradas enormes con relativa celeridad y sin sobrecargar el consumo de memoria del pc.

Detallando la solución técnica del sistema de BI

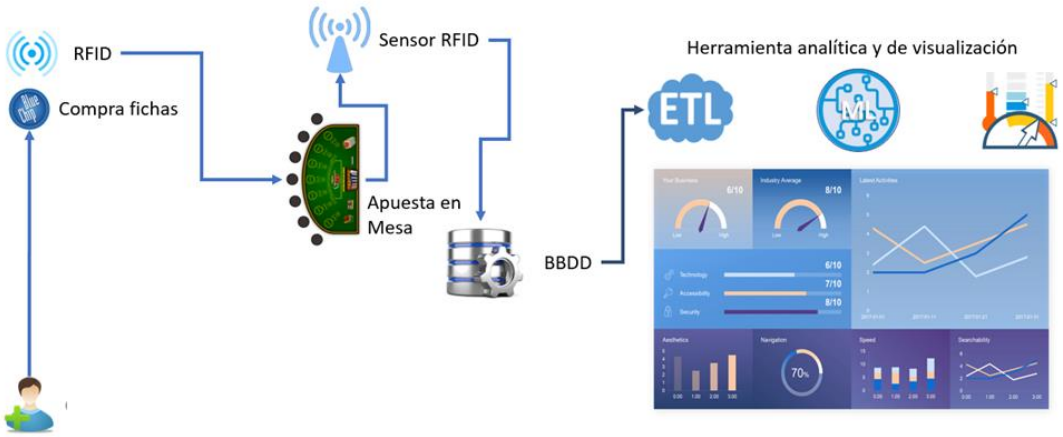


Ilustración 30 - Solución técnica de BI, fuente: propia

Al mirar la ilustración arriba presentada, se entiende que el proceso de obtención de los registros, o bien la data comienza con la entrada del cliente en el recinto de apuestas. Donde el mismo al momento de comprar fichas es registrado por un número de cliente único e irrepetible el cual también se le otorga en una tarjeta de identificable por tecnología RFID descrita anteriormente. Y donde también las fichas, o en su defecto, el valor en euros de dichas fichas compradas, son asignadas al Id del cliente en cuestión.

Al llegar a la mesa equipada con el sensor de RFID, se envía una señal para identificar el Id del cliente y sus transacciones de forma automática. Ahora bien, una vez los datos son recolectados y guardados en la nube, comienza el proceso que involucra a los algoritmos en base a los cuales se perfila el tipo de cliente, y si tiene patrones que se consideran fraudulentos (tal y como se había definido anteriormente).

Una vez concluido este paso, siguen los pasos de extracción, transformación y carga. El cual se divide en tres fases:

- **Fase 1: Base de datos en la nube del casino**

En esta fase se pretende obtener el monto ganado del casino por cliente, el score (o monto apostado por el cliente) y datos relevantes del casino y su funcionalidad.

- **Fase 2: Tabla resultante del análisis del algoritmo de segmentación por tipo de cliente**

De este se quiere el tipo de cliente en base al comportamiento.

- **Fase 3: Tabla resultante del análisis del algoritmo de patrones fraudulentos**

Identificar el Id del cliente que tiene patrones fuera de lo normal y que los hace fraudulentos y que, por tanto, el casino debe observar con detenimiento. También identificar la mesa donde juega y otros datos transaccionales.

Es, por tanto, que en la fase 1 se procede a la conexión de las tablas en la nube referentes a la base de datos del casino. En la siguiente ilustración se muestra el procedimiento de conexión de la tabla de

clientes. Por lo enorme de la base de datos que tiene más de 40 tablas relacionales, optamos por mostrar el proceso ETL, para esta sola tabla que se puede escalar a las otras de forma similar.

En la tabla arriba mostrada, se puede apreciar que los comandos están escritos en lenguaje M, que es propio de la herramienta Power BI Desktop. En el mismo se realizan los siguientes pasos:

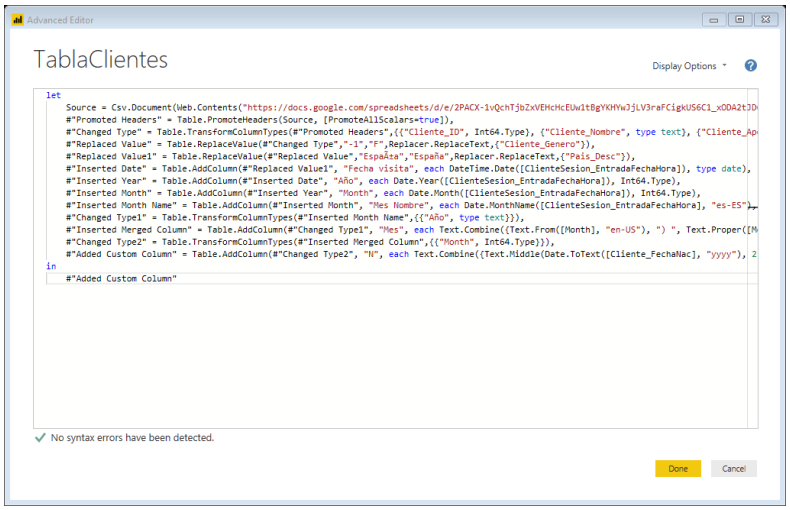


Ilustración 31 - Ingesta, transformación y carga de datos tabla de clientes

- Definición de la fuente de datos a ser consultada, en caso la tabla almacenada en la web.
- Se promueve la primera fila a los campos de la tabla
- Se cambian los tipos de datos según las necesidades Power BI permite inferir de forma automática los tipos de datos, que permite que el usuario no tenga que preocuparse por cambiar él mismo los tipos de datos.
- Y el resto de las transformaciones son para añadir datos de fecha (día, mes y año), para ser usados en los segmentadores.

En las fases 2 y 3, se plantea básicamente la misma solución, la cual es crear unas carpetas identificadas con los nombres Segmentación y fraude respectivamente, en las cuales cada vez que se haga un batch se añada un nuevo archivo de Excel, el cual al refrescar el cuadro de mando pueda incorporar los nuevos datos.

En lo que respecta al cálculo de los montos se crearon medidas, forma en la que se representan las métricas extraídas de la data. A continuación, se detallan las métricas propuestas y como se pueden obtener:

- El monto ganado o perdido por el casino, este se obtiene de la diferencia entre la suma del monto total que el cliente compra, menos la suma del monto total que posee el cliente. Sabemos que, si el monto es negativo, el valor absoluto son pérdidas para el casino y el caso contrario son ganancias.
- El porcentaje de ganancias (rentabilidad Casino), es la razón de la suma del monto total que posee el cliente, entre la suma del monto total que el cliente compra en fichas. Sabemos que, si el monto es negativo, el valor absoluto son pérdidas para el casino y el caso contrario son ganancias.
- La cantidad de clientes se obtiene del conteo de la cantidad total de Id del cliente.
- Las alertas de fraudes vendrían dadas por la cantidad de clientes con el estatus del fraude = 1.

Visualización de los datos

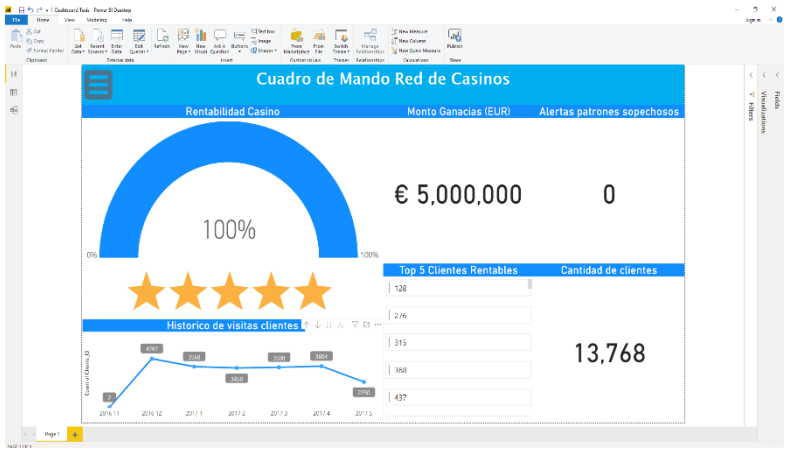


Ilustración 32 - Cuadro de mando casinos

La visualización propuesta en primera instancia posee un cuadro de mando donde podemos observar las métricas como la rentabilidad del casino, el monto ganado o perdido medido en euros, las alertas por posible fraude, la tendencia histórica de las visitas, la cantidad de clientes que han visitado y los cinco clientes que más ganancias generan para el casino.

Esta vista es preliminar, existen un universo de posibilidades, ya que, se puede escalar a cualquier otro casino solo agregando algunas visualizaciones y/o métricas. En lo que respecta a la segmentación de los datos, ha sido incluido un botón de menú en la parte superior del tablero de control y en el mismo se puede segmentar la data por diferentes factores.

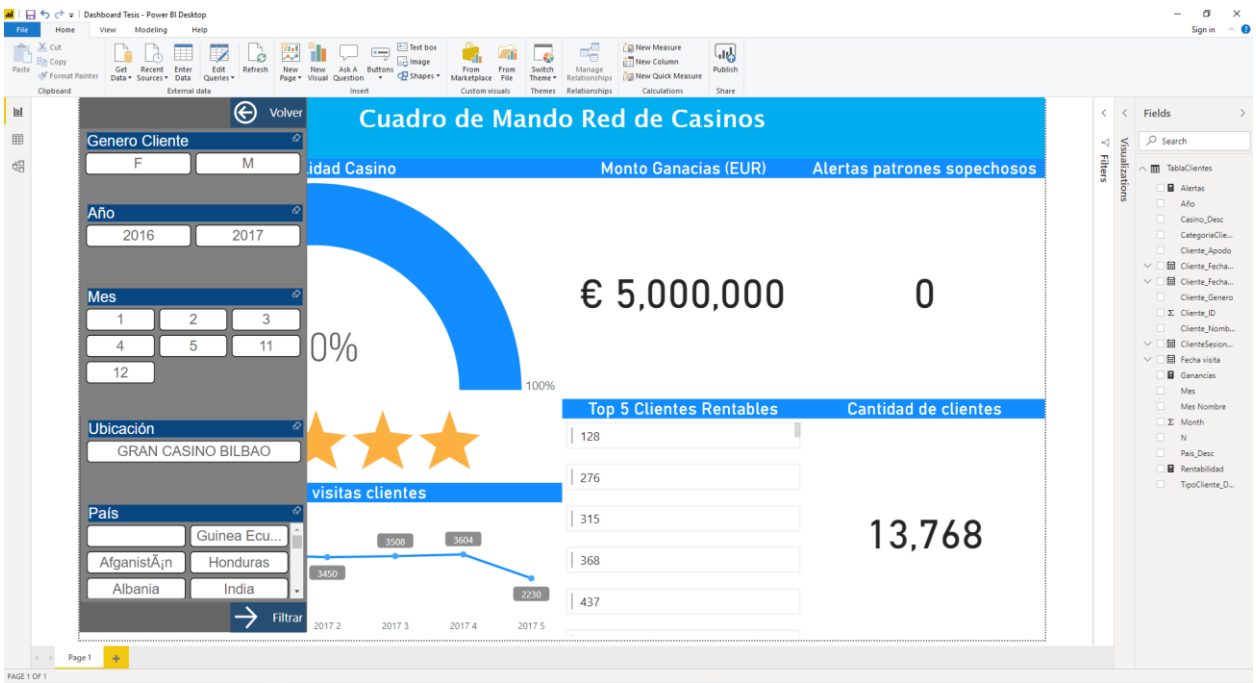


Ilustración 33 - Cuadro de mando casinos filtros

Visualización de los datos para jefes de sala

Se creó un segundo cuadro de mando donde se agregaron datos y consultas más específicas, orientadas a la toma de decisiones por parte de los directivos. Este cuadro de mando costa de 4 visualizaciones. La primera de ellas con datos provenientes de vistas de base de datos BDMarketing, donde los directivos podrán visualizar el histórico de ganancias en las fechas que se implementaron las promociones, lo cual servirá como apoyo al departamento de marketing para implantar nuevas promociones en fechas efectivas

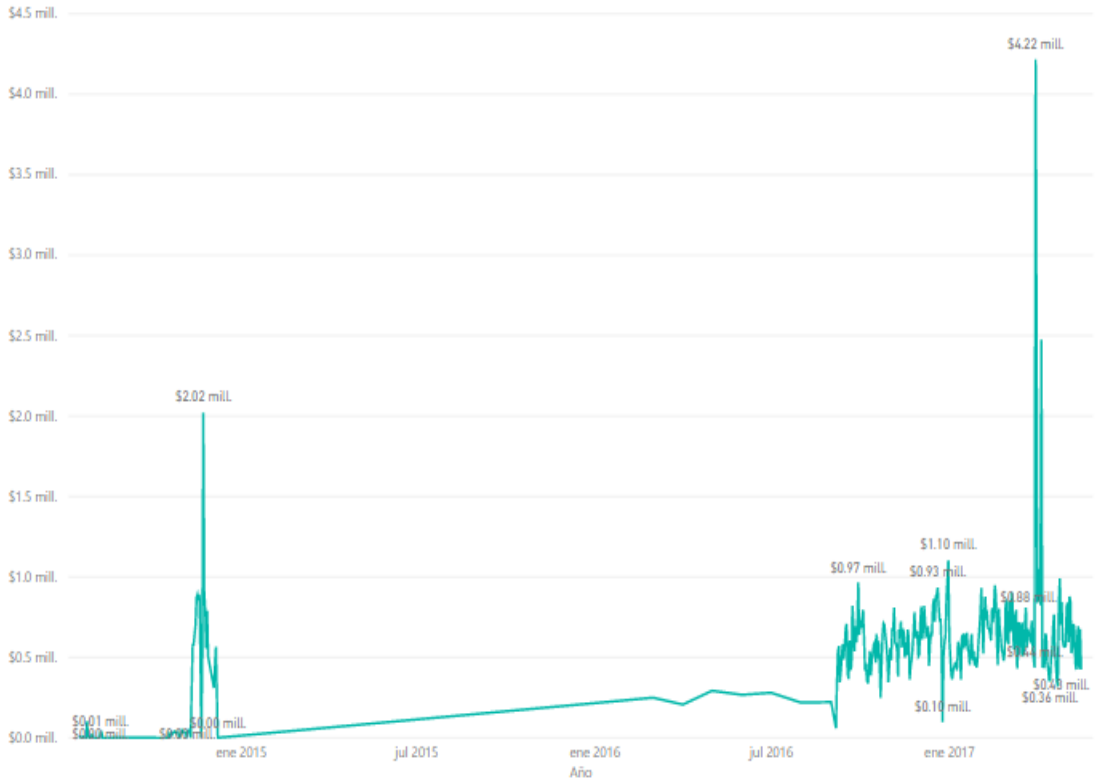
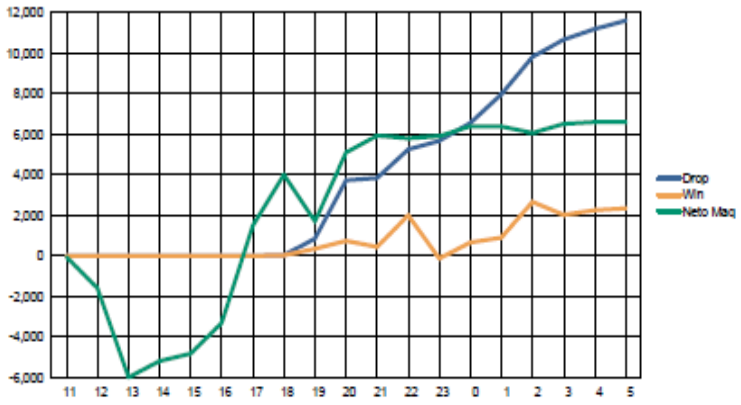


Ilustración 34 - Visualización de Promociones

La segunda visualización está basada en un resumen del día de trabajo, donde la gerencia podrá identificar la eficiencia de las cortesías y las ganancias que se obtuvieron. Aquí se nos muestra un top 10 de los clientes que más beneficios le dejaron al casino en la jornada de jugadas y cuanto se gastó con ellos; además, podemos visualizar un flujo de jugadas por cada hora del día.



SALAS	JUEGOS	DROP	RESULTADO	GASTOS
GENERAL	BJ, PG, PK, RA	945,00	-945,00	0,00
GENERAL	RA	1.820,00	-720,00	16,00
GENERAL	BJ	620,00	-620,00	0,00
GENERAL	BJ	610,00	-610,00	3,50
GENERAL	RA	540,00	-540,00	0,00
GENERAL	RA	390,00	-390,00	0,00
GENERAL	BJ	300,00	-300,00	0,00
GENERAL	RA	300,00	-300,00	0,00
GENERAL	BJ	200,00	-200,00	0,00
GENERAL	RA	200,00	-200,00	0,00

SCORE SALA	
HORA :	5:00:00 am
DROP MESAS:	11.615,00
WIN MESAS:	2.342,25
HOLD :	20 %

DATOS CIERRE	
VISITAS:	125
CORTESÍAS:	180,50

Ilustración 35 - Visualización Resumen diario de Ganancias/Pérdidas

Dentro de las visualizaciones corporativas se crearon dos resúmenes, uno semanal y otro mensual. El resumen semanal calcula la cantidad de clientes que visitan los casinos por día de la semana, lo que ayudará a poder manejar el personal en las mesas y a aplicar mejor las campañas publicitarias para captar más clientes los días donde el flujo es bajo.

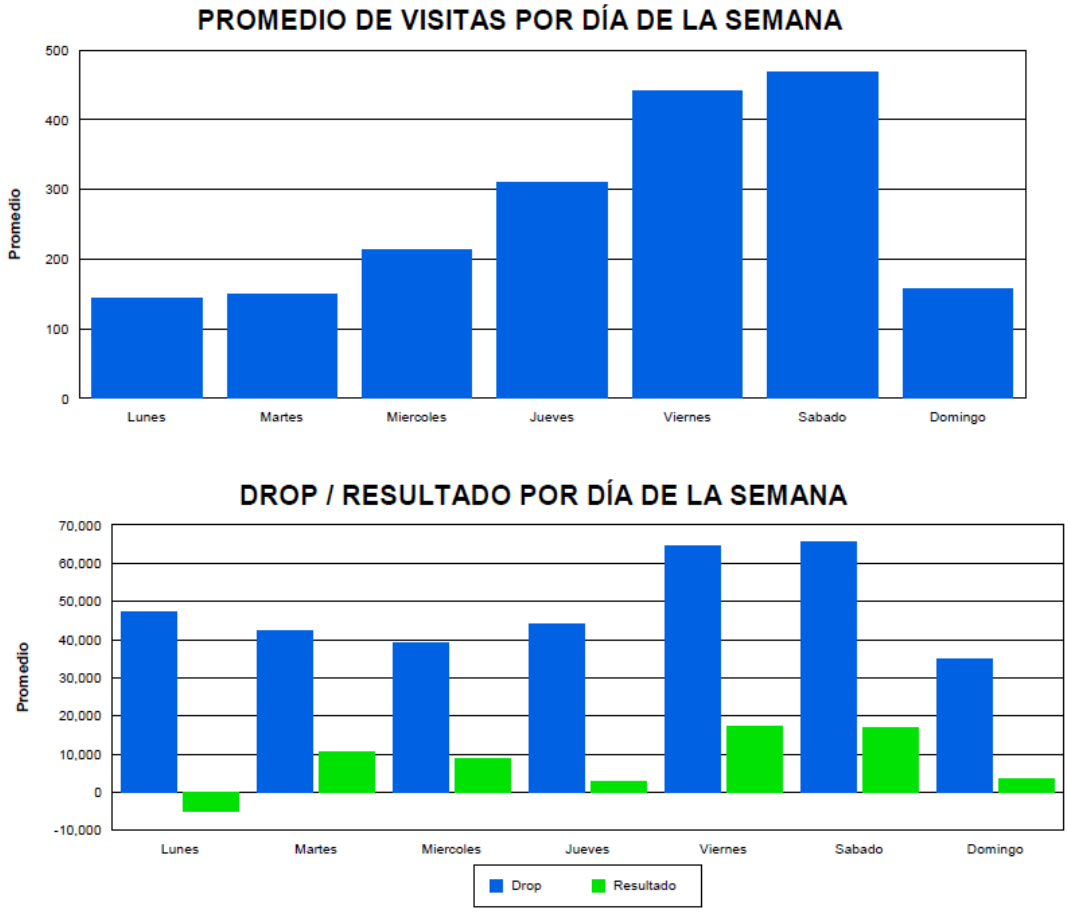


Ilustración 36 - Visualización Resumen Semanal

Por último, creamos una visualización que hace un resumen mensual con el total de apuestas (Drop) vs la ganancia generada (Resultado) esperanzadas cada día del mes.

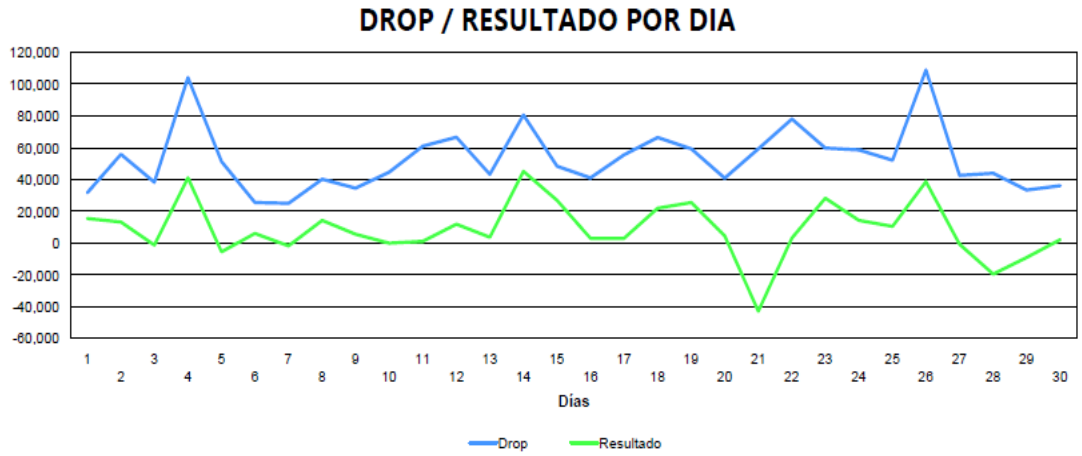


Ilustración 37 - Visualización Resumen Mensual

10. OPTIMIZACIÓN DE LOS RESULTADOS

Después de un proceso de validación e investigación, se llegó a la conclusión de que la situación actual de los casinos pertenecientes a la cartera de clientes de INMABUIN no es el escenario más factible. Los datos de los clientes, tanto sus jugadas como el recorrido por las instalaciones, se registran de manera manual y no se lleva registro exacto de todos los datos necesarios para calcular la rentabilidad de estos, no se cuenta con un proceso de identificación de clientes que sea eficiente y a la vez la alta gerencia no cuenta con herramientas necesarias para realizar un buen proceso de toma de decisiones.

Con la implementación de las soluciones informáticas de BI e IoT que se le ofrece a la empresa INMABUIN, los casinos de su organización se verán óptimamente beneficiados, aumentando considerablemente sus ingresos, dando un mejor servicio al cliente, reduciendo costes de producción y de personal; apoyando, además, a un mejor proceso de captura de datos, evitando la obtención de información erróneo y aumentando el registro de procesos a casi un 100%.

10.1 Detalle de Beneficios

10.1.1 Tangibles

- Gracias a la clasificación de clientes en base a patrones de jugada obtenidos del modelo de segmentación, los casinos podrán incrementar la fidelidad de estos y asignar las cortesías adecuadas a cada uno. Se incrementará en más de un 80% la fidelidad de los clientes al negocio, lo que provocará el aumentando el flujo de jugadas, ya que el cliente se sentirá a gusto en las instalaciones y permanecerá más tiempo en el establecimiento.
- Con la implementación de la tecnología IoT se reducirán los errores en captura de información. Además, los datos registrados serán más exactos y aumentará el registro de estos de un 20% a un 95%, ayudando a la empresa a tener información más detallada de sus clientes.
- Se reducirán los fraudes con la herramienta de predicción de fraudes, lo que ayudará al departamento de seguridad con alertas de comportamientos en tiempo real, evitando pérdidas al negocio.
- Reducción de personal para registro de datos manuales.
- Evitar pérdidas detectando fraudes en tiempo real.

10.1.2 Intangibles

- Como ya mencionamos anteriormente, al implantar el proyecto IoT se podrá tener un repositorio de datos con información más real, datos más limpios, ordenados y con mayor cantidad de registros, lo que permitirá disponer de una mejor fuente de información.
- Con toda la información disponible en las bases de datos se crearán cuadros de mando donde la alta gerencia y directivos tendrán una mejor herramienta para soporte a la toma de decisiones en tiempo real, con información actualizada, precisa, íntegra y disponible en cualquier momento.

10.1.3 Estratégicos

- Al implementar las soluciones BI, el departamento de marketing podrá identificar las estrategias adecuadas para constituir campañas publicitarias y aumentar la captación de clientes.
- Se podrán conocer todas las características que describen a los clientes rentables y asignar cortesías adecuadas a estos.

- Herramientas para la toma de decisiones basadas en hechos históricos de la compañía.

10.2 Análisis Financiero

Luego de detallados los beneficios que aporta la implementación de soluciones BI, procedemos a realizar el análisis financiero, para así demostrar con números y cifras reales los beneficios económicos y tangibles que podrá llegar a tener la empresa con miras al futuro.

ANÁLISIS RENTABILIDAD DEL PROYECTO						
	AÑO 0	AÑO 1	AÑO 2	AÑO 3	AÑO 4	AÑO 5
INVERSIÓN						
Tecnología IOT por mesa de	13,959.00 €					
Recursos de personal	27,000.00 €					
Tarjetas y fichas RFID	182.00 €					
Instalación en Mesas Adicionales	7,959.00 €					
TOTAL INVERSIONES	49,100.00 €	- €	- €	- €	- €	- €
INGRESOS/BENEFICIOS						
Servicios y soporte INABUIM		22,000.00 €	22,000.00 €	22,000.00 €	22,000.00 €	22,000.00 €
TOTAL INGRESOS/BENEFICIOS	- €	22,000.00 €	22,000.00 €	22,000.00 €	22,000.00 €	22,000.00 €
GASTOS						
Servicios Almacenamiento en la Nube 1 TB (1.000 GB)		4,800.00 €	4,800.00 €	4,800.00 €	4,800.00 €	4,800.00 €
Licencia de Power BI Pro		107.89 €	107.89 €	107.89 €	107.89 €	107.89 €
TOTAL GASTOS	- €	4,907.89 €	4,907.89 €	4,907.89 €	4,907.89 €	4,907.89 €
FLUJO DE CAJA OPERATIVO	- 49,100.00 €	17,092.11 €	17,092.11 €	17,092.11 €	17,092.11 €	17,092.11 €
VALOR ACTUAL	- 49,100.00 €	15,538.28 €	14,125.71 €	14,125.71 €	14,125.71 €	12,841.55 €
ACUMULADO	- 49,100.00 €	- 33,561.72 €	- 19,436.01 €	- 5,310.30 €	8,815.41 €	21,656.96 €
VAN:	€ 106,395.40					
TIR:	21.85%					
PAYBACK	3 años, -5 meses					

Tabla 7 - Análisis económico

Como podemos notar en la figura anterior, la mayor inversión se presenta en el año 0, momento de implantación del proyecto. Cabe destacar que la distribución de esta inversión radica en compra de hardware y tecnologías para la implementación de IoT y el recurso humano para su instalación. En los años siguientes, esta inversión no será necesaria ya que las tarjetas RFID y el hardware tienen una larga vida útil, que va aproximadamente de 5 a 7 años.

En cuanto al software, se cobrará a los usuarios por la implementación de modelos y cuadros de mando una iguala en soportes y servicios con valor de 22.000 euros por cada año que utilicen nuestras tecnologías, lo que implicará la principal fuente de ingreso. Los gastos adicionales que implica la ejecución de software radican en los servicios de almacenamiento de la nube para que los datos estén actualizados en tiempo real y la utilización de licencias de Power BI Pro, lo que permitirá a los usuarios la exportación de los cuadros de mando a los servicios web.

10.2.1 Flujo Neto de Efectivo (FNE)

El flujo neto de efectivo no es más que el monto de efectivo que se va a generar en nuestra empresa mediante la venta de nuestra tecnología a los clientes. Según nuestro análisis financiero en el año 0 nuestro FNE será de -49.100 euro, esto es debido a que el costo de inversión se realizará a los inicios de la implementación y el retorno de la inversión no se verá reflejado hasta culminar el primer año, en el cual se obtendrá un flujo neto de efectivo de 17,092.11 y se mantendrá estable con el mismo monto hasta el año 5

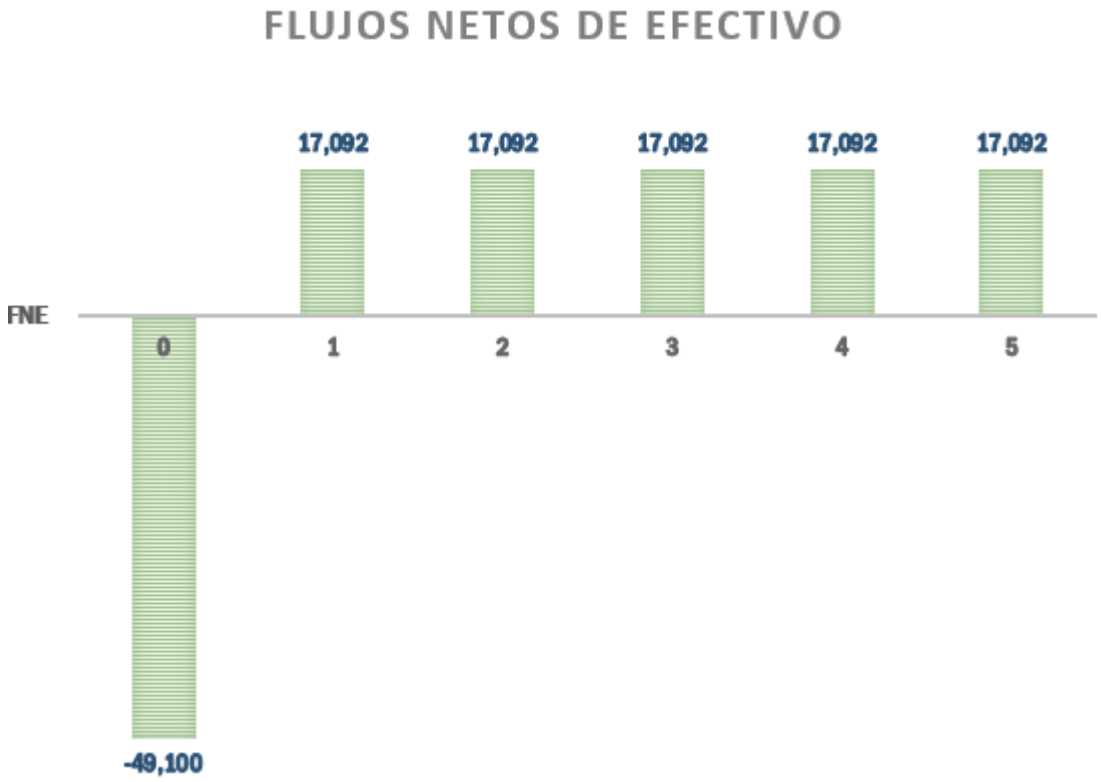


Ilustración 38 - Flujo Neto de Efectivo (FNE)

10.2.2 Valor Actual Neto (VAN)

En los proyectos de tecnología existe un alto riesgo en pérdida cuando se habla de rentabilidad, es por esto que a mayor riesgo mayor rentabilidad esperada. Mediante el cálculo del VAN podemos calcular en base a un interés definido (en este caso el 15%), el cual será la ganancia que se obtendrá al momento de implementar un proyecto y realizar una inversión.

En nuestro análisis económico-financiero, nuestro VAN es mayor que 0, con un total de 106.395.40 euros, lo que hace que el proyecto sea rentable y satisfaga la tasa esperada de un 15%, es decir, que se obtendrán ganancias superiores a las esperadas.

10.2.3 Tasa Interna de Retorno (TIR)

El TIR nos representa el porcentaje, ya sea de benéfico o pérdida, que vamos a obtener al realizar una inversión. En un proyecto de tecnología el TIR esperado debe ser mayor al 14%. En nuestro caso el TIR obtenido al realizar el análisis financiero es de 21.85%, lo que confirma, al igual que la métrica de VAN, que las soluciones tecnológicas que proponemos serán más que viables y proporcionarán los beneficos esperados a nuestra empresa.

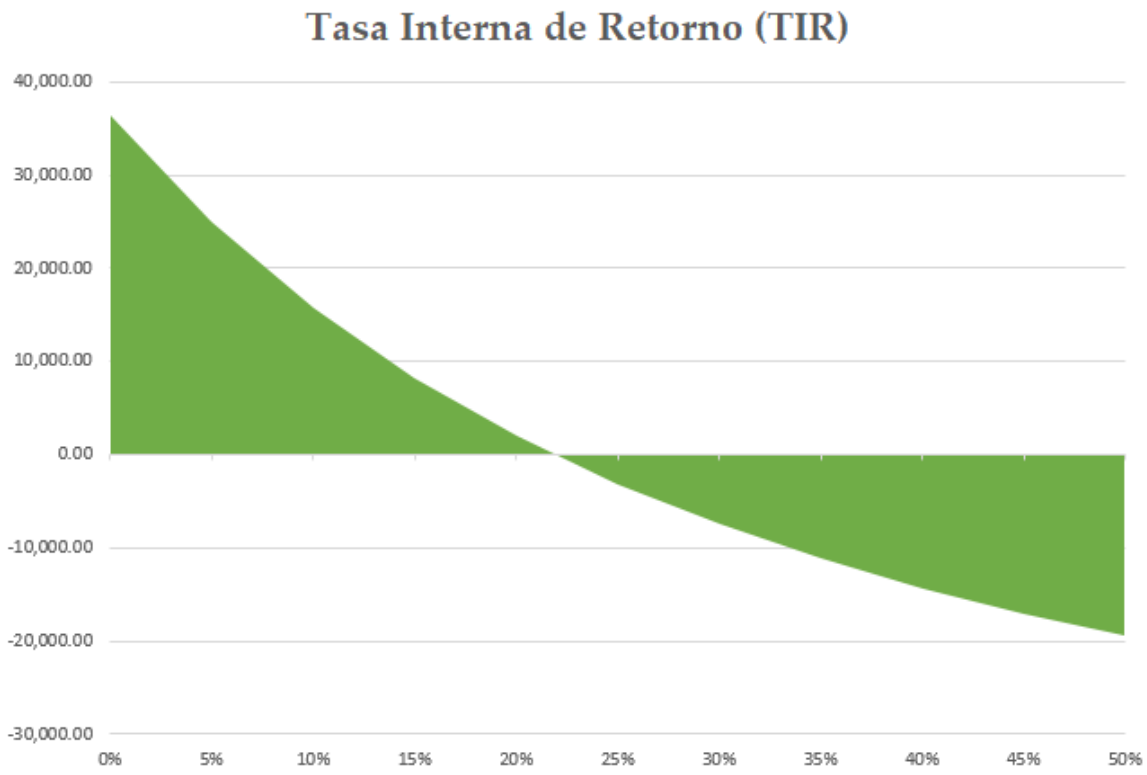


Ilustración 39 - Tasa Interna de Retorno (TIR)

10.2.4 Pay Back

El Pay Back es la métrica que nos indica qué tiempo se tardará en recuperar la inversión inicial, recordando que la nuestra fue de 49.100, la cual se va a recuperar mediante el flujo de caja en un total de 3 años y 5 meses.

11. BIBLIOGRAFÍA Y RECURSOS

Ley Orgánica 3/2018, de 5 de diciembre, de Protección de Datos Personales y garantía de los derechos digitales. De la fuente https://www.boe.es/buscar/pdf/2018/BOE-A-2018-16673-consolidado.pdf

Hernández Sampieri, R., Fernández Collado, C., & Baptista Lucio, M. d. (2014). Metodología de la Investigación. México D.F.: McGRAW-HILL / INTERAMERICANA EDITORES, S.A. DE C.V.

Fundamentos de marketing Philip kotler
https://issuu.com/issuesoto/docs/fundamentos_de_marketing_-_philip_k

Cuadro Mágico de Gartner: plataformas de análisis e inteligencia de negocios, febrero 2019,
https://softwareyhardware.com/software/cuadrante-magico-de-gartner-2019-analitica-e-inteligencia-de-negocio-bi/

Especificaciones Power BI PRO
https://powerbi.microsoft.com/en-us/

12. ANEXOS

Anexo 1 - Entrevistas

1. Entrevista #01 - Ana Medrano (Gerente)
2. Entrevista #02 - Pedro Polanco (Encargado de Seguridad)
3. Entrevista #03 - Marleni Acosta y Sandra Santana (Croupier)

Entrevista #01	
Nombre Entrevistado:	Ana Medrano
Cargo:	Gerente

1.1 ¿Se conoce la cantidad de clientes que visitan los casinos?
La verdad es difícil conocer la cantidad de personas que visitan el casino, ya que, aunque anotemos cuantos entran. Si uno de los clientes sale a tomar una llamada, no hay forma de saber si quien entra nueva vez es un cliente nuevo o bien un cliente que vuelve a entrar a las instalaciones.
1.2 ¿Es posible segmentar el tipo de clientes en base a su comportamiento dentro de los casinos?
Técnicamente no se puede. Aun no es muy difícil, el poder saber este tipo de cosas. Los clientes dan una identidad falsa, nombres números cedula falsa, para que no se tengan informaciones de ellos.
1.3 Podría señalar que tan de acuerdo del 1 al 5 donde 1 representa poco de acuerdo y 5 muy de acuerdo. ¿Es efectivo el proceso de identificación de clientes? 1
1.4 ¿Cómo es el método de identificación de clientes?
Se lleva un registro manual, donde se coloca una característica visual del cliente, ejemplo: Cliente del tshirt rojo con sombrero.
1.5 ¿Cómo se anotan las jugadas, ganancias y/o pérdidas de los clientes de los casinos?
Se anotan por mesas, de forma manual con una hoja de control.
2.1 ¿Es posible determinar el patrón de apuestas por cliente con la metodología actual?
No.
3.1 ¿Se realizan campañas para atraer clientes de forma periódica?
SI. Solo de forma interna, es decir, que se publican y promocionan dentro del casino.
3.2 ¿Qué criterios se usan determinar cuándo hacer campañas?
Días feriados. Por temporadas (Madres, padres, etc.).
3.3 ¿Qué criterios se usan determinar cuándo hacer campañas?
No tenemos actualmente como medir el desempeño de las campañas.
4.1 ¿Es posible determinar el nivel de experiencia del cliente?
Si, son usadas actualmente para quejas y sugerencias. Se registran manuales.
6.1 ¿Se conoce el ROI por mesa? En caso de ser afirmativo ¿Cómo se calcula?
Si. Se lleva un control de lo que ha entrado en la mesa, y se realiza una sumatoria de todas las ganancias y pérdidas por mesa.
6.2 ¿Se conoce el ROI por cliente? En caso de ser afirmativo ¿Cómo se calcula?
No se conoce el ROI por cliente.

6.3 ¿Cuánto tiempo de análisis conlleva analizar la rentabilidad por cliente y por mesa?
No se toma mucho tiempo, es sencillo.
7.1 ¿Poseen algún sistema de visualización de fraudes, en base al comportamiento sospechoso de juego de los clientes?
No. Se realiza de forma manual.
7.2 En caso de ser cierto ¿Cómo funciona dicho sistema?
Los gestos corporales del cliente, es lo que se utiliza para detectar fraudes. A los clientes nuevos se les da seguimiento más de cerca. Se rotan las personas que atienden cada mesa.
13.1 ¿Qué tanto ayuda la asignación de cortesías a los clientes para aumentar la rentabilidad del negocio?
Desconocen esta información.
13.2 ¿Cuáles puntos se toman en cuenta para conceder cortesías a cada cliente?
Dependiendo el nivel de apuesta de los clientes. Si sus apuestas son muy elevadas, la cortesía también es elevada. Solo se les aplica cortesía a los clientes conocidos.
13.3 ¿Se lleva un control o registro de estas cortesías?
Si, se lleva un control para poder llevarlo a los gastos operativos.
13.4 De ser así ¿Cuáles datos se almacena sobre estas?
Se almacena el producto y el monto.

Entrevista #02

Nombre Entrevistado:	Pedro Polanco
Cargo:	Encargado Seguridad

1.1 ¿Se conoce la cantidad de clientes que visitan los casinos?
Tienen un número estimado. Dado que no pueden determinar la cantidad de clientes viene.
1.2 ¿Es posible segmentar el tipo de clientes en base a su comportamiento dentro de los casinos?
Si no se puede determinar el dato anterior, menos se puede determinar este.
1.3 Podría señalar que tan de acuerdo del 1 al 5 donde 1 representa poco de acuerdo y 5 muy de acuerdo. ¿Es efectivo el proceso de identificación de clientes?
1
1.4 ¿Cómo es el método de identificación de clientes?
El cliente se acerca al casino a partir de tarjeta de fidelización. Los clientes importantes no se identifican usualmente.
1.5 ¿Cómo se anotan las jugadas, ganancias y/o pérdidas de los clientes de los casinos?
No es posible determinar el nivel de ganancias y/o pérdidas.

2.1 ¿Es posible determinar el patrón de apuestas por cliente con la metodología actual?
No.
4.1 ¿Es posible determinar el nivel de experiencia del cliente?
Realizan encuestas de satisfacción.
4.2 ¿Crees que sería útil para la mejora continua, tanto de empleados como del negocio, implementar alguna técnica de encuestas al cliente para saber su experiencia en el casino?
Claro sería de gran utilidad.
5.1 ¿Es cierto que los datos se registran de forma manual? En caso de ser Sí ¿Cuáles datos se registran?
Si. En una hoja impresa timbrada. Primero por los croupier, luego consolidado por encargados de mesas y más luego por encargados del área de juegos. En base a los datos que se registran en las mesas de juego se registra cuanto se compra de fichas y cuanto se ganan los apostadores.
7.1 ¿Poseen algún sistema de visualización de fraudes, en base al comportamiento sospechoso de juego de los clientes?
Si, manual. No analítica.
7.2 En caso de ser cierto ¿Cómo funciona dicho sistema?
Los gestos corporales del cliente.
7.3 ¿El sistema de seguridad posee dashboard con la capacidad de generar alertas ante comportamientos sospechosos?
No tienen dashboard que muestre alertas.
7.4 Para el casino se podría señalar que tan de acuerdo del 1 al 5 donde 1 representa poco de acuerdo y 5 muy de acuerdo. ¿Qué tan fácil de usar dicha herramienta de seguridad?
4
7.5 ¿Cuáles datos crees que se pueden tomar en cuenta para pronosticar algún fraude?
Todo mediante observación, no es tan difícil ya que cuentan con entrenamiento sobre comportamiento no hablado.
10.1 ¿El casino posee actualmente un cuadro de mando con información sobre jugadas sospechosas?
No tienen
10.3 ¿Cómo se mide actualmente?
Todo mediante observación.

Entrevista #03	
Nombre Entrevistado:	Marleni Acosta y Sandra Santana
Cargo:	Croupier

1.1 ¿Se conoce la cantidad de clientes que visitan los casinos?
No tenemos conocimiento al respecto.
1.2 ¿Es posible segmentar el tipo de clientes en base a su comportamiento dentro de los casinos?
Ese dato esta mas allá de nuestro conocimiento
5.1 ¿Es cierto que los datos se registran de forma manual? En caso de ser Sí ¿Cuáles datos se registran?
Solo anotan registrando las ventas a través de una Tablet. Tomando en consideración el valor de las fichas. Por ejemplo, 3 fichas de 500 pesos.
5.2 Para el casino se podría señalar que tan de acuerdo del 1 al 5 donde 1 representa poco de acuerdo y 5 muy de acuerdo. ¿Qué tan fácil es registrar los datos de los clientes?
3

Anexo 2 - Código fuente de Modelos Python

1. Modelo #01 – Fraude
2. Modelo #02 – Segmentación de Clientes

Modelo #1 - Fraude

Analítica de operaciones (fraude)

1 - Selección de variables y reducción dimensional

In [248]:

```
import pandas as pd
import seaborn as sns
import matplotlib.pyplot as plt
import matplotlib.pyplot as plt1
import matplotlib.pyplot as plt2
import numpy as np
from sklearn.model_selection import train_test_split
from sklearn.linear_model import SGDClassifier
from sklearn.preprocessing import StandardScaler
from sklearn import preprocessing

%matplotlib inline

RANDOM_SEED = 42
n_dim = 15
plt.style.use('bmh')

df = pd.read_csv('./ClientesFraude5.csv')
obj_df1 = df.select_dtypes(include=['object']).copy()
obj_df =
obj_df1.drop('ScoreCliente_Sesion',1).drop('ScoreCliente_HoraEntrada',
1).drop('ScoreCliente_HoraSalida', 1).drop('ScoreClienteDet_FechaHora',
1).drop('CategoriaCliente_Desc', 1)
df_encoding = pd.get_dummies(obj_df,
columns=['Cliente_Genero','Pais_Desc','TipoCliente_Desc','TipoMoneda_Desc','U
suario_DescID'])
num_df = df.select_dtypes(include=['int64','float64']).copy()
df_num = pd.concat([df_encoding,num_df], axis=1).drop('Cliente_No',1)
df_features=df_num
columns_to_norm = ['ScoreClienteDetMov_Monto']
min_max_scaler = preprocessing.MinMaxScaler()
df_features[columns_to_norm]=min_max_scaler.fit_transform(df_features[columns
_to_norm])
tt = df_features.describe().transpose()
df_features.head()
```

Out[248]:

	Cliente_Genero_F	Cliente_Genero_M	Pais_Desc_España	Pais_Desc_República Dominicana	TipoCliente_Desc_GENERAL	TipoCliente_Desc_ORIENTAL	TipoMoned c_DROP
0	0	1	0	1	0	1	
1	0	1	0	1	0	1	

2	0	1	0	1	0	1
3	0	1	0	1	0	1
4	0	1	0	1	0	1

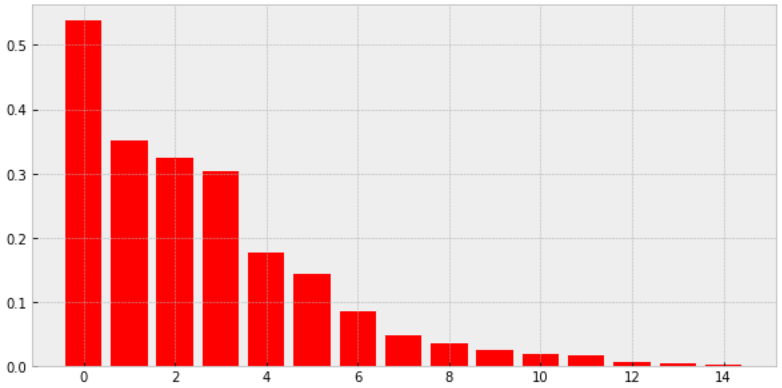
5 rows × 34 columns

2 - Aplicación de la técnica de PCA

In [249]:

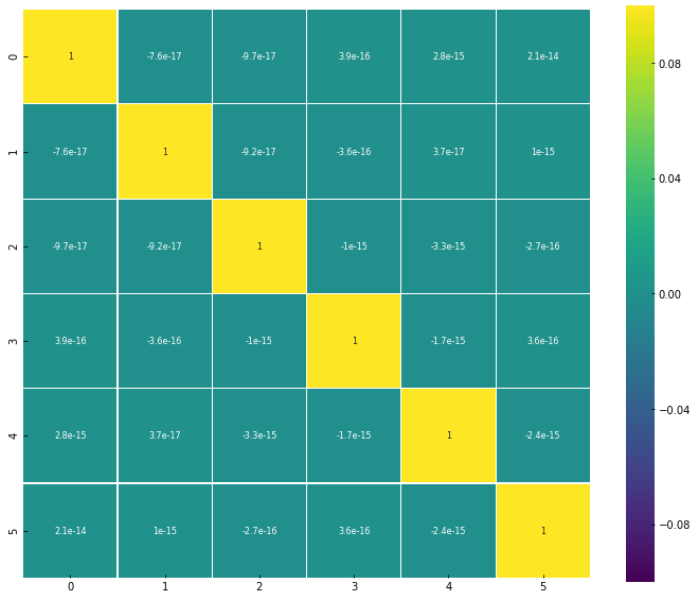
```
from sklearn.decomposition import PCA as sklearnPCA

y=df_features.Fraude
x=df_features.drop('Fraude',axis=1)
sklearn_pca = sklearnPCA(n_components=n_dim, whiten=True)
sklearn_pca.fit(x)
features_pca = pd.DataFrame(data = sklearn_pca.transform(x))
plt.figure(figsize=(10, 5))
rects1 = plt.bar(np.arange(n_dim),sklearn_pca.explained_variance_, color='r')
print(sklearn_pca.explained_variance_)
```

```
[0.53745261 0.35217635 0.32450603 0.30450015 0.1761588  0.14322633
 0.0857638  0.04761733 0.03518407 0.02631894 0.02019986 0.01789117
 0.00759065 0.00535415 0.00179051]
```



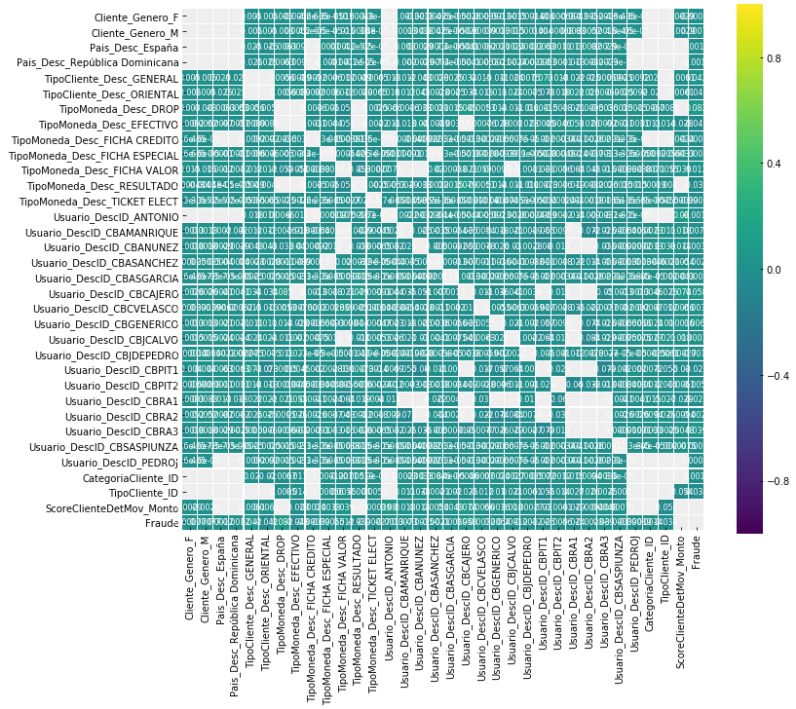
In [250]:

```
features_pca = features_pca.iloc[:,0:6]
features_pca2 = features_pca.values[:,0:6]

corr_pca = features_pca.corr()
plt.figure(figsize=(12, 10))
sns.heatmap(corr_pca, cmap='viridis', vmax=0.1, vmin=-0.1, linewidths=0.1,
annot=True, annot_kws={"size": 8}, square=True);
```



In [251]:

```
corr_base = df_features.corr()
plt.figure(figsize=(12, 10))
sns.heatmap(corr_base[(corr_base <= 0.1) & (corr_base >= -0.1)],
cmap='viridis', vmax=1.0, vmin=-1.0, linewidths=0.1, annot=True,
annot_kws={"size": 8}, square=True);
```

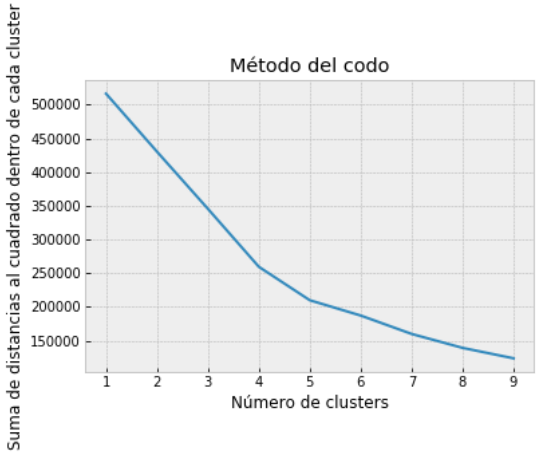


3 - Clustering

In [252]:

```
from mpl_toolkits.mplot3d import Axes3D
from sklearn.metrics import silhouette_score
from sklearn.cluster import KMeans

res = []
N_max=10
N_min=1
```

```
for i in range(N_min, N_max):
    kmeans = KMeans(n_clusters = i, init = 'k-means++', max_iter = 300,
n_init = 10, random_state = 0)
    kmeans.fit(features_pca2)
    res.append(kmeans.inertia_)

plt.plot(range(N_min, N_max), res)
plt.title('Método del codo')
plt.xlabel('Número de clusters')
plt.ylabel('Suma de distancias al cuadrado dentro de cada cluster')
plt.show()
```



In [253]:

```
kmeans = KMeans(n_clusters = 4, init = 'k-means++', max_iter = 300, n_init =
10, random_state = 0)
kmeans.fit(features_pca2)

Z = kmeans.predict(features_pca2)
x = np.transpose(features_pca2)
y = (df_features['Fraude'].values)
z= np.array([Z]).T
xx =np.hstack((x.T,np.array([y]).T))
all = np.hstack((xx,z))
alldf = pd.DataFrame(all)
alldf.head()
```

Out[253]:

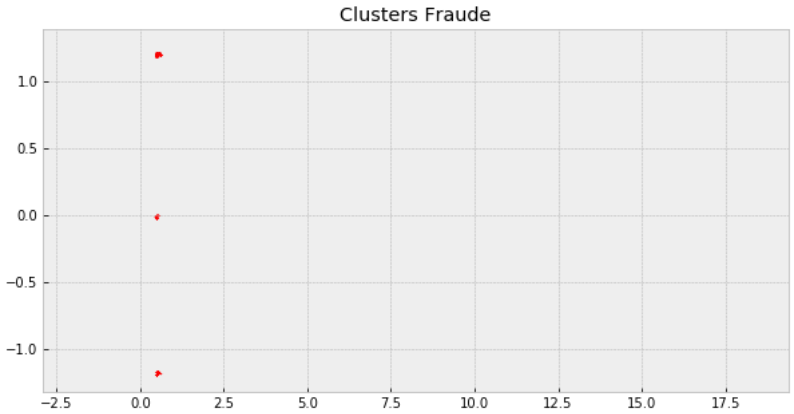
	0	1	2	3	4	5	6	7
0	15.167601	1.263800	0.000336	-0.419641	4.412999	54.925804	0.0	3.0
1	15.159809	-1.126581	-0.037414	-0.436231	4.410947	54.934552	0.0	3.0
2	15.167473	1.261592	0.000520	-0.419532	4.412852	54.924030	0.0	3.0
3	15.159784	-1.127023	-0.037377	-0.436209	4.410918	54.934197	0.0	3.0
4	15.159580	1.259145	0.000622	-0.419143	4.406143	54.824030	0.0	3.0

In [254]:

```
n_fraude = 6
n_cluster = 7
alldfgood0 = alldf.loc[alldf[n_fraude] == 0].loc[alldf[n_cluster] == 0]
alldffraud0 = alldf.loc[alldf[n_fraude] == 1].loc[alldf[n_cluster] == 0]
alldfgood1 = alldf.loc[alldf[n_fraude] == 0].loc[alldf[n_cluster] == 1]
alldffraud1 = alldf.loc[alldf[n_fraude] == 1].loc[alldf[n_cluster] == 1]
alldfgood2 = alldf.loc[alldf[n_fraude] == 0].loc[alldf[n_cluster] == 2]
alldffraud2 = alldf.loc[alldf[n_fraude] == 1].loc[alldf[n_cluster] == 2]
alldfgood3 = alldf.loc[alldf[n_fraude] == 0].loc[alldf[n_cluster] == 3]
alldffraud3 = alldf.loc[alldf[n_fraude] == 1].loc[alldf[n_cluster] == 3]

plt1.figure(figsize=(10, 5))
plt1.title('Clusters Fraude')
plt1.scatter(alldffraud[:,0], alldffraud[:,1],alpha = 1, c='red', s=1)
plt1.scatter(alldfgood[:,0], alldfgood[:,1], alpha=0, c='green', s=1)
```

Out[254]:

```
<matplotlib.collections.PathCollection at 0x1b933780b8>
```

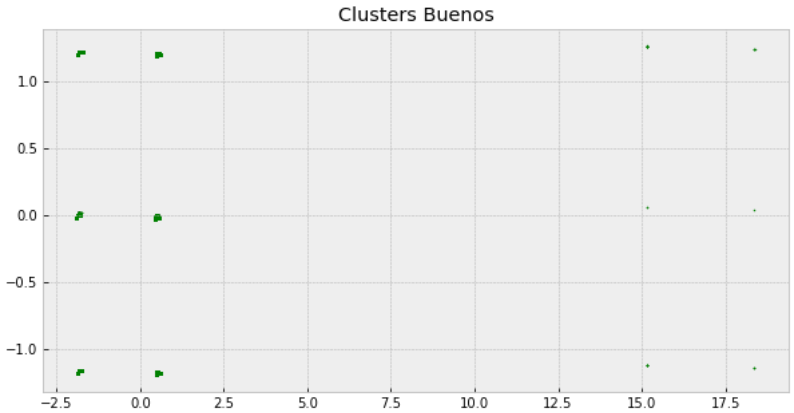


In [255]:

```
plt2.figure(figsize=(10, 5))
plt2.title('Clusters Buenos')
plt2.scatter(alldffraud[:,0], alldffraud[:,1],alpha = 0, c='red', s=1)
plt2.scatter(alldfgood[:,0], alldfgood[:,1], alpha=1, c='green', s=1)
```

Out[255]:

```
<matplotlib.collections.PathCollection at 0x1b9345e780>
```



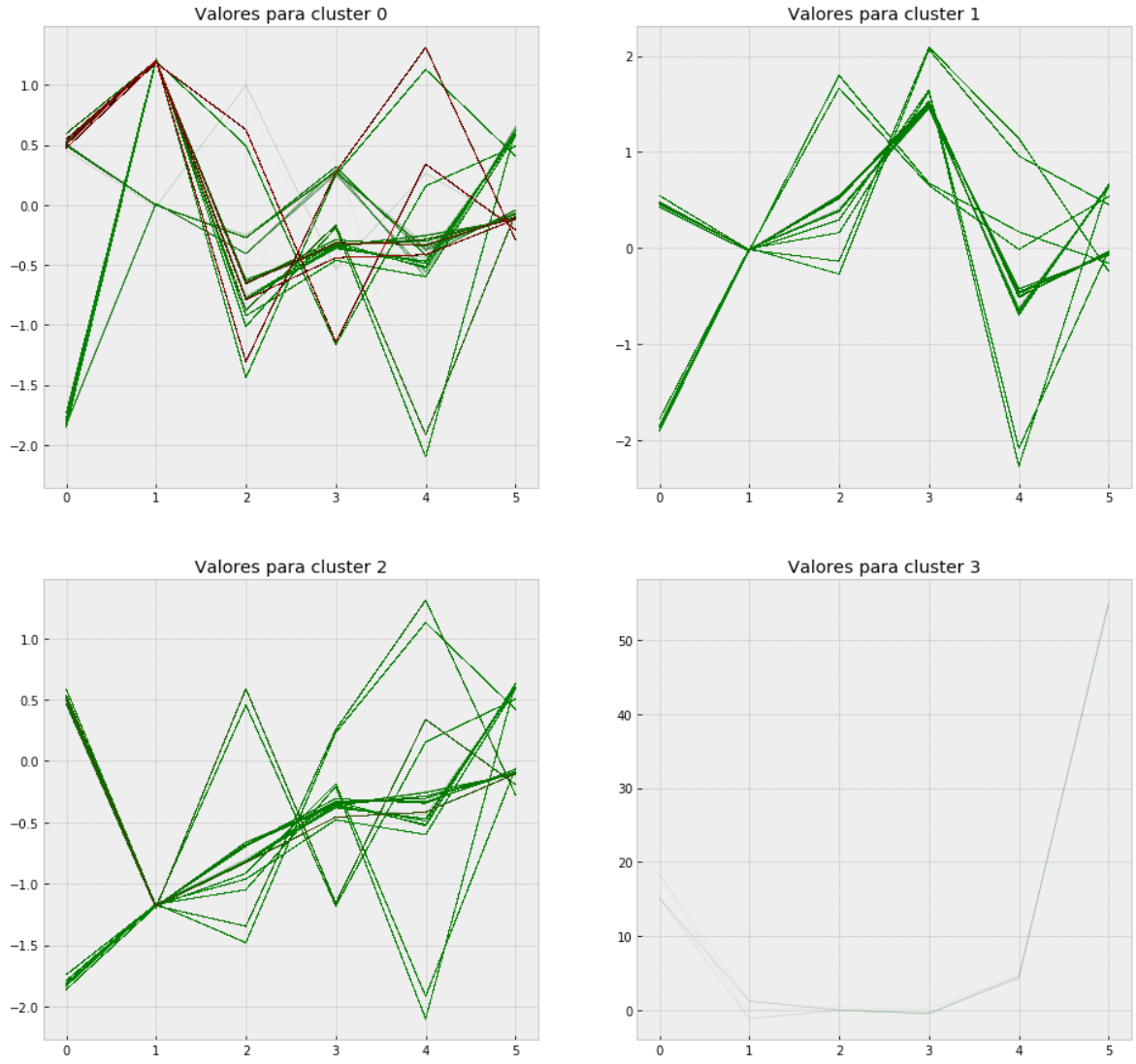
In [256]:

```
plt.figure(figsize=(16, 32))
plt.subplot(4, 2, 1)
if (alldfgood0.values[:,0:n_fraude].shape[0] > 0) :
plt.plot(np.transpose(alldfgood0.values[:,0:n_fraude]),alpha=0.1,
color='green',linewidth=0.1)
```

```
if (alldffraud0.values[:,0:n_fraude].shape[0] > 0) :
plt.plot(np.transpose(alldffraud0.values[:,0:n_fraude]),
alpha=0.1,color='red',linewidth=0.1)
plt.title("Valores para cluster 0")

plt.subplot(4, 2, 2)
if (alldfgood1.values[:,0:n_fraude].shape[0] > 0) :
plt.plot(np.transpose(alldfgood1.values[:,0:n_fraude]),alpha=0.1,
color='green',linewidth=0.1)
if (alldffraud1.values[:,0:n_fraude].shape[0] > 0) :
plt.plot(np.transpose(alldffraud1.values[:,0:n_fraude]),alpha=0.1,
color='red',linewidth=0.1)
plt.title("Valores para cluster 1")

plt.subplot(4, 2, 3)
if (alldfgood2.values[:,0:n_fraude].shape[0] > 0) :
plt.plot(np.transpose(alldfgood2.values[:,0:n_fraude]),alpha=0.1,
color='green',linewidth=0.1)
if (alldffraud2.values[:,0:n_fraude].shape[0] > 0) :
plt.plot(np.transpose(alldffraud2.values[:,0:n_fraude]),alpha=0.1,
color='red',linewidth=0.1)
plt.title("Valores para cluster 2")

plt.subplot(4, 2, 4)
if (alldfgood3.values[:,0:n_fraude].shape[0] > 0) :
plt.plot(np.transpose(alldfgood3.values[:,0:n_fraude]),
alpha=0.1,color='green',linewidth=0.1)
if (alldffraud3.values[:,0:n_fraude].shape[0] > 0) :
plt.plot(np.transpose(alldffraud3.values[:,0:n_fraude]),alpha=1,
color='red',linewidth=0.1)
plt.title("Valores para cluster 3")
```

Out[256]:

```
Text(0.5, 1.0, 'Valores para cluster 3')
```



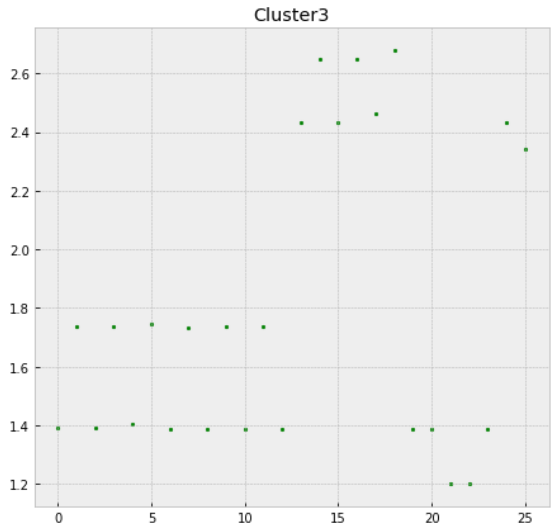
In [257]:

```
alldf['centers']=alldf[n_cluster]
cluster_number = alldf['centers'].values.astype('int64')
centers = kmeans.cluster_centers_
positions = alldf.iloc[:,0:(n_fraude)].values
d=np.sqrt(((positions-centers[cluster_number])**2).sum(axis=1))
alldf = pd.DataFrame(np.hstack((alldf.values,np.array([d]).T)))

alldfgood0 = alldf.loc[alldf[n_fraude] == 0].loc[alldf[n_cluster] ==
0][n_cluster+2]
alldffraud0 = alldf.loc[alldf[n_fraude] == 1].loc[alldf[n_cluster] ==
0][n_cluster+2]
alldfgood1 = alldf.loc[alldf[n_fraude] == 0].loc[alldf[n_cluster] ==
1][n_cluster+2]
alldffraud1 = alldf.loc[alldf[n_fraude] == 1].loc[alldf[n_cluster] ==
1][n_cluster+2]
alldfgood2 = alldf.loc[alldf[n_fraude] == 0].loc[alldf[n_cluster] ==
2][n_cluster+2]
alldffraud2 = alldf.loc[alldf[n_fraude] == 1].loc[alldf[n_cluster] ==
2][n_cluster+2]
alldfgood3 = alldf.loc[alldf[n_fraude] == 0].loc[alldf[n_cluster] ==
3][n_cluster+2]
```

```
alldffraud3 = alldf.loc[alldf[n_fraude] == 1].loc[alldf[n_cluster] ==
3][n_cluster+2]
plt.figure(figsize=(16, 32))

plt.subplot(4, 2, 1)
plt.title("Cluster0")
plt.scatter( range(0,alldfgood0.size),alldfgood0, c='green', s=7)
plt.scatter( range(0,alldffraud0.size),alldffraud0, c='red', s=7)

plt.subplot(4, 2, 2)
plt.title("Cluster1")
plt.scatter( range(0,alldfgood1.size),alldfgood1, c='green', s=7)
plt.scatter( range(0,alldffraud1.size),alldffraud1, c='red', s=7)

plt.subplot(4, 2, 3)
plt.title("Cluster2")
plt.scatter( range(0,alldfgood2.size),alldfgood2, c='green', s=7)
plt.scatter( range(0,alldffraud2.size),alldffraud2, c='red', s=7)

plt.subplot(4, 2, 4)
plt.title("Cluster3")
plt.scatter( range(0,alldfgood3.size),alldfgood3, c='green', s=7)
plt.scatter( range(0,alldffraud3.size),alldffraud3, c='red', s=7)
```

Out[257]:

```
<matplotlib.collections.PathCollection at 0x1bdd778eb8>
```



4 - Matriz de confusión

In [258]:

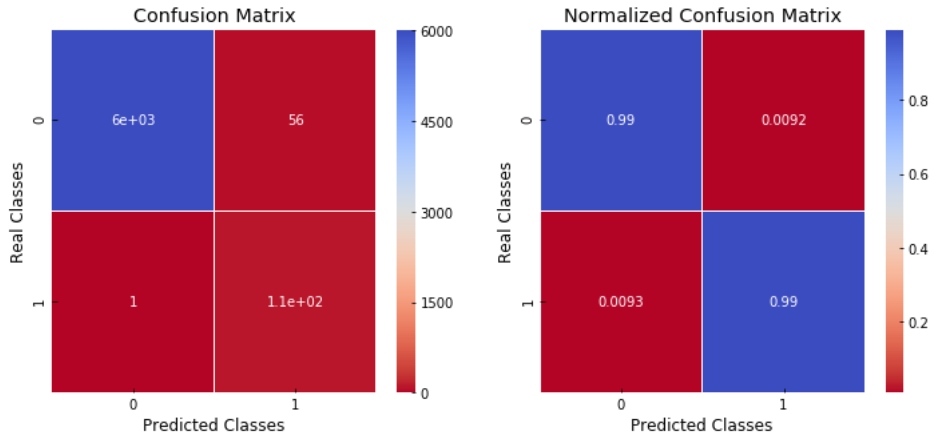
```
from sklearn.metrics import confusion_matrix, classification_report, auc,
precision_recall_curve, roc_curve
from sklearn.model_selection import train_test_split
from sklearn.ensemble import RandomForestClassifier
from sklearn.utils import shuffle
from sklearn.model_selection import cross_val_score
from sklearn.datasets import make_blobs
from sklearn.ensemble import RandomForestClassifier
from sklearn.ensemble import ExtraTreesClassifier
from sklearn.tree import DecisionTreeClassifier
from sklearn.ensemble import GradientBoostingClassifier

def plot_confusion_matrix(y_test, pred):
    y_test_legit = y_test.value_counts()[0]
    y_test_fraud = y_test.value_counts()[1]

    cfn_matrix = confusion_matrix(y_test, pred)
    cfn_norm_matrix = np.array([[1.0 /
y_test_legit,1.0/y_test_legit],[1.0/y_test_fraud,1.0/y_test_fraud]])
    norm_cfn_matrix = cfn_matrix * cfn_norm_matrix
```

```
    fig = plt.figure(figsize=(12,5))
    ax = fig.add_subplot(1,2,1)
    sns.heatmap(cfn_matrix,cmap='coolwarm_r',linewidths=0.5,annot=True,ax=ax)
    plt.title('Confusion Matrix')
    plt.ylabel('Real Classes')
    plt.xlabel('Predicted Classes')

    ax = fig.add_subplot(1,2,2)

sns.heatmap(norm_cfn_matrix,cmap='coolwarm_r',linewidths=0.5,annot=True,ax=ax
)

    plt.title('Normalized Confusion Matrix')
    plt.ylabel('Real Classes')
    plt.xlabel('Predicted Classes')
    plt.show()

    print('---Classification Report---')
    print(classification_report(y_test,pred))

alldf.rename(columns={n_fraude: 'Fraude'}, inplace=True)
```

Cluster 0

In [259]:

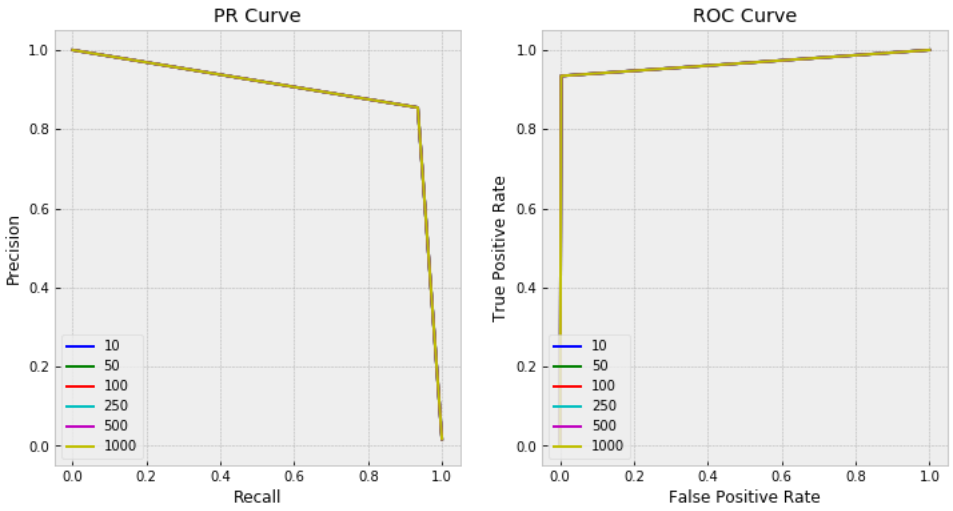
```
features_pca= alldf[alldf[n_cluster] == 0].drop([n_cluster], axis=1)
X_train, X_test = train_test_split(features_pca, test_size=0.2,
random_state=RANDOM_SEED)
Y_train = X_train['Fraude']
X_train = X_train.drop(['Fraude'], axis=1)
Y_test = X_test['Fraude']
X_test = X_test.drop(['Fraude'], axis=1)
rf =RandomForestClassifier(n_estimators=100, max_depth=None,  random_state=0)
rf.fit(X_train, Y_train)
Y_pred = rf.predict(X_test)

Train_Data= pd.concat([X_train, Y_train], axis=1)
X_1=Train_Data[Train_Data["Fraude"]==1 ]
X_0=Train_Data[Train_Data["Fraude"]==0]
X_0=shuffle(X_0,random_state=42).reset_index(drop=True)
X_1=shuffle(X_1,random_state=42).reset_index(drop=True)
ALPHA=1
X_0=X_0.iloc[:round(len(X_1)*ALPHA),:]
data_d=pd.concat([X_1, X_0])

Y_d=data_d['Fraude']
X_d=data_d.drop(['Fraude'],axis=1)
rf =RandomForestClassifier(n_estimators=100, max_depth=None, random_state=0,
n_jobs=-1)
rf.fit(X_d, Y_d)
Y_test_predicted=rf.predict(X_test)
plot_confusion_matrix(Y_test, Y_test_predicted)
```




```
---Classification Report---
              precision    recall  f1-score   support

         0.0       1.00      0.99      1.00      6076
         1.0       0.65      0.99      0.79       107

   micro avg       0.99      0.99      0.99      6183
   macro avg       0.83      0.99      0.89      6183
weighted avg       0.99      0.99      0.99      6183
```

In [260]:

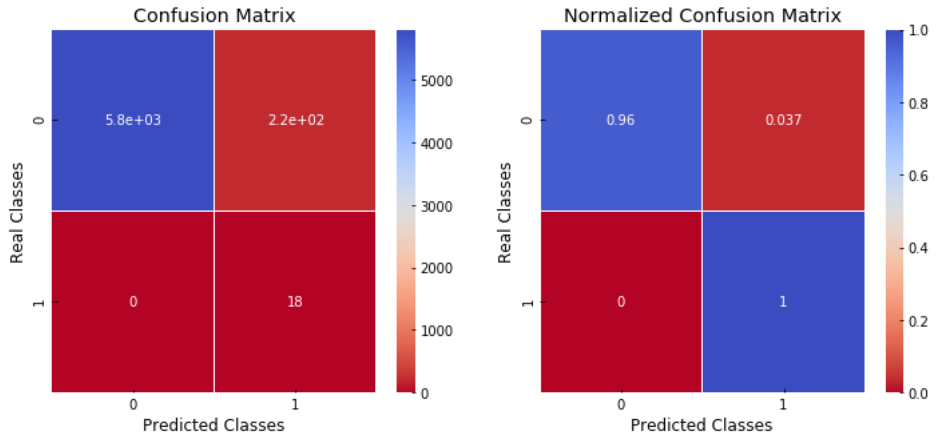
```
fig = plt.figure(figsize=(12,6))
ax1 = fig.add_subplot(1,2,1)
ax1.set_xlim([-0.05,1.05])
ax1.set_ylim([-0.05,1.05])
ax1.set_xlabel('Recall')
ax1.set_ylabel('Precision')
ax1.set_title('PR Curve')
ax2 = fig.add_subplot(1,2,2)
ax2.set_xlim([-0.05,1.05])
ax2.set_ylim([-0.05,1.05])
ax2.set_xlabel('False Positive Rate')
ax2.set_ylabel('True Positive Rate')
ax2.set_title('ROC Curve')

for n_est,k in zip([10, 50, 100, 250, 500,1000],'bgrcmy'):
    RandomForestClassifier(n_estimators=n_est, max_depth=None,
random_state=0, n_jobs=-1)
    rf.fit(X_train,Y_train)
    y_pred = rf.predict(X_test)
    p,r,_ = precision_recall_curve(Y_test, y_pred)
    tpr,fpr,_ = roc_curve(Y_test, y_pred)
    ax1.plot(r,p,c=k,label=n_est)
    ax2.plot(tpr,fpr,c=k,label=n_est)

ax1.legend(loc='lower left')
ax2.legend(loc='lower left')
plt.show()
```



Cluster 2

In [261]:

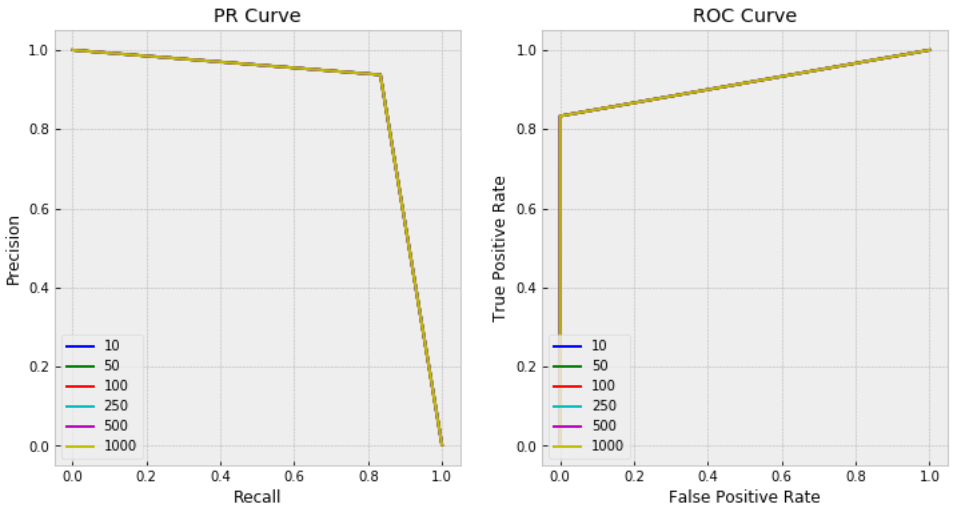
```
features_pca= alldf[alldf[n_cluster] == 2].drop([n_cluster], axis=1)
X_train, X_test = train_test_split(features_pca, test_size=0.2,
random_state=RANDOM_SEED)
Y_train = X_train['Fraude']
X_train = X_train.drop(['Fraude'], axis=1)
Y_test = X_test['Fraude']
X_test = X_test.drop(['Fraude'], axis=1)
rf =RandomForestClassifier(n_estimators=100, max_depth=None,  random_state=0)
rf.fit(X_train, Y_train)
Y_pred = rf.predict(X_test)

Train_Data= pd.concat([X_train, Y_train], axis=1)
X_1=Train_Data[Train_Data["Fraude"]==1 ]
X_0=Train_Data[Train_Data["Fraude"]==0]
X_0=shuffle(X_0,random_state=42).reset_index(drop=True)
X_1=shuffle(X_1,random_state=42).reset_index(drop=True)
ALPHA=1
X_0=X_0.iloc[:round(len(X_1)*ALPHA),:]
data_d=pd.concat([X_1, X_0])

Y_d=data_d['Fraude']
X_d=data_d.drop(['Fraude'],axis=1)
rf =RandomForestClassifier(n_estimators=100, max_depth=None, random_state=0,
n_jobs=-1)
rf.fit(X_d, Y_d)
Y_test_predicted=rf.predict(X_test)
plot_confusion_matrix(Y_test, Y_test_predicted)
```




```
---Classification Report---
              precision    recall  f1-score   support

         0.0       1.00      0.96      0.98      6027
         1.0       0.08      1.00      0.14        18

   micro avg       0.96      0.96      0.96      6045
   macro avg       0.54      0.98      0.56      6045
weighted avg       1.00      0.96      0.98      6045
```

In [262]:

```
fig = plt.figure(figsize=(12,6))
ax1 = fig.add_subplot(1,2,1)
ax1.set_xlim([-0.05,1.05])
ax1.set_ylim([-0.05,1.05])
ax1.set_xlabel('Recall')
ax1.set_ylabel('Precision')
ax1.set_title('PR Curve')
ax2 = fig.add_subplot(1,2,2)
ax2.set_xlim([-0.05,1.05])
ax2.set_ylim([-0.05,1.05])
ax2.set_xlabel('False Positive Rate')
ax2.set_ylabel('True Positive Rate')
ax2.set_title('ROC Curve')

for n_est,k in zip([10, 50, 100, 250, 500,1000],'bgrcmy'):
    RandomForestClassifier(n_estimators=n_est, max_depth=None,
random_state=0, n_jobs=-1)
    rf.fit(X_train,Y_train)
    y_pred = rf.predict(X_test)
    p,r,_ = precision_recall_curve(Y_test, y_pred)
    tpr,fpr,_ = roc_curve(Y_test, y_pred)
    ax1.plot(r,p,c=k,label=n_est)
    ax2.plot(tpr,fpr,c=k,label=n_est)

ax1.legend(loc='lower left')
ax2.legend(loc='lower left')
plt.show()
```


PR Curve
Precision
Recall
10
50
100
250
500
1000
ROC Curve
True Positive Rate
False Positive Rate

In [ ]:

Modelo #02 - Segmentación de Clientes

Segmentación de Clientes para Casinos

Proyecto Final-GRUPO 2

Master en Business Intelligence y Big Data, Grupo 1, 2018/2019

In [71]:

```
# Importacion de librerias
import pandas as pd
import numpy as np
import matplotlib.pyplot as plt
import seaborn as sb
import seaborn as sns
from sklearn.cluster import KMeans
from sklearn.metrics import pairwise_distances_argmin_min
from sklearn.model_selection import train_test_split
from sklearn.metrics import silhouette_score
from mpl_toolkits.mplot3d import Axes3D
from mpl_toolkits.mplot3d import Axes3D
from sklearn.decomposition import PCA
from sklearn.preprocessing import StandardScaler
from sklearn.linear_model import SGDClassifier
from sklearn.preprocessing import StandardScaler

%matplotlib inline
plt.rcParams['figure.figsize'] = (16, 9)
plt.style.use('ggplot')
```

Tratamiento de datos

In [2]:

```
#Importamos los datos y confirmamos el contenido
dfClientes = pd.read_csv("ScoreCliente.csv", sep=",", error_bad_lines=False,
index_col=False, dtype='unicode')
dfClientes.head()
```

Out[2]:

	Cliente_ID	Cliente_Nombre	Cliente_FechaNac	Cliente_FechaAlta	Cliente_Genero	TipoCliente_Desc	Pais_ID	CategoriaCliente_Desc

0	2	Cliente2	1900-01-01	2014-05-20	M	VIP	72	B
1	2	Cliente2	1900-01-01	2014-05-20	M	VIP	72	B
2	2	Cliente2	1900-01-01	2014-05-20	M	VIP	72	B
3	4	Cliente4	1985-09-01	2014-05-29	M	VIP	72	B
4	4	Cliente4	1985-09-01	2014-05-29	M	VIP	72	B

In [3]:

```
#Cambiar valores NaN por -1
dfClientes = dfClientes.fillna(-1)
dfClientes.head()
```

Out[3]:

	Cliente_ID	Cliente_Nombre	Cliente_FechaNac	Cliente_FechaAlta	Cliente_Genero	TipoCliente_Desc	Pais_ID	CategoriaCliente_Desc
0	2	Cliente2	1900-01-01	2014-05-20	M	VIP	72	B

1	2	Cliente2	1900-01-01	2014-05-20	M	VIP	72	B
2	2	Cliente2	1900-01-01	2014-05-20	M	VIP	72	B
3	4	Cliente4	1985-09-01	2014-05-29	M	VIP	72	B
4	4	Cliente4	1985-09-01	2014-05-29	M	VIP	72	B

In [9]:

```
#Cambiar tipos de Datos y borrar celdas inecesarias
dfClientes = dfClientes.drop(['ScoreCliente_ID', 'ScoreClienteDet_ID'],
axis=1)
dfClientes[['Cliente_ID', 'Mesa_ID' ,'Pais_ID']] = dfClientes[['Cliente_ID',
'Mesa_ID' , 'Pais_ID']].astype(int)
dfClientes[['ScoreClienteDetMesa_Drop', 'ScoreClienteDetMesa_Win',
'ScoreClienteDetMov_Monto']] = dfClientes[['ScoreClienteDetMesa_Drop',
'ScoreClienteDetMesa_Win','ScoreClienteDetMov_Monto']].astype(float)
dfClientes[['Cliente_FechaNac', 'Cliente_FechaAlta',
'ScoreCliente_HoraEntrada', 'ScoreCliente_HoraSalida' ,
'ScoreClienteDet_FechaHora']] =
pd.to_datetime(dfClientes[['Cliente_FechaNac', 'Cliente_FechaAlta',
'ScoreCliente_HoraEntrada', 'ScoreCliente_HoraSalida' ,
'ScoreClienteDet_FechaHora']].stack()).unstack()
dfClientes.dtypes
```

Out[9]:

```
Cliente_ID                          int32
Cliente_Nombre                     object
Cliente_FechaNac           datetime64[ns]
Cliente_FechaAlta          datetime64[ns]
Cliente_Genero                     object
TipoCliente_Desc                   object
Pais_ID                             int32
CategoriaCliente_Desc              object
ScoreCliente_HoraEntrada   datetime64[ns]
ScoreCliente_HoraSalida    datetime64[ns]
Usuario_Nombre                     object
ScoreClienteDet_FechaHora  datetime64[ns]
Moneda_Desc                        object
Mesa_ID                             int32
ScoreClienteDetMesa_Drop          float64
ScoreClienteDetMesa_Win           float64
TipoMoneda_Desc                    object
ScoreClienteDetMov_Monto          float64
Cliente_Win                         int32
dtype: object
```

In [10]:

```
#Se agrega una columna con 0 si el cliente perdio y 1 si el cliente gano en
base al monto generado
dfClientes['Cliente_Win'] = (dfClientes['ScoreClienteDetMov_Monto'] >=
1).astype(int)
dfClientes.head()
```

Out[10]:

	Cliente_ID	Cliente_Nombre	Cliente_FechaNac	Cliente_FechaAlta	Cliente_Genero	TipoCliente_Desc	Pais_ID	CategoriaCliente_Desc
0	2	Cliente2	1900-01-01	2014-05-20	M	VIP	72	B
1	2	Cliente2	1900-01-01	2014-05-20	M	VIP	72	B
2	2	Cliente2	1900-01-01	2014-05-20	M	VIP	72	B
3	4	Cliente4	1985-09-01	2014-05-29	M	VIP	72	B
4	4	Cliente4	1985-09-01	2014-05-29	M	VIP	72	B

In [11]:

```
#Estadisticas de datos
print('Estructura de los datos', dfClientes.shape)
dfClientes.describe()
Estructura de los datos (688977, 19)
```

Out[11]:

	Cliente_ID	Pais_ID	Mesa_ID	ScoreClienteDetMesa_Drop	ScoreClienteDetMesa_Win	ScoreClienteDetMov_Monto	Cliente_Win
count	688977.000000	688977.000000	688977.000000	688977.000000	688977.000000	6.889770e+05	688977.000000
mean	1383.520614	132.745666	39.753183	603.017516	-330.430765	1.940762e+02	0.389252
std	3248.050126	62.496145	2.585997	5354.559676	5171.334285	8.334126e+03	0.487581

min	1.000000	1.000000	1.000000	-500.000000	-800000.000000	-1.200000e+06	0.000000
25%	101.000000	71.000000	38.000000	-1.000000	-405.000000	-5.000000e+01	0.000000
50%	107.000000	72.000000	39.000000	50.000000	-1.000000	0.000000e+00	0.000000
75%	307.000000	196.000000	41.000000	550.000000	-1.000000	4.850000e+02	1.000000
max	17567.000000	248.000000	55.000000	800000.000000	500000.000000	1.200000e+06	1.000000

In [12]:

```
#Resumen por categorias
print("Resumen Tipo Cliente")
print(dfClientes.groupby('TipoCliente_Desc').size())
print("Resumen Genero")
print(dfClientes.groupby('Cliente_Genero').size())
print("Resumen Mesa")
print(dfClientes.groupby('Mesa_ID').size())
print("Ganancia Cliente")
print(dfClientes.groupby('Cliente_Win').size())
```

```
Resumen Tipo Cliente
TipoCliente_Desc
GENERAL     628483
ORIENTAL     59215
PREMIUM        288
VIP            991
dtype: int64
Resumen Genero
Cliente_Genero
F    337821
M    351156
dtype: int64
Resumen Mesa
Mesa_ID
1         212
2         140
3          96
4          48
5           7
6          20
7          33
8          31
9          26
10         24
11         58
12         17
13         17
```

```
14         59
15        159
16         61
17         87
18         94
19         87
20        114
21        157
22          4
23         30
24         37
28        650
29        208
30         51
31         82
32         74
33         92
34         75
35        126
36        132
37        302
38     254593
39     162921
40      40534
41     103314
42      12258
43      60926
44       3846
45      47094
47         24
54         20
55         37
dtype: int64
Ganancia Cliente
Cliente_Win
0    420791
1    268186
dtype: int64
```

In [13]:

```
#linealizar variables categoricas (aquellas que tengan el tipo object)
dfClientesObj = dfClientes.select_dtypes(include=['object']).copy()
dfClientesObj.head()
```

Out[13]:

	Cliente_Nombre	Cliente_Genero	TipoCliente_Desc	CategoriaCliente_Desc	Usuario_Nombre	Moneda_Desc	TipoMoneda_Desc
0	Cliente2	M	VIP	B	JEFE DE MESA	DOLAR USA	DROP
1	Cliente2	M	VIP	B	JEFE DE MESA	DOLAR USA	RESULTADO
2	Cliente2	M	VIP	B	JEFE DE MESA	DOLAR USA	DROP

3	Cliente4	M	VIP	B	JEFE DE MESA	DOLAR USA	DROP
4	Cliente4	M	VIP	B	JEFE DE MESA	DOLAR USA	RESULTADO

In [14]:

```
#Se creara una nueva columna para cada valor de variables de categoria lo que nos ayudara a poder filtrar mejor la informacion
dfClientesCat = pd.get_dummies(dfClientesObj,
columns=['Cliente_Genero','TipoCliente_Desc','CategoriaCliente_Desc','TipoMoneda_Desc','Moneda_Desc'])
print('Estructura de los datos', dfClientesCat.shape)
dfClientesCat.head()
```

```
Estructura de los datos (688977, 23)
```

Out[14]:

	Cliente_Nombre	Usuario_Nombre	Cliente_Genero_F	Cliente_Genero_M	TipoCliente_Desc_GENERAL	TipoCliente_Desc_ORIENTAL	TipoCliente_Desc_PREMIUM	TipoCliente_Desc_VIP
0	Cliente2	JEFE DE MESA	0	1	0	0	0	1
1	Cliente2	JEFE DE MESA	0	1	0	0	0	1
2	Cliente2	JEFE DE MESA	0	1	0	0	0	1
3	Cliente4	JEFE DE MESA	0	1	0	0	0	1
4	Cliente4	JEFE DE MESA	0	1	0	0	0	1

5 rows × 23 columns

In [15]:

```
#Unimos el nuevo DataFrame a las demas variables y creamos un DF nuevo
dfDemasDatos = dfClientes.select_dtypes(include=['int32','float64',
'datetime64']).copy()
dfClientesFinal = pd.concat([dfDemasDatos, dfClientesCat], axis=1)
dfClientesFinal.head()
```

Out[15]:

	Cliente_ID	Cliente_FechaNac	Cliente_FechaAlta	Pais_ID	ScoreCliente_HoraEntrada	ScoreCliente_HoraSalida	ScoreClienteDet_FechaHora	Mesa_ID

0	2	1900-01-01	2014-05-20	72	2014-07-17 16:27:47.387	2014-07-17 10:00:00.000	2014-07-17 10:00:00.000	2
1	2	1900-01-01	2014-05-20	72	2014-07-17 16:27:47.387	2014-07-17 10:00:00.000	2014-07-17 10:00:00.000	2
2	2	1900-01-01	2014-05-20	72	2014-07-17 16:27:47.387	2014-07-17 10:00:00.000	2014-07-17 10:00:00.000	2
3	4	1985-09-01	2014-05-29	72	2014-07-17 15:32:42.240	2014-07-18 11:34:24.030	2014-07-18 11:19:26.727	1
4	4	1985-09-01	2014-05-29	72	2014-07-17 15:32:42.240	2014-07-18 11:34:24.030	2014-07-18 11:19:26.727	1

5 rows × 35 columns

In [16]:

```
dfClientes.dtypes
```

Out[16]:

```
Cliente_ID                          int32
Cliente_Nombre                     object
Cliente_FechaNac           datetime64[ns]
Cliente_FechaAlta          datetime64[ns]
Cliente_Genero                     object
TipoCliente_Desc                   object
Pais_ID                             int32
CategoriaCliente_Desc              object
ScoreCliente_HoraEntrada   datetime64[ns]
ScoreCliente_HoraSalida    datetime64[ns]
Usuario_Nombre                     object
ScoreClienteDet_FechaHora  datetime64[ns]
Moneda_Desc                        object
Mesa_ID                             int32
ScoreClienteDetMesa_Drop          float64
ScoreClienteDetMesa_Win           float64
TipoMoneda_Desc                    object
ScoreClienteDetMov_Monto          float64
Cliente_Win                         int32
dtype: object
```

In [17]:

```
#Descripcion de los datos enteros
tt = dfClientesFinal.describe().transpose()
tt[(tt['max']>1) | (tt['min']< -1)]
```

Out[17]:

	count	mean	std	min	25 %	50 %	75 %	max
Cliente_ID	688977.0	1383.520614	3248.050126	1.0	101.0	107.0	307.0	17567.0
Pais_ID	688977.0	132.745666	62.496145	1.0	71.0	72.0	196.0	248.0
Mesa_ID	688977.0	39.753183	2.585997	1.0	38.0	39.0	41.0	55.0
ScoreClienteDetMesa_Drop	688977.0	603.017516	5354.559676	-500.0	-1.0	50.0	550.0	800000.0
ScoreClienteDetMesa_Win	688977.0	-330.430765	5171.334285	-800000.0	-405.0	-1.0	-1.0	500000.0
ScoreClienteDetMov_Monto	688977.0	194.076221	8334.126447	-1200000.0	-50.0	0.0	485.0	1200000.0

In [18]:

```
dfClientesFinal.head()
```

Out[18]:

	Cliente_ID	Cliente_FechaNac	Cliente_FechaAlta	Pais_ID	ScoreCliente_HoraEntrada	ScoreCliente_HoraSalida	ScoreClienteDet_FechaHora	Mesa_ID
0	2	1900-01-01	2014-05-20	72	2014-07-17 16:27:47.387	2014-07-17 10:00:00.000	2014-07-17 10:00:00.000	2
1	2	1900-01-01	2014-05-20	72	2014-07-17 16:27:47.387	2014-07-17 10:00:00.000	2014-07-17 10:00:00.000	2
2	2	1900-01-01	2014-05-20	72	2014-07-17 16:27:47.387	2014-07-17 10:00:00.000	2014-07-17 10:00:00.000	2
3	4	1985-09-01	2014-05-29	72	2014-07-17 15:32:42.240	2014-07-18 11:34:24.030	2014-07-18 11:19:26.727	1

4	4	1985-09-01	2014-05-29	72	2014-07-17 15:32:42.240	2014-07-18 11:34:24.030	2014-07-18 11:19:26.727	1

5 rows × 35 columns

In [19]:

```
#DF Solo para datos numericos
dfClientesNum =
dfClientesFinal.select_dtypes(include=['int32','float64','uint8']).copy()
dfClientesNum.head()
```

Out[19]:

	Cliente_ID	Pais_ID	Mesa_ID	ScoreClienteDetMesa_Drop	ScoreClienteDetMesa_Win	ScoreClienteDetMov_Monto	Cliente_Win	Cliente_Genero_F
0	2	72	2	-1.0	-1.0	100.0	1	0
1	2	72	2	-1.0	-1.0	100.0	1	0
2	2	72	2	-1.0	-1.0	100.0	1	0
3	4	72	1	-1.0	-1.0	100.0	1	0
4	4	72	1	-1.0	-1.0	100.0	1	0

5 rows × 28 columns

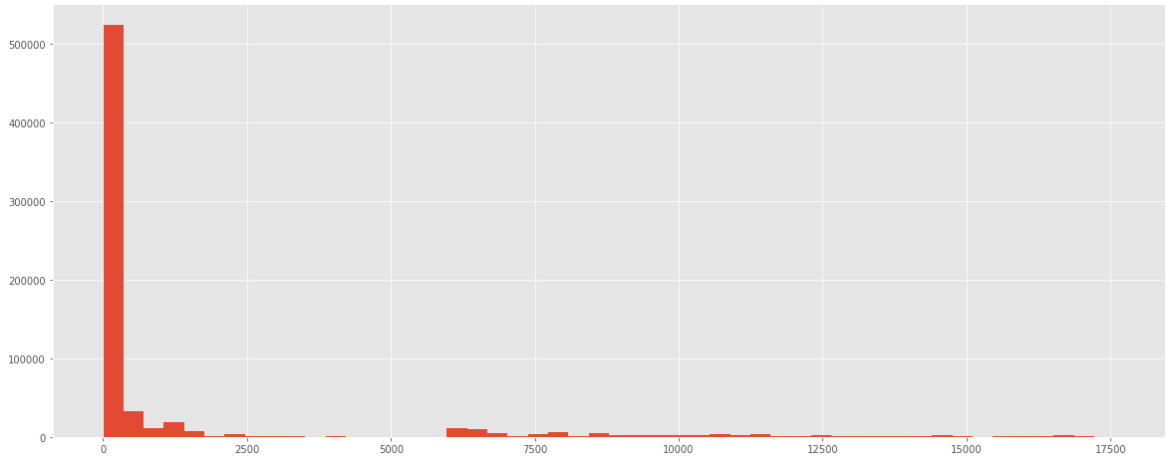
In [20]:

```
#Dibujar grafica para identificar el comportamiento de las variables
plt.figure(figsize=(20,8))
plt.hist(dfClientesNum.Cliente_ID, bins=50)
```

Out[20]:

```
(array([5.24977e+05, 3.30820e+04, 1.05620e+04, 1.91770e+04, 7.88200e+03,
        7.11000e+02, 4.11600e+03, 1.49100e+03, 1.26200e+03, 6.49000e+02,
        3.61000e+02, 4.78000e+02, 1.94000e+02, 1.36000e+02, 1.72000e+02,
        1.60000e+01, 1.87000e+02, 1.08540e+04, 1.03350e+04, 5.36000e+03,
        1.20400e+03, 2.99600e+03, 5.54300e+03, 1.26900e+03, 5.20400e+03,
        2.84300e+03, 1.93700e+03, 2.18800e+03, 2.53300e+03, 2.74900e+03,
        3.77300e+03, 2.02500e+03, 3.69600e+03, 8.89000e+02, 1.07300e+03,
        1.86600e+03, 1.41400e+03, 1.06400e+03, 1.22100e+03, 1.06500e+03,
        7.87000e+02, 2.71400e+03, 1.31100e+03, 1.58000e+02, 5.45000e+02,
        1.61600e+03, 6.90000e+02, 1.67000e+03, 5.85000e+02, 3.47000e+02]),
 array([1.000000e+00, 3.523200e+02, 7.036400e+02, 1.054960e+03,
        1.406280e+03, 1.757600e+03, 2.108920e+03, 2.460240e+03,
        2.811560e+03, 3.162880e+03, 3.514200e+03, 3.865520e+03,
        4.216840e+03, 4.568160e+03, 4.919480e+03, 5.270800e+03,
        5.622120e+03, 5.973440e+03, 6.324760e+03, 6.676080e+03,
        7.027400e+03, 7.378720e+03, 7.730040e+03, 8.081360e+03,
        8.432680e+03, 8.784000e+03, 9.135320e+03, 9.486640e+03,
```

```
        9.837960e+03, 1.018928e+04, 1.054060e+04, 1.089192e+04,
        1.124324e+04, 1.159456e+04, 1.194588e+04, 1.229720e+04,
        1.264852e+04, 1.299984e+04, 1.335116e+04, 1.370248e+04,
        1.405380e+04, 1.440512e+04, 1.475644e+04, 1.510776e+04,
        1.545908e+04, 1.581040e+04, 1.616172e+04, 1.651304e+04,
        1.686436e+04, 1.721568e+04, 1.756700e+04]),
 <a list of 50 Patch objects>)
```

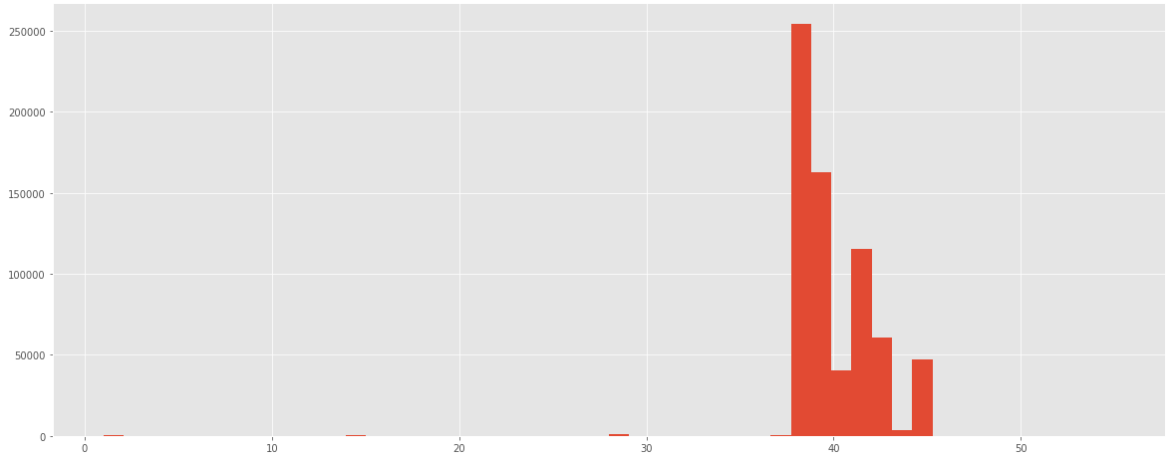


In [21]:

```
plt.figure(figsize=(20,8))
plt.hist(dfClientesNum.Mesa_ID, bins=50)
```

Out[21]:

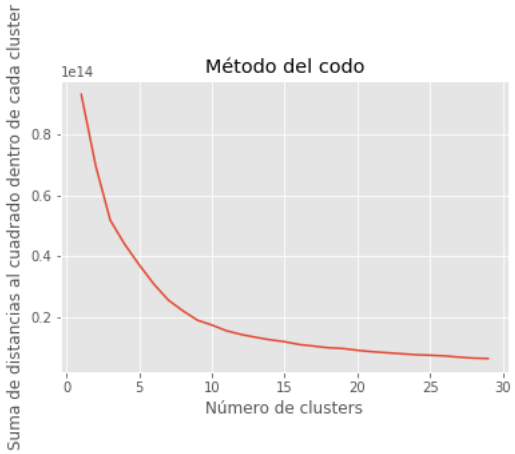
```
(array([3.52000e+02, 9.60000e+01, 4.80000e+01, 7.00000e+00, 2.00000e+01,
        3.30000e+01, 3.10000e+01, 2.60000e+01, 2.40000e+01, 5.80000e+01,
        1.70000e+01, 1.70000e+01, 2.18000e+02, 6.10000e+01, 8.70000e+01,
        9.40000e+01, 8.70000e+01, 1.14000e+02, 1.57000e+02, 4.00000e+00,
        3.00000e+01, 3.70000e+01, 0.00000e+00, 0.00000e+00, 0.00000e+00,
        8.58000e+02, 5.10000e+01, 8.20000e+01, 7.40000e+01, 9.20000e+01,
        7.50000e+01, 1.26000e+02, 1.32000e+02, 3.02000e+02, 2.54593e+05,
        1.62921e+05, 4.05340e+04, 1.15572e+05, 6.09260e+04, 3.84600e+03,
        4.70940e+04, 0.00000e+00, 2.40000e+01, 0.00000e+00, 0.00000e+00,
        0.00000e+00, 0.00000e+00, 0.00000e+00, 0.00000e+00, 5.70000e+01]),
 array([ 1.  ,  2.08,  3.16,  4.24,  5.32,  6.4 ,  7.48,  8.56,  9.64,
        10.72, 11.8 , 12.88, 13.96, 15.04, 16.12, 17.2 , 18.28, 19.36,
        20.44, 21.52, 22.6 , 23.68, 24.76, 25.84, 26.92, 28.  , 29.08,
        30.16, 31.24, 32.32, 33.4 , 34.48, 35.56, 36.64, 37.72, 38.8 ,
        39.88, 40.96, 42.04, 43.12, 44.2 , 45.28, 46.36, 47.44, 48.52,
        49.6 , 50.68, 51.76, 52.84, 53.92, 55.  ]),
 <a list of 50 Patch objects>)
```



K MEANS

In [22]:

```
#Metodo del codo
res = []
N_max=30
N_min=1
for i in range(N_min, N_max):
    kmeans = KMeans(n_clusters = i, init = 'k-means++', max_iter = 300,
n_init = 10, random_state = 0)
    kmeans.fit(dfClientesNum)
    # Obtenemos la suma de las distancias al cuadrado dentro de cada cluster
    res.append(kmeans.inertia_)

plt.plot(range(N_min, N_max), res)
plt.title('Método del codo')
plt.xlabel('Número de clusters')
plt.ylabel('Suma de distancias al cuadrado dentro de cada cluster')
plt.show()
```



In [73]:

```
X = dfClientesNum
X.shape
```

Out[73]:

```
(688977, 28)
```

In [74]:

```
#Utilizaremos 5 cluster para nuestro modelo
kmeans = KMeans(n_clusters=5, init = 'k-means++', random_state = 42)
label = kmeans.fit_predict(X)
centroides = kmeans.cluster_centers_
print(centroides)
```

```
[[ 1.50876698e+03  1.27414689e+02  3.96515248e+01  3.95084255e+02
  -2.12672866e+02 -9.07637824e+02  3.15870603e-01  4.47669531e-01
   5.52330469e-01  9.08534077e-01  8.99309762e-02  2.89428541e-04
   1.24551833e-03  6.17881155e-05  9.98071560e-01  1.41787465e-03
   4.48776839e-04  2.11380395e-05  1.93600052e-01  9.90723653e-02
   7.93489484e-04  9.93634198e-02  1.99526833e-01  3.08288550e-01
   9.93341518e-02  1.55445891e-03  9.95660198e-01  2.78534321e-03]
 [ 3.02777778e+01  7.88888889e+01  2.52777778e+01  5.94444444e+05
  -5.11111111e+05  7.33333333e+05  1.00000000e+00  5.00000000e-01
   5.00000000e-01  0.00000000e+00  9.44444444e-01  5.55555556e-02
  -2.16840434e-19  4.06575815e-20  1.00000000e+00 -6.50521303e-19
  -5.42101086e-20  0.00000000e+00  1.00000000e+00 -4.16333634e-17
  -1.08420217e-19 -4.16333634e-17  0.00000000e+00  5.55111512e-17
  -2.77555756e-17  6.50521303e-19  4.44089210e-16  1.00000000e+00]
```

```
 [ 3.74838710e+01  9.40000000e+01  2.52903226e+01  1.28894758e+05
   7.48705645e+04  3.78274194e+05  1.00000000e+00  4.51612903e-01
   5.48387097e-01 -1.22124533e-15  7.09677419e-01  2.90322581e-01
   1.95156391e-18  8.06451613e-02  9.19354839e-01  1.08420217e-18
   6.50521303e-19  0.00000000e+00  5.16129032e-01  2.77555756e-17
   2.25806452e-01  2.77555756e-17  4.83870968e-02  2.09677419e-01
   5.55111512e-17  1.73472348e-18  7.77156117e-16  1.00000000e+00]
 [ 3.37307692e+01  9.58461538e+01  2.41923077e+01  4.93846154e+05
  -4.41923077e+05 -5.22692308e+05  1.11022302e-16  5.00000000e-01
   5.00000000e-01 -4.44089210e-16  8.46153846e-01  1.53846154e-01
   4.33680869e-19  5.42101086e-20  1.00000000e+00 -2.16840434e-19
  -1.08420217e-19  0.00000000e+00  1.11022302e-16 -5.55111512e-17
  -3.25260652e-19 -5.55111512e-17  0.00000000e+00  1.00000000e+00
  -2.77555756e-17  0.00000000e+00  6.66133815e-16  1.00000000e+00]
 [ 3.42659654e+02  1.77189776e+02  4.06207321e+01  1.90825186e+03
  -1.09409330e+03  9.05494886e+03  9.99702163e-01  8.45476945e-01
   1.54523055e-01  9.44006715e-01  5.17694203e-02  1.17780846e-03
   3.04605637e-03  4.06140850e-04  9.95424146e-01  3.38450708e-03
   7.85205643e-04  2.07184259e-17  9.77797634e-01  2.43684510e-03
   9.74738039e-04  1.35380284e-05  1.20488452e-03  1.75723608e-02
   2.39946951e-14  3.15436060e-03  9.88641594e-01  8.20404516e-03]]
```

In [75]:

```
print(label)
label.shape
```

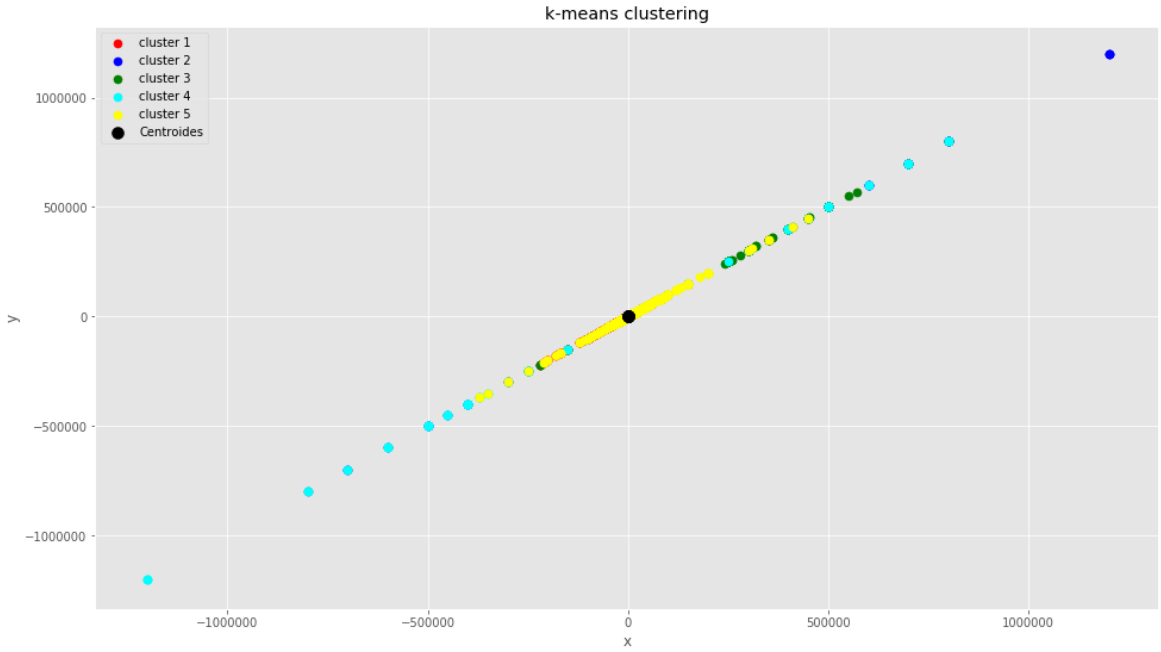
```
[0 0 0 ... 0 0 0]
```

Out[75]:

```
(688977,)
```

In [76]:

```
#Graficar

plt.scatter(X.loc[label==0], X.loc[label==0],s = 50, c= 'red', label='cluster
1')
plt.scatter(X.loc[label==1], X.loc[label==1],s = 50, c= 'blue',
label='cluster 2')
plt.scatter(X.loc[label==2], X.loc[label==2],s = 50, c= 'green',
label='cluster 3')
plt.scatter(X.loc[label==3], X.loc[label==3],s = 50, c= 'cyan',
label='cluster 4')
plt.scatter(X.loc[label==4], X.loc[label==4],s = 50, c= 'yellow',
label='cluster 5')

plt.scatter(centroides[:,0], centroides[:,1], s = 100, c= 'black', label =
'Centroides')

plt.xlabel('x')
plt.ylabel('y')
plt.title('k-means clustering')
plt.legend(loc='best')
plt.show()
```



Matriz de Confusión

In [77]:

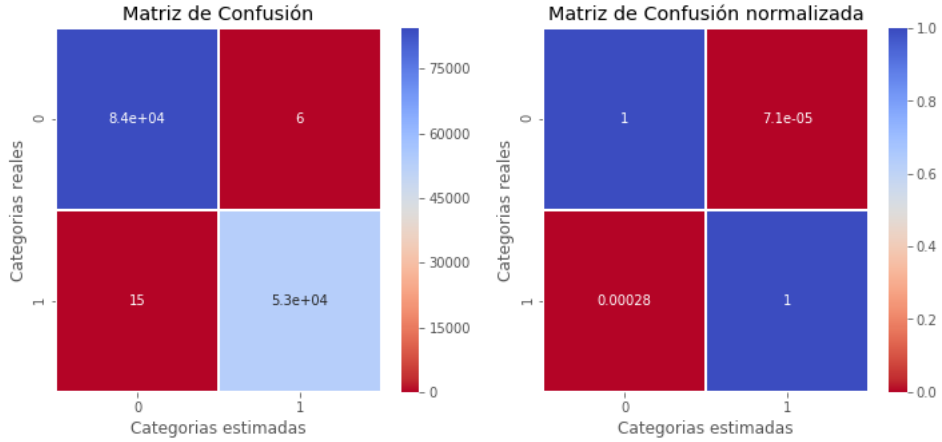
```
#Funcion para visualizar las matrices de confusión
from sklearn.metrics import confusion_matrix, classification_report, auc,
precision_recall_curve, roc_curve
def plot_confusion_matrix(y_test, pred):

    y_test_legit = y_test.value_counts()[0]
    y_test_fraud = y_test.value_counts()[1]

    cfn_matrix = confusion_matrix(y_test, pred)
    cfn_norm_matrix = np.array([[1.0 /
y_test_legit,1.0/y_test_legit],[1.0/y_test_fraud,1.0/y_test_fraud]])
    norm_cfn_matrix = cfn_matrix * cfn_norm_matrix

    fig = plt.figure(figsize=(12,5))
    ax = fig.add_subplot(1,2,1)
    sns.heatmap(cfn_matrix,cmap='coolwarm_r',linewidths=0.5,annot=True,ax=ax)
    plt.title('Matriz de Confusión')
    plt.ylabel('Categorias reales')
    plt.xlabel('Categorias estimadas')

    ax = fig.add_subplot(1,2,2)

sns.heatmap(norm_cfn_matrix,cmap='coolwarm_r',linewidths=0.5,annot=True,ax=ax
)

    plt.title('Matriz de Confusión normalizada')
    plt.ylabel('Categorias reales')
    plt.xlabel('Categorias estimadas')
    plt.show()

    print('---Report de clasificación---')
    print(classification_report(y_test,pred)) #CALCULO DE PRECICION, RECALL Y
F1-SCORE
```

In [78]:

```
#Creacion de datos para test y train
RANDOM_SEED = 42
X_train, X_test = train_test_split(dfClientesNum, test_size=0.2,
random_state=RANDOM_SEED)
Y_train = X_train['Cliente_Win']
X_train = X_train.drop(['Cliente_Win'], axis=1)
Y_test = X_test['Cliente_Win']
X_test = X_test.drop(['Cliente_Win'], axis=1)
```

In [79]:

```
from sklearn import metrics

#MODLO DE REGRESION
sgd_clasificador =SGDClassifier(alpha=0.0001, average=False,
class_weight=None, epsilon=0.1,
       eta0=0.0, fit_intercept=True, l1_ratio=0.15,
       learning_rate='optimal', loss='hinge', max_iter=5, n_iter=None,
       n_jobs=1, penalty='l2', power_t=0.5, random_state=42, shuffle=True,
       tol=None, verbose=0, warm_start=False)

sgd_clasificador.fit(X_train, Y_train) #Entrenar el modelo
Y_train_predicted=sgd_clasificador.predict(X_train) #Entrenamiento
Y_test_predicted=sgd_clasificador.predict(X_test) #TEST

plot_confusion_matrix(Y_test, Y_test_predicted) #LLamada a la funcion para
dibujar matriz
C:\ProgramData\Anaconda3\lib\site-
packages\sklearn\linear_model\stochastic_gradient.py:183: FutureWarning:
max_iter and tol parameters have been added in SGDClassifier in 0.19. If
max_iter is set but tol is left unset, the default value for tol in 0.19 and
0.20 will be None (which is equivalent to -infinity, so it has no effect) but
will change in 0.21 to 1e-3. Specify tol to silence this warning.
  FutureWarning)
```




```
---Report de clasificación---
              precision    recall  f1-score   support

           0       1.00      1.00      1.00     84397
           1       1.00      1.00      1.00     53399

   micro avg       1.00      1.00      1.00    137796
   macro avg       1.00      1.00      1.00    137796
weighted avg       1.00      1.00      1.00    137796
```

In [ ]: