Instituto Tecnológico y de Estudios Superiores de Monterrey
Escuela de Ingeniería y Ciencias
Ingeniería en Ciencias de Datos y Matemáticas

Proyecto Ciencia de Datos

Estudiantes:

Jose Andres Meyer A01366785
Antonio Patjane Ceballos A01657978
Luis Gabriel Martínez Rentería A01651812

Profesores:

Laura Hervert Escobar
María de los Angeles Constantino González

Materia: Análisis de ciencia de datos
Grupo: 3

Monterrey, Nuevo León, México
1 de mayo de 2021

Índice

Resumen

CEMEX Ventures es una compañía de producción industria de cemento que quiere disminuir sus costos sin tener un mayor impacto en el medio ambiente. Para lograr esto se han recaudado datos acerca de diferentes factores que estan involucrados en la producción de cemento, los cuales fueron analizados con diferentes métodos de regresión para encontrar la manera más óptima para utilizar las diferentes energías. Dentro de los modelos que se plantearon están Random Forest Regressor, Support Vector Regressor, entre otros. Para poder encontrar él modelo más óptimo se utilliazaron diferentes métricas para evaluarlos, entre ellas el valor de su R^2, error medio y precisión, dando como resultado Random Forest Regresor. Con un modelo base, se fueron modificando los valores de sus hiperparámetros hasta encontrar los más adecuados. Finalmente se creo una plataforma amigable al usuario, la cual permite visualizar los resultados obtenidos.

Palabras claves: Energía, Ahorro energético, Machine learning, optimización, basura.

1. Introducción

Al contar con un sistema de producción, siempre queremos obtener los mejores resultados con el menor costo posible sin dejar atrás la disminución del impacto ecólogico de la creación de nuestro producto. Para esto es necesario entender el funcionamiento de nuestra sistema de producción al igual que los datos que se obtienen de la misma. Tomando en cuenta los factores anteriores y aplicándolos a nuestro caso, debemos de comparar el costo de las energías y el impacto que ambas tienen, de manera que siga siendo redituable sin perjudicar al ambente a corto, mediano o largo plazo.

2. Problema

CEMEX Ventures tiene distintas problemáticas que lo limitan de obtener la mayor cantidad de ingresos, por lo que tienen un dilema entre usar lo recursos indispensable y minimizar los gastos por los costo de las energía durante el proceso de manufactura, los cuales son necesarios para cumplir los estándares de calidad establecidos por la empresa. El principal desafío al cual se enfrentan es encontrar la forma de optimizar el uso y gasto de sus energías. Conociendo las ventajas y desventajas durante el proceso de manufactura al igual que sus costos, podemos observar que el precio de la energía calórica es aproximadamente 0.724 veces lo de la energía eléctrica, sin embargo, la emisiones de carbono emitidas por la energía calórica representan un riesgo para los alrededores de la planta y para el clima a nivel mundial.

2.1. Valoración de la situación

Actualmente la compañía CEMEX Venture cuenta con dos principales fuentes de energía, las cuales son calórica y eléctrica. Esta genera ciertos costos en la compañía, los cuales son considerados obsoletos e ineficientes, además de generar un impacto en el medio ambiente. Sin embargo, este modelo se ha utilizado por varios años y se teme que se pierda cierta calidad de los productos, la cual es un factor que no solo los diferencia de la competencia, también les permite operar siguiendo las normas establecidas.

Durante el proceso de manufactura se utiliza tanto diesel, alimentando la antigua maquinaria, como energía eléctrica, para modernos dispositivos. La maquinaria moderna es más eficiente (produce más unidades) en la cadena de producción, pero la electricidad es más costosa que el diesel. Ambos tipos de maquinaria se pueden emplear, de forma combinada, en el proceso de fabricación de los soportes metálicos. La calidad de los soportes metálicos que se fabrican es evaluada por un equipo en fábrica, y depende de la estabilidad y control de todos los procesos involucrados en toda cadena de fabricación.

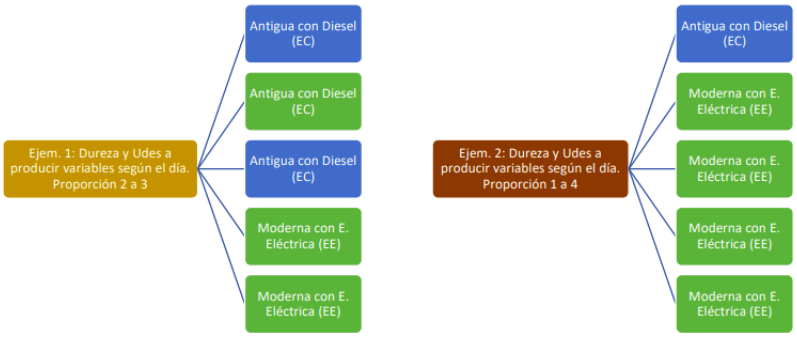

Figura 1: Posibles combinaciones del uso de maquinaria

3. Antecedentes

3.1. Antonio - Multiobjective control architecture to estimate optimal set points for user comfort and energy saving in buildings

La optimización del uso de energía puede tener diferentes enfoques, ya sea en la producción o en la climatización de edificios. En ambos casos se debe de encontrar un equilibrio entre las necesidades y las posibles mejoras por lo que se debe de encontrar un conjunto óptimo, el cual puede estar compuesto por puntos o factores. Uno de los usos directos en el ahorro de energía es mediante cambios en la arquitectura de un edificio. Estos se realizan en un conjunto de puntos específicos, obtenidos en simulaciones, que representan los cambios más óptimos, donde manteniendo el confort de los usuarios y se puede disminuir el consumo hasta en un rango de 7-9 %[Martell, 2020] [3]

3.2. Luis - Modelado de regresión para el consumo de electricidad empresarial: una comparación de la red neuronal recurrente y sus variantes

Hacer un análisis de las energías que se usan ayuda a las empresas a que tengan una mejor eficiencia en consumo y producción. En los últimos años las RNN y sus variantes han revolucionado el análisis de datos para el pronóstico de series de tiempo, aunque no todas las variantes de esta funcionan bien para proponer

un uso eficiente del consumo de electricidad en las empresas. Entre las variantes de RNN están:

- **Estándar**
- Basado en memoria a corto plazo y largo plazo **(LSTM)**
- Basado en unidades recurrentes con compuerta **(GRU)**

De estos métodos, los que han mostrado los mejores resultados son los LSTM y GRU son los que mejor se adaptaron para hacer las predicciones, aunque GRU es un modelo mucho más sencillo.[Bai,2021] [1]

3.3. Jose - Historia de la energía: desde los comustibles fósiles a la energía nuclear

La habilidad de transformar la energía ha sido de suma importancia para la existencia del ser humano. La energía nos permite que ciertas acciones sean puestas en marcha y aunque existen diferentes fuentes, desde su descubrimiento en la revolución indusrtial, seguimos usando la misma, Siendo esta los combustibles fósiles, los cuales no son inagotables, por lo que es necesario pensar en energías alternativas. Una de las mejores alternativas es la energía nuclear, ya que se ha demostrado que es una de las más poderosas. Un ejemplo de esto es que otorga más del 11 % de la electricidad en el mundo con un impacto mínimo en el ambiente.[Kelkar,2015] [2]

3.4. Luis - Is it worth generating energy with garbage? Defining a carbon tax to encourage waste-to-energy cycles

La generación de basura en los últimos años ha ido en aumento de manera alarmante, por lo que se han propuesto diferentes soluciones de que hacer con esto. Los sistemas de conversión de residuos en energía han sido algunas de las soluciones adoptadas, pero llevan la carga de la emisión de gases contaminantes. Para regular esto, se han hecho acuerdos internacionales en los que se ha creado el concepto de impuesto de carbono, de manera que se puede medir el nivel de emisiones de dióxido de carbono que se imponen a los sistemas de generación de electricidad por quema de combustibles. Este concepto aplicado a la quema de basura ayuda a que empresas no puedan abusar de la quema de desechos o de combustibles en procesos que lo requieran.[Mollica, 2020] [4]

4. Objetivos

4.1. Objetivos del negocio

Conociendo las necesidades de CEMEX Ventures, nuestro objetivo es encontrar el modelo de regresión más adecuado para la optimización de las energías disponibles sin aumentar el costo de producción tomando en cuenta que este debe de seguir siendo rentable.Esto presenta un beneficio al medio ambiente, a la compañía a mediano y a largo plazo. Este proyrcto no solo permite una mejora a la imagen pública, también puede disminuir sus costos de producción cuando haya avances a la obtención y distribución de energías limpias.

4.2. Objetivos de Data Mining

Nuestro primer objetivo en la minería de datos es obtener más información de todos los factores que influyen en la producción de cemento. Con estos coonocimientos procederemos a obtener el uso de energía habitual dentro de sus operacione. El siguiente objetivo en nuestra lista es obtener el rango de datos más útil para ser introducidos y probados en nuestros modelos. Con esta limpieza podremos obtener nuestro objetivo principal, obtener el mejor modelo de regresión. Complementando el objetivo anterior, debemos de escoger el mejor ajuste para nuestro modelo, para que este optimizado.

5. Enfoque

5.1. Plan de proyecto

Nuestros pasos a seguir para la realización del proyecto son: el entendimiento de los datos proporcionados, obtención de la distibución y correlación de los parámetros, definición de rangos para la obtención de datos óptimos, limpieza y estandarizacón de datos, implementación y prueba de modelos, obtención de los mejores parámetros para el modelo seleccionado y la creación de una plataforma donde se puedan visualizar los resultados obtenidos.

6. Método

El primer paso que se realizó fue la carga de las librerías necesarias para el correcto funcionamiento del código. Teniendo este requisito, lo segundo con el código fue cargar la base de datos con la librería de pandas, la cual incluye herramientas muy útiles de análisis de datos. Después con la base de datos ya cargada se eliminaron los casos en los que:

- Se tuvieran valores nulos
- Se tuvieran valores que no deberían de ser posibles (tasa de procucción igual a cero, entre otros)

Una vez eliminados estos datos se hizo un scatter plot de las variables más importantes para así visualizar su comportamiento y los outliers, dándonos una primera idea de los rangos que se iban a utilizar para la limpieza.Como siguiente paso se aplicó un boxplot a cada una de las variables para así definir los rangos que se iban a tomar en cuenta para el ajuste de variables.[Fig 2] El método aplicado fue que aquellas variables que estuvieran muy alejadas del valor mínimo del rango establecido, estsos casos fueron mínimos, se eliminaron y aquellas variables que se ecnontraran un poco más abajo del rango establecido se ajustarían al valor mínimo de nuestro rango, lo mismo para los valores mayores.

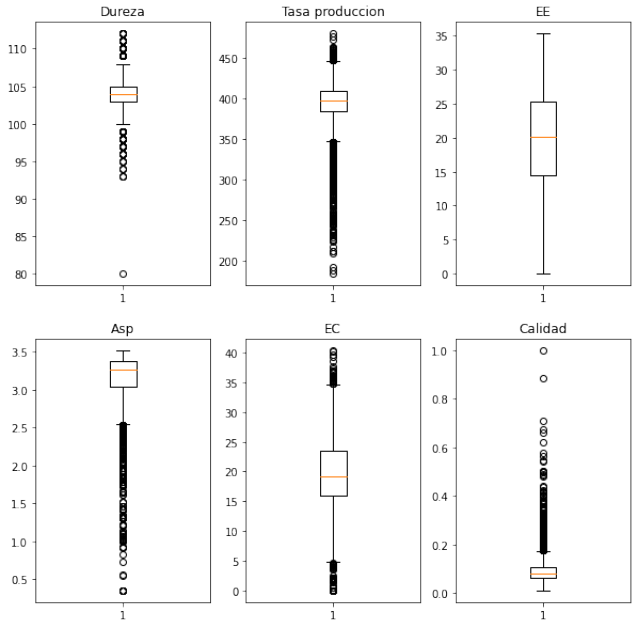

Figura 2: Box plot de las variables

Los rangos que se establecieron para las variables son los siguientes:

- Dureza = [100, 108]
- Tasa de producción = [270, 450]
- Asp = [1.5, 3.5]
- Calidad = (0, 0.5] (*Los valores de calidad cero ya habían sido eliminados anteriormente*)
- *Las variables Energia Eléctrica y Energía calórica no fueron modificadas*

Se hizo un plot de todas las variables contra las diferentes energías con el fin de encontrar patrones o

grupos más claros para después poder trabajar sobre estos.[Fig 3] De la misma manera se busco la matriz de correlación y se gráfico como mapa de calor.[Fig 4]

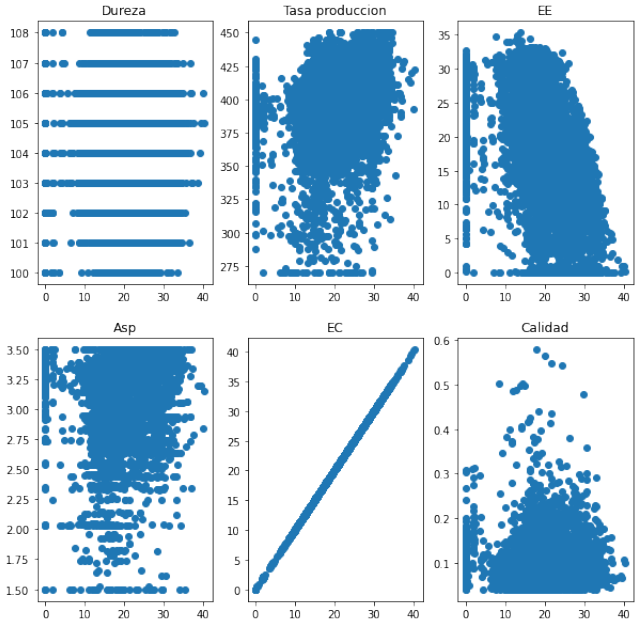

Figura 3: Gráfica de puntos de la Energía Calórica vs las demás variables

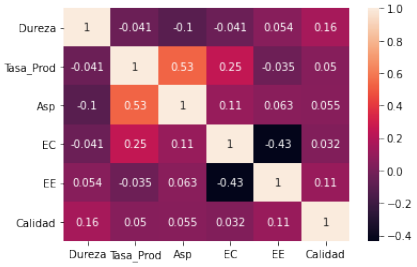

Figura 4: Mapa de calor de correlación de las variables

Al identificar los patrones dentro de los datos e identificar qu la correlación entre las variables es muy baja, decidimos estandarizar los datos con el modelo de mínimo y máximo. Este consiste en encontrar el valor más pequeño y más grande de cada columna, que serán los límites en una escala de 0 a 1 [Fig 5], para después realizar los cálculos a cada dato dentro de las columnas del data set.

$$x_{scaled} = \frac{x - x_{min}}{x_{max} - x_{min}}$$

Figura 5: Fórmula de estandarización min/max

Con una mayor correlación comenzamos a implementar y probar nuestros modelos de regresión. Para

esto dividimos el dataset en características y objetivos. Dentro de las carcterísticas se encuentran tasa de producción, aspiración, calidad y dureza, mientras que en objetivos esta energía calórica y energía electrica. De la misma manera resalizamos una subdivisión en entrenamineto y prueba, de esta manera el modelo se familiariza con los resulrtados que deseamos obtener y realiza los cálculos correspondientes.

Los modelos que utilizamos y comparamos fueron:

- Support Vector Regressor
- Decision Tree Regressor
- Random Forest Regressor
- X G Boost

De estos obtuvimos su R^2, MAE, MSE y RMSE, para tener un parámetro de comparación, donde al obtener sus resultados el que tiene un mejor ajueste es el Random Forest Regressor con una diferencia de 0.01 al siguiente modelo. Esta pequeña diferencia nos hizo decidir este modelo y buscar los mejores parámetros para seguir mejorando su eficacia. Esto fue mediante un GridSearch, donde se introducen los hiperparámetros del modelo y un rango en el cual quieren ser probados. Estos hiperparámetros fueron profundidad máxima, número de características y número de estimadores, con rangos de 1 a 9.

Ya con un modelo optimizado, procedimos finalmente al desarrollo de un medio donde se puedan visualizar los datos de manera agradable para el usuario. En esta se puede introducir la calidad, tasa de producción, aspiración y dureza desada, devolviendo los valores de energía necesaria para su realización, al igual que el costo de la producción, además de una gráfica en 3D, dondé se puede esa configuración.

7. Resultados

Parámetro	Resulltado
R^2	0.142
MAE	0.137 (0.003)
MSE	0.03408
RMSE	0.1845950529066626

Cuadro 1: Resultados de pruebas al modelo Support Vector Regressor

Parámetro	Resulltado
R^2	-0.573
MAE	0.184 (0.004)
MSE	0.06297
RMSE	0.25182234358630967

Cuadro 2: Resultados de pruebas al modelo Decision Tree Regressor

Parámetro	Resulltado
R^2	0.184
MAE	0.134 (0.003))
MSE	.032377
RMSE	0.17990285467758116

Cuadro 3: Resultados de pruebas al modelo Random Forest Regressor

Parámetro	Resulltado
R^2	0.14457980096612993
MAE	0.967 (0.722))
MSE	0.033887
RMSE	0.18406251261430653

Cuadro 4: Resultados de pruebas al modelo X G Boost

Como se puede ver, el modelo con el mejor ajuste a nuestros datos es el Random Forest Regressor, por lo que mediante la variación de sus hiperparámetros lo optimizaremos. La variación de sus hiperparámetros es probar con los valores de 1 a 9, en máxima profundidad, número de características y número de estimadores, al igual que las combinaciones entre ellos. LLegando a la conclusión que la mejor combinación es: max depth: 9, max features: 'auto', n estimators': 15 con un resultado de 0.185667804167262.

8. Discusión

Al implementar cuatro modelos, podemos ver que cada uno se adapta de manera diferente a los datos. Sin embargo, hay que encontrar parámetros para poder definir cual es el mejor, por lo que se hicieron los cálculos para obtener el error absoluto, el error cuadrado de la media y la raiz del error cuadrado de la media. Sin emargo, con estas mediciones, podemos ver que no existe una correlación directa entre los datos proporcionados y el resultado esperado. Estas mediciones cuentan con valores muy bajor¿s en uno de los

factores más determinantes en un modelo de regresión, el coeficiente de determinación, el cual muestra la correlación directa de los datos al modelo. Esta es menor al 20 %, cuando los estándares aceptados son mayores al 90 % para concluir que el modelo es adecuado.

9. Conclusiones

Conclusión de Luis Gabriel Martínez: Dentro de los resultados que obtuvimos como equipo, a pesar de ser pocos y en parte tener que trabajar más, los resultados fueron los esperados por cada uno de los miembros, el producto final cumple con los requisitos y más, aunque con un poco más de tiempo se hubiera podido hacer un mejor resultado (no por problema de la materia, más bien que esta unidad formativa durara más de 5 semanas). Dentro de lo que la palabra reto conlleva, siento que este proyecto cumple con todos los requisitos, ya que nos impulsa a aprender por nuestra cuenta, buscar nuevas maneras de hacer las cosas y empujar nuestos límites más allá de lo aprendido de un libro. El apoyo por parte del equipo docente fue más que solo apoyo y confianza. Aunque en cuanto a los resultados que nos da el modelo final, aunque el modelo está lejos de ser bueno, esto no es gracias a una mala enseñanza de la materia o un mal trabajo, es gracias al material que se nos proporcionó que el error que nos da el modelo es tan bajo. Este tipo de trabajo es cada vez más cercano a lo que nos vamos a dedicar y con las herramientas que aprendimos a lo largo de este curso se pueden hacer muchas más cosas. En cuanto a mi trabajo dentro del equipo quedo muy contento con el resultado final, aunque si me hubiera gustado implementar más herramientas como lo aprendido en SQL o más trabajo con git y github para el trabajo de la app web.

Conclusión Antonio Patjane: Comparando los valores que obtuvimos como resultado de los errores en cada modelo, el que mas se adecua utilizando dos variables de salida es el de Random Forest Regressor y al optimizarlo, sus resultados son aún mejors. Sin emargo el que se adecué mejor a los demás no equivale a que deba de ser utilizado. Esto se debe a sus bajos resultados comparados con el estándar aceptado por la comunidady la poca correlación que existe entre sus variables. Una manera de solucionar este problema y encontrar un modelo que permita estimar de manera adecuada una regresión es cambiar las variables medidas a unas con mayor correlación.

Conclusión Jose Andres Meyer: A pesar de que fue un reto bastante dificl, creo que aprendí todo lo que tenía que aprender, ha sido uno de los retos que más me han exigido pero todo ha valido la pena, apesar de que al principio tuvimos un problema y nos quedamos siendo 3 solo en el equipo, todos contribuimos mucho y por ende aprendimos demasiado, agradezco a mis compañeros por todo el esfuerazo que pusieron ya que esto demuestra el esfuerzo de los 3 en el proyecto, Este proyecto ha sido unos de los mejores en los que he participado, aparte de que aprendí mucho más de lo que esperaba y a pesar del nivel que traía al principio del curso, me voy muy contento con todo lo que realice y lógranos con nuestro equipo, aprendí algo que sinceramente me va a servir en un futuro lo cual lo podré usar como profesionista en un mañana, Aprendí sobre los modelos de machine learning así como entrenarlos y la limpieza de datos, que fue de las cosas

que más me costaron trabajo pero de la misma forma aprendí bastante de ellos y es algo que me emociona bastante, el trabajo en equipo a pesar de que éramos 3 pero tuvimos un desempeño bastante bueno y creo no es por presumir pero hicimos un excelente trabajo los 3 a pesar de todas las dificultades que traíamos

Referencias

[1] Yun Bai y col. "Regression modeling for enterprise electricity consumption: A comparison of recurrent neural network and its variants." En: *International journal of electrical power energy systems*. PA. 2021. URL: http://0-search.ebscohost.com.biblioteca-ils.tec.mx/login.aspx?direct=true&db=edsbl&AN=vdc.100116371492.0x000001&lang=es&site=eds-live&scope=site.

[2] N Kelkar. "HISTORIA DE LA ENERGÍA: DESDE LOS COMBUSTIBLES FÓSILES A LA ENERGÍA NUCLEAR." En: *MOMENTO - Revista de Física*. 2015, PA. URL: https://repositorio.unal.edu.co/handle/unal/67367.

[3] M. Martell y col. "Multiobjective control architecture to estimate optimal set points for user comfort and energy saving in buildings." En: *ISA transactions*. 2020, pág. 454. URL: http://0-search.ebscohost.com.biblioteca-ils.tec.mx/login.aspx?direct=true&db=edsbl&AN=vdc.100114020299.0x000001&lang=es&site=eds-live&scope=site.

[4] Gustavo José Gonçalves Mollica y José Antonio Perrella Balestieri. "Is it worth generating energy with garbage? Defining a carbon tax to encourage waste-to-energy cycles." En: *Applied thermal engineering*. 2020. URL: http://0-search.ebscohost.com.biblioteca-ils.tec.mx/login.aspx?direct=true&db=edsbl&AN=vdc.100114018767.0x000001&lang=es&site=eds-live&scope=site.

que sin ellos esto hubiera sido imposible, me encta mand1.51.5

10. Anexo

10.1. Aportes al código

Luis. Lo que yo realice, fue definir nuevas variables de costo, costo ponderado, ee ponderada y ec ponderada, de la misma forma realice una grafica en la cual presentamos una interpretación del trabajo realizado, igualmente reduje el dataset para poder trabajar con los datos que presentan la mejor calidad y el menor costo ponderado.

Jose. Lo que yo realice fue ayudar a definir el dataset para test y para prueba, de igual forma definir las variables de respuesta y las explicatorias, tuvimos un problema con nuestro dataset y tuvimos que reajustar los índices de nuestra base ya que nos estaba arrojando datos NaN el cual no nos permitía seguir adelante

para poder empezar a realizar los modelos. Hice los modelos de SVR, DTR, XGboost y Random Forest (Los 3 intregrantes del equipo apoyamos en esto) y por ultimo lo que realice fue sacar los errores (de igual forma aquí apoyamos todos), de la misma forma aporte en la realizacion de nuestro min max en el cual tuvimos varios problemas, asi mismo en el grid search y randomize search para saber cual era nuestro mejor modelo y saber cual iba a ser utilizado

Antonio. Mi aporte a este proyecto ha sido la limpieza del dataset, donde se eliminaron los valores nulos y atípicos. Ya con estos, se realizaron los boxplot, histogramas, mapa de calor y gráfica de puntos. Al ver la correlación de los datos y su funcionamiento, llegamos a la conclusión que debíamos de transformarlos. Utilicé el método de mínimo y máximo para realizar este cambio. Finalmente, apoyé con la implementacioón de los modelos y la obtención de sus errores. Con apoyo de mis compañeros, también aporté a la optimización del modelo seleccionado.

10.2. Aportes al Reporte

Antonio: Mis aportes al reporte escrito fueron la creación del abstract, escritura de la introducción, complementación a la valoración de la situación y objetivos, método y discusión.

Luis: Mi aporte al reporte escrito fue la escritura de la metodología, la complementación de la discusión y el plan de proyecto.

Jose: Mi aporte al proyecto fue la escritura de los objetivos del negocio y data mining, la complementación a la metodología y los resultados obtenidos.

De la misma manera, todos participamos en los antecedentes, conlusiones y bibliografía, aportando textos de manera individual y participando en pláticas para la creación de una conclusión grupal.