ESTADÍSTICA MULTIVARIADA AVANZADA

ANTEPROYECTO

PRESENTADO POR:

Edgar Leandro Jiménez Jaimes

Jorge Luis Renteria Roa

DOCENTE:

Tomás Olarte Hernandez

UNIVERSIDAD EAFIT

MEDELLÍN

MAESTRÍA EN CIENCIAS DE LOS DATOS Y ANALÍTICA

MARZO DE 2020

1. Pregunta de investigación y objetivos:

Las diversas técnicas de aprendizaje durante el tiempo se han estado incorporando en la solución de retos organizacionales dado su alto componente computacional, pese a esto, una vista a priori de las soluciones permite identificar que son las grandes empresas quienes han implementado la mayoría de estas capacidades en sus procesos diarios. Si esta vista se enfoca en el sector gastronómico el panorama es aún más desolador pues aquí conviven negocios con un alto grado de informalidad acostumbrados a llevar su operación de forma convencional, apalancada en el conocimiento de pocas personas (dueños y/o chef) y procesos construidos desde lo empírico. No hay que alejarse mucho de la realidad para evidenciar que preguntas como *¿Cuánto se vendió del producto X ayer? ¿Cuanto insumo Y nos queda para mañana?* y la sensación de desperdicios o pérdidas hacen parte de la cotidianidad de los negocios.

Actualmente controlar los procesos de estos negocios gastronómicos suele tener un costo demasiado alto (tanto operacional como económico), pues es necesario considerar muchas variables y estar al tanto de muchos detalles para garantizar un día a día medianamente óptima. Uno de los principales dolores se encuentra alrededor del inventario. Conocer cuáles fueron las salidas reales de un producto, con cuál particularidad y en qué momento del día ha sido un reto constante en los pequeños y medianos restaurantes pues su tamaño comercial no les permite afrontar costos de supervisión y/o automatización de procesos.

En este trabajo se busca identificar si es factible desarrollar un modelo robusto que identifique el tipo de producto en un conjunto de imágenes de dos clases (dos tipos de pizza) apoyados en el material visto en clase y como motivación personal en fuentes de aprendizaje externas relativas a técnicas de visión computacional y aprendizaje profundo. En la implementación de sistemas de análisis de imágenes existen muchos componentes a tener en cuenta para garantizar un modelo correcto, buscando acotar el alcance, se dará enfocará únicamente en inferencia gráfica, por lo tanto, se hace uso algoritmos de aumentación de datos para generar nuevos registros propios con enfoques ligeramente diferentes. Con base a la imposibilidad de garantizar que las pizzas siempre van a llegar en la misma posición, con el mismo tamaño, a la misma distancia y bajo las mismas condiciones de luz se considera indispensable iterar cada uno de los registros entre cada una de estas propiedades.

En resumen, el objetivo del trabajo es evaluar cuán factible es la aplicación de técnicas de aprendizaje automático convencional y aprendizaje profundo para el manejo de inventarios de materias primas y control de ventas en los negocios gastronómicos. Para esto, es indispensable implementar algoritmos de aprendizaje supervisado básicos como Regresión logística, Random Forest y Gradient Boosting; entrenar redes neuronales de tipo Fully connected, red neuronal convolucional propia y red neuronal convolucional pre-entrenada resnet de keras; presentar un comparativo de los aportes de cada forma de aprendizaje para solucionar problemas de estudio y análisis de imágenes y entregar resultados base que permitan escalar y estabilizar la solución en producción para ser implementada en los procesos rutinarios del negocio.

Pregunta de investigación

¿Cuán aplicables son las técnicas de aprendizaje de máquinas convencionales y aprendizaje profundo para el manejo de materias primas y control de ventas en establecimientos de comercio gastronómicos?

Objetivo general:

Identificar la aplicabilidad de técnicas de aprendizaje de máquinas convencional y aprendizaje profundo para el manejo de materias primas y control de ventas en establecimientos de comercio gastronómicos.

Objetivos específicos:

- Implementar algoritmos de aprendizaje supervisado convencional y de ensamble como regresión logística, gradient boosting y random forest.
- Entrenar redes neuronales fully connect, convolucional propia y convolucional pre-entrenada.
- Presentar un análisis comparativo de los aportes de cada tipo de aprendizaje para solucionar problemas de estudio y análisis de imágenes.
- Entregar resultados base que permitan escalar y estabilizar la solución en producción.

2. Metodología de investigación: Resumen de la metodología a usar y los métodos de análisis propuestos.

Metodología a usar:

La normalización del proceso de hallar conocimiento en los datos ha convocado esfuerzos desde finales de los años 90, durante este tiempo se han planteado metodologías que buscan alinear la ejecución de los proyectos analíticos con los objetivos del negocio sin descuidar los elementos técnicos que deben tenerse en cuenta. Dentro de esas metodologías se encuentra ASUM – DM planteada por IBM la cual por medio de cinco categorías busca abordar de manera holística la salida a producción de un sistema analítico partiendo desde la necesidad del negocio. Para este proyecto se trabajará con este enfoque, buscando una aplicación real a un problema de negocio.

Metodología ASUM-DM ("Towards an Improved ASUM-DM Process Methodology for Cross-Disciplinary Multi-organization Big Data & Analytics Projects," 2018)

Analicemos algunos puntos de esta metodología y la aplicación en este proyecto.

- *Entendimiento del negocio:* En un ambiente de competencia abierta y mercados dinámicos estar a la vanguardia de las herramientas tecnológicas y reducir operatividad es una obligación para todas aquellas empresas que buscan sobrevivir, por ello, como se expuso anteriormente existe una necesidad actual de negocio de reducir pérdidas y costos operativos en cuanto a la producción de una materia prima y su inventario y se plantea la implementación de sistemas de análisis de imágenes que permita automatizar el proceso de control y organizar producción, cuentas e inventario.

- *Acercamiento Analítico:* A través de diferentes modelos de aprendizaje automático y de visión computacional, modelos de redes neuronales profundas, se busca asignar realizar una clasificación de imágenes para identificar un tipo puntual de materia prima, en este caso pizza pepperoni o pizza hawaiana.

- *Datos necesarios:* Los datos empleados para realizar esta modelación son las imágenes de los ambos tipos de pizza y que se encuentre respectivamente con su label de la clase a la que pertenece

- *Captura de datos:* La captura de datos, en este proyecto se realizó de manera manual, no se profundizó en un trabajo que permitiera realizar esta captura de manera automática, sino que se enfocaron esfuerzos en el estudio y la modelación de este tipo de problemas de clasificación de imágenes mediante aprendizaje profundo.

- *Preparación y entendimiento de datos:* Lo que respecta a la preparación, como se hará énfasis más adelante es un proceso iterativo que se va ajustando a medida que se entrenan los modelos mientras que el entendimiento de datos es claro ya que tanto el problema como la naturaleza de los datos son claros

- *Modelamiento:* Para el modelamiento se presentará a continuación las técnicas que se van a estudiar y desarrollar en este proyecto. (Ver métodos de análisis propuestos)

- *Evaluación:* Métricas de evaluación: Se trabajará con métricas de clasificación binaria, tales como accuracy, precisión, recall, f1-score y curva roc.

- *Despliegue:* Este proyecto no incluye despliegue a productivo como tal, se van a estructurar y entrenar los modelos, estos serán guardados en formatos h5 para posteriormente ser cargados y utilizados.

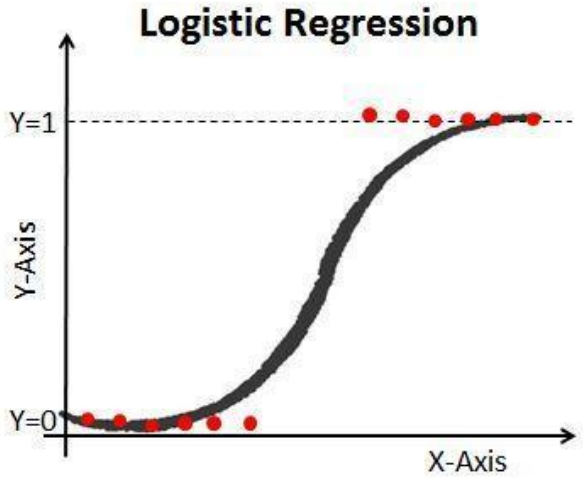
Métodos de análisis propuestos:

Los métodos propuestos en este proyecto, entendiendo que se está trabajando en clasificación de imágenes, se enfocan dos propuestas. La primera utilizando métodos de aprendizaje estadístico o de máquinas tradicional o clásico, utilizando la librería de sklearn de python empleando tres modelos sugeridos en clase y por otra parte, utilizando métodos avanzados como es el aprendizaje profundo (Deep Learning por sus siglas en inglés), que emulan el comportamiento del cerebro humano y que es estudiado los últimos años en problemas de clasificación de imágenes, predicción de series de tiempo y procesamiento del lenguaje natural.

Como se mencionó anteriormente, este proyecto se enfoca en el estudio de dos enfoques, el primero y que se presenta a continuación es realizando una clasificación de imágenes utilizando modelos de aprendizaje de máquinas mediante la librería de sklearn de python, el segundo enfoque será presentado posteriormente con sus detalles correspondientes.

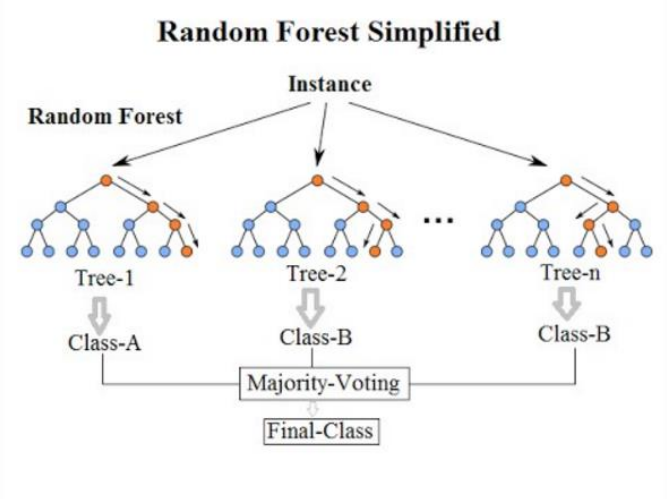
En ese orden de ideas, se utilizaron tres modelos de ML, teniendo en cuenta la sugerencia del docente, para este problema de clasificación, los cuales se presentan a continuación:

1. **Regresión logística:** La regresión logística es un método estadístico que se utiliza comúnmente para predecir clases binarias. El resultado final es la probabilidad de que ocurra un evento binario utilizando una función logit. La función sigmoidea, también llamada función logística, proporciona una curva en forma de 'S' que puede tomar cualquier número de valor real y asignarlo a un valor entre 0 y 1. Un número que resulte como salida de esta función por encima de 0.5, por ejemplo, podría ser asignado al evento 1, y menor a este ser asignado al evento 0. A continuación se presenta gráficamente la función



Regresión Logística, función sigmoidea. (“Understanding Logistic Regression inPython" 2019)

2. **Random Forest:** Los bosques aleatorios son técnicas de ensamble basados en el concepto de árbol. Si se supone que la relación entre las variables predictoras y la respuesta se puede modelar correctamente con un árbol de decisión, es muy probable que dentro del proceso de bagging sigamos escogiendo las mismas variables para particionar las observaciones en todos los modelos. Esto conlleva a que todos los árboles no sean independientes uno de los otros porque tendrán los mismos nodos y valores, por lo tanto el promedio de los resultados será menos exitoso al tratar de reducir la varianza en el ensamblaje.

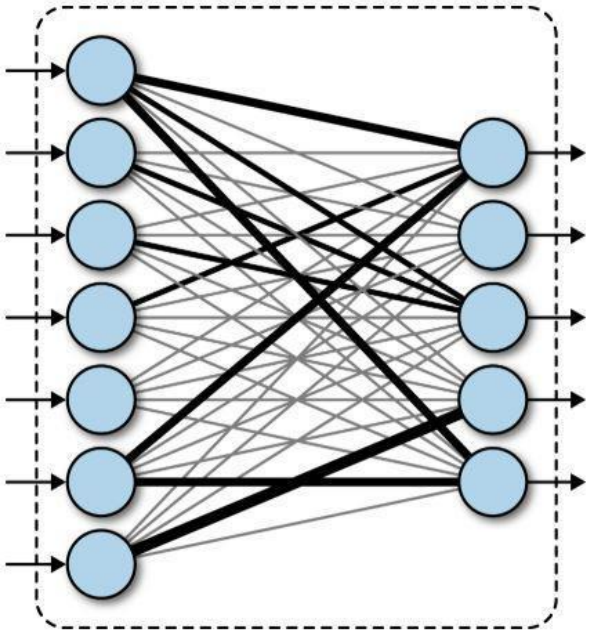


Gráfica de funcionamiento de bosques aleatorios. (Koehrsen, 2017)

3. **Gradient Boosting**: Esta técnica de aprendizaje automático es ampliamente utilizada para problemas de clasificación,como resultado se obtiene un modelo en forma de un conjunto de modelos de predicción débiles, normalmente se tratan árboles de decisión. GB construye este modelo de una forma escalonada al igual que otros métodos de boosting, y los generaliza permitiendo la optimización de una función de pérdida ("Gradient boosting", 2017)

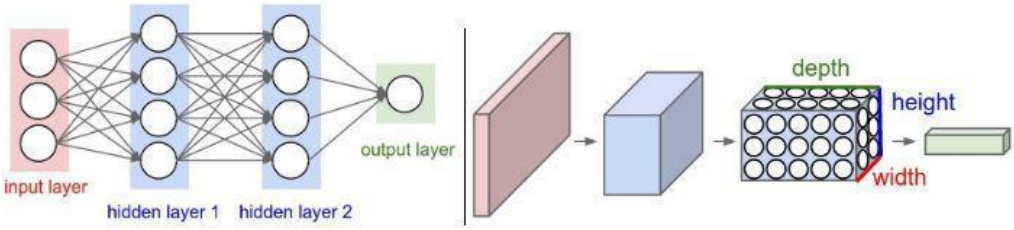
Ahora bien, vamos a enfocarnos en técnicas de modelación utilizando aprendizaje profundo, se presentarán tres modelos de redes neuronales las cuales son empleadas en el desarrollo de este proyecto:

1. **Red neuronal Fully Conected**: Este primer modelo se estudió y adaptó al problema con base en una primera asesoría del docente y un libro de jupyter que fue suministrado. Este tipo de red neuronal consta de una serie de capas completamente conectadas como se presentará una imagen a continuación. Los nodos en redes completamente conectadas se denominan comúnmente "neuronas". Este tipo de red se utilizan para miles de aplicaciones. La principal ventaja de estas redes es que son "independientes de la estructura". Es decir, no es necesario hacer suposiciones especiales sobre la entrada (por ejemplo, que la entrada consiste en imágenes o videos).



Red neuronal totalmente conectada. ("TensorFlow for Deep Learning," 2014)

2. **Red Neuronal Convolucional con arquitectura propuesta:** Las redes neuronales convolucionales suponen explícitamente que las entradas son imágenes, lo que permite codificar ciertas propiedades en la arquitectura. Esto hace que la función forward sea más eficiente de implementar y reduce enormemente la cantidad de parámetros en la red. Las redes neuronales convolucionales aprovechan el hecho de que la entrada consiste en imágenes y limitan la arquitectura de una manera más sensata. En particular, a diferencia de una red neuronal normal, las capas de un ConvNet tienen neuronas dispuestas en 3 dimensiones: ancho, alto, profundidad. A continuación.



Red neuronal convolucional. ("CS231n Convolutional Neural Networks for Visual Recognition," 2014)

3. **Red Neuronal Convolucional con arquitectura predefinida (Resnet de Keras):** ResNet , que es abreviatura de Residual Networks es una red neuronal clásica utilizada como columna vertebral para muchas tareas de visión por computadora. ResNet usa la conexión de omisión para agregar la salida de una capa anterior a una capa posterior. Esto ayuda a mitigar el problema del gradiente de fuga. Este modelo de red neuronal se carga desde Keras y se utilizará para el problema de clasificación en curso. ("Deep Residual Learning for Image Recognition," 2015)

3. Datos y análisis previo.

Utilizaremos información propia recolectada en campo en su mayoría en “VULCANO PIZZERIA” ubicada en la ciudad de Medellín, está compuesta por dos clases de pizzas (Pepperoni y Hawaina).

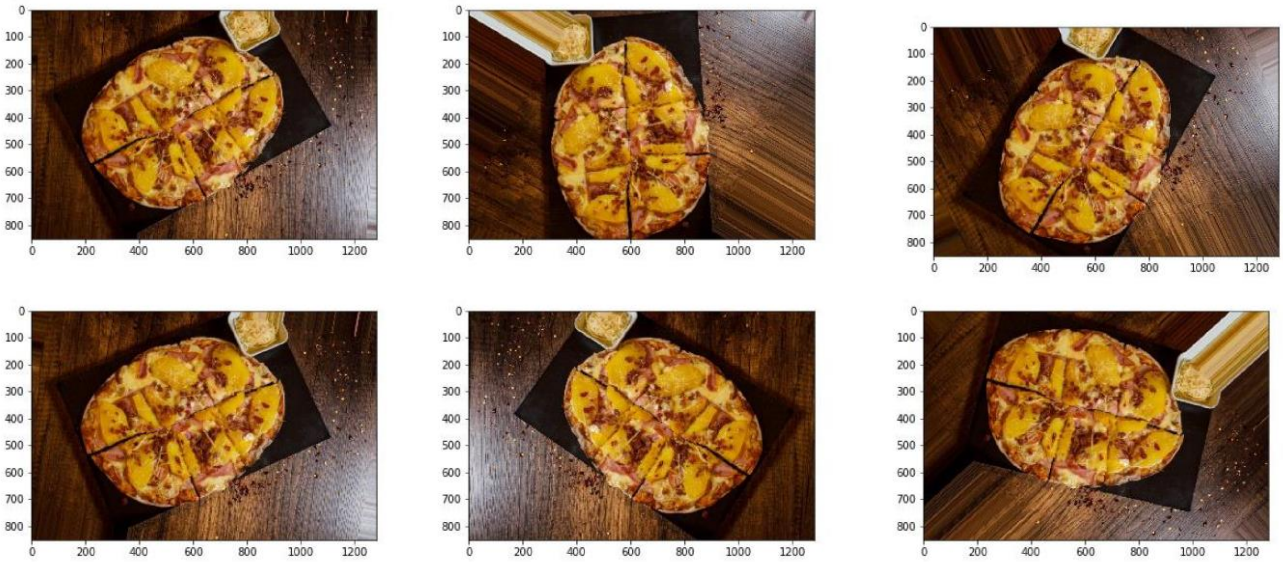
Los datos iniciales son aproximadamente 100 imágenes únicas desestructuradas que en esencia lucen de la siguiente forma:

Preprocesamiento y generación de los datos

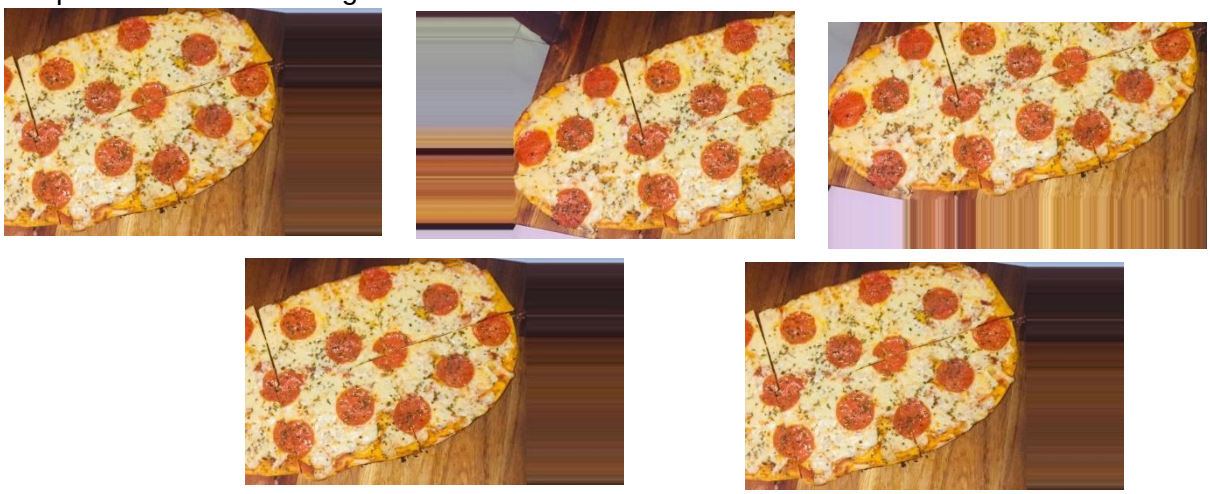
Con base a las imágenes anteriores, implementamos técnicas de generación de imágenes modificando las características inherentes a ellas como lo son:

Zoom

Rotación de la imagen



Desplazamiento de la imagen



Luminosidad

Con la función descrita, se generan 2.868 imágenes más (una tasa de crecimiento de datos del 2.800% aproximadamente) considerando la futura clasificación entre los datos de entrenamiento, validación y testeo del modelo.

Resultados preliminares

En este anteproyecto, se presentará los resultados preliminares de los modelos de aprendizaje estadístico o de máquinas, con algunas consideraciones, la primera de ellas comprendiendo que este tipo de problemas de clasificación de imágenes representa una alta complejidad debido al entorno que se exponen este tipo de problemas, no obstante, se han venido estudiando mejores configuraciones y arquitecturas de los modelos de redes neuronales profundas, y las investigaciones han venido avanzando. Este sin duda es un mundo con una amplia gama de investigación, por su complejidad al seleccionar las variables de entradas (inputs), capas ocultas, números neuronas por capa, funciones de activación por neurona, y más profundamente hablando, la función de optimización o funciones, dependiendo del problema.

Presentamos algunos resultados preliminares de los modelos de aprendizaje estadístico, con varias anotaciones:

- Los resultados presentados a continuación no son el desempeño final, puesto que se ha tenido dificultad para la ingesta de todas las imágenes al Google Colab, estos resultados son obtenidos utilizando aproximadamente 150 imágenes de cada pizza (utilizando las generadas con el data augmentation), es decir, un total de 305 imágenes para ser puntuales. Se está explorando diferentes opciones para poder realizar el entrenamiento de los modelos utilizando las casi 1400 imágenes que se tienen por cada tipo de pizza. Por eso estos resultados son preliminares.

- Los resultados preliminares que se presentan a continuación aún tienen pendiente el proceso de optimización de hiper-parametros, estos están ya codificados y probados, pero no se realizó esta optimización debido a que no es un resultado final.

- Se revisará que los desempeños en validación y test sean coherentes, en este anteproyecto y utilizando toda la metodología descrita anteriormente los resultados en desempeño dan un poco mayor en testeo, esto es una tarea futura a revisar y se espera agendar espacio de asesoría para entender si este comportamiento es normal y como arreglar en caso de que no lo sea.

- La partición de los datos para el entrenamiento de los modelos se realiza en los siguiente 3 conjuntos y porcentajes:

    - Train (60%)
    - Validation (20%)
    - Test (20%)

El primer método que se implementó fue el modelo de regresión logística. Este modelo de regresión se utilizó con los parámetros por defecto que trae la librería, para posteriormente

realizar procesos de optimización de hiper-parámetros (que no se incluye en esta entrega de anteproyecto). Los resultados en accuracy son los siguientes

- El desempeño de la regresión logística en validación es de 0.81
- El desempeño de la regresión logística en test es de 0.85

El segundo método que se implementó fue un random forest. Al igual que para el primer modelo, la primera corrida se utilizó con los parámetros por defecto que trae la librería, para posteriormente realizar procesos de optimización de hiperparametros (que no se incluye en esta entrega de anteproyecto). Los resultados en accuracy son los siguientes

- El desempeño del random forest en validación es de 0.80
- El desempeño del random forest en test es de 0.88

El último método que se implementó fue un GradientBoosting. Al igual que los anteriores modelos, la primera corrida se utilizó con los parámetros por defecto que trae la librería, para posteriormente realizar procesos de optimización de hiperparametros (que no se incluye en esta entrega de anteproyecto). Los resultados en accuracy son los siguientes

- El desempeño del random forest en validación es de 0.86
- El desempeño del random forest en test es de 0.93

Ahora vamos a presentar una comparación del desempeño preliminar de los modelos con otras métricas de clasificación binaria:

	Precisión	Recall	F1-Score
Pizza 1	0,88	0,73	0,8
Pizza 2	0,78	0,9	0,84
Accuracy			0,82

Regresión logística sin optimización de hiper-parametros

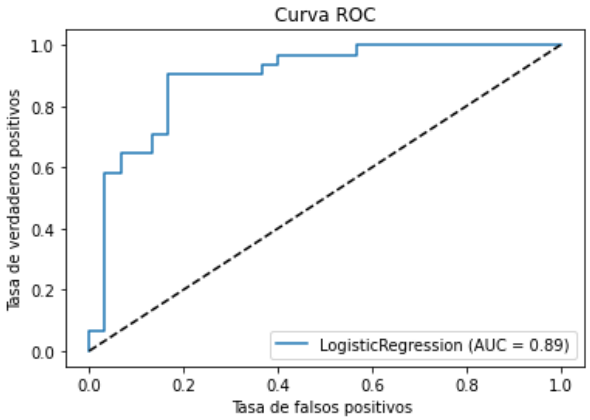
	Precisión	Recall	F1-Score
Pizza 1	0,85	0,73	0,79
Pizza 2	0,77	0,87	0,82
Accuracy			0,8

Random Forest sin optimización de hiper-parametros

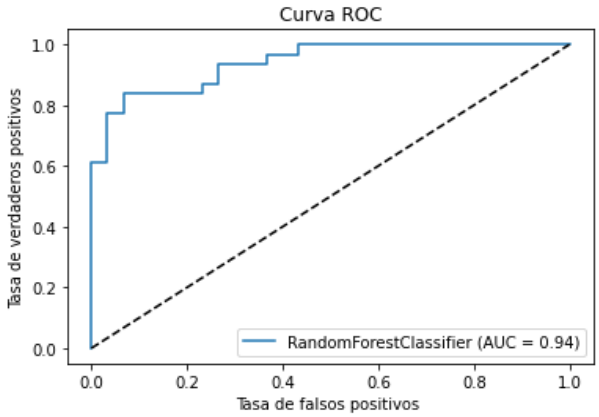
	Precisión	Recall	F1-Score
Pizza 1	0,87	0,87	0,87
Pizza 2	0,87	0,87	0,87
Accuracy			0,87

GradientBoosting sin optimización de hiper-parametros

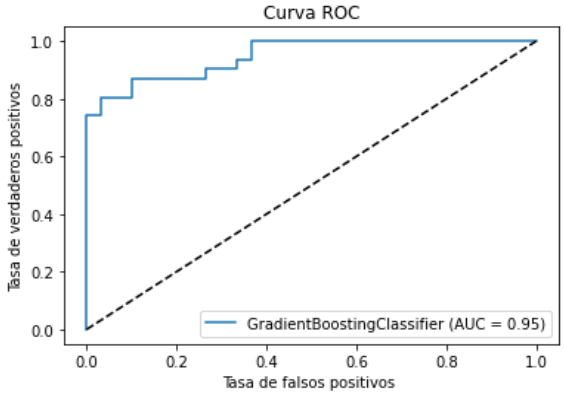
Ahora vamos a presentar las curvas ROC y el AUC para los tres modelos clásicos, estos son resultados preliminares, teniendo en cuenta las consideraciones expuestas anteriormente:



Regresión logística sin optimización de hiper-parametros



Random Forest sin optimización de hiper-parametros



GradientBoosting sin optimización de hiper-parametros

Se observa entonces que los tres métodos tienen desempeños parcialmente parecidos, siendo por supuesto el gradient boosting el que mejor AUC presenta. Sin embargo, estos resultados no reflejan completa y finalmente el desempeño de los modelos, ya que no se ha utilizado todas las imágenes disponibles debido a un problema de procesamiento que se presentó en Google Colab, esto se está estudiando para lograr procesar y entrenar con todos los datos disponibles y tener resultados sobre los cuales se puedan concluir más certeramente. En los libros de jupyter de esta modelación ya se tiene programada para los modelos las fases de instancia, entrenamiento, y optimización de sus hiper-parámetros.

Así mismo, en el repositorio que se expone al final de este anteproyecto, se encuentra el libro de jupyter de redes profundas, donde los modelos 1 y 3 han sido construidos, probados en pequeños datos, pero que no se presentan en este anteproyecto ya que son resultados muy preliminares. El modelo 2 que es la construcción de una arquitectura desde cero, en este momento se encuentra en una etapa de estudio y diseño, por lo tanto, se excluye en la presentación de resultados preliminares.

4. (15%) Plan (diagrama Gantt o Pert)

A continuación, se presenta el detalle de las actividades que deberán ser llevadas a cabo para la finalización satisfactoria de este proyecto:

Task Name	Duración	Comienzo	Fin	% Línea Base	% completado	% Var	Predecesoras
Cronograma Proyecto Clasificación de Imágenes	**69 días**	**13/02/2020 8:00**	**19/05/2020 17:00**	**44%**	**60%**	**-16%**	
Captura de imágenes	51 días	13/02/2020 8:00	23/04/2020 17:00	53%	70%	-17%	
Asignación de label	51 días	13/02/2020 8:00	23/04/2020 17:00	53%	70%	-17%	
Lectura y estructuración de pipeline	5 días	20/02/2020 8:00	26/02/2020 17:00	100%	100%	0%	2CC+5 días
Data preprocessing	56 días	20/02/2020 8:00	7/05/2020 17:00	39%	60%	-21%	3CC+5 días
Modelación y calibración Machine Learning	26 días	24/02/2020 8:00	30/03/2020 17:00	77%	90%	-13%	5CC+2 días
Modelación y calibración Deep Learning	33 días	23/03/2020 8:00	6/05/2020 17:00	0%	15%	-15%	5CC+22 días
Hto: Entrega Anteproyecto	0 días	23/03/2020 8:00	23/03/2020 8:00	100%	100%	0%	6FC-6 días
Análisis de resultados	4 días	7/05/2020 8:00	12/05/2020 17:00	0%	0%	0%	6;7
Conclusiones	2 días	13/05/2020 8:00	14/05/2020 17:00	0%	0%	0%	9
Entrega final	3 días	15/05/2020 8:00	19/05/2020 17:00	0%	0%	0%	10;4;3;6;7;5;2
Entrega Documento Fina	0 días	19/05/2020 17:00	19/05/2020 17:00	0%	0%	0%	11;8

Se realiza una rápida descripción de estas actividades en el proyecto en curso:

- **Captura de imágenes:** Las imágenes de las pizzas se toman de manera manual, por lo cual es un proceso operativo que inició desde el comienzo del proyecto y se extiende a lo largo de casi 2 meses y medio. La dinámica es que se va entrenando modelos y luego se vuelven a entrenar con más imágenes.
- **Asignación de label:** Esta actividad consiste en asignar qué tipo de pizza es cada imagen, también es manual y se va realizando a medida que se obtengan más

imágenes. Para esto se desarrolló un script que nos permite automatizar parte del proceso.

- **Lectura y estructuración de pipeline:** Esta actividad consiste en construir el proceso de carga e ingesta de información hacia el ambiente donde se van a construir los modelos de ML y DL.

- **Data preprocessing**: En este proyecto, se entiende la data preprocessing como un proceso iterativo que se va ajustando a medida que se entrenan los modelos, es decir, se evalúa el desempeño de los modelos y en caso de que sea necesario se realizará transformaciones y ajustes los datos, buscando con ello mejorar las métricas de clasificación seleccionadas.

- **Modelación y calibración Machine Learning:** En este punto, se abordan tres modelos de aprendizaje de máquinas, durante esta etapa se realiza la calibración de parámetros de cada uno de los modelos seleccionados y se presentan la evolución en desempeño respecto al entrenamiento inicial.

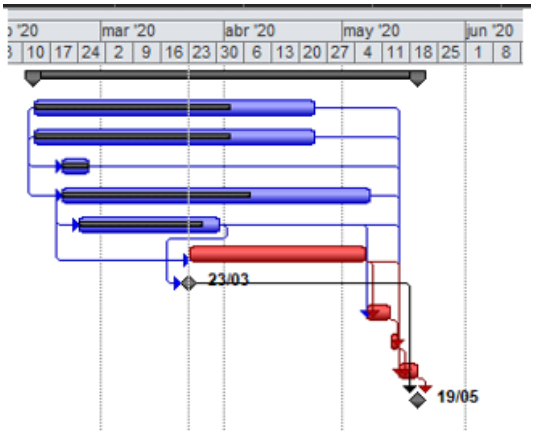
- **Modelación y calibración Deep Learning:** Esta etapa, es sin duda la más retadora e importante del proyecto, pues es donde se aplican técnicas de aprendizaje profundo para la clasificación de imágenes. Se proponen tres modelos al igual que en la etapa pasada, entendiendo que en esta la complejidad en la estructuración de las arquitecturas y el tiempo de ejecución es superior. Hay que mencionar también que requiere de un estudio de este tipo de técnicas y lectura constante para la configuración de los parámetros.

- **Análisis de resultados:** Finalizando las etapas de modelación y revisión conjunta de las métricas, se realiza un análisis de los resultados que se presentaron en el proyecto durante las etapas de modelación, desde el enfoque de ML como el DL. En esta etapa se da estructura al informe y presentación final.

- **Conclusiones**: Luego de realizar un análisis detallado de las técnicas y resultados, y teniendo el documento de presentación final avanzado, se concluye sobre la pregunta de investigación que se abordó en este proyecto y cuáles son los hallazgos e hitos más representativos del trabajo.

- **Entrega Final:** Se presenta el proyecto en clase con los compañeros y profesor, se realiza entrega de documento final.

Avance al día de entrega del anteproyecto. 23 de marzo de 2020:



El proyecto hoy, 23 de marzo, presenta una holgura de siete días, es decir, el porcentaje completado de las actividades es superior al de la Línea Base del proyecto (60% completado vs 44% esperado), esto se explica dado el avance que se tiene desde 3 frentes: El data preprocessing, avance significativo en la modelación de machine learning y un avance en la estructuración de las arquitecturas de las redes neuronales profundas.

Es de anotar que la fase más importante precisamente es la modelación de las redes profundas, por lo cual, estos días de holgura se pueden perfectamente ser empleados cuando se tengan que calibrar parámetros y ajustar las arquitecturas que se están entrenando. Como se mencionó anteriormente, el preprocesamiento de los datos es un proceso iterativo que va acompañando las dos etapas de la modelación, por lo cual se puede requerir más adelante realizar tareas en este sentido para ajustar la modelación en Deep Learning.

En conclusión, se han logrado avances significativos en el proyecto, actualmente se está trabajando en la realización de las tareas en tiempo empleado, pero es importante tener presente la etapa que se viene en la modelación de Deep Learning ya que podría consumir parte del tiempo de holgura o incluso agotarlo hasta requerir más.

5. Implicaciones éticas:

La existencia de los controles está ligada al origen del hombre, atribuible principalmente a la ausencia de confianza, dichos controles han significado durante todo el tiempo la vulneración de la privacidad y en algunos casos, señales de “irrespeto” hacia los controlados. En este caso, se ha buscado cambiar la forma en la cual se revisan las ventas y producción de la pizzería utilizando técnicas de menor fricción, pese a ello, la mera existencia de una cámara allí implica ruptura de confianza por parte de los empleados y la sensación de poca privacidad dentro del negocio. Considerando una evolución avanzada del proyecto se tendrían impactos en el número de empleos (o al menos en el rol que se desempeña) dentro de la pizzería, pues ya no sería necesario contratar a una persona que lidere el consumo y abastecimiento de la materia prima.

6. Aspectos legales y comerciales:

Uno de los grandes problemas que encuentran los restaurantes durante su fase de crecimiento es la pérdida de control sobre los inventarios y dada la naturaleza del negocio, constantemente se presentan situaciones que dan lugar a malos manejos de la materia prima o la contabilización de las ventas. Esta solución apunta directamente a disminuir dichas situaciones de manera muy económica (baja inversión) y basado en su gran componente tecnológico se presenta como un servicio capaz de ser replicado en cualquier negocio de alimentos sin importar su tamaño.

Desde el ámbito legal no existen problemas para la puesta en producción pues los datos son información propia que se recopila en cada uno de los establecimientos que requieran implementar la solución. Esto, sin dejar de lado que se precisa registrar legalmente la actividad comercial y proteger la propiedad intelectual.

7. Retos y trabajo futuro

Durante este trabajo se ha evidenciado que la clasificación de imágenes es un problema con retos en modelación interesantes, desde la definición de los modelos, la construcción de las arquitecturas, hasta el procesamiento y capacidad de cómputo.

Dentro de las dificultades a las que enfrentadas es la capacidad de almacenamiento y cómputo para entrenar los modelos, incluyendo la gran cantidad de tiempo que se emplea cada que se planea realizar una corrida de un modelo para ver resultados preliminares y hacer ajustes, esto sin duda es un desafío si se quisiera aumentar el alcance de este proyecto como un modelo de negocio que se encuentre en productivo y que pueda ser escalable.

Pese a que existen varias actividades pendientes se destacan las siguientes por su relevancia: Capacidad de realizar el entrenamiento de los modelos con todas las imágenes (buscando una solución cloud o un equipo de alto rendimiento), volver a entrenar los modelos clásicos de ML y realizar optimización de hiperparametros, y la configuración y entrenamiento de los modelos 1 y 3 de deep learning y la construcción de la arquitectura del modelo 2. En este punto se realizará la comparación de todos los métodos.

8. Repositorio de Github.

Se trabajó con libros de jupyter en Google Colab y se crearon copias al siguiente repositorio de Github:

https://github.com/eljimenezj/CM0868--Estadistica-Multivariada-Avanzada

Carpeta con las imágenes:
https://drive.google.com/drive/folders/1p3bgn9OOU0alH9KHvfPfj_4mwOTrZuXL?usp=sharing

9. Bibliografía

(Tutorial) Understanding Logistic REGRESSION in PYTHON. (2019, December 16). DataCamp Community. https://www.datacamp.com/community/tutorials/understanding-logistic-regression-python.

CS231n Convolutional Neural Networks for Visual Recognition. (2014). CS231n Convolutional Neural Networks for Visual Recognition. https://cs231n.github.io/convolutional-networks/

Deep Residual Learning for Image Recognition. (2015). arXiv.org. https://arxiv.org/abs/1512.03385
Gradient boosting. (2017, September 29). Wikipedia, la enciclopedia libre. Retrieved March 23, 2020, from https://es.wikipedia.org/wiki/Gradient_boosting

Koehrsen, W. (2017, December 27). *Random Forest Simple Explanation*. Medium. https://medium.com/@williamkoehrsen/random-forest-simple-explanation-377895a60d2d

TensorFlow for Deep Learning. (2014.). O'Reilly Online Learning. https://www.oreilly.com/library/view/tensorflow-for-deep/9781491980446/ch04.html

Towards an Improved ASUM-DM Process Methodology for Cross-Disciplinary Multi-organization Big Data & Analytics Projects. (2018). SpringerLink. https://link.springer.com/chapter/10.1007/978-3-319-95204-8_51