Informe Ejecutivo – Ciencia de datos aplicada: Insights para el Taller 2

Emil Rueda, cod: 202415281
Camilo Morillo, cod: 202015224

1. **Resumen**: El objetivo del proyecto era diseñar un modelo de predicción automatizado para el supermercado inteligente que permita registrar productos en tiempo real, reduciendo costos y mejorando la experiencia del cliente. Los resultados obtenidos se analizaron en términos de precisión del modelo y generación de valor económico.
2. **Insights**:
    a. El conjunto de datos original era pequeño (2,640 imágenes), por lo que se aplicaron técnicas de aumento de datos, generando un conjunto final de 7,920 imágenes.
    b. El conjunto de datos estaba desbalanceado, se observó que algunas clases como manzanas, jugos o leche tienen una mayor cantidad de datos en contra posición de Jengibre, espárragos o papayas, lo que podría haber afectado la precisión de los modelos. Esto se puede observar en la figura 1.

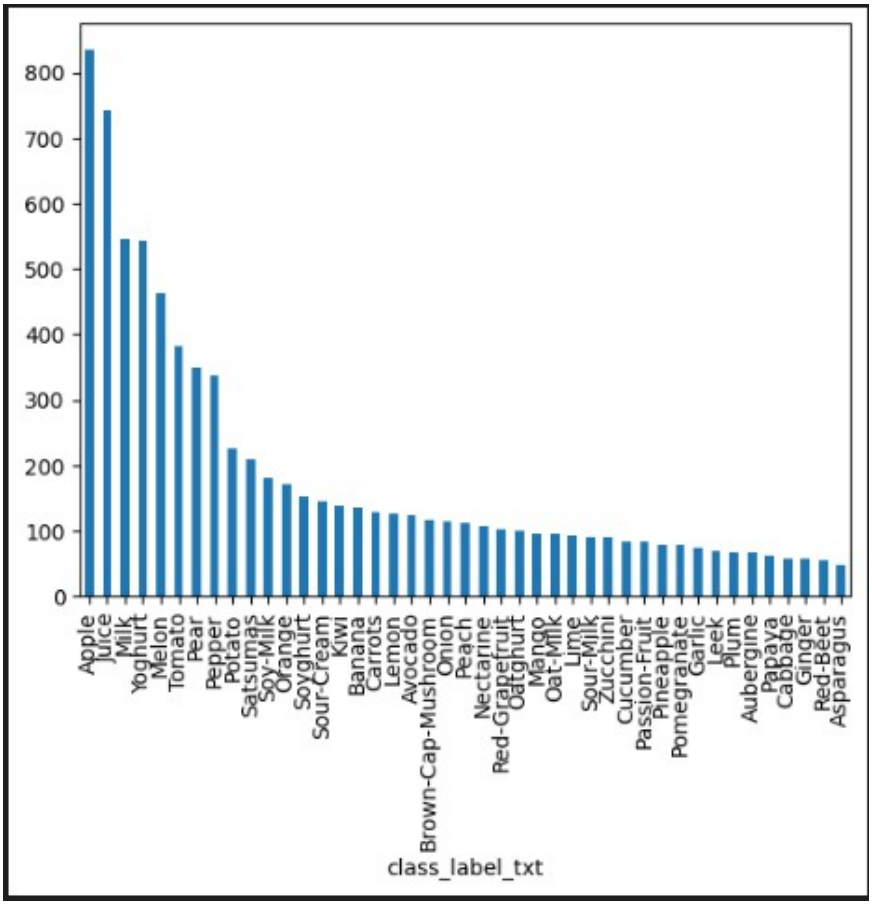

Figura 1. Distribución de imágenes por clase.

    c. Redes neuronales convolucionales (CNN): Precisión de validación del 12%. Se puede observar que el modelo generalizan de una manera más adecuada para las clases con porcentajes altos de datos en el conjunto de entrenamiento, sin embargo, para las clases con pocas imágenes no lo hace bien.
    d. Random Forest: Precisión de validación del 24%. Se repite el mismo patrón de desbalance mencionado anteriormente, lo que genera errores en la generalización.
    e. Errores del modelo: Los cálculos estimaron un costo diario de $18,500 USD debido a errores en el registro automatizado.
    f. Registro manual: Los costos actuales asociados al registro manual son significativamente menores.
    g. El ROI calculado fue de -1176%, indicando una pérdida considerable si se implementa el modelo actual.
3. **Analisis de Viabilidad**: Los resultados muestran que: la baja precisión de los modelos no permite alcanzar los niveles de ahorro esperados, los costos diarios de los errores del modelo superan ampliamente los costos del registro manual y el conjunto de datos desbalanceado y pequeño afecta negativamente el rendimiento de los modelos.
4. **Recomendaciones**:
    a. No implementar el modelo actual: Dados los altos costos de error y el ROI negativo, implementar este modelo sería económicamente inviable.
    b. Mejorar la precisión del modelo: incrementar el tamaño del conjunto de datos recolectando más imágenes reales y generando una mayor cantidad de datos aumentados con diferentes técnicas como rotaciones o traslaciones. Realizar un pre-procesamiento exhaustivo de las imágenes para eliminar fondos, imágenes con varias clases en la misma foto e incluso considerar la eliminación de imágenes con varios elementos pues esto puede afectar la capacidad de generalización del modelo. Aplicar técnicas para balancear las clases dentro del conjunto de datos. Por último, se recomienda experimentar con modelos más complejos como redes neuronales profundas pre-entrenadas (e.g., ResNet, EfficientNet).
    c. Evaluar alternativas: Recolectar una mayor cantidad de datos, hacer un preprocesamiento más robusto, generar aumentación de datos con técnicas más complejas y realizar una prueba piloto con modelos pre-entrenados para mejorar la precisión antes de considerar una implementación a gran escala y continuar utilizando el registro manual mientras se desarrollan modelos más robustos.
5. **Conclusión**: El modelo actual no es adecuado para reemplazar el sistema de registro manual debido a su baja precisión y los costos asociados a los errores. Sin embargo, mejorar el conjunto de datos y probar modelos avanzados pueden ser estrategias clave para alcanzar el objetivo de automatización en el futuro. Se recomienda no proceder con la implementación hasta alcanzar un modelo con una precisión y un ROI que justifiquen su uso.