



# Elevando la satisfacción con IA

Informe de clasificación



## Índice de contenido

Introducción.....	3
Curva ROC-AUC de los modelos.....	4
Curva ROC.....	4
AUC (Área Bajo la Curva).....	4
Gráficas.....	5
Elección del modelo.....	9
Conclusiones.....	11
Conclusión General.....	12



# Introducción

En el presente informe, se analizará el rendimiento de un modelo de machine learning desarrollado para predecir la satisfacción de los pasajeros en vuelos comerciales. Este modelo tiene como objetivo predecir la satisfacción del cliente en función de las respuestas ingresadas en el test de satisfacción proporcionado por F5 Airlines.

- Se entrenan y validan los siguientes modelos con validación cruzada donde previamente se han buscado los hiperparámetros para ajustar la ejecución de los modelos: Random Forest, KNN, Logistic Regression, SVM.
- En los casos de Random Forest, SVM y Logistic Regression, para agilizar la búsqueda de los hiperparámetros, se realiza dicha búsqueda con una muestra de 10000 casos. Luego se entrena y valida el modelo con el sistema de validación cruzada con el 100% de los datos. La parte de validación se hace con el 20% de dicha muestra.
- En todos los casos se aplica la validación cruzada con cinco "vueltas" (fold) excepto en el caso del modelo de SVM donde se realizan sólo tres para agilizar su ejecución.
- También se entrena y valida un modelo que ensambla los modelos random forest, logistic regression y KNN y que llamamos (EM)





# Curva ROC-AUC de los modelos

## Curva ROC

La curva ROC es un gráfico que muestra el desempeño de un modelo de clasificación binaria a medida que se varía el umbral de decisión. En el eje X, se grafica la **Tasa de Falsos Positivos (FPR)**, y en el eje Y, la **Tasa de Verdaderos Positivos (TPR)** o sensibilidad.

- **Tasa de Falsos Positivos (FPR):** Es el cociente entre el número de falsos positivos y el total de verdaderos negativos. Es decir, la proporción de casos negativos que el modelo clasifica incorrectamente como positivos.
- **Tasa de Verdaderos Positivos (TPR)** (Sensibilidad o Recall): Es el cociente entre el número de verdaderos positivos y el total de verdaderos positivos más falsos negativos. Representa la proporción de casos positivos que el modelo clasifica correctamente.
- La curva ROC traza el TPR contra el FPR para diferentes umbrales de clasificación, lo que permite visualizar cómo cambia el rendimiento del modelo según el umbral.

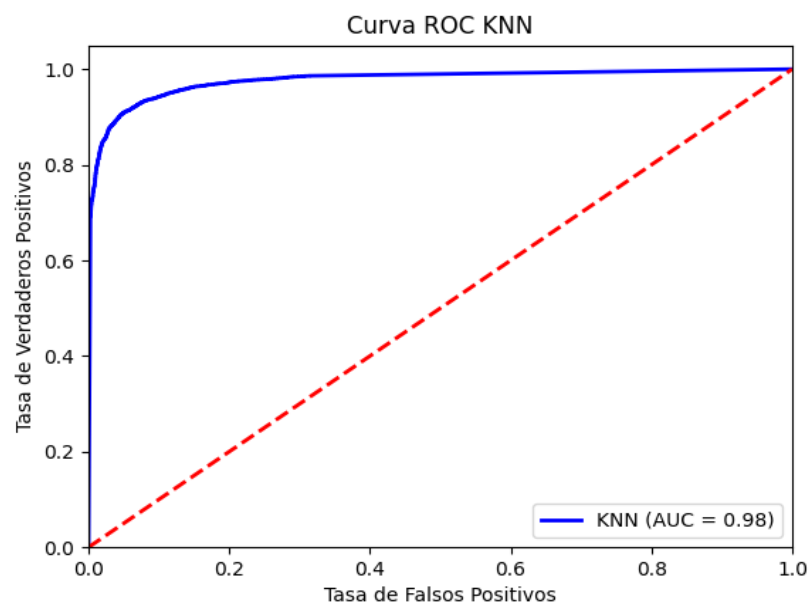
## AUC (Área Bajo la Curva)

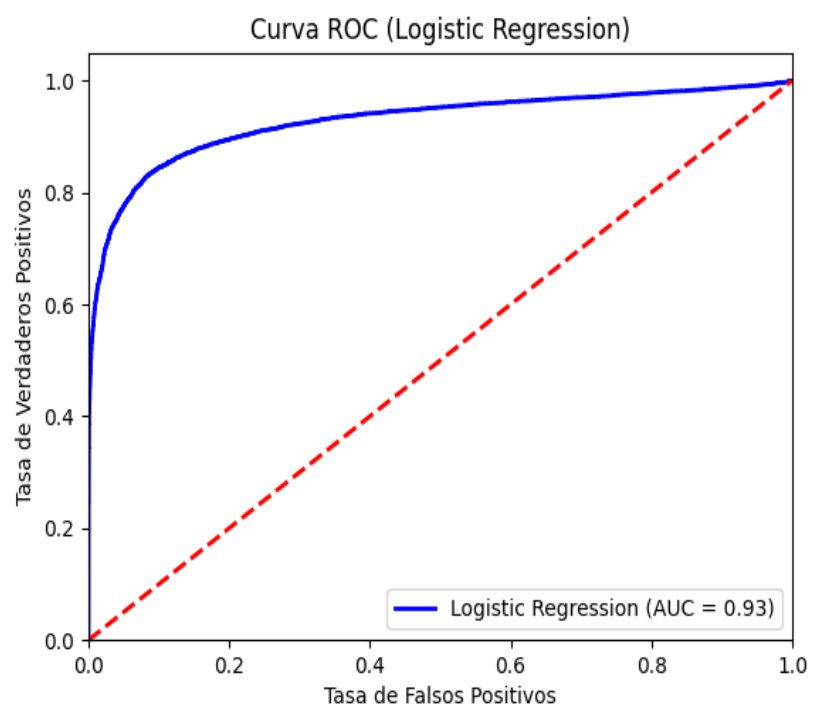
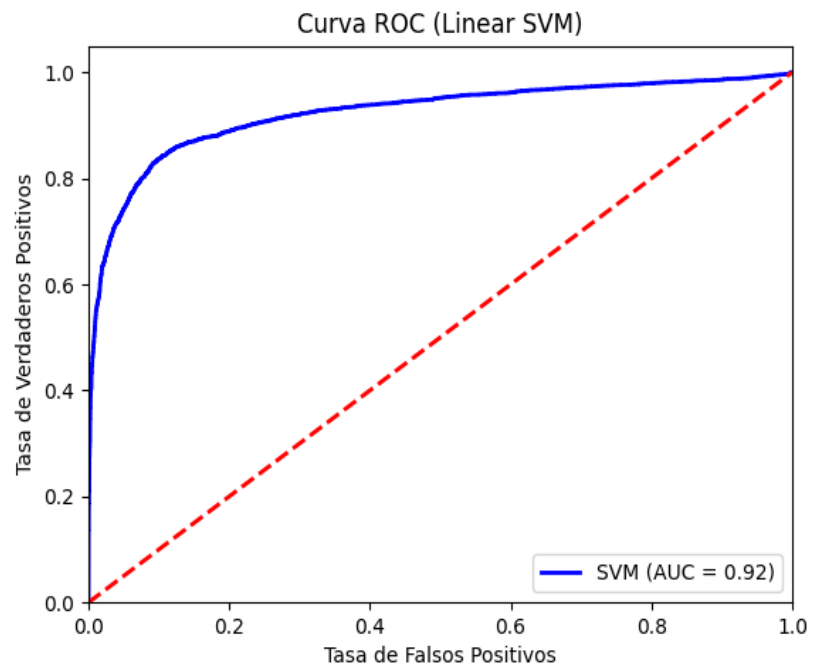
El **AUC (Area Under the Curve)** es un valor numérico que mide el área bajo la curva ROC. Este valor oscila entre 0 y 1, y refleja la capacidad general del modelo para separar las clases positivas y negativas. Interpretación del AUC:

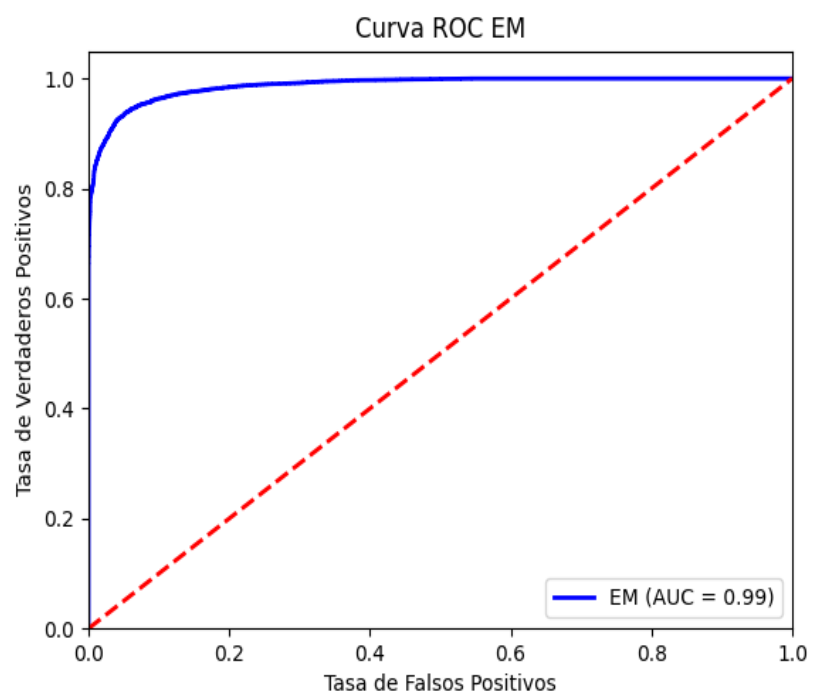
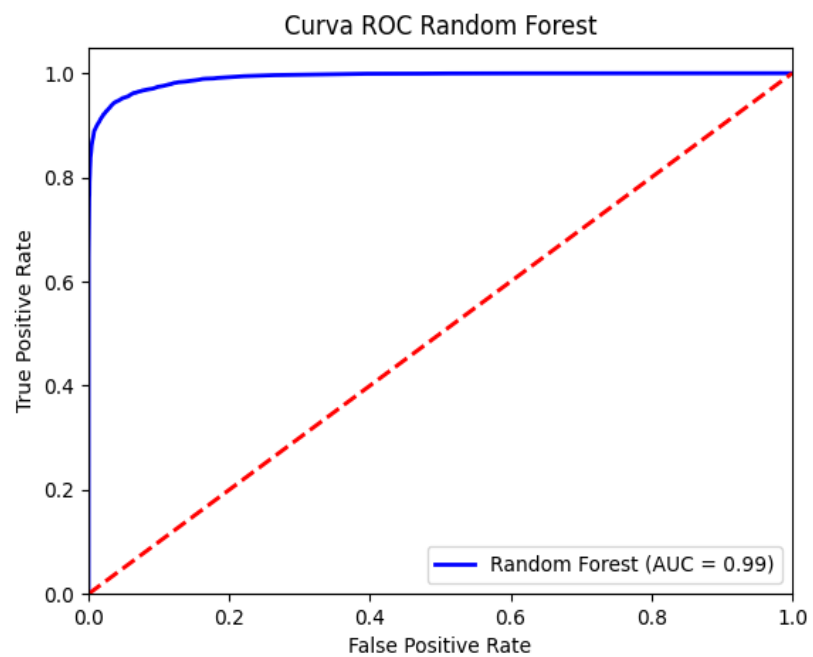
- **AUC = 1:** El modelo tiene una capacidad perfecta para clasificar correctamente todas las clases.

- **AUC = 0.5:** El modelo no tiene capacidad de discriminación; es equivalente a hacer una clasificación aleatoria.
- **AUC < 0.5:** El modelo tiene un rendimiento peor que la clasificación aleatoria, lo que indica que está clasificando las clases al revés (las negativas como positivas y viceversa).

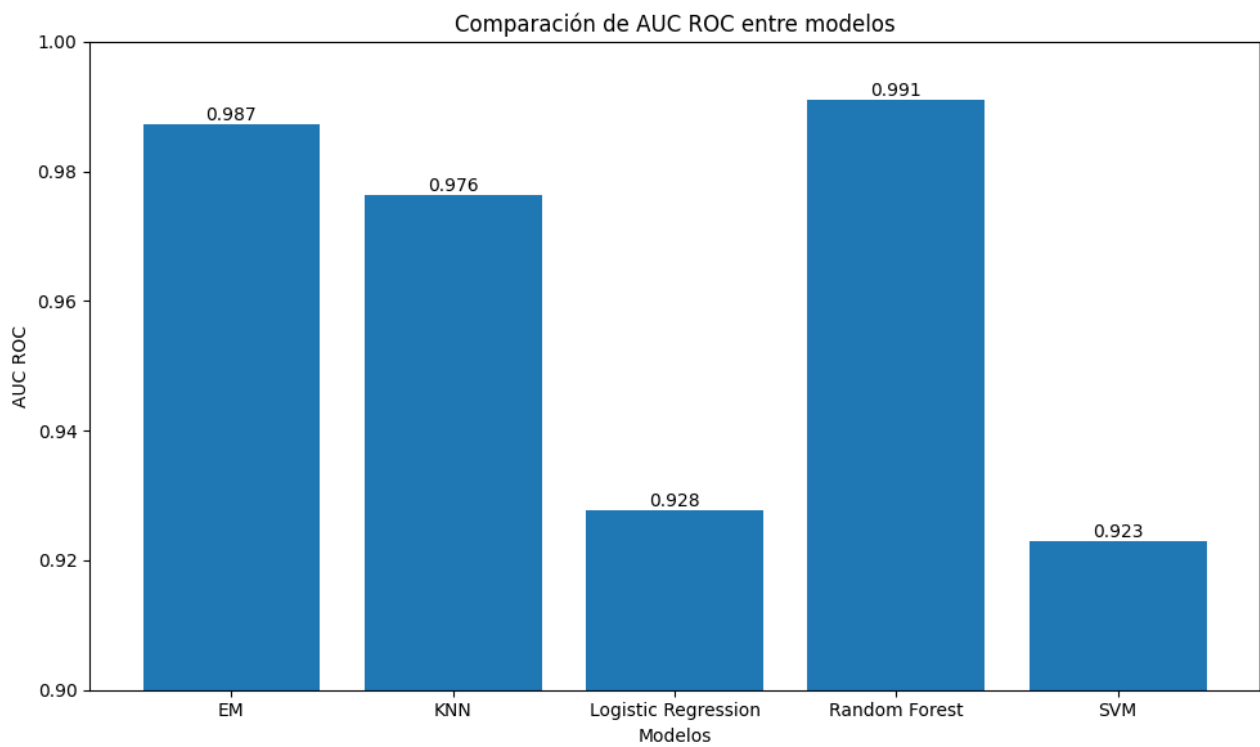
## Gráficas





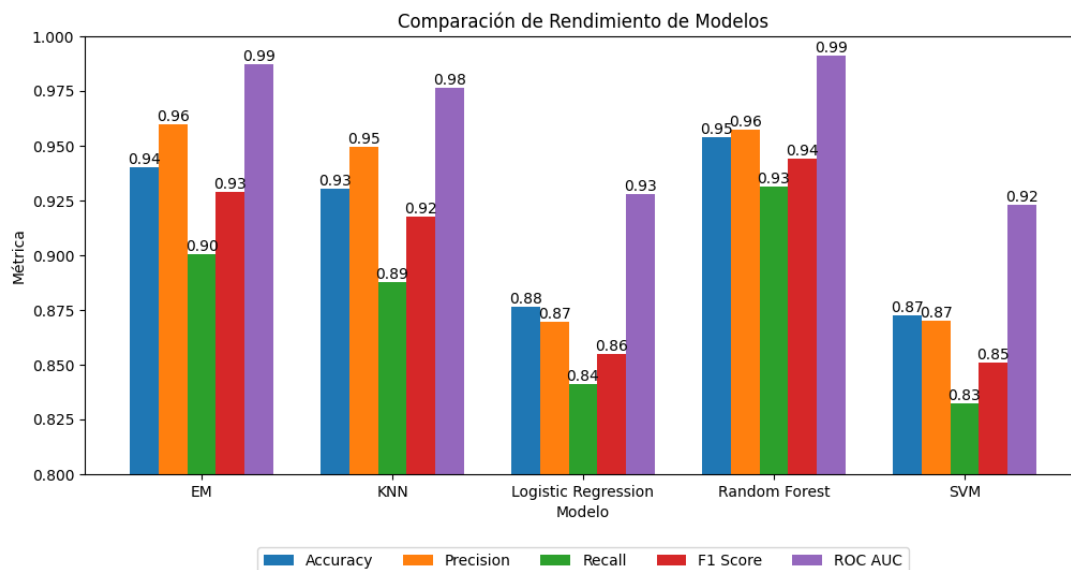


En todos los casos los valores de la curva roc-auc son muy altos. Especialmente en el modelo Random Forest.





# Elección del modelo



Nos inclinamos por utilizar el modelo de Random Forest para una aplicación predictiva, por los siguientes motivos:

## 1. Rendimiento superior:

- El Random Forest tiene el valor más alto de **Accuracy (0.9537)**, lo que significa que clasifica correctamente el 95.37% de todas las instancias.
- Tiene el mejor ROC AUC (0.9909), indicando una excelente capacidad para distinguir entre las clases.
- Muestra el F1 Score más alto (0.9451), lo que refleja un buen equilibrio entre precisión y exhaustividad.

## 2. Equilibrio entre Precision y Recall:

- El Random Forest tiene la **Precision más alta (0.9573)** y el **segundo Recall más alto (0.9311)**, lo que indica un buen equilibrio entre falsos positivos y falsos



negativos.

### 3. Robustez y generalización:

- Los Random Forests son conocidos por su capacidad de manejar datos complejos y no lineales, así como por su resistencia al sobreajuste.
- Los hiperparámetros seleccionados (300 árboles, profundidad ilimitada) sugieren un modelo complejo pero bien ajustado.

### 4. Comparación con otros modelos:

- Supera consistentemente a los otros modelos en todas las métricas, excepto en Precision, donde está muy cerca del mejor (EM).
- Aunque el modelo EM tiene una Precision ligeramente superior, el Random Forest lo supera en todas las demás métricas, especialmente en Recall y F1 Score.

### 5. Interpretabilidad y flexibilidad:

- Aunque no es tan interpretable como un árbol de decisión simple, los Random Forests ofrecen medidas de importancia de características que pueden ser útiles para entender el modelo.
- Son flexibles y pueden manejar tanto variables numéricas como categóricas sin necesidad de escalado.

### 6. Consideraciones prácticas:

- Los Random Forests suelen ser más rápidos en la predicción que modelos como SVM, lo cual es importante para aplicaciones en tiempo real.
- Son relativamente fáciles de implementar y mantener en producción.



# Conclusiones

El análisis comparativo de los modelos de **machine learning** revela que el **Random Forest** es el modelo que ofrece el mejor rendimiento general para predecir la satisfacción de los pasajeros. Este modelo ha superado a otros como **KNN**, **Logistic Regression**, y **SVM** en la mayoría de las métricas clave:

1. **Rendimiento superior:** El modelo **Random Forest** tiene el mayor **Accuracy** (95.37%), el mejor **ROC AUC** (0.9909), y el más alto **F1 Score** (0.9451), demostrando su capacidad para clasificar con precisión tanto las clases positivas como negativas.
2. **Equilibrio entre Precision y Recall:** Con una **Precision** de 0.9573 y un **Recall** de 0.9311, el **Random Forest** mantiene un buen balance entre falsos positivos y negativos, lo que lo hace efectivo en escenarios donde ambas métricas son importantes.
3. **Robustez y Generalización:** **Random Forest** es conocido por manejar datos complejos y no lineales, mostrando resistencia al sobreajuste y adaptándose bien a diferentes tipos de datos sin necesidad de un preprocesamiento extenso.
4. **Comparación con otros modelos:** Aunque el modelo **EM (Ensamblado)** tiene una **Precision** ligeramente mayor, el **Random Forest** supera al EM en otras métricas como **Recall** y **F1 Score**, lo que lo convierte en una mejor opción general para predicción.
5. **Consideraciones prácticas:** **Random Forest** es más rápido en predicción en comparación con **SVM**, lo que lo hace adecuado para aplicaciones en tiempo real y escenarios donde el tiempo de respuesta es crucial.



## Conclusión General

El modelo **Random Forest** es la mejor opción para una aplicación predictiva debido a su excelente rendimiento en todas las métricas evaluadas, su capacidad de generalización, y su facilidad de implementación. A pesar de que otros modelos pueden ofrecer ventajas menores en áreas específicas, **Random Forest** se destaca por su robustez y versatilidad, lo que lo convierte en la opción preferida para esta tarea.