

# Predicción de popularidad de canciones K-Pop

Trabajo de Fin de Master

Master Universitario en Ciencia de Datos



Alumna: Paula de Jaime de Toro

Tutor: Sergio Trilles Oliver

Profesor: Albert Solé Ribalta

ENERO 2020

# Índice

1. Contexto y motivación
2. Objetivos
3. Datos
4. Modelos
5. Evaluación
6. Conclusión

# Contexto y motivación

- El K-pop es música pop surcoreana
- Producto importante de Corea del Sur
- Agencias de entretenimiento (SM, YG, JYP, etc.)
- Reclutamiento y entrenamiento de artistas
- *Hit Song Science y Music Information Retrieval*
- Interés personal y curiosidad
- No se han encontrado trabajos o proyectos iguales








# Objetivos

- **Predecir** si canciones surcoreanas serán populares en Corea del Sur
- **Investigar** el estado del arte para conocer los métodos que previamente se han usado en este tipo de problemas.
- Encontrar fuentes de **datos** y adquirir estos. Preprocesar datos y obtener un dataset con datos válidos.
- Generar **modelos** predictivos y ajustar estos para lograr mejores resultados.
- Comparar y **evaluar** la eficiencia de múltiples modelos predictivos.

# Datos: Fuentes de datos

- *Gaon Digital Chart*
- Septiembre 2020 hasta Enero 2010
- Granularidad mensual
- Top 25% y bottom 25%
- Spotify API
- Títulos en inglés o coreano
- Artistas en coreano, inglés o romanizado
- DBKpop

Title / Artist	가온지수	Production
 VVS (Feat. JUSTHIS) (Prod. GroovyRoom) 미란이 (Mirani) , 먼치맨 , Khundi Panda , ...	25,121,948	제작 Stone Music Entertainm... 유통 지니뮤직
 All I Want For Christmas Is You Mariah Carey   Merry Christmas (Deluxe ...	24,009,950	제작 Columbia/Legacy 유통 Sony Music
 밤하늘의 별들(2020) 경서   밤하늘의 별들(2020)	19,944,510	제작 공의엔진 유통 카카오 M
 Dynamite 방탄소년단   Dynamite	19,797,027	제작 빅히트 엔터테인먼트 유통 Dreamus
 Santa Tell Me Ariana Grande   Santa Tell Me	19,050,565	제작 Republic Records, a div... 유통 Universal Music

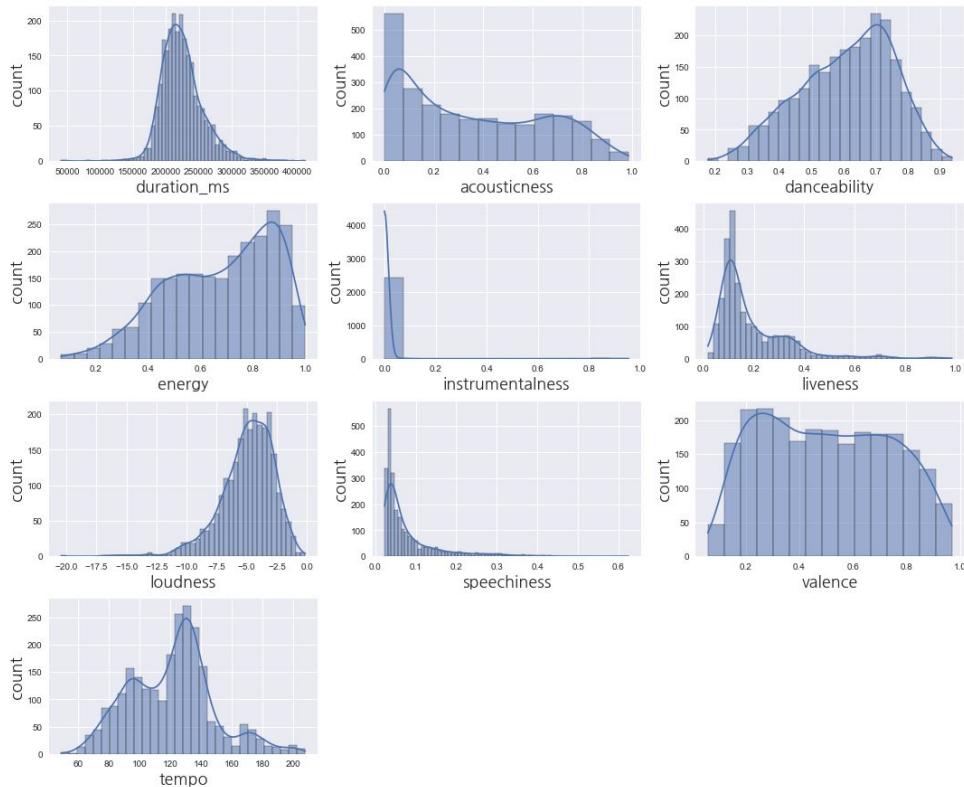
현 아 = Hyuna

## Datos II: Creación del *dataset*

- Problemas de codificación
- Limpieza de datos
  - Artistas internacionales
  - Colaboraciones
  - Bandas sonoras
  - Programas de televisión
  - Transformar el artista en inglés o romanizado
- Se aprende a leer coreano
- Spotify API + Google
- 67% canciones no encontradas → manualmente
- 3629 a 3070 canciones

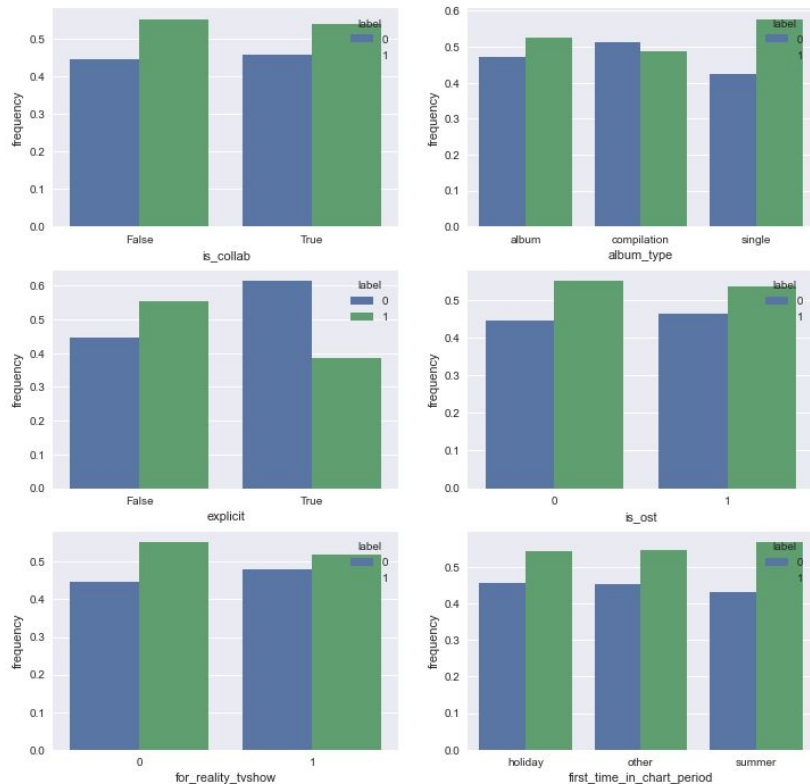
title	singer	album
8만원 (2015. 3.)	블랙넛(black nut)	II II II

# Datos III: Análisis



- QQ-Plot
- Shapiro-Wilk Test

## Datos III: Análisis

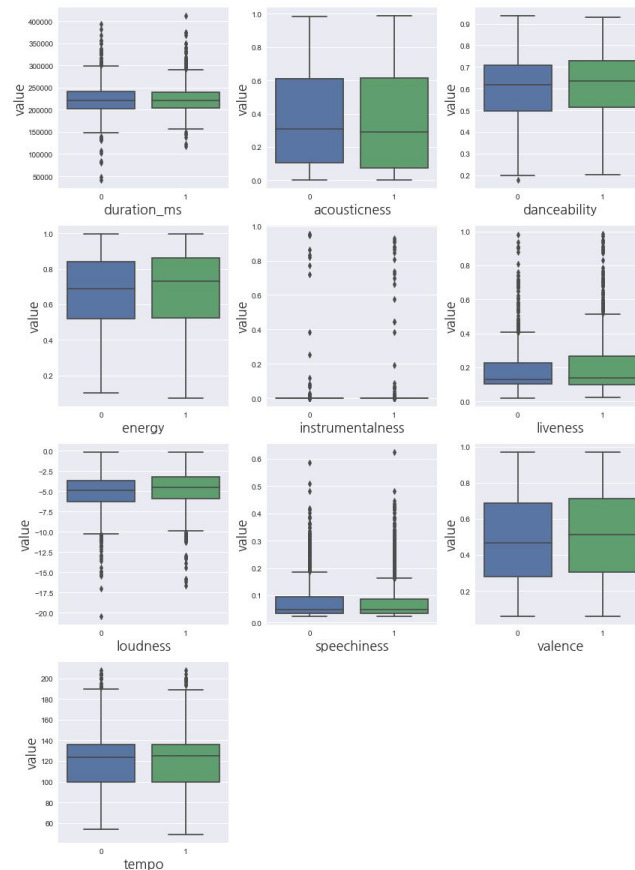


- La popularidad parece independiente del número de artistas
- Los single tienen más canciones populares
- Las compilaciones tienen más canciones no populares
- Hay más canciones explícitas no populares



# Datos III: Análisis

- Análisis visual
- No se observa mucha diferencia en los valores numéricos
- Valores atípicos



## Datos III: Análisis

- Las canciones populares son más bailables y tienen más energía que las no populares
- La música K-Pop varía en el tiempo (*ANOVA*)
- Correlación alta entre *loudness* y *energy*
- De los atributos categóricos sólo el tipo de álbum no es independiente de la clase (*Chi-Square Independence*)
- Hay diferencias significativas entre las canciones populares y no populares para las variables numéricas (*Kruskal-Wallis*)

# Modelos

- Problema de clasificación supervisada
- Problema balanceado
- Maximizar *precision*
- Dividir datos en train y test (*hold-out*)
- Validación mediante *3-fold cross-validation*
- *Grid search*
- Curvas de aprendizaje
- Data leakage
- Pipelines
- Transformaciones propias
- Baseline

# Modelos II

- Atributos **numéricos**
  - *Standardization*
  - *RobustScaler*
- Atributos **categoricos**
  - Agrupación de valores (*low entropy of categorical attributes*)
  - *One Hot Encoding*
  - *Label Encoding*

# Modelos III

- **Logistic Regression**

- Regularización (*Lasso, Ridge o Elasticnet*)
- Fuerza regularización
- No es necesario escalar los datos
- Se prueba:
  - Transformaciones logarítmicas, raíz cuadrada, polinomios e interacciones entre atributos

# Modelos IV

- **K-NN**

- Número vecinos ( $k$ )
- Pesos (*uniform* y *distance*)

- **SVC**

- Tipo de *kernel* (*poly*, *rbf* y *linear*)
- Fuerza regularización
- Coeficiente *gamma*

- **Random Forest**

- Número de árboles
- Número de atributos
- Profundidad de los árboles
- Número mínimo de instancias para que un nodo sea considerado hoja
- No hace falta escalar los datos

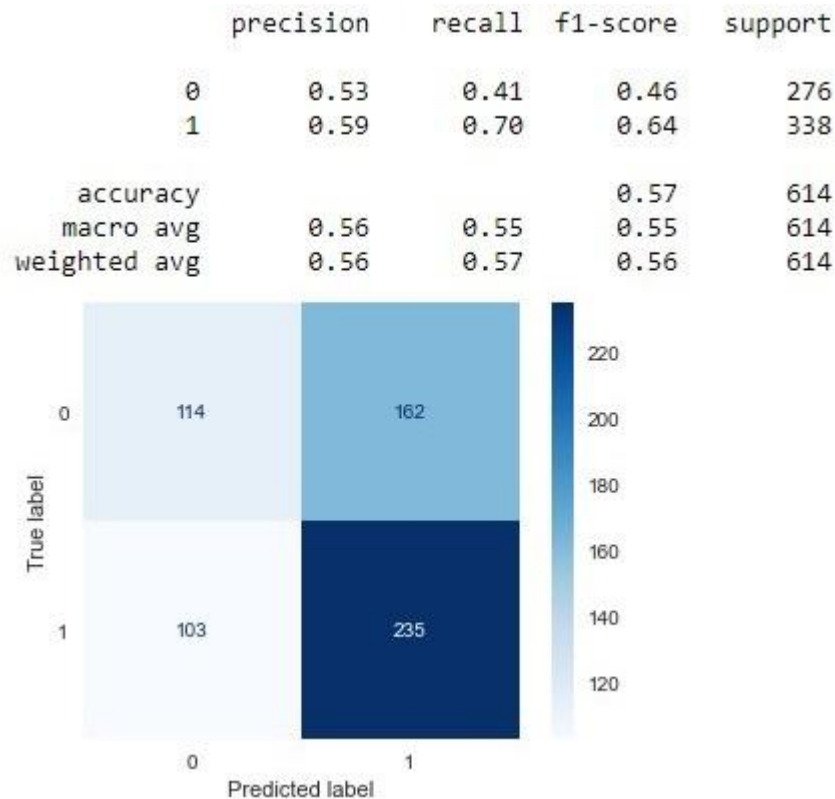
# Evaluación

- Se eligen los mejores modelos
- Los *baseline* no son mejores que un modelo aleatorio
- Algunos sufren de *high bias* y otros de *overfitting*
- Resultados bastante malos
- Los resultados de los cuatro modelos son muy parecidos



# Evaluación II

- El modelo elegido es el K-NN
- Es el que más *precision* junto con *f1* consigue





# Evaluación III

- *Permutation Importance*
- Se exploran los atributos que más han contribuido a la predicción
- No hay ningún atributo extremadamente importante

Weight	Feature
0.0315 ± 0.0275	danceability
0.0239 ± 0.0113	loudness
0.0238 ± 0.0327	first_time_in_chart_year
0.0226 ± 0.0157	acousticness
0.0191 ± 0.0244	tempo
0.0162 ± 0.0176	valence
0.0145 ± 0.0155	liveness
0.0138 ± 0.0204	first_time_in_chart_month
0.0136 ± 0.0060	album
0.0116 ± 0.0071	stone music entertainment
0.0104 ± 0.0090	카카오 m
0.0103 ± 0.0050	sm entertainment
0.0098 ± 0.0051	지니뮤직
0.0097 ± 0.0082	is_collab
0.0096 ± 0.0053	other
0.0083 ± 0.0059	is_ost
0.0080 ± 0.0086	other
0.0074 ± 0.0185	energy
0.0074 ± 0.0108	single
0.0070 ± 0.0148	speechiness
0.0046 ± 0.0141	duration_ms
0.0039 ± 0.0119	summer
0.0036 ± 0.0076	other
0.0035 ± 0.0135	holiday
0.0034 ± 0.0029	for_reality_tvshow
0.0018 ± 0.0012	stone music entertainment
0.0015 ± 0.0000	explicit
0.0011 ± 0.0016	compilation
-0.0005 ± 0.0052	instrumentalness
-0.0007 ± 0.0023	yg entertainment
-0.0012 ± 0.0012	jyp entertainment

# Conclusión

- Objetivos cumplidos
- Presencia de dificultades
- Aprendidas nuevas técnicas o métodos
- Decepción por los resultados de los modelos

# Conclusión

## Algunas líneas futuras

- Web app interactiva para el usuario
- Buscar canciones no populares en otro sitio
- Utilizar datos semanales y no mensuales
- Utilizar más atributos con información de los artistas

# Fin.

*¿Alguna pregunta?*