# Credit Risk Analysis Predictor

---

## Abstract:

In the evolving music industr y, predic ting song success is crucial. This research explores m achine lear ning models—logistic regression, decision trees, suppor t vec tor m achines (SVM), r andom forests, and gr adient boosting—to uncover their effec tiveness in discer ning hit songs. The study aims to provide insights for

infor m ed decision-m aking in the music dom ain.

## Introduction:

As the music industr y adapts, m achine lear ning em erges as a tool for forecasting song success. This study assesses logistic regression, decision trees, SVM, r andom forests, and gr adient boosting to unr avel their strengths in identifying hit songs.

## Objective:

Evaluate and compare m achine lear ning models in predic ting song success, offer ing a nuanced understanding of their effec tiveness.

## Methodology:

The research employs logistic regression for baseline modeling, decision trees for inter pretability, SVM for nuanced classification, r andom forests for ensem ble lear ning, and gr adient boosting for enhanced

predic tive power.

## Structure:

The paper proceeds with an over view of the dataset, explor ator y data analysis, and detailed analyses of logistic regression, decision trees, SVM, r andom forests, and gr adient boosting. A compar ative analysis

follows, concluding with key findings and implications. This research contr ibutes insights at the intersec tion of music and m achine lear ning, shaping the landscape of predic tive modeling in song success.

## Dataset:

The dataset utilized is the " Spotify Hit Predic tor Dataset (1960-2019)," sourced from Kaggle. This dataset contains var ious attr ibutes extr ac ted from tr acks using Spotify's Web API. The tr acks are categor ized as either '1' (indicating a 'Hit') or '0' (indicating a 'Flop') based on specific cr iter ia set by the dataset's author.

This dataset lends itself to the creation of a classification model aim ed at predic ting whether a given tr ack will achieve 'Hit' status or not. The dataset encompasses a r ange of features, including Tr ack nam e, Ar tist, URI, Danceability, Energy, Key, Loudness, Mode, Speechiness, Acousticness, Instrum entalness, Liveness, Valence, Tempo, Dur ation_ms, Tim e_Signature, Chorus_Hit, Sec tions, and the Target var iable.

## SMART questions:

. What are the top three audio features strongly associated with hit songs?

. How much improvem ent in accur acy, precision, and recall was achieved in the SVM RBF ker nel model after hyper par am eter tuning compared to the initial model?

. How much did the accur acy and inter pretability of the decision tree model improve after pruning compared to the initial deep tree

. What specific per for m ance m etr ics (e.g. , accur acy, AUC) can be used to assess the effec tiveness of the r andom forest model with default par am eters, and how did it compare to other models?

. What recomm endations can be m ade regarding the choice of m achine lear ning model for predic ting song success based on the compar ative analysis of SVM, decision trees, and r andom forests?

The questions were developed to address specific aspec ts of our analysis and decision-m aking process. We aim ed to understand the key audio features influencing song success and to selec t the best m achine lear ning model for predic tion. Each question was tailored to extr ac t precise infor m ation relevant to our

research objec tives. For example, we asked about the top hit-associated audio features to pinpoint influential fac tors, quantified per for m ance improvem ents to gauge the impac t of optim iz ation, and compared m etr ics to evaluate model effec tiveness. These questions were designed to guide our investigation and ensure that we obtained relevant insights to m ake infor m ed decisions.

## Exploratory Data Analysis:

Explor ator y Data Analysis (EDA) was conduc ted to gain insights into the dataset's statistics, struc ture, and the significance of different feature values. Dur ing this process, skewed data in sever al features were

nor m alized. Additionally, through univar iate analysis and the F-test, we identified the most influential features.

From our compar ative analysis, we m ade intr iguing obser vations. There was a significant contr ast in

loudness between hit and flop songs, with hit songs having a substantially higher loudness level of 3.56 units when compared to flop songs

Additionally, the " key" feature appears to have a relatively high p-value (0.755) in the F-test, suggesting it m ay not be statistically significant in distinguishing between hit and flop songs.

---

## Logistic Regression Modeling:

This sec tion analyzes the per for m ance of the logistic regression model used for predic ting hit songs.

### Model Performance Metrics

- Accuracy: 0.84

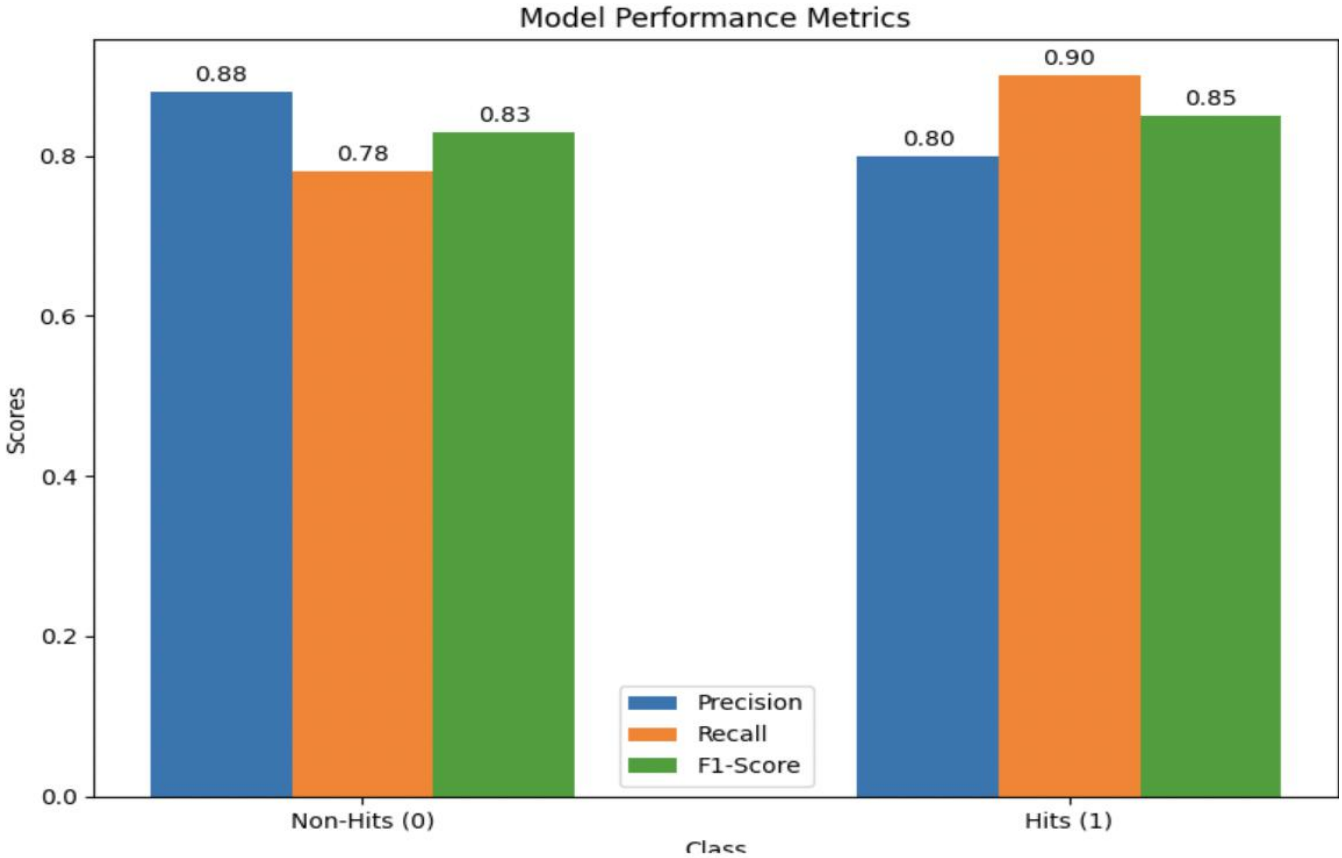
Hits:

- Precision: 0.80
- Recall: 0.90
- F1-score: 0.85
- Suppor t: 588

Non-hits:

- Precision: 0.88
- Recall: 0.78
- F1-score: 0.83
- Suppor t: 580

### Performance metrics :




### Interpretation

- The accur acy of the model is 84% , indicating that it cor rec tly predic ts 84% of the songs.
- The recall for hits is 90% , m eaning that the model cor rec tly identifies 90% of the hit songs.
- The precision for hits is 80% , signifying that 80% of the songs predic ted as hits are ac tually hits. The F1-score for hits is 85% , a har monic m ean of precision and recall, indicating a good
- balance between the two m etr ics.
- The confusion m atr ix shows that the model has more true negatives (cor rec tly predic ted non-hits) than false positives (incor rec tly predic ted hits). This suggests that the model is better at identifying non-hits than hits.
- The AUC of 0.91 indicates that the model is good at distinguishing between hits and non-hits.

### Observations

The logistic regression model per for ms well on this task, with an accur acy of 84% and a high recall for hits (90% ). However, the model is better at identifying non-hits than hits, as evidenced by the higher precision for non-hits (88% ) and the confusion m atr ix. This suggests that there m ay be som e room for improvem ent in the model's ability to identify hit songs.

---

## SVM Modeling Introduction:

In the initial phase of our SVM modeling for music classification using the Spotify dataset, our focus was on understanding the role of different ker nels and the regular iz ation par am eter (C) in the SVM model.

#### Kernel Selection:

SVM allows for var ious ker nels, including linear, Radial Basis Func tion (RBF), Sigmoid, and Polynom ial. The choice of ker nel deter m ines how the model captures complex patter ns within the data. In this context, we explored the impac t of different ker nels to identify the one that best suits our music classification task.

#### Regularization Parameter (C):

The regular iz ation par am eter (C) in the SVM model is crucial for controlling the balance between fitting the tr aining data well and gener aliz ing to new, unseen data. A large C value results in a sm aller m argin of the hyper plane, which can lead to over fitting, while a sm all C value allows for a larger m argin, r isking under fitting.

#### Significance to Problem Statement:

Understanding the role of ker nel selec tion and regular iz ation in SVM modeling is essential for addressing the challenges posed by diverse music features. This knowledge will guide us in refining the SVM model to achieve optim al classification per for m ance and provide insights into the significant features influencing music success.

---

## Linear Kernel:

In the second phase of our SVM modeling, we focused on employing a linear ker nel to build the initial SVM model. This step aim ed to assess the model's per for m ance using this simpler for m of the ker nel.

#### Model Performance Metrics:

The linear ker nel SVM model was tr ained and evaluated on the Spotify dataset, compr ising var ious audio features. Key per for m ance m etr ics were computed to gauge the model's effec tiveness in predic ting music success.

- Train Accuracy: The model exhibited a Tr ain Accur acy of 0.82, indicating its proficiency in lear ning from the tr aining set.
- Test Accuracy: The Test Accur acy m aintained a high value of 0.82, suggesting robust gener aliz ation to unseen music tr acks.
- Precision: Precision, representing the accur acy of positive predic tions, stood at 0.83.
- Recall: The Recall m etr ic, m easur ing sensitivity to positive instances, reached 0.82.
- F1 Score: The balanced F1 Score of 0.82 emphasized the har monious tr ade-off between precision and recall.

#### Challenges Identified:

Despite the impressive accur acy m etr ics, a closer exam ination revealed a lim itation in recall (0.82), indicating challenges in identifying positive instances. This obser vation prompted fur ther investigation and steps to address the recall lim itation.

#### Observations:

The evaluation of the linear ker nel model provides valuable insights into its initial per for m ance on music classification. Recogniz ing the recall lim itation is crucial for refining the model and achieving a more

balanced predic tion of music success, aligning with our over arching goal of developing an effec tive classification model for the Spotify dataset.

---

## RBF Kernel:

Building on the insights gained from the linear ker nel model, we progressed to implem ent the Radial Basis Func tion (RBF) ker nel in our SVM model. The tr ansition aim ed to enhance the model's complexity and capture more intr icate patter ns within the Spotify dataset.

### Model Performance Metrics:

The SVM model with the RBF ker nel underwent r igorous evaluation, and key per for m ance m etr ics were assessed to gauge its effec tiveness in predic ting music success.

- Train Accuracy: The RBF ker nel SVM model showcased an improved Tr ain Accur acy of 0.86, highlighting its proficiency in lear ning from the tr aining set.
- Test Accuracy: The Test Accur acy m aintained a high value of 0.85, indicating robust gener aliz ation to unseen music tr acks.
- Precision: Precision, representing the accur acy of positive predic tions, reached a comm endable value of 0.85.
- Recall: The Recall m etr ic, m easur ing sensitivity to positive instances, achieved a notewor thy value of 0.85.
- F1 Score: The balanced F1 Score of 0.85 emphasized the har monious tr ade-off between precision and recall.

### Possible Reasons for Improvement:

The adoption of the RBF ker nel proved effec tive in captur ing complex patter ns within the dataset, contr ibuting to the notable enhancem ents obser ved in var ious per for m ance m etr ics.

### Observations:

The refined model with the RBF ker nel signifies a str ategic step towards achieving higher accur acy and captur ing intr icate relationships within the Spotify dataset. This aligns with our over arching goal of developing an effec tive classification model to discer n music success fac tors.

---

## Hyperparameter Tuning using Grid Search CV:

Recogniz ing the pivotal role of hyper par am eters in SVM models, we employed the Gr idSearchCV technique to m eticulously fine-tune the model. This process aim ed to identify the optim al com bination of hyper par am eters for our RBF ker nel SVM model.

### GridSearchCV Process:

A systematic exploration of hyperparameter space was conducted, considering various combinations to ascertain the most favorable settings for our SVM model.

**Numerical Insights:**

- **Optimal Parameters:**
  - **Regularization Parameter (C): 1**
  - **Kernel: RBF**
  - **Gamma: 0.1**
- **Accuracy: The optimized model achieved an accuracy of 0.85, signifying a well-balanced classification capability.**
- **Precision: Precision, denoting accurate positive predictions, reached 0.81, maintaining a favorable balance.**
- **Recall: The Recall metric exhibited a noteworthy value of 0.90, showcasing the model's heightened sensitivity to positive cases.**

### Possible Reasons for Improvement:

**The meticulous tuning of hyperparameters addressed model intricacies, leading to an optimal balance between precision and recall. The selected parameters reflect a strategic configuration for improved model performance.**

### Observations:

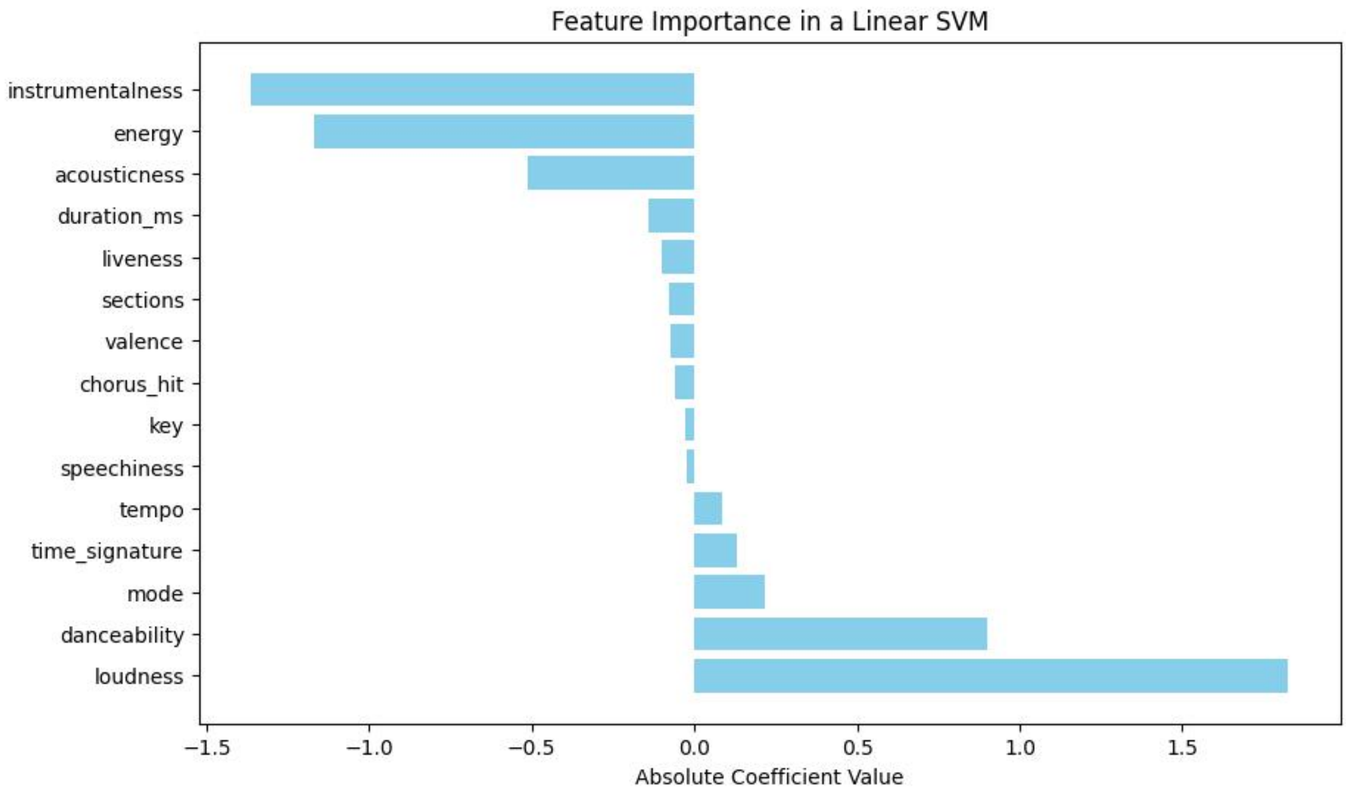
**The hyperparameter optimization step adds a layer of sophistication to our SVM model, ensuring it operates at peak efficiency. By identifying the optimal settings, we enhance the model's ability to discern intricate patterns within the Spotify dataset, contributing to our broader objective of accurate music success prediction.**

---

## Feature Importance:

**Moving beyond hyperparameter optimization, we delved into understanding feature importance to streamline our SVM model. By identifying and focusing on the most significant predictors, we aimed to create a more interpretable and efficient model.**

### Feature Importance Analysis:




**Showing the top 3 positive and negative cor related features and their impac t on the Outcom e.**

| Feature | Correlation | Explanation |
|---|---|---|
| Instrum entalness | Negative | Music with higher instrum entalness is more likely to be flop. |
| Energy | Negative | Music with higher energy is more likely to be a flop. |
| Acousticness | Negative | Music with higher acousticness is more likely to be flop. |
| Loudness | Positive | Music with higher loudness is more likely to be a hit song. |
| Danceability | Positive | Music with more danceability is more likely to be a hit song. |
| Mode | Positive | Music with a specific mode is more likely to be a hit song. |

- **Identifying Crucial Predictors: Lever aging the coefficients of the best-per for m ing SVM model, we conduc ted a feature impor tance analysis.**
- **Top Predictors: Loudness, instrum entalness, energy, danceability, and acousticness em erged as highly significant predic tors.**

**Num er ical Insights:**

- **Simplified Model Per form ance: Developing a simplified model using only the top 5 features m aintained comm endable accur acy m etr ics.**
  - **Accuracy: Slight decrease from 0.85 to 0.82, reflec ting a tr ade-off for a more inter pretable and efficient model.**
  - **Precision: Minim al decrease from 0.85 to 0.82 indicates robust positive predic tions.**
  - **Recall: Maintenance of high recall (0.81) suggests effec tive identification of positive cases.**
  - **F1 Score: The har monic m ean reflec ts the balanced nature of precision and recall.**

#### Observations:

The emphasis on significant features aim ed at creating a more inter pretable model without comprom ising predic tive power. By focusing on the top predic tors, we refined the model's complexity while retaining its ability to m ake accur ate predic tions. This step aligns with the broader goal of developing a predic tive model for music success, providing valuable insights to stakeholders in the music industr y.

---

## SVM Margin Visualization

#### Introduction:

In this phase, we aim ed to comprehend the impac t of different m argin complexities on classification boundar ies. Due to the challenge of visualiz ing an n-dim ensional SVM hyper plane, we employed Pr incipal Component Analysis (PCA) to reduce the dataset to two dim ensions. This allowed us to visualize the SVM m argin in the context of a two-dim ensional plane and gain insights into the model's behavior.

#### Margin Visualization:

- Default Margin: Visualized the default SVM m argin, illustr ating the initial decision boundar y.
- Weak Margin: Explored a scenar io with a weaker m argin, representing a more per m issive model.
- Strong Margin: Exam ined the impac t of a stronger m argin, reflec ting a more restr ic tive model.

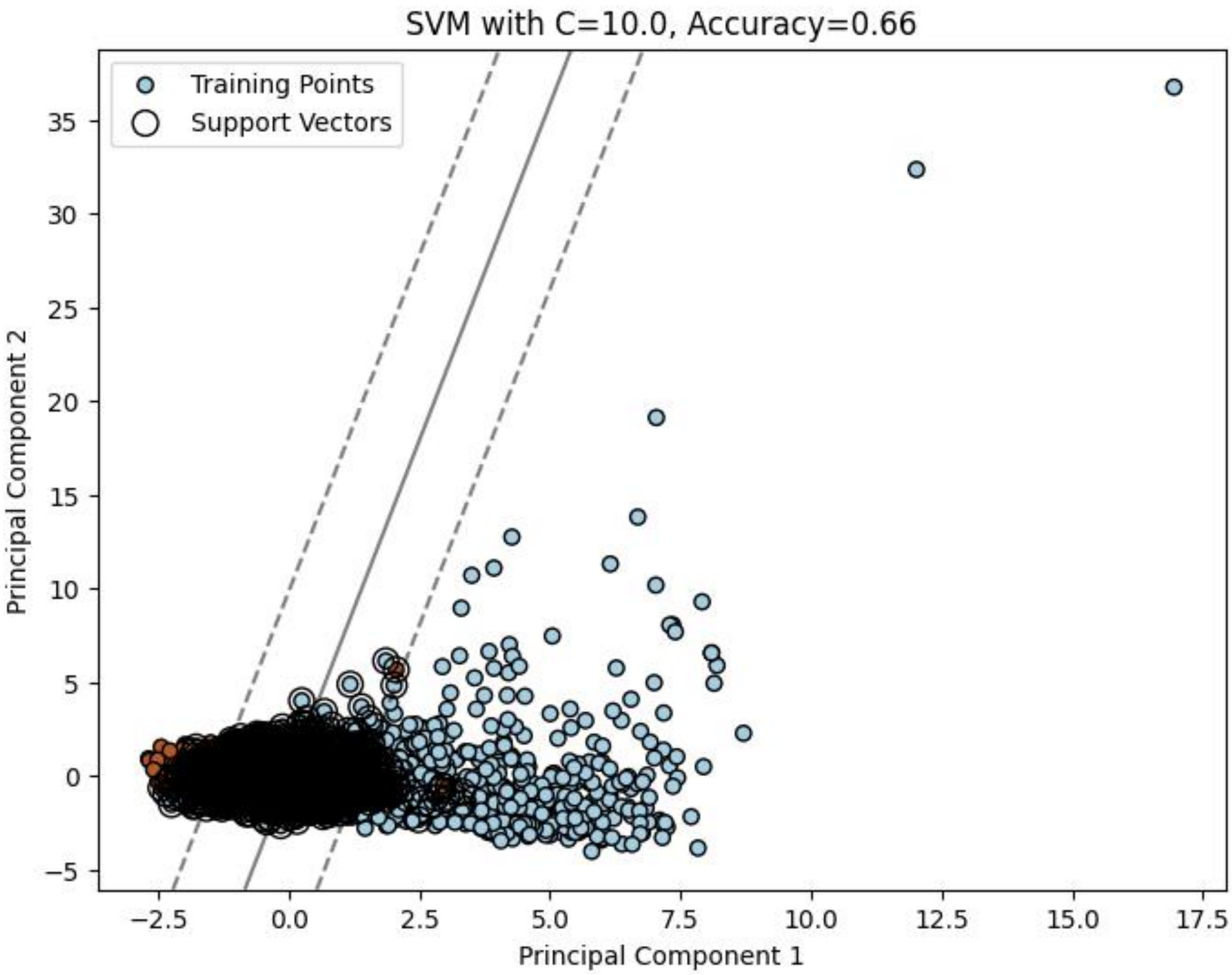
Obser vations:

- All three m argins, default, weak, and strong, exhibit sim ilar accur acies in the reduced two-dim ensional space. This unexpec ted consistency r aises questions about the impac t of m argin complexity on model per for m ance.

Possible Reasons for Sim ilar Accuracies:

- Data Distr ibution: The nature of the dataset, even in a reduced space, m ight not distinc tly favor a stronger or weaker m argin, resulting in compar able accur acies.

Here is the SVM Margin for one of the scenar io:




#### Conclusion:

Understanding the nuances of SVM m argin visualiz ation provides insights into the model's behavior under different scenar ios. This comprehension is vital for selec ting an optim al m argin complexity that aligns with real-world use cases.

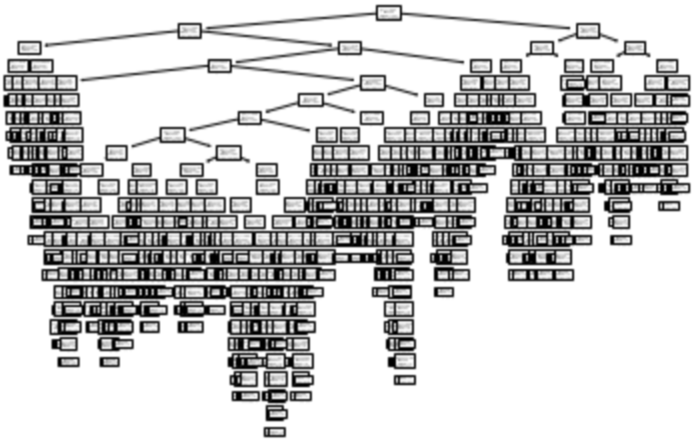
By visualiz ing SVM m argins, we gain a deeper understanding of how our model adapts to var ying complexities. This understanding is essential for m aking infor m ed decisions about model selec tion and deploym ent, ensur ing the model's robust per for m ance in music classification applications.

---

## Decision Trees

#### Generic Model

We first created a gener ic model with default par am eters which resulted in a tree of depth 24. The accur acy is just 75% .



| | Feature Importance |
|---|---|
| instrumentalness | 0.359622 |
| danceability | 0.116929 |
| acousticness | 0.077506 |
| duration_ms | 0.074855 |
| loudness | 0.069296 |
| speechiness | 0.051714 |
| valence | 0.051269 |
| energy | 0.049218 |
| liveness | 0.038237 |
| tempo | 0.033713 |
| chorus_hit | 0.029688 |
| key | 0.022076 |
| sections | 0.013483 |
| mode | 0.008791 |
| time_signature | 0.003604 |

**This helps to answer our first SMART question (" What audio features (e.g. acoustics, instrum entals, valence, tempo) are most strongly associated with hit songs in this dataset?"). Instrum entalness and**

**Danceability are pr im ar y features that are strongly associated with song's success.**

**Param eters**

- **m ax_depth**
- **m ax_leaf_nodes**
- **m ax_features**
- **m in_sample_split m**
- **in_sample_leaf**
- **m in_impur ity_decrease bootstr ap**
-

### Pruned Tree

**Then, we chose m ax_depth as par am eter and tr ied to find the depth with m aximum accur acy using a FOR loop. The result showed m aximum accur acy at m aximum depth = 3.**

**We build a tree with m ax depth 3, and these were the results:**

In conclusion, the classification model with depth=3 demonstr ates a comm endable over all per for m ance with an accur acy of 83% . The precision values for both classes indicate a high level of cor rec tness in predic tions, with 88% for class 0 and 79% for class 1. The model exhibits strong recall, par ticularly

notewor thy for class 1 at 90% , signifying its effec tiveness in identifying positive instances. The F1-scores, balancing precision and recall, are 0.82 for class 0 and 0.84 for class 1. The m acro and weighted aver ages, both at 0.84, reinforce the consistency of the model's per for m ance across classes, consider ing the

dataset's class im balance. While accur acy provides an over all m easure, stakeholders should weigh the impor tance of precision and recall based on the specific goals of the classification task. Over all, these

results suggest a robust model, but fur ther consider ation of the application context is advised to align with task-specific objec tives.

This pr ac tice could be done with other par am eters as well.

After this, we used Bagging Ensem ble technique - Random Forests.

---

## Random Forests

### Default Parameters

Dur ing tr aining, r andom forests (also known as r andom choice forests) gener ate a huge num ber of decision trees to use as an ensem ble lear ning approach for classification, regression, and other problems. The output of a r andom forest is the class selec ted by the vast m ajor ity of trees, which is useful for solving

classification issues. Decision trees m ay over fit their tr aining data, although r andom decision forests m itigate this problem . Random forests are more effec tive than decision trees in most cases.

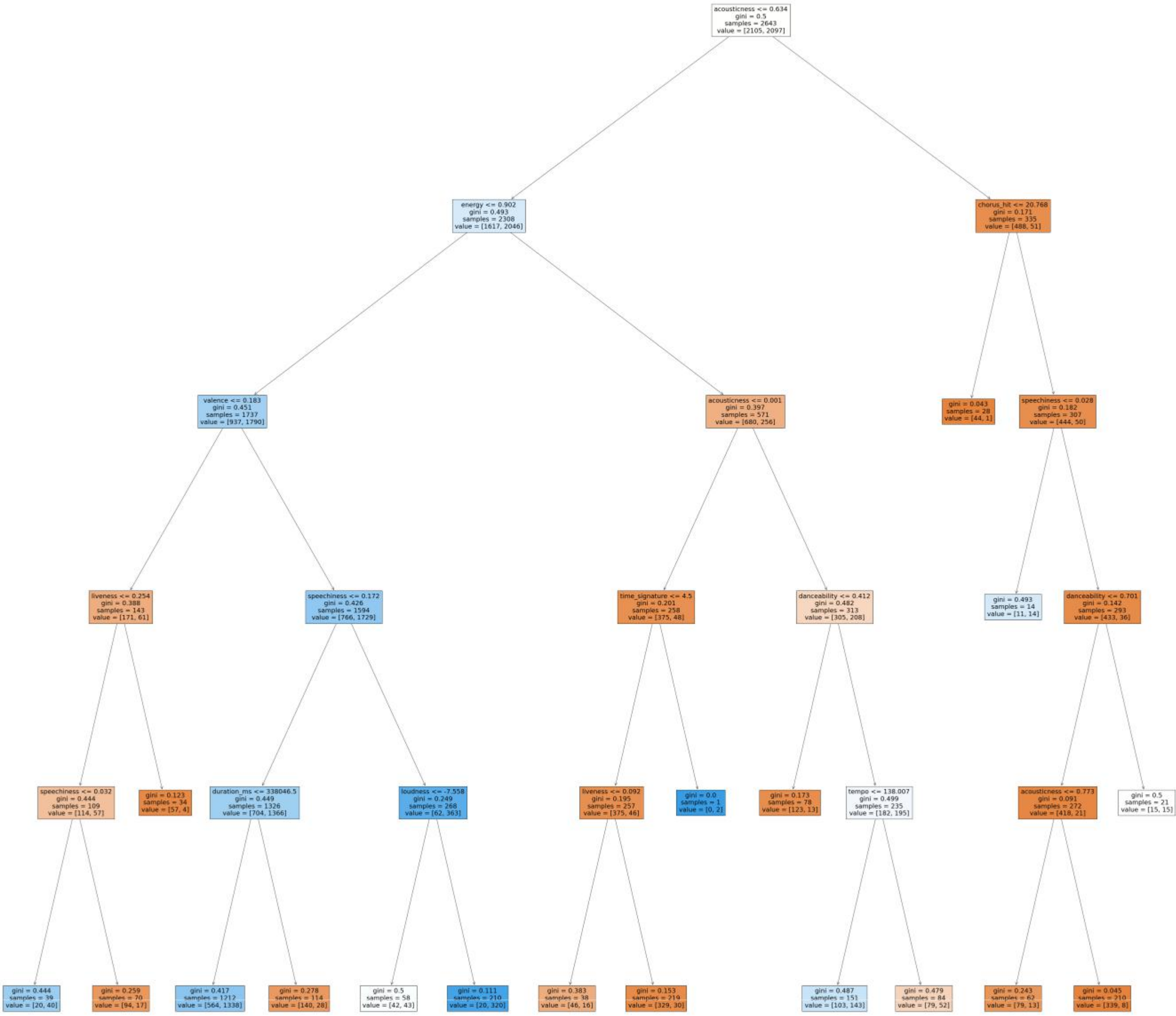
| | precision | recall | f1-score | support |
|---|---|---|---|---|
| 0 | 0.86 | 0.82 | 0.84 | 580 |
| 1 | 0.83 | 0.87 | 0.85 | 588 |
| accuracy | | | 0.85 | 1168 |
| macro avg | 0.85 | 0.85 | 0.85 | 1168 |
| weighted avg | 0.85 | 0.85 | 0.85 | 1168 |

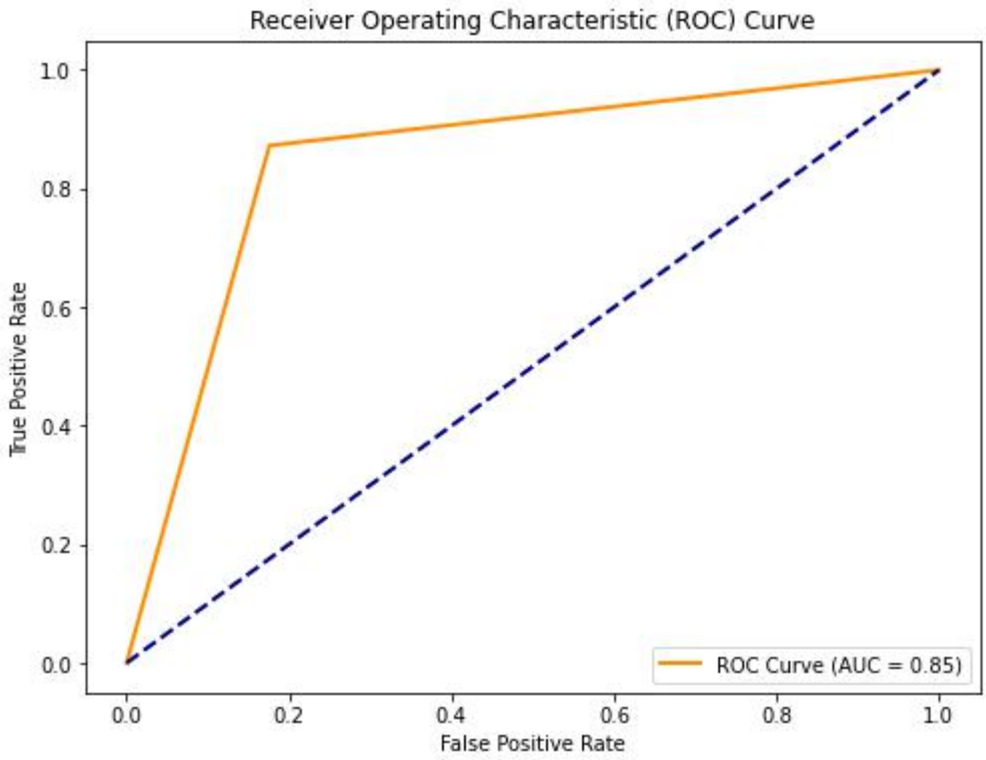
With default par am eters and 50 estim ator trees, we build r andom forest model. The classification model showcases comm endable per for m ance with an over all accur acy of 85% . It effec tively balances precision and recall for both classes, indicating its ability to cor rec tly identify instances while m inim iz ing false

positives and negatives. The model's consistent per for m ance is reflec ted in the m acro-aver age and weighted-aver age m etr ics, both at 85% .

### Cross Validation using RandomSearchCV

**We also per for m ed cross validation using Random Search. And finally we built the final model with the best par ams found r andom gr id.**






The Area Under the Receiver Oper ating Char ac ter istic Cur ve reflec ts the model's ability to distinguish

between positive and negative instances, with a higher AUC value suggesting super ior separ ation between the classes. In this case, the AUC of 0.83 signifies a robust predic tive capacity, where the model

consistently assigns higher probabilities to positive instances compared to negative instances.

---

## Conclusion

- Each model was scrutinized for its strengths and lim itations, enabling a nuanced understanding of their effec tiveness.
- This detailed analysis of SVM var iants, hyper par am eter optim iz ation, feature impor tance, and model complexity visualiz ation provides a comprehensive understanding of the Spotify dataset's classification task. The findings contr ibute insights into the nuanced dynam ics of SVM per for m ance, facilitating infor m ed decision-m aking in music classification applications.
- Moving to decision trees, a gener ic model with default par am eters revealed a deep tree with lim ited accur acy. Pruning the tree at a depth of 3 significantly improved over all per for m ance, achieving an accur acy of 83% with balanced precision and recall. Fur ther explor ation involved r andom forests, where default par am eters and cross-validation using Random SearchCV demonstr ated solid accur acy m etr ics and an AUC of 0.83, indicating robust predic tive capacity.
- In conclusion, the research contr ibutes a nuanced understanding of var ious m achine lear ning models' per for m ance in predic ting song success where all three models per for m sim ilar with an accur acy of over 85% . While each model exhibited strengths, the choice of the most suitable model depends on the specific goals and char ac ter istics of the music dataset.

## References:

. Lei, C. , Deng, J. , Cao, K. , Xiao, Y. , Ma, L. , Wang, W. , Ma, T. , & Shu, C. (2019). A compar ison of r andom forest and suppor t vec tor m achine approaches to predic t coal spontaneous com bustion in gob. Fuel, 239, 297-311.

. Fawcett, T. (2003). ROC Gr aphs: Notes and Pr ac tical Consider ations for Researchers (Technical Repor t No. HPL-2003-4). HP Labor ator ies.

. Segal, M. R. (2004). Machine Lear ning Benchm ar ks and Random Forest Regression (Technical Repor t). Center for Bioinfor m atics & Molecular Biostatistics, University of Califor nia, San Fr ancisco.

. Phan, T. N. , & Kappas, M. (2017). Compar ison of Random Forest, k-Nearest Neighbor, and Suppor t Vec tor Machine Classifiers for Land Cover Classification Using Sentinel-2 Im ager y.

. Kavzoglu, T. , Bilucan, F. , & Teke, A. (2020, Novem ber ). Compar ison of Suppor t Vec tor Machines, Random Forest, and Decision Tree Methods for Classification of Sentinel-2A Im age Using Different Band Com binations.

. Ribeiro, M. T. , Singh, S. , & Guestr in, C. (2016). " Why should I trust you?": Explaining the predic tions of any classifier. In Proceedings of the ACM SIGKDD Inter national Conference on Knowledge Discover y and Data Mining (pp. 1135–1144).

. Uddin, S. , Khan, A. , Hossain, M. E. , & Moni, M. A. (2019, Decem ber 21). Compar ing different super vised m achine lear ning algor ithms for disease predic tion.

. Demšar, J. (2006). Statistical compar isons of classifiers over multiple data sets. Jour nal of Machine Lear ning Research, 7, 1–30.