UNIVERSIDADE SÃO JUDAS TADEU

DOMINIC NASCIMENTO LAUBHOUET - RA: 822.141.082
EDUARDO VIEIRA DE JESUS - RA: 823.123.930
JULIA MARIA SILVA SIRINO - RA: 822.138.749
MARCOS VINÍCIUS PEREIRA PINTO - RA: 823.122.151
NICOLE LUANA DA SILVA - RA: 822.167.650

ENTREGA 04 – DOCUMENTAÇÃO DO PROJETO - MODELO DE MACHINE LEARNING

Inteligência Artificial

ENTREGA PARTE DO TRABALHO ACADÊMICO A3 REFERENTE À UC DE INTELIGÊNCIA ARTIFICIAL

SÃO PAULO
2023

DOMINIC NASCIMENTO LAUBHOUET - RA: 822.141.082
EDUARDO VIEIRA DE JESUS - RA: 823.123.930
JULIA MARIA SILVA SIRINO - RA: 822.138.749
MARCOS VINÍCIUS PEREIRA PINTO - RA: 823.122.151
NICOLE LUANA DA SILVA - RA: 822.167.650

ENTREGA 04 – DOCUMENTAÇÃO DO PROJETO - MODELO DE MACHINE LEARNING

Inteligência Artificial

Entrega 04 (quatro) do componente de trabalho acadêmico referente à atividade avaliativa A3 apresentado à UC de Inteligência Artificial como parte dos requisitos necessários para pontuação no cálculo final da nota da disciplina

Professor(a): Evandro Catelani Ferraz e Renato Alexandre de Medeiros
Disciplina: Inteligência Artificial
Turma: CCP1AM-BUC1-3198699

São Paulo
2023

RESUMO

A força motriz das Inteligências Artificiais são os dados, que representam um papel fundamental no seu uso, funcionamento e treinamento.

Em vista disso, é inerente a conclusão que a qualidade do modelo de Inteligência Artificial está diretamente proporcional à qualidade dos dados usados durante seu desenvolvimento. Dados mais precisos, relevantes, consistentes e completos tornam modelos de IA mais eficientes e exatos em uma ampla gama de cenários.

Logo, escolher um bom *database* para ser usado durante o desenvolvimento do software de Inteligência Artificial é uma etapa essencial para garantir sua qualidade, usabilidade e capacidade de ser atualizado.

Portanto, nós, como grupo, escolhemos uma base de dados que não só esteja de acordo com os objetivos que buscamos com o modelo de IA, mas que também cumpram requisitos de excelência para que o software desenvolvido atinja os objetivos que estabelecemos para nosso projeto.

ABSTRACT

The driving force of Artificial Intelligence is data, which represents a fundamental role in its use, operation and training.

Given this, it is inherent to conclude that the quality of the Artificial Intelligence model is directly proportional to the quality of the data used during its development. More accurate, relevant, consistent, and complete data makes AI models more efficient and accurate across a wider range of scenarios.

Therefore, choosing a good database to be used during the development of Artificial Intelligence software is an essential step to guarantee its quality, usability and ability to be updated.

Hence, we, as a group, chose a database that not only complies with the objectives we seek with the AI model, but also meets excellence requirements so that the software developed achieves the objectives we set for our project.

SUMÁRIO

INTRODUÇÃO

É sabido que a Inteligência Artificial, em especial o campo de estudo do Aprendizado de Máquina (*Machine Learning* - ML) são extremamente dependentes de dados e de sua qualidade. Diante disso, entender a organização e o que cada dado representa dentro de um *database* é essencial para julgar não só sua integridade, mas o quão útil ele é para ser usado no treinamento e desenvolvimento de modelos de IA e ML.

As Tabelas são um dos meios mais comuns de se organizar dados e são empregadas em inúmeros implementações de ML e IA e, para verificar sua qualidade, é necessário entender que tipo de dado cada coluna e o que eles representam.

Tendo em vista isso será efetuada nas seções subsequentes uma análise profunda de todas as tabelas e suas respectivas colunas assim como da variável *target* se será prevista com base nas interações entre os outros dados.

DOCUMENTAÇÃO DE TABELAS E COLUNAS

A *database* selecionada é composta de 3 tabelas diferentes, todas no formato de Valores Separados por Vírgulas (.csv), que garante simplicidade e legibilidade aos dados, sendo elas; *train.csv*; *test.csv* e *submission.csv*.

A tabela *train.csv* é a principal tabela do database, contendo todos os dados necessários para treinar o modelo de Machine Learning, que inclui 17.996 linhas distribuídos em 17 colunas, sendo elas:

- **Artist Name** *(Nome do Artista)* – Coluna que armazena dados em formato *string* a qual representa o(s) nome(s) do(s) artista(s) que participaram da faixa musical;
- **Track Name** *(Nome da Faixa)* – Coluna que armazena dados em formato *string* e representa o nome da faixa musical cuja informações estão reunidas na linha correspondente;
- **Popularity** *(Popularidade)* – Coluna cujos dados são armazenados em formato *int* e transmitem a ideia do quão aceita uma música é recebida pelo público, onde maiores valores representam maior popularidade;
- ***Danceability*** *(Capacidade de Dança)* – Coluna que armazena dados em formato *float* que com base em características como ritmo, andamento, estabilidade e batida da música mede o quão adequada ela é para ser dançada;
- **Energy** *(Energy)* – Coluna que retém dados no formato *float* e refere-se a sua intensidade e entusiasmo, onde valores numéricos mais altos são traduzidos como músicas mais empolgantes, rápidas e barulhentas, enquanto valores mais baixos representam músicas mais calmas, lentas e suaves;
- **Loudness** *(Volume)* – Coluna que armazena dados em formato *float* e que indica a intensidade do som, se aquela música é mais alta ou baixa;
- **Mode** *(Modo)* – Coluna que armazena dados em formato *int* e indica se uma faixa tem modalidade maior ou menor com base no tipo de escala da qual deriva seu conteúdo melódico. É representado por números inteiros, sendo 1 (maior) e 0 (menor);
- **Speechiness** *(Intensidade da Fala)* – Coluna que armazena dados em formato *float* e detecta a presença de palavras em uma faixa, indicando se a

música é mais falada/cantada – se a música tiver esse valor numérico alto, ela terá letras e vocais mais sobressaídos;

- **Acousticness** *(Intensidade Acústica)* – Coluna que armazena dados em formato *float* representando se a música tem mais elementos acústicos do que elementos eletrônicos;
- **Instrumentalness** *(Intensidade Instrumental)* – Coluna que armazena dados em formato *float* e representa se a música é mais instrumental, onde valores maiores significando instrumentos serem mais ressaltados;
- **Key** *(Classe de Nota)* – Coluna que armazena dados em formato *float* que representa a "chave musical" em que a música é performada, seguindo a notação padrão de tom musical;

Classificação de Tom Musical

Classificação Númerica de Tonalidade	Contrapartes tonais	Solfejo
0	C (também B ♯ , D 𝄫)	do
1	C♯ , D ♭ (também B 𝄪)	
2	D (também C 𝄪, E 𝄫)	ré
3	D♯ , E ♭ (também F 𝄫)	
4	E (também D 𝄪, F ♭)	mi
5	F (também E ♯ , G 𝄫)	fa
6	F♯ , G ♭ (também E 𝄪)	
7	G (também F 𝄪, A 𝄫)	sol
8	G♯ , A ♭ _	
9	A (também G 𝄪, B 𝄫)	la
10, t ou A	A♯ , B ♭ (também C 𝄫)	
11, e ou B	B (também A 𝄪, C ♭)	si

Figura 1.1 - Notação Padrão de Tom Musical

- **Liveness** *(Vivacidade)* – Coluna que armazena dados em formato *float* se refere à detecção da presença de público na gravação, onde valores mais altos indicam que há maior probabilidade que a faixa foi gravada ao vivo e valores mais baixos indicam menor probabilidade;
- **Valence** *(Valência)* – Coluna que armazena dados em formato *float* representando se a música é mais positiva ou negativa. Altas valências indicam maior alegria, enquanto uma música com baixa valência será mais triste;

- **Tempo** *(BPM)* – Coluna que armazena dados em formato *float* que representa o BPM (Batidas por Minuto) de uma música;
- **Duration_in min/ms** – Coluna cujos dados armazenados em *float* representam a duração de uma música em milissegundos;
- **Time_signature** – Coluna que armazena dados em formato *int* que descreve a estrutura rítmica da música, especificando quantas batidas há em cada compasso da música. Por exemplo, uma faixa com assinatura de tempo 5 significa uma estrutura rítmica de 5/4;
- **Class** – Coluna que armazena dados em formato *int* os quais representam o gênero musical que determinada música pertence.

Já a tabela *test.csv* é uma tabela com valores não rotulados, contendo as mesmas colunas, com os mesmos nomes e mesmos tipos de dados da tabela *train.csv*; com exceção da coluna *class* (que foi removida da tabela), totalizando 7.713 linhas em 16 colunas. O objetivo principal contendo é testar o modelo de machine learning desenvolvido em dados “não vistos” por ele antes.

Por fim, a tabela *submission.csv* também possui 7.713 linhas, porém, em 11 colunas. Por se tratar de um database criado para uma competição Hackathon, a tabela possui algumas informações sobre como enviar os resultados obtidos com a predição da tabela test.csv; onde as primeiras sete linhas armazenam uma matriz identidade e as restantes armazenam apenas valores *int* iguais a 0.

O cabeçalho de coluna dessa tabela possui um índice, que relaciona o valor *int* da coluna *class* com um gênero musical que o database permite o treinamento de um modelo de IA:

- **Acoustic/Folk** – 0;
- **Alt Music** – 1;
- **Blues** – 2;
- **Bollywood** – 3;
- **Country** – 4;
- **HipHop** – 5;
- **Indie/Alt** – 6;
- **Instrumental** – 7;
- **Metal** – 8;
- **Pop** – 9;
- **Rock** – 10;

DOCUMENTAÇÃO DA VARIÁVEL TARGET

A variável *target* (ou variável dependente) pode ser descrita como a variável que deseja ser prevista ou estimada a partir de um modelo de Machine Learning, chamada de *target* à medida que é o alvo e interesse dentro de um conjunto de dados.

A partir de algoritmos de Machine Learning, é possível realizar manipulações matemáticas sob os dados para encontrar padrões e/ou fórmulas matemáticas que sejam capazes de descrever as relações entre os dados e possam ser usadas para estimar valores desconhecidos, gerando então um modelo de IA baseado em Machine Learning.

A variável target do database escolhido para treinar os modelos de machine learning será a variável ***class***, armazenada na tabela *train.csv* que é responsável por rotular o gênero musical que uma faixa se encaixa.

Ela possui 11 possíveis resultados que são influenciados pelas outras variáveis (independentes).

Como cada gênero musical possui sua própria assinatura - características, que, independentemente da faixa, são semelhantes por todo o estilo (como por exemplo, músicas pop possuem normalmente um BPM entre 110 a 150). Isso torna a escolha de ***class*** como variável target extremamente apropriada, tendo em vista que a database possui um amplo conjunto de faixas dos mais diversos tipos, rotuladas e com as métricas que permitem encontrar padrões de cada categoria.

DOCUMENTAÇÃO DAS TRANSFORMAÇÕES

Para garantir que a *database* esteja nos padrões de qualidade desejados, é necessário analisar as suas tabelas e verificar se é preciso transformar os dados que armazenam com o objetivo de adequá-la para o uso no treinamento, em um processo de **Limpeza de Dados**, à medida que dados brutos podem conter erros que afetam a precisão do modelo de Machine Learning, levando a previsões incorretas.

É necessário localizar e corrigir dados duplicados; dados irrelevantes; exceções; dados ausentes e erros estruturais.

Na *database* escolhida há 3 tabelas, das quais serão usadas somente 2 - *trains.csv* e *test.csv* (a tabela *submission.csv* não será usada diretamente, apenas o seu cabeçalho, que relaciona os valores da variável *class* com gêneros musicais).

Das colunas da tabela *train.csv* e *test.csv*, serão excluídas as que armazenam o nome do artista (*Artist Name*) e o nome da faixa (*Track Name*), à medida que não contêm valores numéricos e não apresentam repetição frequente o suficiente para serem consideradas significativas. Além disso, músicas com nomes idênticos podem pertencer a gêneros diferentes, como "*Karma*" de Taylor Swift e "*Karma*" de Alicia Keys e um único artista pode criar músicas de gêneros diversos, como as faixas "*Would You - Acoustic*" e "*Sometimes it Rains in LA*" de The Vamps.

A coluna *Popularity* também será desconsiderada em ambas as tabelas, a medida que, mesmo representando um valor numérico, é extremamente volátil, sujeita a mudanças significativas com o passar do tempo (o que pode torná-la imprecisa se comparada com os valores atuais, já que o *database* foi publica em 2021). Além disso, músicas dentro do mesmo gênero podem apresentar amplas variações em termos de popularidade (como por exemplo “*Follow Through*” de David Kennedy com 64% de popularidade e “*Full Throttle*” de Ben Schuller com 33%), tornando-a uma métrica inadequada para o desenvolvimento e treinamento de um modelo de Machine Learning.

A coluna *duration_in min/ms* também será desprezada em ambas as tabelas, à medida que a database já conta com uma vasta gama de características das faixas, os gêneros musicais que serão previstos não possuem grandes diferenças na duração média de suas músicas e músicas específicas como *All To Well (10*

Minutes Version) da Taylor Swift podem não ser classificadas de maneira correta pelo algoritmo treinado com os valores de duração.

Logo, após a retirada das 4 colunas (*Artist Name, Track Name, Popularity* e *duration_in min/ms*), restam 13 colunas (*danceability, energy, key, loudness, mode, speechiness, acousticness, instrumentalness, liveness, valence, tempo, time_signature* e *Class*) na tabela train.csv e 12 colunas na tabela test.csv (*Class* não existe nessa tabela).

No entanto, ainda será necessário imputar dados, à medida que existem valores faltantes nas colunas *key* e *instrumentalness* em ambas as tabelas. Em *train.csv*, estão ausentes 2014 valores na coluna *key* e 4377 na coluna *instrumentalness*, enquanto na tabela *test.csv*, há 808 valores faltantes em *key* e 1909 em *instrumentalness*.

Para corrigir esse problema na tabela *train.csv*, dados faltantes na coluna *key* serão imputados usando as modas de cada gênero musical – por exemplo, uma linha cuja coluna *class* tenha valor 5 e *key* tenha valor ausente será preenchida com a moda dos valores de *key* das outras músicas da mesma *class* 5. Na coluna de *instrumentalness*, será adotado um procedimento semelhante, usando a média dos valores de *instrumentalness* dos gêneros musicais para realizar a imputação dos dados faltantes.

Já para o caso da tabela *test.csv*, como os dados não são rotulados por uma coluna *class*, as linhas onde *key* e *instrumentalness* forem nulas serão descartadas e desconsideradas, à medida que a tabela não está ligada diretamente com o treinamento do modelo de IA e sim com sua validação usando dados além do conjunto de treino rotulado.

1° MODELO - ÁRVORE DE DECISÃO

Para classificar os gêneros musicais com base em seus dados numéricos que representam seus aspectos essenciais das músicas, usar **Árvores de Decisão** é uma decisão pertinente. Essas características são importantes, mas sua relação não linear e influência variável na classificação demandam uma abordagem flexível.

As árvores de decisão, ao considerarem esses atributos de forma hierárquica, podem mapear os padrões existentes entre elas de maneira eficiente.

O algoritmo de árvore de decisão tem estrutura visual que recorda a de um fluxograma, o que permite fácil entendimento e visualização.

Seu funcionamento é análogo à uma estrutura de tomada de decisões, tornando-a acessível, conseguindo lidar tanto com problemas de regressão como de classificação.

Ela se estabelece em **nós** (*decision nodes*) que se relacionam entre si por uma hierarquia – há dois tipos de nós: o **nó-raiz** (*root node*) que é o mais importante e os **nós-folha** (leaf nodes). No contexto do Machine Learning, a raiz é um dos atributos da base de dados e as folhas são a classe ou valor que busca como resultado.

Esses nós são ligados por caminhos, chamados de ramos, escolhidos a partir das condições e conferências realizadas pelos nós.

Figura 2.1 - Exemplo simples de árvore de decisão com 2 nós e 4 ramos.

Nos ramos, encontram-se regras semelhantes às estruturas de controle "*if-else*" (por exemplo, se a **X > 15**, o caminho será para o **nó 1**; caso contrário, o caminho será rumo o **nó 2**).

Uma árvore de decisão é um algoritmo chamado de recursivo (conceito que define funções que chamam a si mesmo), ou seja, repete o mesmo padrão sempre na medida em que entram novos níveis de profundidade – O algoritmo se invoca em cada nó, tentando encontrar a melhor divisão de dados.

Essa melhor divisão de dados é a tarefa que a árvore de decisão precisa cumprir com objetivo de encontrar que nós serão encaixados em cada posição, qual será o nó raiz, etc.

Para realizar esses cálculos e obter resultados satisfatórios, uma abordagem comum é usar a **"impureza" dos dados** – quanto mais alta a impureza, menor a uniformidade dos dados – junto ao **ganho de informações**, que mede a mudança na entropia quando usa-se um atributo para dividir os dados – quanto maior o ganho da informação, melhor o atributo usado e maior a redução da incerteza sobre a classe que se quer prever.

Primeiro, se calcula a impureza da classe saída (seja por meio da entropia ou do índice de Gini), calcula-se o ganho de informações para cada atributo analisado e escolhe-se o atributo cuja métrica é maior; repetindo o processo, criando novos nós e ramos.

No modelo desenvolvido para o database escolhido leva como critério de impureza o índice de Gini; a divisão de cada nó é baseada na melhor divisão possível com base na impureza; a profundidade máxima da árvore é 10 (ou seja, o comprimento mais longo de uma raiz até uma folha é 10 – ou seja, o número máximo de decisões consecutivas de decisões é 10); o número mínimo de amostras necessárias para formar uma folha é 1 enquanto o número mínimo de amostras para se dividir um nó interno é de 10. Por fim, o estado aleatório é de 42, que define a aleatoriedade do estimador.

```
clf = DecisionTreeClassifier(criterion = 'gini', splitter = 'best', max_depth = 10, min_samples_leaf = 1, min_samples_split = 10, random_state = 42)
```

2.2 - Código que define o modelo de árvore de decisão

O modelo de IA possui precisão de 52.28%

FONTES BIBLIOGRÁFICAS

Figura 1.1 – Pitch class. Disponível em: <https://en.wikipedia.org/wiki/Pitch_class> - Acesso em 25 out. 2023.

Figura 2.1 – Como funciona o algoritmo de Árvore de Decisão (Decision Tree). Disponível em: <https://didatica.tech/como-funciona-o-algoritmo-arvore-de-decisao/>. Acesso em 15 nov. 2023.

Figura 2.2 - VINÍCIUS, Marcos “DeVinc1”. PROJETO A3 - Inteligência Artificial e Machine Learning Disponível em: <https://github.com/DeVinc1/PROJETO-A3-Inteligencia-Artificial-e-Machine-Learning/blob/main/ENTREGA%2004/Modelo%20de%20Predi%C3%A7%C3%A3o%20de%20G%C3%AAnero%20Musical%20-%20%C3%81rvore%20de%20Decis%C3%A3o.ipynb>. Acesso em: 15 nov. 2023.

Music Genre Classification. Disponível em: <https://www.kaggle.com/datasets/purumalgi/music-genre-classification> Acesso em: 25 out. 2023.

SPOTIFY. Web API Reference | Spotify for Developers. Disponível em: <https://developer.spotify.com/documentation/web-api/reference/get-audio-features>. Acesso em: 25 out. 2023.

PARSONS, C. O que é um Modelo de Machine Learning? | Blog da NVIDIA. Disponível em: <https://blog.nvidia.com.br/2021/09/28/o-que-e-um-modelo-de-machine-learning/>. Acesso em: 31 out. 2023.

Entenda: Aprendizado Supervisionado ou Não Supervisionado. Disponível em: <https://didatica.tech/aprendizado-supervisionado-ou-nao-supervisionado/>. Acesso em: 31 out. 2023.

KIM, M. The secret math behind feel-good music. Disponível em: <https://www.washingtonpost.com/news/to-your-health/wp/2015/10/30/the-mathematical-formula-behind-feel-good-songs/>. Acesso em: 31 out. 2023.

O que é Limpeza de dados? — Limpeza de dados explicada — AWS. Disponível em: <https://aws.amazon.com/pt/what-is/data-cleansing/>. Acesso em: 1 nov. 2023.

SACRAMENTO, G. Árvore de decisão: entenda esse algoritmo de Machine Learning. Disponível em: <https://blog.somostera.com/data-science/arvores-de-decisao>. Acesso em: 15 nov. 2023.