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 02 – PARTE INICIAL DA DOCUMENTAÇÃO DO PROJETO

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 02 – PARTE INICIAL DA DOCUMENTAÇÃO DO PROJETO

Inteligência Artificial

Entrega 02 (dois) 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.

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;

FONTES BIBLIOGRÁFICAS

Figura 1.1 – Pitch class. Disponível em: <https://en.wikipedia.org/wiki/Pitch_class> - Acesso em 25/10/2023.

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.