UNIVERSIDADE
FEDERAL RURAL
DE PERNAMBUCO

Vinicius Mateus Mendonça da Silva

Uso de Machine Learn para classificação de lançamentos financeiro
Estudo comparativo entre modelo AutoML e Redes MLP

Recife

Outubro de 2022

Vinicius Mateus Mendonça da Silva

Uso de Machine Learn para classificação de lançamentos financeiro
Estudo comparativo entre modelo AutoML e Redes MLP

Artigo apresentado ao Curso de Bacharelado em Sistemas de Informação da Universidade Federal Rural de Pernambuco, como requisito parcial para obtenção do título de Bacharel em Sistemas de Informação.

Universidade Federal Rural de Pernambuco – UFRPE
Departamento de Estatística e Informática
Curso de Bacharelado em Sistemas de Informação

Orientador: Cleviton Monteiro

Recife

Dezembro de 2021

Uso de Machine Learn para classificação de lançamentos financeiro - Estudo comparativo entre modelo AutoML e Redes MLP

Vinicius Mateus Mendonça da Silva [1], **Cleviton Monteiro** [2]

[1]Departamento de Estatística e Informática – Universidade Federal Rural de Pernambuco
Rua Dom Manuel de Medeiros, s/n, - CEP: 52171-900 – Recife – PE – Brasil

[vmms16@gmail.com, cleviton.moteiro@ufrpe.br]

Resumo. *O estudo desse trabalho visa auxiliar as empresas na sua gestão financeira gerando modelos baseados em Machine Learning para classificação de lançamentos financeiros. Com auxilio de bibliotecas desenvolvidas na linguagem Python, foi possível realizar o treinamento de modelos de AutoML e Redes Neurais Multilayer Perceptron responsáveis pela classificação dos dados. Com resultados acima de 85% nas métricas de Accuracy, Recall, F-measure e Precision para ambos os modelos, a utilização dos mesmo trás a possibilidade de uma melhor gestão dos lançamento financeiro com menos esforço.*

Abstract. *The study of this work aims to help companies in their financial management by generating models based on Machine Learning to classify financial releases. With the help of libraries developed in the Python language, it was possible to train AutoML models and Multilayer Perceptron Neural Networks responsible for data classification. With results above 85% in the metrics of Accuracy, Recall, F-measure and Precision for both models, using them brings the possibility of better management of financial releases with less effort.*

1. Introdução

A abertura de micro e pequenas empresas (MPEs) no Brasil vem crescendo desde o final de 2019. Em 2021, mais de 3 milhões de MPEs foram formalizadas, dados levantados pelo Serviço Brasileiro de Apoio às Micro e Pequenas Empresas (SEBRAE), afirmando que apesar de muitas pessoas terem migrado por necessidade para o empreendedorismo, houve uma grande busca devido ao surgimento de novas oportunidades.

Apesar das oportunidades muitas empresas brasileiras foram impactadas negativamente com a pandemia. Segundo o Instituto Brasileiro de Geografia e Estatísticas (IBGE) em pesquisa realizada em um período quinzenal entre os meses de julho e agosto de 2020, 34.5% das empresas foram impactadas de forma negativa ocasionando um grande desafio de gestão para o setor financeiro das mesmas.

Em estatísticas fornecida também pelo IBGE para o ano de 2014, aproximadamente 14% das empresas brasileiras encerram suas atividades com menos de um ano e 60% com menos de cinco anos de atividade (SEBRAE, 2014), sendo um dos fatores a má gestão financeira e a falta de capital de giro.

Alguns dos pontos mais comuns que podemos verificar na má gestão das finanças de uma empresa é a falta de estimativa de despesas e o lapso no controle de recebimentos. A estimativa do quanto e quais são as despesas fixas da empresa em conjunto com

as estatísticas do quanto ele recebe por seu produto e quais são seus principais clientes, oferecem um grande conhecimento do estado da empresa, possibilitando assim mais possibilidades na tomada de decisões.

Para auxiliar e melhorar a gestão financeira, muitas empresas utilizam software de gestão para controle de suas despesas e contas a receber. Os sistemas conhecidos como Enterprise Resource Planning (ERP) ou Planejamento de Recursos das Empresas, fornecem rastreamento e visibilidade global da informação de qualquer parte da empresa e de sua Cadeia de Suprimento, o que possibilita a tomada de decisões inteligentes (CHOPRA e MEINDL, 2003).

Os sistemas de ERP vem evoluindo com o passar dos anos, operações que antes eram feira de forma manual pelo usuários, passaram a ser otimizadas e automatizadas com a utilização de ferramentas de inteligência artificial e de aprendizado de máquina, também conhecidas como Machine Learning(ML) [SILVA 2022]. Diversos setores da empresa como chats, monitoramento para manutenções preventivas e no setor de vendas e marketing passaram a utilizar da tecnologia.

Diariamente são inseridos uma grande quantidades de dados nos sistemas de gestão como contas de água, energia, impostos, pagamentos, recebimento de honorários entre diversas categorias. A inserção e classificação desses dados devem ser feitas de forma confiável para que possa ser tomadas decisões de investimento, gerir resultados entre outras. Em vários casos essas informações são lançadas e classificadas de maneira manual ou por meio da importação de arquivos, sendo feita por um usuário de maneira repetitiva.

A utilização de modelos de ML para a automação de lançamentos financeiros trás como oportunidade minimizar a ocorrência de erros para essa atividade, evitando que um usuário possa fazer uma lançamento em uma conta contábil não correspondente. Além de reduzir o esforço de preenchimento de dados este fator aumenta a chance do usuário utilizar o sistema.

Uma característica dos dados importados é a desorganização das informações que dificultam a compreensão pelo usuário. Esse fator também prejudica o desenvolvimento de modelos tradicionais de ML pois passam apresentam maior complexidade, custo computacional e recursos de tempo para seu treinamento. Como alternativa para otimizar o desenvolvimento e automatizar toda parametrização necessárias surgiu o Aprendizado de Máquina Automatizado(AutoML) [HE, X. et al. 2009].

Esse trabalho tem como objetivo principal o desenvolvimento de modelos utilizando técnicas de ML para auxiliar a classificação de lançamento financeiros em sistemas de gestão e a comparação entre algoritmos de Redes MLPs[BRAGA, A. P. et al. 2000] e AutoML[HE, X. et al. 2009].

2. Trabalhos Relacionados

O aprendizado de máquina pode ser encontrado na intersecção entre as disciplinas de ciências da computação, engenharia e estatística, não sendo limitados a essas três áreas, tendo diversos usos em politica, geociências e nas mais diversas áreas. [HARRINGTON 2012]

Ethem Alpaydın et al.(2012) conceitua o aprendizado de máquina como sendo

a programação para otimizar o desempenho de critérios usando dados de exemplo ou de experiencias anteriores. Nessa programação temos um modelo definido ao qual são passados parâmetros, onde a aprendizagem é feita a partir da execução de um algoritmo que otimiza esses parâmeros ao utilizar os dados de treinamento. [ALPAYDIN 2012]

Em seu trabalho Monteiro e seus colegas [MONTEIRO, C.V.F et al. 2018], utiliza algoritmos de Regressão Logísticas e Naive Beyes com técnicas de seleção de atributos para realizar de classificação de lançamentos como água, energia, salários entre outra. Os resultado apresentados pelo autor se mostraram positivos tendo um bom percentual para as métricas utilizadas para validação dos modelos os modelos desenvolvidos.

Na Figura 1 podemos ver o extrato utilizado por Monteiro para realizar o tratamento e classificação dos lançamentos. A classificação de cada lançamento é representada na coluna categoria baseadas em função das informações fornecidas pelas demais.

Data de compensação	Descrição na conta corrente	Detalhamento	Documento	Valor	Categoria
18/12/2015	474 Transferência Agendada	17/12 0232 24819-3 ELARALDO COLOIA	23,200,000,024,819	-1043	13º Salário
18/12/2015	474 Transferência Agendada	17/12 1814 29640-6 BRUNO RAMOS C	181,400,000,029,640	-412	13º Salário
18/12/2015	474 Transferência Agendada	17/12 1833 25866-0 ANTONIO JOSE	183,300,000,025,866	-995	13º Salário
18/12/2015	474 Transferência Agendada	17/12 1850 13503-8 LARA LUCIA XAV	185,000,000,013,503	-133	13º Salário
18/12/2015	474 Transferência Agendada	17/12 3613 22477-4 RENATO CELSO S	361,300,000,022,477	-894	13º Salário
18/12/2015	474 Transferência Agendada	17/12 3613 44149-X JAMESSON RICARDO	361,300,000,044,149	-995	13º Salário
18/12/2015	474 Transferência Agendada	17/12 3613 48388-5 UBIRATAN CLEY	361,300,000,048,388	-995	13º Salário
18/12/2015	474 Transferência Agendada	17/12 3613 56685-3 RIVALDO GONCALV	361,300,000,056,685	-996	13º Salário
20/11/2015	470 Transferência on line	20/11 4357 58893-8 ITALO JORGE DA	664,357,000,058,893	-3209.54	13º Salário
29/09/2015	470 Transferência on line	29/09 1509 15539-X JOSÉ SANTO	661,509,000,015,539	-93.87	13º Salário
30/11/2015	474 Transferência Agendada	30/11 0232 24819-3 ELARALDO COLOIA	23,200,000,024,819	-1392	13º Salário
30/11/2015	474 Transferência Agendada	30/11 1814 29640-6 BRUNO RAMOS C	181,400,000,029,640	-373	13º Salário
30/11/2015	474 Transferência Agendada	30/11 1833 25866-0 ANTONIO JOSE	183,300,000,025,866	-1319	13º Salário
30/11/2015	474 Transferência Agendada	30/11 1850 13503-8 LARA LUCIA XAV	185,000,000,013,503	-75	13º Salário
30/11/2015	474 Transferência Agendada	30/11 3613 22477-4 RENATO CELSO S	361,300,000,022,477	-1099	13º Salário
30/11/2015	474 Transferência Agendada	30/11 3613 44149-X JAMESSON RICARDO	361,300,000,044,149	-1319	13º Salário
30/11/2015	474 Transferência Agendada	30/11 3613 48388-5 UBIRATAN CLEY	361,300,000,048,388	-1318	13º Salário
30/11/2015	474 Transferência Agendada	30/11 3613 56685-3 RIVALDO GONCALV	361,300,000,056,685	-1319	13º Salário

Figura 1. Extrato de planilha eletrônica utilizada por Monteiro
[MONTEIRO, C.V.F et al. 2018]

No trabalho ”AutoML for Multi-Label Classification: Overview and Empirical Evaluation”[WEVER, M. et al. 2021], Wever busca avaliar o desempenho de algoritmos de AutoML para a classificação de dados que podem apresentar mais de uma classificação, como exemplo, um imagem pode conter mais de uma classificação baseado nos elementos que podem ser encontradas nela.

Para o tratamento de dados textuais Ferreira[FERREIRA 2019] apresenta varias técnicas para extração de informação, uma delas é a tokenização que consiste em segmentar um texto em unidade básicas. Estrategia para identificação de palavras chaves que podem auxiliar na classificação dos dados.

Por sua vez, as Redes Neurais Artificiais(RNA), algoritmo que sera utilizado nesse trabalho, são modelos matemáticos de aprendizagem de máquina inspirados na estrutura neural do cérebro de seres inteligentes,ou seja, aprendendo, errando e realizando descobertas.

São sistemas paralelos distribuídos compostos por unidades de processamento simples conhecidas como nós, que calculam determinadas funções matemáticas. Esses nós são dispostos em uma ou mais camadas e interligadas por um grande número de de conexões.[BRAGA, A. P. et al. 2000]

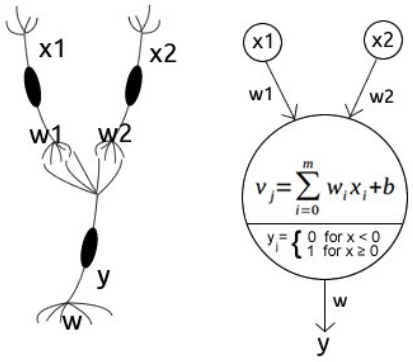

Figura 2. Exemplo comparativo de neurônio biológico
[KUROSE, J. F. and ROSS, K. W. 2005]

Existem muitos tipos de arquitetura para RNAs, a mais simples apresentam duas camadas, uma de entrada e uma de saída, essa ultima responsável pela classificação dos dados. As mais complexas apresentam camadas intermediarias entre as camadas de entrada e saída chamadas de camadas escondidas ou ocultas.

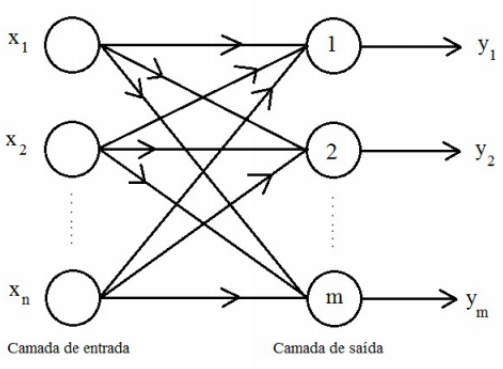

Figura 3. Arquitetura de redes com duas camadas
[KUROSE, J. F. and ROSS, K. W. 2005]

Os algoritimos de AutoML estão sendo usados em diversas áreas trazendo a facilidade na modelagem de modelos de ML. Tendo um pipeline consistente e composto pelas etapas de preparação dos dados, engenharia de recursos, geração de modelos e validações a utilização do mesmo torna-se uma vantagem para pessoas que não apresentam grande conhecimento na área[SILVA 2019].

Apesar de facilitar a geração de modelos Herbst[HERBST 2022] levanta algumas limitações de customização e para a utilização do AutoML em questão relacionada a processamento de texto, imagem, video e voz. Além das já citadas o algoritmo também evita a etapa de preparação adequada dos dados.

O estudo desenvolvido nesse trabalho utiliza algoritmos supervisionados para classificação de lançamentos financeiros em determinadas categorias pertencentes ao plano de contas contábeis da empresa, que esta previamente configurado no ERP pelos usuários do setor financeiro das empresas.

3. Materiais e Métodos

A metodologia utilizada para desenvolvimento foi a Cross Industry Standard for Data Mining (CRISP-DM) muito utilizada na área de mineração de dados. Em seu trabalho Shearer[SHEARER 2000] divide o método CRISP-DM em 6 etapas distintas:

1. **Entendimento do negócio**: Fase destinada a compreensão do projeto, o problema ou oportunidade que se quer resolver ou garantir, quais passos terão que ser dados para que isso ocorra, quais metas terão que ser atingidas e se irá agregar valor a solução.
2. **Entendimento dos dados**: Fase de busca, coleta e analise das informações disponíveis para desenvolvimento da solução. Nessa fase temos a compreensão dos dados, seus pontos chaves e pontos de dificuldade.
3. **Preparação dos dados**: Fase ao qual é feita a limpeza dos dados brutos, construindo uma base de dados consistentes através de vários processos, que sera utilizada na modelagem da solução.
4. **Modelagem**: Fase de desenvolvimento da solução. É nesse momento que os modelos são gerados, treinados e corrigidos.
5. **Avaliação**: Fase de testes, onde os modelos são validados ao serem submetidos a avaliações mais rigorosas, visando garantir a solução mais precisa e estável que garanta o objetivo a ser cumprido.
6. **Implantação**: Fase em que a solução é aplicada a ambiente não controlado ou de produção. Nessa fase a solução deve ser monitorada, realizar analise de desempenho e validar possíveis melhorias.

Serão abordadas as primeiras 5 etapas, com o intuito de criar modelos que alcançassem os objetivos específicos. O trabalho realizado em cada etapa serão descritos nas suas respectivas seções a seguir.

3.1. Entendimento do negócio

A empresa que cedeu os dados é proprietária de um sistema de gestão de projetos, equipes e financeiro e tem o intuito de melhorar uma funcionalidade no seu sistema que atualmente não esta atendendo as expectativas. A funcionalidade corresponde a classificação de lançamentos financeiro em categorias previamente cadastradas no banco de dados do sistemas, como exemplo podemos citar categorias como água, luz, impostos e muitas outras.

A importação dos dados são feitos via dois diferentes tipos de arquivos com extensões .txt e .ofx, que são fornecidos pelos bancos, sendo inseridos pelos clientes. O sistema tem a função de identificar os dados dos lançamentos e exibir em tela para que o usuário possa classificar de acordo com o plano de contas contábeis da empresa e confirmar sua inserção.

Foi desenvolvido em momento anterior um modelo de aprendizado de máquina baseado no trabalho de Monteiro [MONTEIRO, C.V.F et al. 2018] que visa melhorar a experiencia do usuário, realizando uma pré-classificação dos dados e exibindo uma sugestão de categorização. Após validação da classificação os dados são inseridos no sistema, os quais serão utilizados em treinamentos futuros.

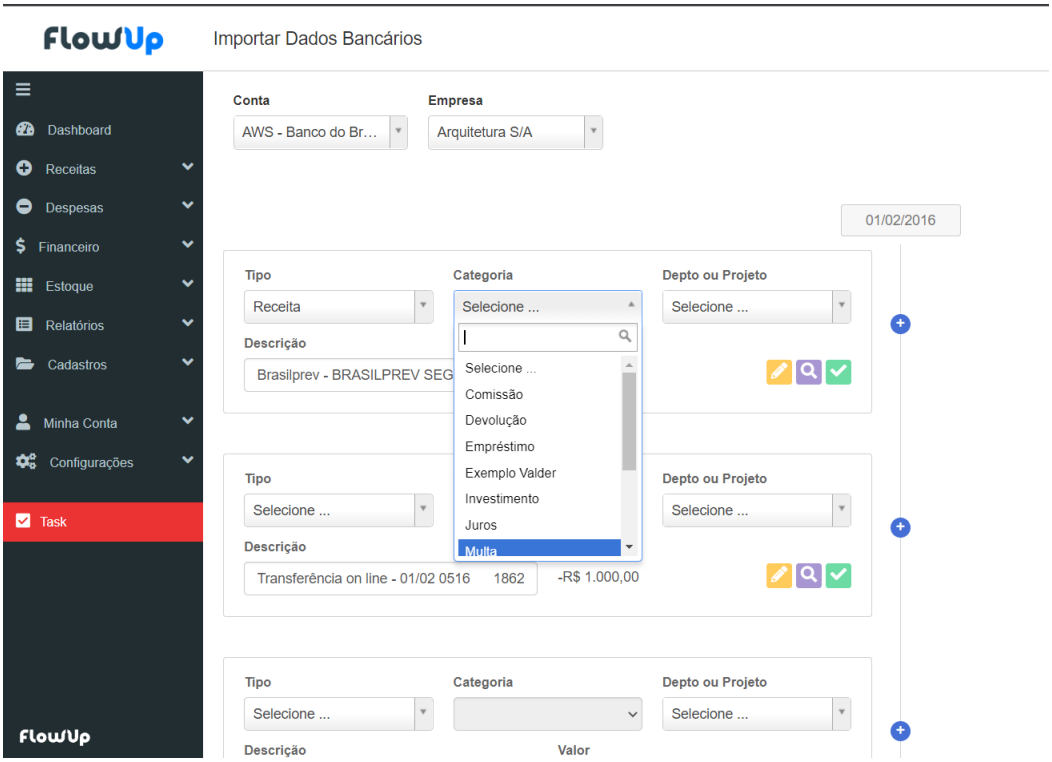

Figura 4. Exemplo de inserção e classificação de lançamentos financeiros
Fonte: [FLOWUP 2022]

Observando que os modelos gerados pelo algoritmo atual não apresenta categorizações precisas, surgiu a oportunidade de gerar novos modelos que tinham melhor desempenho em suas classificações.

3.2. Entendimento dos dados

Os dados utilizados foram os mesmo utilizados por Monteiro [MONTEIRO, C.V.F et al. 2018] em seu trabalho, sendo extraídos 3 *datasets* de duas empresas diferentes do banco de dados das empresas. Um da empresa A (EA) e dois da empresa B referentes a duas contas (EBC1, EBC2) que foram cedidas para analise,. Os *datasets* possuem 8 atributos, apresentados a seguir:

- **Categoria**: Representa o id da categoria do plano de contas da empresa.
- **Data de compensação**: Data ao qual o lançamento é referente
- **Descrição**: Descrição resumida sobre o tipo de operação executada na conta bancaria.
- **Detalhamento**: Informações adicionais sobre a operação. Pode conter data, nomes de pessoas e pessoas jurídicas, documentos, etc...
- **Documento**: Número do documento referentes ao tipo de transação.
- **Descrição na conta corrente**: Informações textuais com o tipo de operação e identificadores.
- **Valor**: Valor do lançamento em reais, sendo positivo para credito e negativo para débitos.

O período dos dados coletados variam de acordo com a empresa e o tempo que ela passou a utilizar o sistema de importação de lançamentos financeiros, fazendo com que

Data de compensação	Descrição na conta corrente	Detalhamento	Documento	Valor
18/12/15	474 Transferência Agendada	17/12 0232 24819-3 ELARALDO COLOIA	23,200,000,024,819	-1043
18/12/15	474 Transferência Agendada	17/12 1814 29640-6 BRUNO RAMOS C	181,400,000,029,640	-412
18/12/15	474 Transferência Agendada	17/12 1833 25866-0 ANTONIO JOSE	183,300,000,025,866	-995
18/12/15	474 Transferência Agendada	17/12 1850 13503-8 LARA LUCIA XAV	185,000,000,013,503	-133
18/12/15	474 Transferência Agendada	17/12 3613 22477-4 RENATO CELSO S	361,300,000,022,477	-894
18/12/15	474 Transferência Agendada	17/12 3613 44149-X JAMESSON RICARDO	361,300,000,044,149	-995
18/12/15	474 Transferência Agendada	17/12 3613 48388-5 UBIRATAN CLEY	361,300,000,048,388	-995
18/12/15	474 Transferência Agendada	17/12 3613 56685-3 RIVALDO GONCALV	361,300,000,056,685	-996
20/11/15	470 Transferência on line	20/11 4357 58893-8 ITALO JORGE DA	664,357,000,058,893	-3209.54
29/09/15	470 Transferência on line	29/09 1509 15539-X JOSÉ SANTO	661,509,000,015,539	-93.87
30/11/15	474 Transferência Agendada	30/11 0232 24819-3 ELARALDO COLOIA	23,200,000,024,819	-1392
30/11/15	474 Transferência Agendada	30/11 1814 29640-6 BRUNO RAMOS C	181,400,000,029,640	-373
30/11/15	474 Transferência Agendada	30/11 1833 25866-0 ANTONIO JOSE	183,300,000,025,866	-1319
30/11/15	474 Transferência Agendada	30/11 1850 13503-8 LARA LUCIA XAV	185,000,000,013,503	-75
30/11/15	474 Transferência Agendada	30/11 3613 22477-4 RENATO CELSO S	361,300,000,022,477	-1099
30/11/15	474 Transferência Agendada	30/11 3613 44149-X JAMESSON RICARDO	361,300,000,044,149	-1319
30/11/15	474 Transferência Agendada	30/11 3613 48388-5 UBIRATAN CLEY	361,300,000,048,388	-1318

Figura 5. Amostra do dataset
Fonte: Autor

a quantidade de dados disponíveis para análise também variasse.A quantidade e período pode ser encontrados na Tabela 1.

Empresa	Periodo coletado	Quantidade de instâncias
Empresa A (EA)	Setembro/2015 à Dezembro/2015	350
Empresa B (EBC1)	Setembro/2015 à Março/2016	524
Empresa B (EBC2)	Dezembro/2015 à Dezembro/2016	396

Tabela 1. Dados históricos coletados

Assim para melhor entendimento dos dados de cada uma das empresas analisadas foi realizada distribuição da quantidade de instâncias por categoria, tendo cada uma de suas categorias identificadas em sistema. A seguir pode ser observados essa distribuição para cada categoria nas figuras de 5 e 6.

Ao analisar as distribuições por categoria nas figuras 7 e 8 é possível observar que os datasets das empresas EA e EBC2 apresentam maior frequência em poucas categorias como Transferências, Despesas e Salário. Por apresentar muitas categorias como Água e Aluguel com numero de lançamentos pouco frequentes caracterizando o conjunto de dados como desbalanceado. O desbalanceamento dos dados podem interferir no treinamento do modelo, como solução foi utilizada a técnica de Oversampling, que consiste em replicar as categorias que apresentam menor frequência [ERTEKIN, S. et al. 2007] [CHAWLA 2009]. A técnica foi utilizada fazendo com que as categorias menores que 10% da maior fossem replicadas aleatoriamente para 10%.

Categorias que apresentam número de instâncias inferiores a 3 são lançamentos com periodicidade muito espaçadas como impostos anuais, semestrais ou trimestrais enquanto as que apresentaram uma grande quantidade representam lançamentos com periodicidade mensais, quinzenais e que muitos lançamentos daquele tipo podem ser inseridos nesse período, como o pagamento de contribuintes ou o recebimento de serviços prestados.

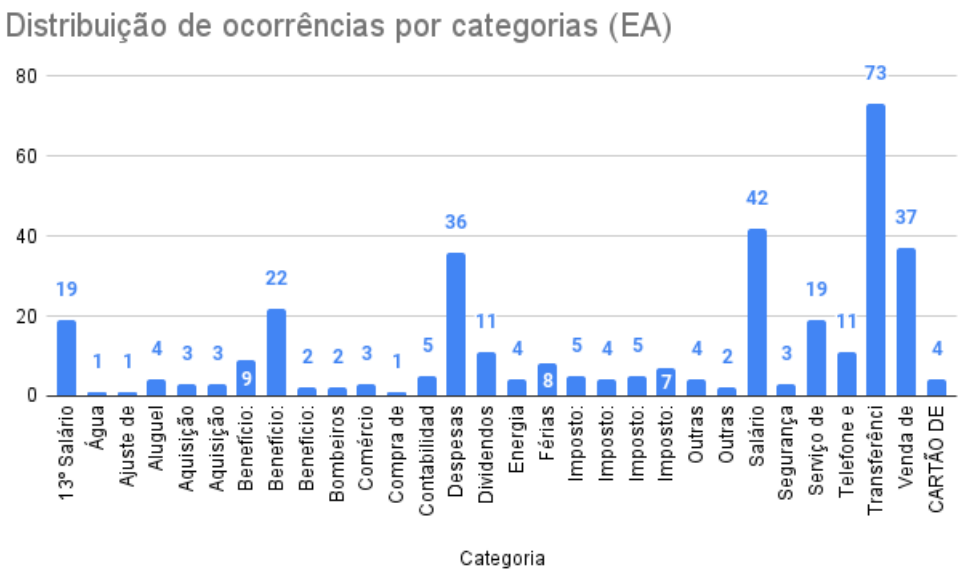

Figura 6. Distribuição Empresa A

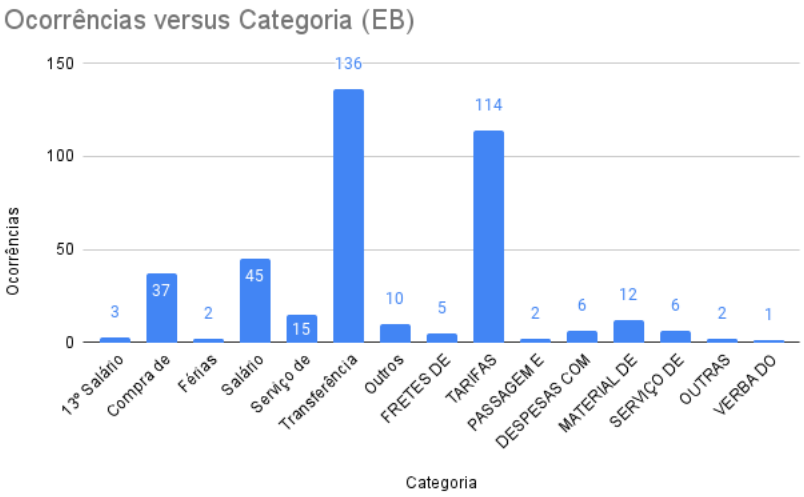

Figura 7. Distribuição Empresa B Conta 2

3.3. Preparação dos dados

Esta fase corresponde a preparação dos dados, sendo abordadas técnicas e ferramentas utilizadas para construção do *dataset* final que serão utilizados na modelagem das redes MLPs e modelo de AutoML. O elaboração dos scripts para realizar o pré-processamento dos dados foram desenvolvidos na linguagem de programação Python, com o auxilio da biblioteca Pandas, a qual é voltada para manipulação e analise de dados, muito utilizada para manipular tabelas e séries temporais.

A partir dos atributos citados no sessão anterior foi iniciado o pré-processamento, com o objetivo final de criar um *dataset* binário. Para isso foi utilizado o método One-hot Encode, também conhecida como dummy, que consistem na representação de dados catégoricos, que não possuem um relacionamento ordinal, em uma nova variável binária[BROWNLEE 2017]. Essa representação foi realizada buscando ter maior granularidade dos dados para que novos padrões possam ser identificados e auxiliar na modelagem.

Avaliando quais informações poderiam ser retiradas, o primeiro atributo tratado foi o *Valor* que representa o valor a ser creditado ou debitado em conta, dando origem a dois novos atributos que são representados no *dataset* como duas novas colunas binarias. O segundo tratamento ao atributo foi a construção de categorias de valores baseado na amplitude de intervalos calculados utilizando-se do maior e menor valor do conjunto de dados, sendo cada categoria correspondente a uma coluna do conjunto final. Os valores foram identificados e classificados em cada categoria através de uma representação binaria.

Tomando como exemplo o primeiro valor de - R$ 500,00 da Tabela 2, o mesmo é tratado como sendo um debito, tendo seu valor na coluna debit como verdadeiro. Utilizando seu valor absoluto identificamos em qual faixa de valor ele se encontra inserido, sendo a faixa de 0 á 1586 (aic_0_1586) como verdadeira e as demais faixas como falsa.

Valor	credit	debit	aic_0_1586	aic_1587_3173	aic_3174_4760
-R$ 500,00	0	1	1	0	0
-R$ 1.048,98	0	1	1	0	0
R$ 1.894,00	1	0	0	1	0
R$ 3.489,00	0	1	0	0	1

Tabela 2. Transformação de valores em dataset binário

Dos atributos tipo data foram criados novos que representam o mês e a semana ao qual o lançamento é referente. Também foram retiradas informações baseado no intervalo de dias do mês, sendo o mês divido em intervalos especifico de dias, e assim identificando se a data esta presente nesse intervalos. Como exemplo é feita a divisão do mês em 15 dias e então é identificada se a pertence a primeira quinzena do mês ou a segunda, esse tratamento foi realizado para os intervalos de 5, 10 e 15 dias.

Para atributos de *Detalhamento* e *Descrição na conta corrente*, por apresentarem informações que não possuem um padrão definido, podendo conter datas, nomes próprios de pessoas físicas e jurídicas, diferentes tipos de documentos, foi utilizada a técnica de tokenização. Essa técnica é comum no processamento de linguagem natural, sendo um

método que separa o texto em partes menores para criar um vocabulário que será utilizado para realizar o treinamento do modelo. A geração dos tokens foi feita a partir da divisão das informações por espaço em branco, onde cada token é uma palavra, data ou documento.

Com a criação do vocabulário foi realizada a busca no conjunto por nomes próprios de pessoas físicas para enriquecer o dataset final, criando um novo atributo para indicar a presença ou não. A busca foi realizada ao comparar cada token com a base de dados do Instituto Brasileiro de Geografia e Estatística(IBGE) que contem dados histórico dos nomes de pessoas físicas assim como suas possíveis abreviações. Os dados foram coletados da API da própria instituição em sua versão 2.0.0 que contém todo os nomes registrados até o ano de 2010.

Também foi contemplada a busca e identificação de Cadastro de Pessoa física (CPF) e Cadastro Nacional de Pessoa jurídica (CNPJ), tendo novos atributos para identificação da presença de ambos. A identificação foi feita por algoritmos disponibilizados pelo site da Receita Federal.

O tratamento para o atributo *Documento* foi realizados em duas etapas. A primeira etapa consistiu em verificar a quantidade de caracteres numéricos estava presentes, tendo um novo atributo para cada quantidade distinta encontrada, ou seja, se mais de um documento tiver 5 atributos numéricos apenas um novo atributo representando o tamanho 5 é criado. O segundo tratamento quebrar os documentos em parte menores através de seu separador, ponto ou virgula, a partir dessa quebra 5 grupos foram gerados e os documentos foram classificados de acordo com o a quantidade de subpartes geradas, onde, caso o documento for quebrado em 2 partes o mesmo seria classificado como grupo 2.

Para finalizar o tratamento e enriquecimento dos dados dando origem ao *dataset* final foi feita a normalização dos dados com o intuito de evitar que houvesse uma variação grande entre os intervalos dos dados. A normalização consiste em um método aplicados aos dados que transforma seus valores para serem representados em um determinado intervalo, no caso entre 0 e 1.

3.4. Modelagem e avaliação dos modelos

Foram gerados 9 modelos para cada algoritmo, totalizando 18 modelos gerados. Os modelos foram gerados a partir das bibliotecas feita em python sendo essas Auto-Sklearn e Keras para AutoML e Redes Neurais Multilayer Perceptron respectivamente. Os modelos baseados em Redes Neurais foram gerados a partir de 3 configurações que tiveram como variáveis o número de camadas escondidas e a presença de taxa dropout de 0.5 (Tabela 2), com todos os modelos tendo uma redução da quantidade de nós por camada de 0.5. Os modelos gerados a partir de AutoMl foram baseados no tempo disponível para geração dos mesmos (Tabela 3).

Para validação dos treinamentos foi utilizada a técnica da validação cruzada, também conhecida como cross-validation. Esse tipo de validação consiste em que dividir o conjunto de dados em um número definido de partes, que serão utilizadas para treinamento do modelo em iterações igual ao número de partes definidas. Após a seleção e treino foi feita a classificação a partir do dataset de testes onde foram geradas as métricas de Accuracy, Recall, F-measure e Precision para analise dos resultados[SHALEV-SHWARTZ, S. and BEN-DAVID, S. 2014].

Modelo	Número de camadas	Dropout	Taxa de aprendizagem
Modelo 1	1	0	0.001
Modelo 2	1	0	0.001
Modelo 3	1	0	0.001
Modelo 4	2	0	0.001
Modelo 5	2	0	0.001
Modelo 6	2	0	0.001
Modelo 7	3	0.5	0.001
Modelo 8	3	0.5	0.001
Modelo 9	3	0.5	0.001

Tabela 3. Configurações dos modelos de Redes Neurais

Modelo	Tempo de tarefa (min)
Modelo 1	5
Modelo 2	5
Modelo 3	5
Modelo 4	10
Modelo 5	10
Modelo 6	10
Modelo 7	15
Modelo 8	15
Modelo 9	15

Tabela 4. Configurações dos modelos de AutoML

4. Resultados

Nesta sessão serão apresentados os resultados obtidos para os modelos de AutoML e Redes Neurais apresentados anteriormente, sendo selecionados os melhores entre os 9 gerados para cada algorítimo. Ao utilizar das métricas das métricas de Accuraracy, Recall, F-measure e Prescision foi possível realizar a analise da eficacia de cada modelo na classificação das categorias por empresa.

Metrica	Accuracy		
Modelo — Empresa	EA	EBC1	EBC2
Automl (5min)	93,60%	88,62%	90,33%
AutoML(10min)	96,18%	89,19%	89,73%
AutoML(15min)	95,04%	89,14%	90,63%
Rede MLP (1 camada)	94,83%	85,15%	90,09%
Rede MLP (2 camada)	95,97%	84,36%	91,24%
Rede MLP (3 camadas + dropout)	95,97%	85,27%	92,30%

Tabela 5. Resultados da métrica de Accuracy para os 3 datasets

Ao analisar a Tabela 5 com os resultados de cada modelo para a métrica de Accuracy, que indica o quanto cada um classificou corretamente, podemos observar todos tiveram um bom resultado, com métricas acima de 80% para os 3 datasets em estudo. Ao

comparar os tipos de modelos é notado que a diferença entre os modelos gerados por AutoML e Redes Neurais apresentam taxas bastante similares para os datasets das empresas EA e da EBC2 enquanto para a EBC1 existe um diferença de aproximadamente 4% a mais para os algoritmos de AutoMl.

Metrica	Precision		
Modelo — Empresa	EA	EBC1	EBC2
Automl (5min)	94,31%	88,97%	88,64%
AutoML(10min)	96,23%	89,49%	87,76%
AutoML(15min)	95,20%	89,50%	87,92%
Rede MLP (1 camada)	95,06%	85,32%	89,90%
Rede MLP (2 camada)	96,24%	85,23%	90,41%
Rede MLP (3 camadas + dropout)	96,30%	87,00%	90,99%

Tabela 6. Resultados da métrica de Precision para os 3 datasets

A métrica Precision indica o quão preciso é o modelo, sendo o calculado de quantos verdadeiros positivos que foram classificados estão realmente corretos. A Precision dos modelos apresentada na Tabela 5 nos mostra que para a empresa EA os valores, assim como a Accuracy, são bastante similares, tendo uma diferença mais visível para os datasets da EBC1 e EBC2. Para a empresa EBC1 os modelos de AutoML obtiveram melhor percentual com cerca de 2% á 4% a mais que as Redes Neurais, enquanto para o dataset EBC2 as Redes Neurais apresentaram melhor desempenho com 2% á 3% maiores.

Metrica	Recall		
Modelo — Empresa	EA	EBC1	EBC2
Automl (5min)	93,84%	89,60%	88,94%
AutoML(10min)	96,50%	89,88%	87,99%
AutoML(15min)	95,30%	90,03%	89,17%
Rede MLP (1 camada)	95,22%	85,96%	91,03%
Rede MLP (2 camada)	96,45%	85,01%	89,96%
Rede MLP (3 camadas + dropout)	96,45%	85,46%	91,20%

Tabela 7. Resultados da métrica de Recall para os 3 datasets

Pode ser observado na Tabela 7 o Recall dos modelos tiveram comportamento semelhante a Precision, com pouca diferença entre os eles para o dataset da EA, os modelos de AutoMl com melhor desempenho para o EBC1 e as Redes Neurais sendo melhores para o EBC2.

Sendo o F-measure definido como a média harmônica entre a Precision e o Recall, é uma métrica utilizada para avaliação de conjuntos de dados desbalanceados, como os datasets apresentados nesse trabalho. Por utilizar as métricas anteriormente citadas podemos observar que o desempenho dos modelos apresentou o mesmo comportamento para as métricas de Precision e Recall.

Avaliando as métricas em conjuntos é notado que os melhores desempenhos foram para os datasets das empresas EA e EBC2, possivelmente pelo dataset EBC1 apresentar um maior número de dados e categorias para serem classificadas. Ao observar os dados

Metrica	F-maseure		
Modelo — Empresa	EA	EBC1	EBC2
Automl (5min)	93,77%	88,52%	88,31%
AutoML(10min)	96,27%	88,99%	87,56%
AutoML(15min)	95,08%	89,18%	88,42%
Rede MLP (1 camada)	94,96%	85,05%	90,09%
Rede MLP (2 camada)	96,18%	84,25%	89,60%
Rede MLP (3 camadas + dropout)	96,18%	85,17%	90,60%

Tabela 8. Resultados da métrica de F-measure para os 3 datasets

reais da empresa EBC1 pode ser encontrado dados com data de compensação em branco, podendo ter dificultado a aprendizagem dos modelos para esta dataset. Entre os mesmo tipos de modelos a variação de tempo para treinamento de AutoML não teve uma grande interferência no desempenho dos mesmo, sendo aqueles com tempo de treino de 10 e 15 minutos tendo uma melhor classificação do que o de 5 minutos. Para as redes Neurais a arquitetura com 3 camadas e taxa de dropout teve um resultado superior ou semelhante as demais para os 3 datases.

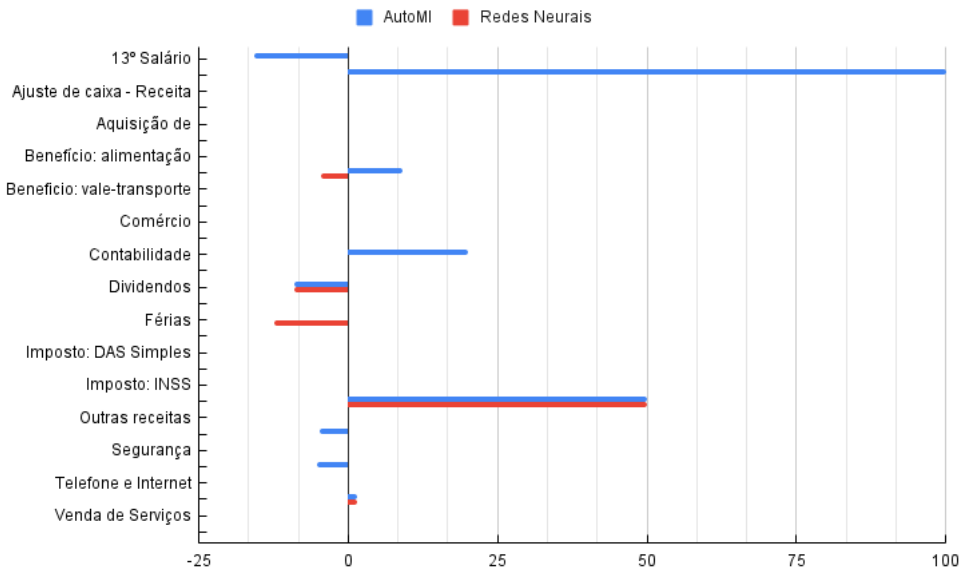

Figura 8. Gráfico de falsos positivos e falsos negativos por categoria para EA

Os gráficos das figura 8 , Figura A.1 e A.2 no material suplementar , representam os erros por categoria para os melhores modelos de AutoML e Redes Neurais, representando em porcentagem a quantidade de verdadeiro negativos e falsos positivos por categoria. É observado na figura 8 que para a categoria Outras receitas ambos os modelo classificaram 50% a mais em relação a quantidade de valores corretos, o mesmo ocorre para a classificação de Dividendos, como os modelos classificando de forma errada aproximadamente 8,4%. Um ponto relevante a ser levantado é que os modelos se comportam de maneira semelhantes, errando em muitas vezes para as mesmas categorias. Esse fato pode ocorrer devido a semelhança dos dados entre categorias, dificultando o aprendizado

de ambos os algoritmos.

As categorias que tiveram um percentagem de erro alto é devido ao fato de pertencerem a categorias que apresentam no máximo 3 instancias no dataset sendo assim havendo um maior impacto na ocorrência de erro. A categoria de Transferências tem um percentual menor devido ao fato de a mesma ser a que apresentar maior número de instancias nos datasets.

Para melhor visualização das classificações correta e incorretas podemos utilizar as matrizes de confusão dos modelos nas figuras 9 e 10, onde sua diagonal principal representa a quantidade de classificações corretas. As matrizes para as empresas EBC1 e ABC2 podem ser encontradas no material suplementar. Segundo a matrix da figura 9 e 10, os modelos tiveram dificuldade na classificação de Beneficio Plano de Saúde (BEN_PLANO) tendo co-relacionados com as categorias de salários, contabilidade e despesas variadas. Essa confusão ocorreu devido ao fato que no dataset original há muitos lançamento com descrição, detalhamento e documento semelhantes, tendo sua maior diferença nos valores de cada lançamento.

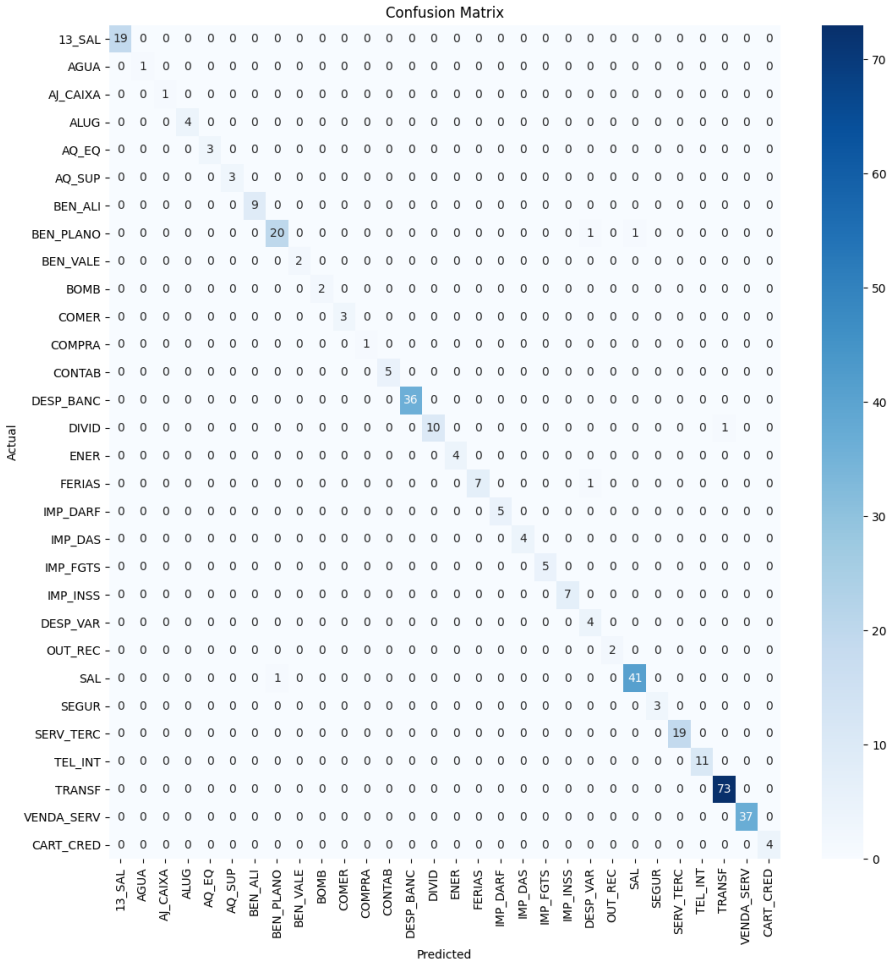

Figura 9. Matriz de confusão do modelo Redes MLP EA

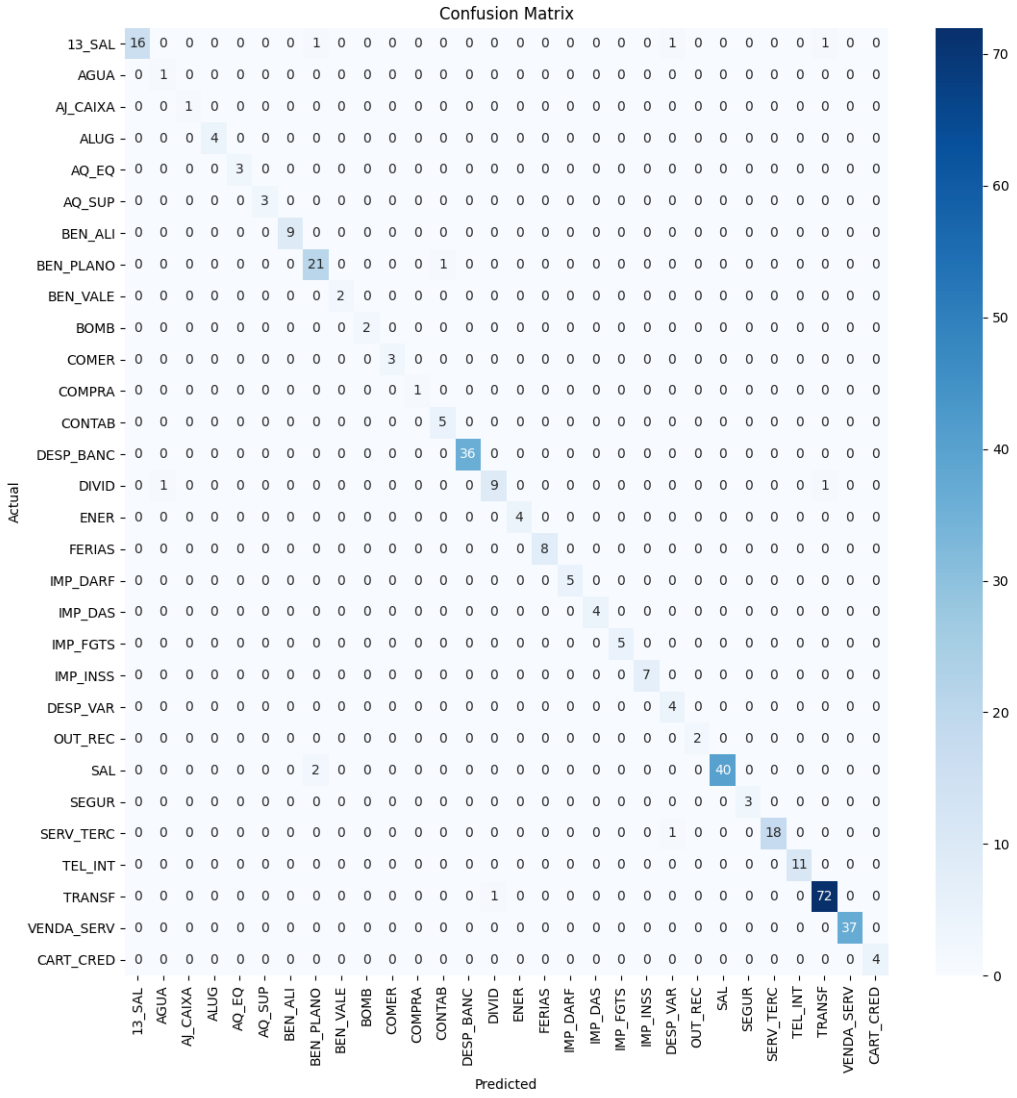

Figura 10. Matriz de confusão do modelo AutoMl EA

Comparando os resultados obtidos com o desenvolvido por Monteiro [MONTEIRO, C.V.F et al. 2018] é visto que ambos os algoritmos de AutoML e RNAs tiveram um desempenho tão bom quanto Regressão Logistica (RL), algoritmo com melhor resultando utilizado em seu trabalho. Com F-measure de 92,8% obtido pela RL para a empresa E1, como pode ser visto na figura 11, os melhores modelos de AutoML e RNAs obtiverem 2 pontos percentuais maior. Para o dataset da empresa EBC1, equivalente a E2C1 no trabalhor de Monteiro, o Automl obteve um F-measure superior aos modelo de de RNAs e os utilizados por Monteiro, tendo um diferença de 4 e 15 pontos percentuais respectivamente. As RNAs obtiveram um desempenho melhor para o dataset da EBC2, equivalente ao E2C2, comparados com o AutoML e a RL, dos dois ultimos tiveram resultados semelhantes com métricas aproximadas de 80%.

	Monteiro e colegas						Autor	
	DataSet	Naive Bayes	Regresão Logistica	SVM	J48	PART	Rendes Neurais	AutoML
Precision	EA	86,40%	93,10%	92,70%	88,60%	85,80%	96,30%	96,23%
	EBC1	63,90%	80,40%	74,70%	72,80%	73,00%	87,00%	89,49%
	EBC2	85,60%	88,80%	87,00%	86,20%	87,20%	90,99%	87,76%
Recall	EA	81,60%	92,70%	92,40%	89,70%	86,40%	96,45%	96,50%
	EBC1	60,40%	80,20%	76,10%	74,70%	73,90%	85,46%	89,88%
	EBC2	83,50%	88,20%	86,90%	86,50%	86,90%	91,20%	87,99%
F-measure	EA	82,40%	92,80%	92,40%	88,90%	85,80%	96,18%	96,27%
	EBC1	59,70%	79,70%	74,70%	73,30%	72,50%	85,17%	88,99%
	EBC2	84,10%	88,30%	86,80%	85,90%	86,90%	90,60%	87,56%
Accuracy	EA	81,57%	92,68%	92,41%	89,70%	86,45%	95,97%	96,18%
	EBC1	60,35%	8018,00%	76,14%	74,74%	73,86%	85,27%	89,19%
	EBC2	83,50%	88,18%	86,95%	86,45%	86,95%	92,30%	89,73%

Figura 11. Tabela de resultados obtidos por Monteiro X Autor
[MONTEIRO, C.V.F et al. 2018]

5. Conclusões e Trabalhos futuros

Visando auxiliar a gestão financeira de pequena e médias empresas, o objetivo desse trabalho foi realizar um estudo comparativo entre dois algorítimos de aprendizado de máquina na classificação de lançamento financeiros. Algorítimos que serão utilizados para mitigar possíveis problemas de categorização de lançamentos financeiros de acordo com o plano de contas e ajudar na tomada de decisão. Os algoritmos utilizados para alcançar essa meta foram o AutoML e Redes Neurais MLP, sendo analisado seus desempenhos e e complexidade no estudo de caso.

Com ambos os algoritmo tendo apresentado métricas próximas a 90% podemos considerar que os dois podem ser utilizados para a tarefa de classificação. Com resultados semelhantes vale a consideração de complexidade para manutenção, tempo de treino e recursos utilizados para o processamento dos mesmos. As Redes Neurais necessitam de maior tempo e ajustes finos nos seus parâmetros para que possa obter um modelo suficientemente bom para realizar as atividades, demandando tempo e maior conhecimento por parte de quem as estão modelando. Quando comparado com o AutoML a complexidade de ajustes é significativamente reduzida e possibilita que um usuário com menos conhecimento possa realizar a criação de modelos.

Comparando com trabalho de Monteiro [MONTEIRO, C.V.F et al. 2018], tivemos um melhoria na classificação dos lançamentos financeiro ao utilizar AutoML e RNAs. Técnicas de seleção de atributos utilizadas por Monteiro podem melhorar o desempenho dos algoritmos utilizados nesse trabalho, sendo recomendado para estudos futuros além da utilização de arquiteturas maisnovas de Redes Neurais como CNN e RNN.

Por fim, podemos constatar o uso de AutoML se torna uma ótima alternativa para a geração de modelos para lançamentos financeiros, visto sua parametrização é feita de forma simples e foi possivel obter resultados semelhantes a de modelagens mais tradicionais. O uso do mesmo em aplicações semelhantes ao do estudo de caso desse trabalho trás um grande beneficio pois é de fácil manutenção e pode aprender sem precisar realizar muitos ajustes.

Referências

ALPAYDIN, E. (2012). *Introduction to Machine Learning*. Massachusetts Institute of Technology, 2 edition.

BRAGA, A. P., CARVALHO, A.P.L., and LUDERMIR, T.B (2000). *Redes Neurais Artificiais: Teoria e Aplicação*. LTC - Livros técnicos e científicos editora S.A., 1 edition.

BROWNLEE, J. (2017). How to one hot encode sequence data in python.

CHAWLA, N. (2009). Data mining for imbalanced datasets: An overview. *Data mining and knowledge discovery handbook.*, page 875–886.

ERTEKIN, S., HUANG, J., BOTTOU, L., and GILES, L (2007). Learning on the border: active learning in imbalanced data classification. *ACM. Proceedings of the sixteenth ACM conference on Conference on information and knowledge management.*, page 127–136.

FERREIRA, H. (2019). Processamento de linguagem natural e classificação de textos em sistemas modulares. Monografia (Bacharelado em Ciência da Computação), Universidade de Brasília, Brasília, Brasil.

FLOWUP (2022). Flowup. Disponivel em: https://www.flowup.me/ .

HARRINGTON, P. (2012). *Machine Learning in action*. Manning Publications Co., 1 edition.

HE, X., ZHAO, K., and CHU, X. (2009). Automl: A survey of the state-of-the-art. *Data mining and knowledge discovery handbook.*, page 875–886.

HERBST, C. (2022). O que é automl e quais são as suas vantagens?

KUROSE, J. F. and ROSS, K. W. (2005). *Redes de computadores e a Internet*. Pearson, 3 edition.

MONTEIRO, C.V.F, DELGADO FILHO, A.J.F., LIMA, R., and SOARES, E (2018). Método supervisionado para categorização automática de lançamentos financeiros. *XIV Encontro Nacional de Inteligência Artificial e Computacional, p. 715–726, 2017.*

SHALEV-SHWARTZ, S. and BEN-DAVID, S. (2014.). Understanding machine learning: From theory to algorithms. *Cambridge university press*.

SHEARER, C. (2000). Tthe crisp-dm model: the new blueprint for data mining. *Journal of data warehousing*, 5(4):13–22.

SILVA, G. A. (2022). Inteligência artificial: como a tecnologia vai ajudar nos erps.

SILVA, L. (2019). Uso de automated machine learning (auto ml) em sistemas de recomendação. Monografia (Bacharelado em Ciência da Computação), Universidade Federal de Campina Grande, Campina Grande - PB, Brasil.

WEVER, M., TORNEDE, A., MOHR, F., and HULLERMEIR, E. (2021). Automl for multi-label classification:overview and empirical evaluation.

A. Material suplementar

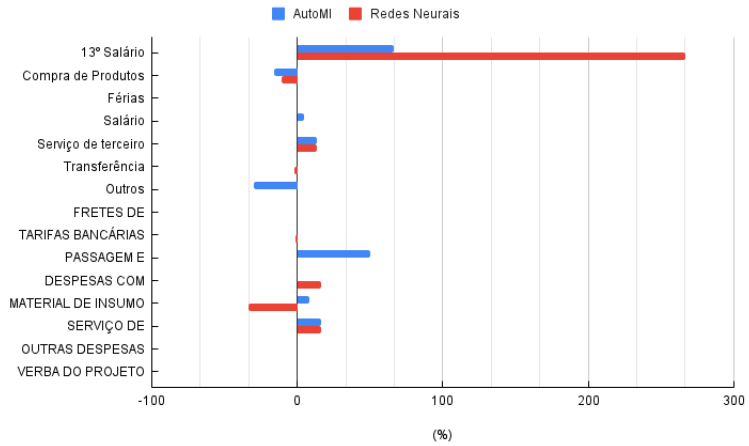

Figura A.1. Gráfico de falsos positivos e falsos negativos por categoria para EBC2

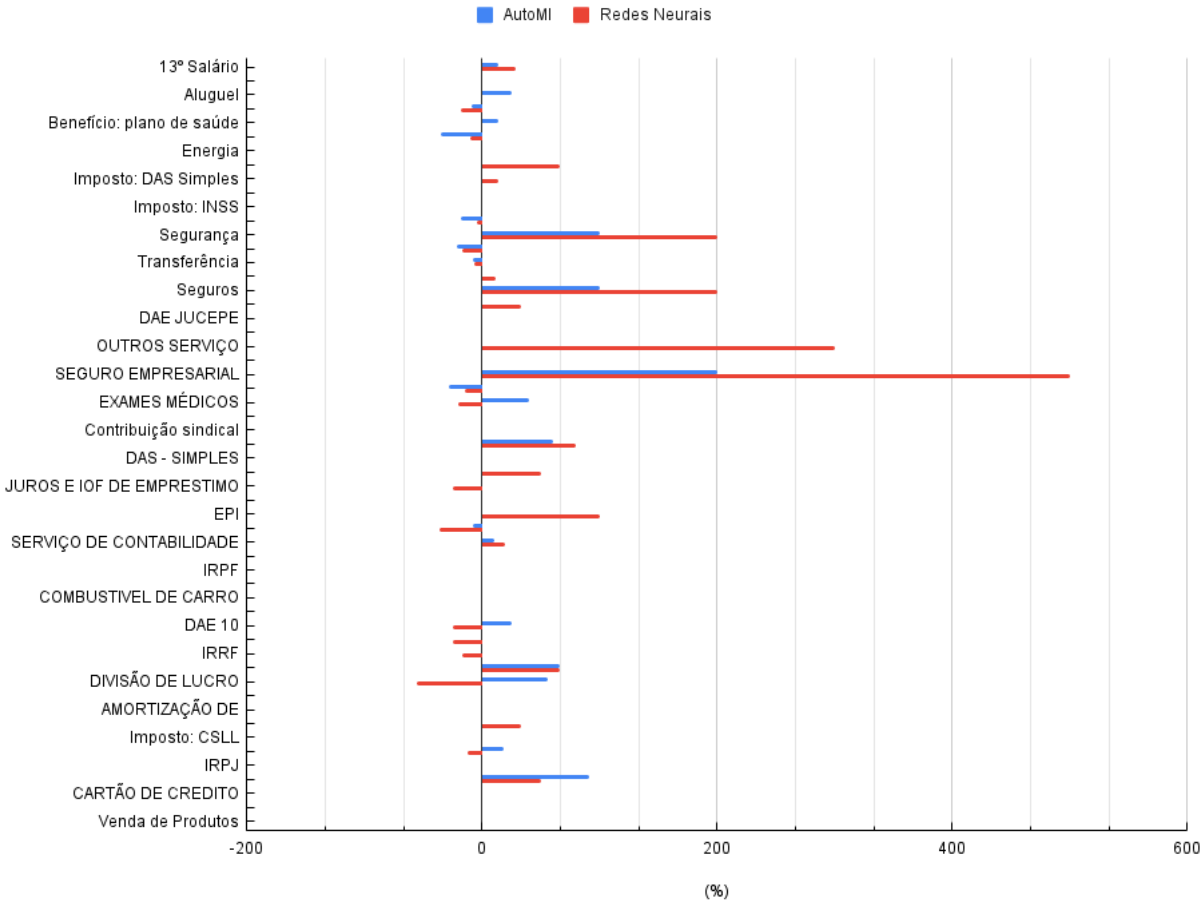

Figura A.2. Gráfico de falsos positivos e falsos negativos por categoria para EBC1

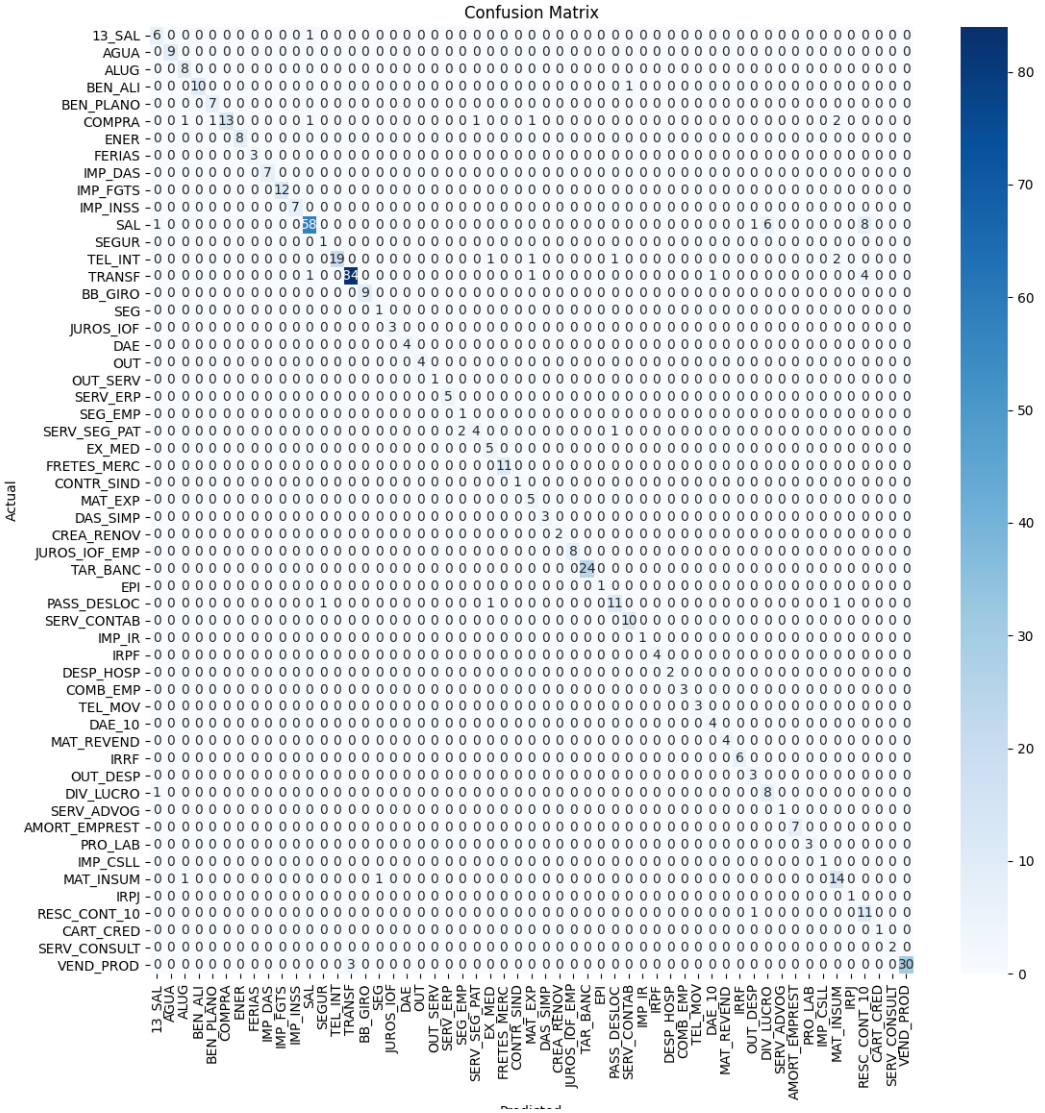

Figura A.3. Matriz de confusão do modelo AutoMl EBC1

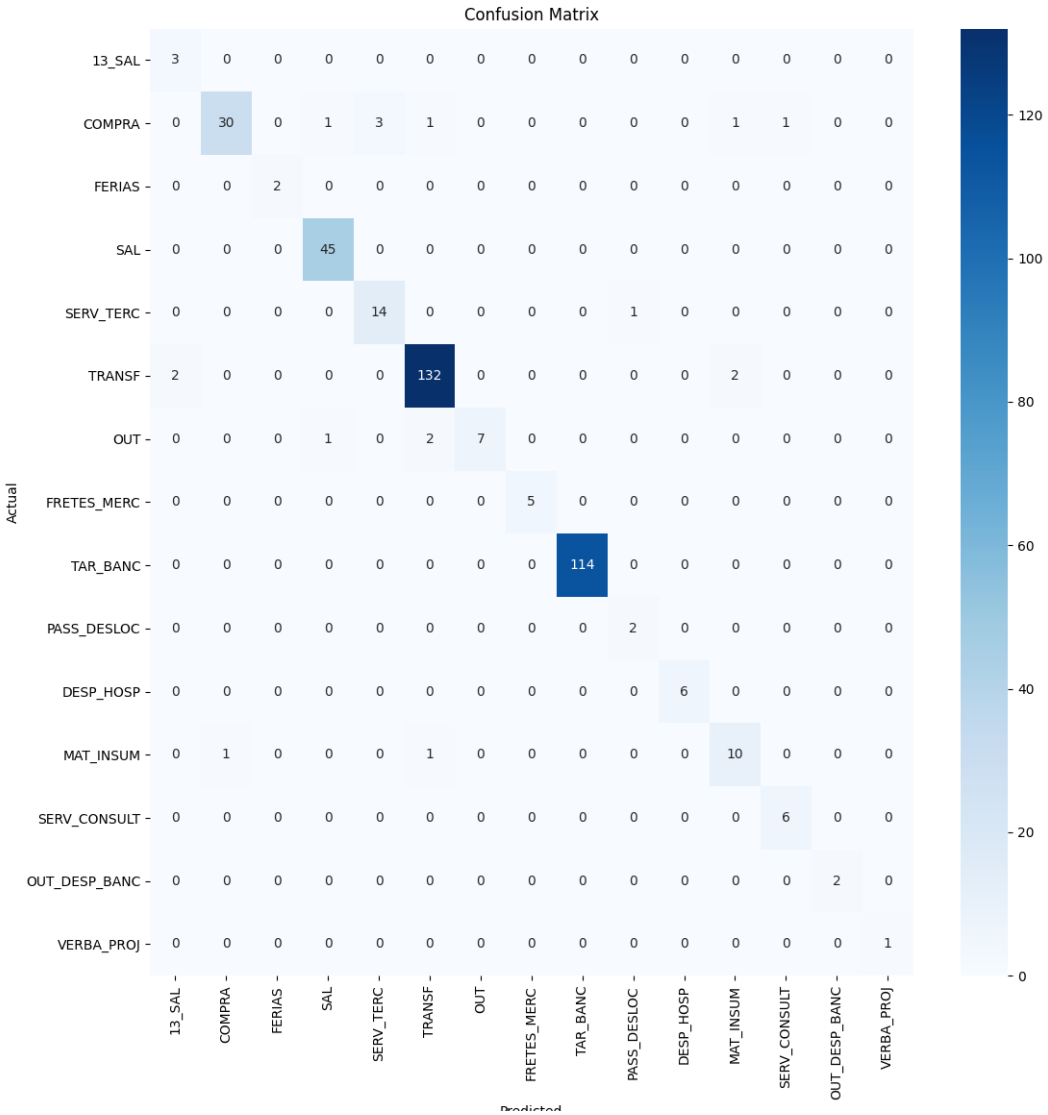

Figura A.4. Matriz de confusão do modelo AutoMl EBC2

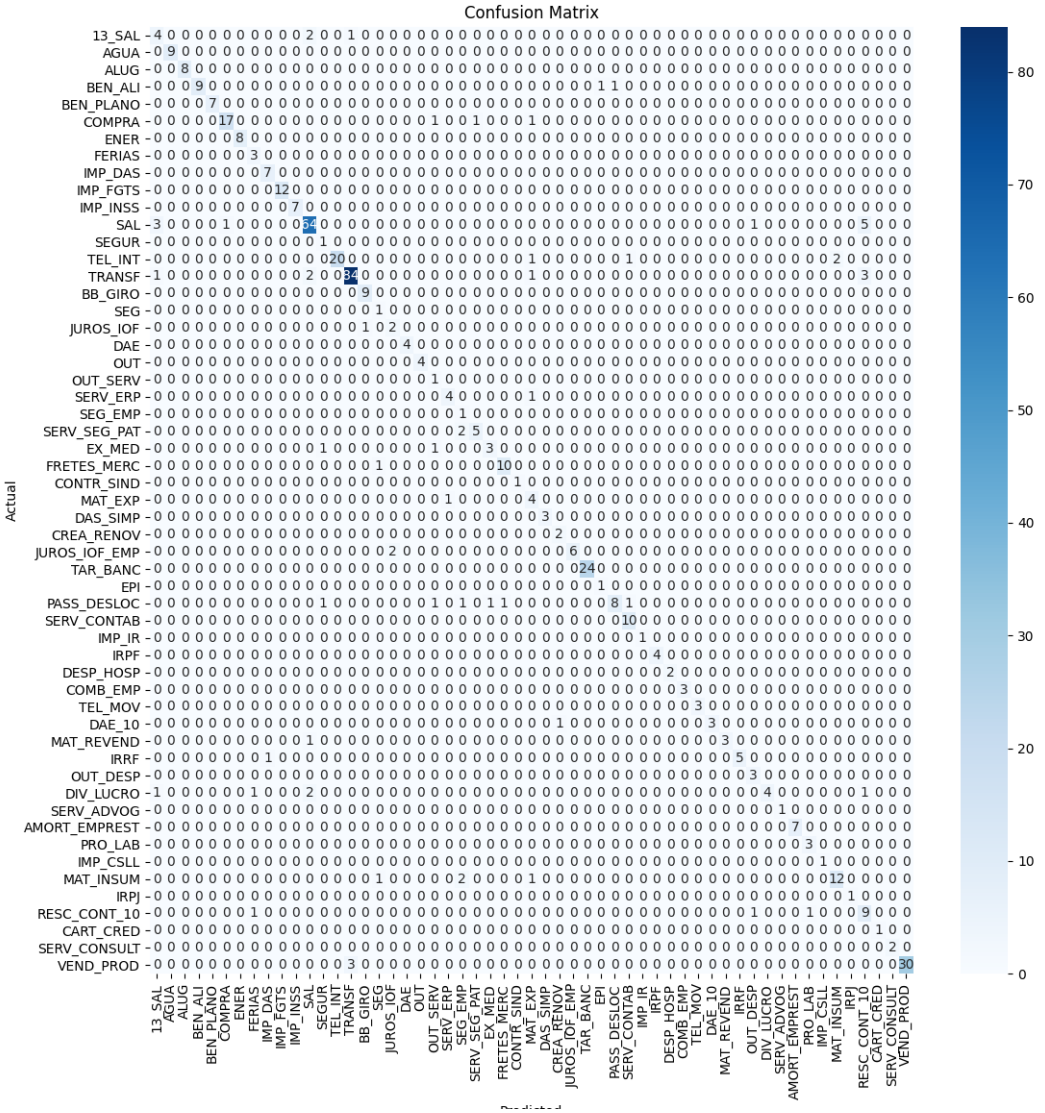

Figura A.5. Matriz de confusão do modelo Redes MLP EBC1

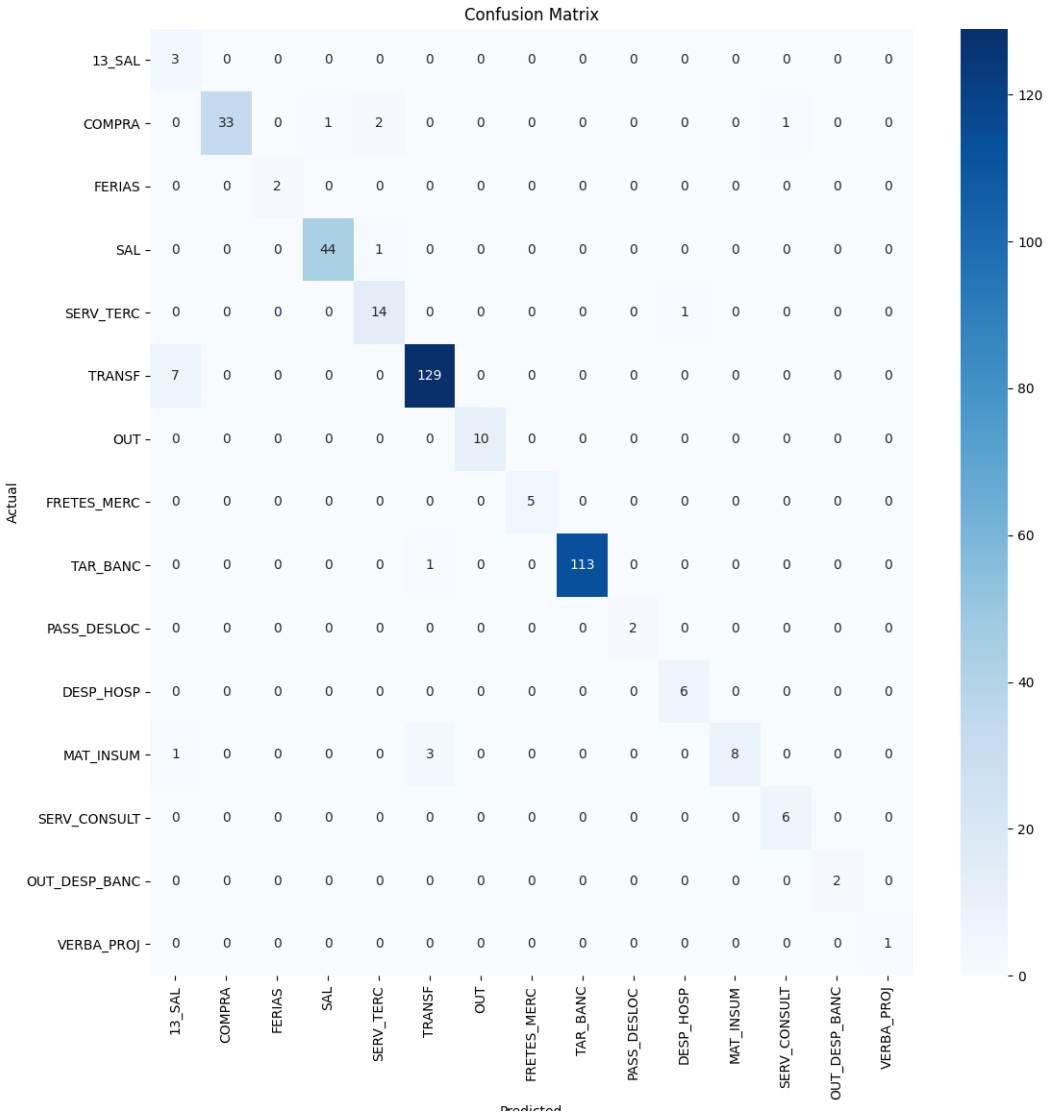

Figura A.6. Matriz de confusão do modelo Redes MLP EBC2