Universidade Federal de Goiás
Instituto de informática
Profa Nádia Félix Felipe da Silva

Relatório da Primeira Competição de Inteligência Computacional

Alunos: Lucas Duarte Sobreira
Tarek Campos Saleh
Diciplina: Inteligência Computacional

Dezembro
2022

Universidade Federal de Goiás
Instituto de Informática
Disciplina: Inteligência Computacional

Relatório

Primeiro Relatório da participação dos Alunos Lucas D. Sobreira e Tarek C. Saleh do Curso Engenharia de Computação da Universidade Federal de Goiás, como requisito parcial para Aprovação da Disciplina Inteligência Computacional.

Alunos: Lucas Duarte Sobreira e Tarek Campos Saleh

Professora: Nádia Félix Felipe da Silva

Dezembro
2022

Conteúdo

1 Resumo

Uma operadora de planos de saúde forneceu precisa decidir se os pedidos devem ou não serem aceitos. Estes pedidos veem de Prestadores que atendem os beneficiários(pacientes). Para reduzir o custo de manter pessoal qualificado o tempo todo para classificar os pedidos em *Autorizados* e *Negados*, o objetivo desta competição era utilizar da Inteligência Computacional para automatizar este processo.
Neste relatório está descrito cada passo realizado nesta competição, começando por a descrição do conjunto de dados, analise do conjunto de dados, os passos específicos da criação do modelo e terminando na analise do resultado final. No resultado final, foi possível concluir que os resultados foram suficientes para um teste teórico, porém poderiam causar diversos problemas em casos de mundo real.

2 Descrição do Conjunto de Dados

O conjunto de dados de teste inicial contava com 31 features, sendo 16 dessas categóricas, 10 numéricas e 5 textos. As features presentes são:

Feature	Tipo	Quantidade	Faltando	Classes
Unnamed: 0	Numérico	227,122	0	227,122
NR_SEQ_REQUISICAO	Numérico	227,122	0	80,699
NR_SEQ_ITEM	Numérico	227,122	0	227,122
DT_REQUISICAO	Numérico	227,122	0	357
DS_TIPO_GUIA	Categórico	227,122	0	3
DT_NASCIMENTO	Numérico	227,112	10	16,557
NR_PRODUTO	Categórico	227,122	0	1
DS_TIPO_PREST_SOLICITANTE	Categórico	227,122	0	12
DS_CBO	Categórico	227,122	0	59
DS_TIPO_CONSULTA	Categórico	10,511	216,611	4
QT_TEMPO_DOENCA	Numérico	266	226,856	17
DS_UNIDADE_TEMPO_DOENCA	Categórico	266	226,856	3
DS_TIPO_DOENCA	Categórico	531	226,591	2
DS_INDICACAO_ACIDENTE	Categórico	209,539	17,583	4
DS_TIPO_SAIDA	Textual	0	227,122	0
DS_TIPO_INTERNACAO	Categórico	59,863	167,259	6
DS_REGIME_INTERNACAO	Categórico	59,863	167,259	3
DS_CARATER_ATENDIMENTO	Categórico	227,122	0	2
DS_TIPO_ACOMODACAO	Categórico	59,781	167,341	8
QT_DIA_SOLICITADO	Numérico	58,995	168,127	34
CD_GUIA_REFERENCIA	Numérico	37,463	189,659	4,610
DS_TIPO_ATENDIMENTO	Categórico	168,045	59,077	13
CD_CID	Textual	131,250	95,872	1,626
DS_INDICACAO_CLINICA	Textual	179,944	47,178	40,428
DS_TIPO_ITEM	Categórico	227,122	0	2
CD_ITEM	Numérico	227,122	0	6,220
DS_ITEM	Textual	227,122	0	6,146
DS_CLASSE	Textual	227,122	0	460
DS_SUBGRUPO	Categórico	227,122	0	72
DS_GRUPO	Categórico	227,122	0	9
QT_SOLICITADA	Categórico	227,122	0	270

Analisando os dados, essas *features* destacam-se de maneira *negativa*:

- **NR_SEQ_REQUISICAO e NR_SEQ_ITEM**: Quantidade de classes muito alta;
- **DS_TIPO_SAIDA, QT_TEMPO_DOENCA e DS_UNIDADE_TEMPO_DOENCA** Baixa amostra de dados em relação ao total;
- **NR_PRODUTO**: Apenas uma classe.

Features que destacam-se de maneira *positiva* são:

- **DS_INDICACAO_ACIDENTE, DS_GRUPO, DS_TIPO_GUIA e DS_TIPO_ITEM**: Alta amostra de dados com poucas classes;
- **DT_NASCIMENTO**: Alta amostra de dados;
- **DS_ITEM, DS_CLASSE, DS_INDICACAO_CLINICA**: Alta amostra de dados e por definição desses itens, tendem a descrever o motivo da requisição.

3 Descrição de atividades

O primeiro passo, por meio da analise do conjunto de dados, foi determinar quais *features* definitivamente não serão utilizadas. Por meio da analise realizada anteriormente, alvos óbvios são *NR_SEQ_REQUISICAO, NR_SEQ_ITEM, DS_INDICACAO_CLINICA, DS_TIPO_SAIDA, QT_TEMPO_DOENCA, DS_UNIDADE_TEMPO_DOENCA e NR_PRODUTO*. Outras *features* que não serão utilizadas, pois serão transformadas em outras *features*, são: *DS_ITEM, DS_CLASSE, DS_INDICACAO_CLINICA*.

O segundo passo foi determinar quais *features* definitivamente serão utilizadas. As *features DS_INDICACAO_ACIDENTE, DS_GRUPO, DS_TIPO_GUIA, DS_TIPO_ITEM e DT_NASCIMENTO* são features que serão utilizadas sem passar por algum processo de transformação além de tratamento de dados vazios. As *features DS_ITEM, DS_CLASSE, DS_INDICACAO_CLINICA* serão utilizadas na forma de *TAMANHO_feature*, visto que são textos não categóricos.

O terceiro passo foi a implementação das transformações e tratamentos necessários, onde foram criadas duas funções, uma para transformar *features* textuais em *features* do tamanho do texto em quantidade de palavras, a outra é uma função para popular dados faltantes e *features*.

O quarto passo foi a utilização do *StandardScaler* e do *OneHotEncoder* em *features* numéricas e categóricas, respectivamente.

O quinto passo foi a criação da tabela de dados, que serão utilizados para treinar o classificador, utilizando as *features* do segundo passo, juntamente com a separação dos dados para o treino (80%) e teste(20%) do classificador. Após a separação dos dados, os classificadores *GaussianNB* e *RandomForestClassifier* eram criados e testados, seguidos da verificação dos resultados e repetindo este passo com tabela de dados sendo alterada, podendo adicionar ou remover *features* que não estão listadas nos primeiros dois passos. Este passo foi repetido até obter bons resultados nos *scores* de *Precisão*, *Recall* e *F1*.

O sexto passo foi a submissão para a plataforma Kaggle para contrapor os resultados obtidos nos dados de treino com os dados de teste disponíveis na competição.

4 Análise dos Resultados

Os resultados final foi obtido utilizando as *features DS_INDICACAO_ACIDENTE, DS_TIPO_ITEM, DS_TIPO_INTERNACAO, DS_TIPO_GUIA, DS_TIPO_ACOMODACAO, DS_TIPO_ATENDIMENTO, DS_TIPO_CONSULTA, DS_TIPO_PREST_SOLICITANTE, DS_GRUPO, DS_REGIME_INTERNACAO, TAMANHO_DS_INDICACAO_CLINICA, TAMANHO_DS_ITEM, TAMANHO_DS_CLASSE, DS_CBO, QT_SOLICITADA, DT_REQUISICAO e DT_NASCIMENTO* no classificador *RandomForest-Classifier*, onde os resultados sobre o conjunto de testes foi de:

	precision	recall	f1-score	support
Autorizado	0.84	0.91	0.87	30832
Negado	0.77	0.63	0.69	14593
accuracy			0.82	45425
macro avg	0.81	0.77	0.78	45425
weighted avg	0.82	0.82	0.82	45425

Figura 1: Resultados do modelo final sobre os dados locais de teste

O resultado obtido no teste disponível na competição foi: 0,68634, onde foi utilizado o *Mean F1-Score* para pontuar

Nestes resultado locais é possível observar que o modelo atingiu resultados decentes para a classificação de um pedido autorizado, com boa precisão e bom recall, entretando, o modelo não atingiu o mesmo nível visto na classificação de autorizados com a classe de negados, principalmente devido ao *Recall*. Assim, podendo concluir que o modelo tende a ter muitos falsos negativos, o que cria um novo grande problema, visto que negar um pedido que deveria ser autorizado pode ser catastrófico, tanto para o paciente final, quanto para a empresa no ponto de vista jurídico. Portanto, podemos concluir que o modelo apresenta uma resposta suficiente para um caso teorico, entretanto, ainda deve ser melhorado em caso de uma aplicação em mundo real.

Bibliografia

AGUIRRE, L. A. Introdução à Identificação de Sistemas, Técnicas Lineares e Não lineares Aplicadas a Sistemas Reais. Belo Horizonte, Brasil, EDUFMG. 2004.