Revista Brasileira de Informática na Educação – RBIE
Brazilian Journal of Computers in Education
(ISSN online: 2317-6121; print: 1414-5685)
https://sol.sbc.org.br/journals/index.php/rbie

Submission: 18/Aug/2025; 1st round notif.: 18/Aug/2025; New version: 18/Aug/2025; 2nd round notif.: 20/Oct/2025;
Camera ready: 20/Oct/2025; Edition review: 20/Oct/2025; Available online: 17/Nov/2025; Published: 17/Nov/2025;

Identificação de tecnologias emergentes de etanol a partir de dados de patentes brasileiras usando ML

Title: Identification of emerging ethanol technologies from Brazilian patent data using ML

Osvaldo Carvalho dos Santos Neto
Escola de Artes, Ciências e Humanidades - USP
ORCID: 0000-0000-0000-0000
osvaldocsantos@usp.br

Gabriel da Silva Simões
Escola de Artes, Ciências e Humanidades - USP
ORCID: 0000-0000-0000-0000
gabriel.s.simoeslp@usp.br

Luan Pereira Pinheiro
Escola de Artes, Ciências e Humanidades - USP
NUSP: 13672471
luanpinheiro@usp.br

Luis Felipe Pinheiro Felisberto
Escola de Artes, Ciências e Humanidades - USP
ORCID: 0000-0000-0000-0000
luis.felipe@usp.br

?abstractname?

Com o grande enfoque dado aos Objetivos de Desenvolvimento Sustentável (ODS), o estudo e uso de tecnologias sustentáveis se torna imprescindível. Um grande aliado nesse desenvolvimento é o etanol, que é um combustível sustentável produzido em larga escala no Brasil. Sabe-se, hoje, que o País tem capacidade para aperfeiçoar as tecnologias em uso e atingir grandes ganhos na produção de etanol, com o advento de algumas tecnologias em desenvolvimento, sem a necessidade de expansão das áreas cultivadas com cana-de-açúcar (Vian, 2022). A partir da análise de patentes brasileiras, faremos uma análise de tecnologias emergentes de etanol utilizando técnicas sofisticadas de aprendizado de máquina semissupervisionado, com aplicação dos algoritmos random forest e support vector machine.

Palavras-chave: *Etanol; Patente; Aprendizado de Máquina; Random Forest; Support Vector Machine; Emergente.*

Abstract

With the strong focus on the Sustainable Development Goals (SDGs), the study and use of sustainable technologies has become essential. A major ally in this development is ethanol, a sustainable fuel produced on a large scale in Brazil. It is now known that the country has the capacity to improve the technologies in use and achieve significant gains in ethanol production with the advent of some technologies under development, without the need to expand the areas cultivated with sugarcane (Vian, 2022). Based on an analysis of Brazilian patents, we will analyze emerging ethanol technologies using sophisticated semi-supervised machine learning techniques, applying random forest and support vector machine algorithms.

Keywords: *Ethanol; Patent; Machine Learning; Random Forest; Support Vector Machine; Emerging; Brazil.*

Cite as: Neto, O. C. S., Simões, G. S., Pinheiro, L. P. & Felisberto, L. F. P. (2025). Identificação de tecnologias emergentes de etanol a partir de dados de patentes usando ML. Revista Brasileira de Informática na Educação.

1 Introdução

O etanol ou álcool etílico ou apenas álcool é uma substância de grande importância na indústria em geral. No setor energético ele se destaca como uma alternativa menos poluente e renovável comparada aos combustíveis fósseis já que, dentre outras vantagens, emite 73% a menos de CO2 que a gasolina (Ferreira, 2009). Segundo a Barros (2021), a indústria alcoolquímica que utiliza o etanol como matéria-prima para a fabricação de produtos poderá vir a substituir a petroquímica, colocando o álcool etílico como uma opção de matéria prima acima do petróleo.

O Brasil é o maior produtor do mundo de cana-de-açúcar e na safra 2020/2021 foi responsável pela produção de 654,5 milhões de toneladas, destinados à produção de 41,2 milhões de toneladas de açúcar e 29,7 bilhões de litros de etanol (CONAB - Companhia Nacional de Abastecimento, 2021) sendo o estado de São Paulo líder na produção no país com 425,6 milhões de toneladas colhidas e 14,7 milhões m³ de etanol produzidos segundo o SEADE - Sistema Estadual de Análise de Dados (2021).

Esses dados colocam o etanol como um produto de extrema importância para o avanço em relação aos Objetivos de Desenvolvimento Sustentável (ODS), em particular o ODS 7: energia limpa e sustentável (ONU - Organização das Nações Unidas, 2023). Sendo assim, o estudo de tecnologias de etanol emergentes desenvolvidas no Brasil se faz relevante no cenário brasileiro e global, além de servir de apoio para o avanço do ODS 9 (indústria, inovação e infraestrutura), especialmente nos tópicos 9.5 e 9.b.

2 Fundamentos Teóricos

Para o desenvolvimento do presente trabalho, alguns fundamentos teóricos são essenciais, tais como o conceito de patentes e conceitos relacionados aos algoritmos de aprendizado de máquina usados na identificação das tecnologias emergentes. Esses fundamentos serão apresentados nesta seção.

Patente é um título de propriedade temporária sobre uma invenção ou modelo de utilidade, outorgado pelo Estado aos inventores (SEBRAE - Serviço Brasileiro de Apoio às Micro e Pequenas Empresas, 2017). Em 2023, o INPI - Instituto Nacional da Propriedade Industrial (2023) registrou o depósito de 27.918 patentes, um crescimento de 2,9% em relação ao período anterior.

Uma predição é o resultado de uma análise que permite inferir previamente conclusões sobre o futuro, sendo essas inferências consideradas valiosas para a indústria para realizar a tomada de decisões de forma a minimizar os riscos e custos, e atingir os objetivos suavemente (LIN, 2021, p.74 citado em (Lee et al., 2022, p. 5)), neste trabalho, será comparado o desempenho de um algoritmo de random forest e de um support vector machine na realização de predições assertivas.

Uma árvore de decisão é um tipo de diagrama hierárquico que ajuda a visualizar etapas, decisões e o possível resultado de cada decisão popular em machine learning para tarefas de classificação e regressão de modelos (IBM - INTERNATIONAL BUSINESS MACHINES, 2023b). Random forest é um algoritmo utilizado para tarefas de classificação e regressão que combina a saída de múltiplas árvores de decisão para alcançar um único resultado (IBM - INTERNATIONAL

BUSINESS MACHINES, 2023a).

Uma máquina de vetores de suporte (SVM) é um algoritmo supervisionado de aprendizado de máquina que classifica dados encontrando uma linha ou hiperplano ótimo que maximiza a distância entre cada classe em um espaço N-dimensional (IBM - INTERNATIONAL BUSINESS MACHINES, 2023c).

A base da possibilidade de inferir quais são as tecnologias promissoras vem da implicação que a rede de patentes conectadas pelas citações se comporta de forma similar a um grafo direcionado evoluindo, cujas conexões representam referências a uma tecnologia anterior como base para criação de uma nova, logo os vértices de origem mais centrais podem ser destacados como fonte de inovação. Logo, algoritmos de aprendizado supervisionado como Random Forest e Support Vector Machine podem ser treinados a partir de dados anteriores para reconhecer e destacar as características desses vértices, de forma a obter as patentes, autores e tecnologias citadas.

Mais especificamente, neste trabalho faremos uso de algoritmos de aprendizado semissupervisionado, que consistem em algoritmos que fazem uso das técnicas de aprendizado supervisionado em um pequeno conjunto de dados, como base para um aprendizado não supervisionado para um conjunto grande de dados (Zhu, 2005). São usados principalmente quando adquirir rótulos para os dados é difícil (e.g. classificação de potencial de patentes a partir de revisão humana).

3 Trabalhos Relacionados

No estudo de Chung et al. (2020), "Early detection of valuable patents using a deep learning model: Case of semiconductor industry", é proposto um modelo de aprendizado profundo combinando CNN e LSTM para extrair características semânticas de patentes, classificando-as em três níveis de valor com base em citações futuras anuais. O modelo apresentou mais de 75% de precisão na identificação de patentes promissoras no setor de semicondutores.

De forma complementar, Kwon e Geum (2020) utilizaram 17 indicadores de patentes e técnicas de machine learning para prever invenções promissoras, destacando que a qualidade da acumulação de conhecimento é o preditor mais relevante para o sucesso das invenções (Kwon & Geum, 2020).

Além disso, o estudo de Park et al. (2021) avaliou patentes de circuitos integrados por meio de uma estratégia multidimensional de indicadores e diferentes modelos de machine learning, verificando que o algoritmo Random Forest alcançou precisão e acurácia superiores a 95% na classificação de patentes de alto valor (Park et al., 2021). No domínio de veículos elétricos, Li et al. (2021) abordaram a previsão de citações futuras como um problema de classificação, utilizando SVM otimizado para identificar patentes altamente citadas e mapear frentes tecnológicas emergentes (Li et al., 2021).

Em patentes de biomedicina têxtil, Zhao et al. (2021) desenvolveram o modelo BioTexVal, integrando BERT e múltiplos algoritmos de machine learning para prever o valor das patentes, alcançando aproximadamente 88% de acurácia ao treinar com 113.428 patentes (Zhao et al., 2021). No contexto brasileiro, Kazmi et al. (2022) investigaram o papel do país no desenvolvimento de tecnologias para produção de etanol de segunda geração por meio da análise de patentes publi-

cadas entre 2006 e 2015 (Kazmi et al., 2022).

No estudo "Forecasting emerging technologies: A supervised learning approach through patent analysis" de Kyebambe et al. (2017) desenvolveu um algoritmo para rotular automaticamente patentes como "emergentes" ou "não emergentes" e usar esses dados para treinar modelos de aprendizado de máquina supervisionado. No entanto, diferente do nosso estudo, o artigo de Kyebambe busca identificar ondas tecnológicas emergentes enquanto nós buscamos identificar uma tecnologia emergente apenas.

Por fim, o trabalho de Park et al. (2020) apresentou uma abordagem semi-supervisionada para identificar tecnologias emergentes, combinando um pequeno conjunto de patentes rotuladas por especialistas com um grande conjunto não rotulado, permitindo rotular automaticamente muitas patentes e facilitar a descoberta de inovações promissoras (Park et al., 2020). Nosso estudo se inspira nesta metodologia, buscando preencher a lacuna existente na análise de patentes recentes de tecnologias emergentes de etanol no Brasil.

4 Metodologia

Para o desenvolvimento do presente estudo foi necessário uma etapa inicial de coleta de dados. Para isso, utilizou-se das bases de dados de patentes do Instituto Nacional da Propriedade Industrial (INPI) e da Organização Mundial da Propriedade Intelectual (WIPO).

Adicionou-se à pesquisa de patentes do INPI a palavra-chave "etanol", que retornou resultados satisfatórios. Limitações na infraestrutura do sítio do INPI impedem a busca por consultas complexas.

Utilizamos os mesmos códigos IPC selecionados por Perrone et al. (2011), uma vez que o foco do trabalho deles é o mesmo do nosso: a identificação e análise de tecnologias associadas à produção de etanol. Assim, aproveitamos a expertise previamente validada para compor nossa estratégia de busca. Além disso, testamos diversas consultas diferentes na WIPO, até chegar na consulta abaixo, que foi a abordagem adotada.

```
IC:(C12P OR C12N OR C10L OR C07C OR A23B) AND
FP:(metanol OR methanol OR sugar OR etanol OR ethanol OR cana OR stover
    OR celulose OR bagasse OR madeira OR wood OR wooden OR cellulose
   OR bagaco OR beterraba OR beet OR sugarcane OR sucrose OR acucar*
   OR melaco OR alcoo* OR alcohol OR bioetanol OR bioethanol OR etilic
   * OR ethyl OR milho OR corn OR soy OR soybean OR soja OR cereal OR
   trigo OR starch OR lignocellulose OR lignocelulose OR palha OR
   residuo* OR biomass OR biomassa)
```

Ambas as bases possuem como retorno de suas pesquisas informações como título da patente, nome do inventor, data de publicação e número de publicação. Os números de publicação extraídos das pesquisas feitas nas bases de dados servem como identificadores globais das patentes. Cada identificador foi inserido em um arquivo de valores separados por vírgula (CSV). A partir do arquivo foi realizado um tratamento nos números de publicação uma vez que eles possuem um formato diferente na plataforma da WIPO que impede a consulta no site do Google Patents usando

os números de publicação diretamente.

Com o tratamento realizado foi criado um web scrapper utilizando a linguagem de programação python para extrair as variáveis necessárias para o treinamento dos modelos de aprendizado de máquina. O scraper foi utilizado para extrair e calcular as variáveis a partir dos dados sobre as patentes em suas respectivas páginas no Google Patents. As variáveis calculadas foram:

1. Número de reivindicações independentes (independent claims);
2. Número de inventores (inventors);
3. Número de citações anteriores (backward citations);
4. Número de imagens da patente (number of patent images);
5. Número de membros da família de patentes (number of patent family members);
6. Número de referências não-patente (number of non-patent references);
7. Número de classificações (number of IPCs and CPCs);
8. Diversidade de tecnologia (Diversity of technology);
9. Citações a termo (forward citations).

Sendo essa última de grande importância quantitativa, pois similar aos algoritmos pioneiros de ranqueamento de páginas, a importância de um artigo pode ser medida através da quantidade de artigos que o citam.

Para o treinamento dos modelos, será utilizado um conjunto de patentes publicadas entre 1995 e 2013. Neste conjunto, os 10% de patentes mais citadas anualmente serão rotuladas como "promissoras" (nossa variável-alvo, pois irá orientar os modelos treinados quanto a identificação de patentes promissoras de um período futuro), e as demais como "não promissoras". A partir do conjunto rotulado, vamos separar 10% para ser o conjunto de testes do treinamento. Dos 90% restantes faremos um treinamento semi-supervisionado composto por 30% do conjunto inicial rotulado e o resto será o conjunto não rotulado. Esse processamento dos conjuntos e das patentes promissoras será feito com Python através da biblioteca de ciência de dados Pandas. Dessa forma, estes dados rotulados e não rotulados serão utilizados para treinar os algoritmos de aprendizado de máquina Random Forest e Support Vector Machine (SVM), sendo o primeiro especialmente relevante para evitar vieses acidentais a respeito das tecnologias destacadas como “promissoras” decorrentes do caso de uma única árvore, e o segundo para uma correta classificação sobre o status das mesmas tecnologias, as performances de ambos serão validadas em cima do conjunto de testes. As métricas utilizadas para medir a performance dos algoritmos serão a Acurácia e o F1-Score. Por fim, com os modelos treinados, utilizaremos o conjunto de patentes publicadas entre 2019 e final de 2024 para prever se essas patentes são promissoras e analisar o cenário atual e futuro do etanol no Brasil.

5 Resultados Preliminares

Ao acessar a plataforma do Instituto Nacional da Propriedade Industrial obtivemos um total de 908 patentes em 18 de agosto de 2025 e na plataforma da WIPO obtivemos outro conjunto de 6242 patentes em 10 de outubro de 2025. O conjunto de patentes obtido do INPI possui informações como número do pedido, data de depósito, título da patente e o código da Classificação Internacional de Patentes (IPC). Já o conjunto de patentes obtido do WIPO possui informações como número de pedido, número da submissão, data da submissão, país, título e IPC.

No entanto, só os dados dos dois conjuntos de patentes não são suficientes para o treinamento dos modelos de inteligência artificial. Por essa razão, desenvolvemos um web scraper para complementar os dados. Os dados buscados para cada patente foram: url, título, data de publicação, citações da patente (patent citations), citador por (cited by), resumo (abstract) , descrição (description), quantidade de imagens, documentos similares (similar documents), aplicações que reivindicam prioridade (application claiming priority), eventos Legais(legal events), conceitos (concepts), inventores (authors), outras linguagens (other languages), worldwide applications?, info e links externos (external links).

6 Discussão e Conclusão

A obtenção dos dados representa a parte mais crítica do estudo, pois dados de má qualidade podem comprometer a acurácia do aprendizado dos algoritmos, impedindo que eles encontrem e classifiquem com precisão a "promissoridade" de uma patente. No entanto, a extração dos dados englobou o cálculo de variáveis que exigiram maior complexidade computacional e o acesso a outras fontes de dados que armazenam informações de patentes para complementar os dados extraídos do Google Patents. Outro fator relevante para a busca de dados em outras plataformas foi o fato de algumas patentes mais antigas não estarem disponíveis facilmente na internet. Todos esses fatores fizeram com que a etapa de coleta de dados precisasse de maior alocação de tempo.

Referências

Barros, T. D. (2021). *Etanol* [Acesso em: 20 ago. 2025.]. https://www.embrapa.br/agencia-de-informacao-tecnologica/tematicas/agroenergia/p-d-e-i/etanol

Chung, J., Kim, H., & Lee, S. (2020). Early detection of valuable patents using a deep learning model: Case of semiconductor industry. *Technological Forecasting and Social Change*. https://www.sciencedirect.com/science/article/pii/S0040162520309720

CONAB - Companhia Nacional de Abastecimento. (2021). *Série Histórica das Safras* (Acesso em: 20 ago. 2025.). https://www.conab.gov.br/info-agro/safras/serie-historica-das-safras

Ferreira, A. L. (2009). *Estudo mostra que etanol de cana emite menos gás carbônico para a atmosfera do que a gasolina* [Acesso em: 20 ago. 2025.]. https://www.embrapa.br/busca-de-noticias/-/noticia/18044516/estudo-mostra-que-etanol-de-cana-emite-menos-gas-carbonico-para-a-atmosfera-do-que-a-gasolina

IBM - INTERNATIONAL BUSINESS MACHINES. (2023a). *O que é random forest?* (Acesso em: 20 ago. 2025.). https://www.ibm.com/br-pt/think/topics/random-forest

IBM - INTERNATIONAL BUSINESS MACHINES. (2023b). *O que é uma árvore de decisão?* (Acesso em: 20 ago. 2025.). https://www.ibm.com/br-pt/think/topics/decision-trees

IBM - INTERNATIONAL BUSINESS MACHINES. (2023c). *O que são SVMs?* (Acesso em: 20 ago. 2025.). https://www.ibm.com/br-pt/think/topics/support-vector-machine

INPI - Instituto Nacional da Propriedade Industrial. (2023, dezembro). Boletim mensal de propriedade industrial: estatísticas preliminares [Boletim publicado pela Presidência, Diretoria Executiva, Assessoria de Assuntos Econômicos (AECON).].

Lee, C.-W., Tao, F., Ma, Y.-Y., & Lin, H.-L. (2022). Development of Patent Technology Prediction Model Based on Machine Learning. *Axioms*, *11*(6), 253. https://doi.org/10.3390/axioms11060253

ONU - Organização das Nações Unidas. (2023). *7 - Energia limpa e acessível* (Acesso em: 20 ago. 2025.). https://brasil.un.org/pt-br/sdgs/7

Perrone, C. C., Appel, L. G., Lellis, V. L. M., et al. (2011). Ethanol: An Evaluation of its Scientific and Technological Development and Network of Players During the Period of 1995 to 2009. *Waste Biomass Valor*, *2*, 17–32. https://doi.org/10.1007/s12649-010-9049-z

SEADE - Sistema Estadual de Análise de Dados. (2021). *São Paulo lidera produção de etanol no país* (Acesso em: 20 ago. 2025.). https://informa.seade.gov.br/sao-paulo-lidera-producao-de-etanol-no-pais/

SEBRAE - Serviço Brasileiro de Apoio às Micro e Pequenas Empresas. (2017). *O que é patente?* (Acesso em: 20 ago. 2025.). https://sebrae.com.br/sites/PortalSebrae/ufs/mt/artigos/o-que-e-patente,af88f8ba5a17a510VgnVCM1000004c00210aRCRD

Vian, C. E. F. (2022). *Etanol* [Acesso em: 20 ago. 2025.]. https://www.embrapa.br/agencia-de-informacao-tecnologica/cultivos/cana/pos-producao/alcool/tecnologias-emergentes/etanol

Zhu, X. (2005). *Semi-Supervised Learning Literature Survey* (rel. técn.). University of Wisconsin-Madison Department of Computer Sciences. http://digital.library.wisc.edu/1793/60444