# Título:

Early delirium detection using machine learning algorithms

Machine learning approach do detect delirium at hospital setting.

# Resumo:

Abstract:

Delirium is a common manifestation of severe acute neuropsychiatric dysfunction, prevalent in hospital settings, which due to the complex multifactorial causes is often underdiagnosed and neglected. Early detection of delirium is a critical concern that can be effectively addressed using machine learning techniques. As such, methods to improve the accuracy of machine learning classification models for the detection of delirium are covered in this document. The aim is to develop and validate a tool for use in a hospital setting to accurately identify delirium during the admission of patients. A database collected at a Portuguese hospital between 2014 and 2016 was used to conduct this experimental research. Available data comprised 511 records and 54 variables, including patient demographics, medications administered, admission category, urgent admission, hospitalization period, history of alcohol abuse and laboratory results. The methodology of the proposed system included data pre-processing, data imbalances processing, feature selection, machine learning classifiers, evaluating the performance of classifiers and the development of a python web-based application. The model achieved consists of 26 predictors assessed during admission to a healthcare facility. Overall, the best overcome were obtained by combining the SelectFromModel method with the logistic regression method, with an AUC-ROC result of 0.833 and AUC-PR of 0.582. Although the prediction model can be enhanced, this approach could be a useful support tool to identify patients at increased risk for delirium in healthcare settings.

Abstract:

Delirium is a common manifestation of severe acute neuropsychiatric dysfunction, prevalent in hospital settings, which due the complex multifactorial causes is often underdiagnosed and neglected. Early detection is a critical issue that can be effectively achieved by machine learning (ML) techniques. Thus in this article, the methods to improve the accuracy of ML classification models for the detection of delirium are investigated. The objective is to develop and validate an instrument for use in the hospital settings to accurately recognize delirium in admitted patients. The dataset under this research experimentation is referred from the data collected from a Portuguese hospital between 2014 and 2016. The data available included 511 records and 54 features, which included patient demographics, medical histories, physiological measurements, medications administered, and lab results. The methodology of the proposed system included: data pre-processing; data imbalance handling; feature selection; machine learning classifiers; classifier’s performance evaluation and development of a web app. ﻿The model achieved consists of 26 predictors assessed at healthcare admission. Overall, the best results were achieved by the combination of SelectFromModel with logistic regression, with AUC-ROC of 0.833 and AUC-PR of 0.582. Although the prediction model can still be improved, this approach might be a useful support tool to identify patients at increased risk of delirium in healthcare settings.

age, history of cognitive impairment, history of alcohol abuse, blood urea nitrogen, admission category, urgent admission, mean arterial blood pressure, use of corticosteroids, and respiratory failure.

Wrapper-based feature selection approach has been used to identify the important features.

The methodology of the proposed system included: Data Pre-processing; Data imbalance handling; Feature Selection; Machine Learning Classifiers; classifier’s performance evaluation.

﻿To develop and validate a model based on data collected from a Portuguese hospital between 2014 and 2016 to predict delirium development.

In this paper, was analysed data collected from a Portuguese hospital between 2014 and 2016. The data available included patient demographics, medical histories, physiological measurements, medications administered, and lab results and 511 records. After data preparation was obtained a dataset with

Considering the imbalance of delirum detection datasets resulted in low classification performance was used Adaptive Synthetic Sampling (ADASYN) method to balance datasets.

Based on the experimental results, the improvement of precision, recall, F1 scores and AUC values after ADASYN is then analyzed. Experiments show that the proposed method can be applied to delirium detection and can effectively improve the classification accuracy of that kind of disease.

making 26 features. The final dataset included 434 records but with a low percentage of success of delirium (27,9%). To handle this was performed ADASYN balancing training strategy. Then was performed a two-stage ML comparison study. Firstly, we consider Random Forest (RF) and the wrapper method as features selection strategy. Then was performed Logistic Regression (LR) and the wrapper method as features selection strategy.

To develop and validate an instrument for use in the hospital settings to accurately recognize delirium in admitted patients.

Overall, the best results were achieved by the combination of SelectFromModel with logistic regression.

multicomponent non-pharmacological risk factor approaches could be the most effective strategy for prevention.

The abstract should briefly summarize the contents of the paper in 15--250 words

# Introdução:

Nas últimas décadas, tem-se registado um aumento considerável da esperança média de vida, no entanto, paralelamente a este envelhecimento da população, tem-se assistido a um acréscimo de doenças crónicas caracterizadas por produzirem elevados graus de incapacidade ((DuGoff, Canudas-Romo, Buttorff, Leff, & Anderson, 2014)) sendo muitas vezes responsáveis por pressões sobre o sistema de saúde.

O envelhecimento biológico a que o indivíduo está sujeito tem uma evolução variável, e nesta degradação natural podem ocorrer inúmeras alterações no normal funcionamento do organismo, destacando-se aqui a deterioração ocorrida a nível cognitivo, esta que é uma alteração bastante comum em pessoas idosas. Pequenas alterações no indivíduo causadas pela mudança de estado da doença, medicamentos, internamento, um défice físico apresentado recentemente ou uma combinação destes podem promover estados de confusão e desorientação. O delirium é uma manifestação comum de disfunção neuropsiquiátrica aguda grave, muito prevalente em ambiente hospitalar, que devido à sua variabilidade de apresentação clínica é frequentemente subdiagnosticado e negligenciado. O delirium pode afectar pessoas de todas as idades, mas afecta predominantemente adultos idosos hospitalizados e está associado não só a um aumento da morbilidade e mortalidade, como também a um aumento do tempo de internamento, bem como a uma deterioração do estado físico e mental do individuo. A capacidade para avaliar o delirium é uma componente essencial na estratégia de avaliação do doente de modo a prevenir ou tratar a ocorrência desta perturbação. Estudos apontam que o diagnóstico precoce e uma abordagem adequada, estão associados a uma redução das taxas de morbilidade e mortalidade associadas ao delirium (Sharon K. Inouye et al., 2014; Mittal et al., 2011).

Nos últimos anos, têm sido desenvolvidos vários instrumentos de avaliação clínica para o delirium, o que representou um importante avanço metodológico no estudo e diagnóstico desta perturbação. Como o delirium pode passar facilmente despercebido aos profissionais de saúde, especialmente em doentes internados em UCI e SU, torna-se importante o uso de ferramentas de rastreio que permitam detetar de forma mais precoce este distúrbio. Não só, pela possível melhoria na qualidade de vida dos pacientes, mas também pela possível contenção de custos relacionados com o tratamento dos doentes. Face a esta necessidade, têm vindo a ser desenvolvidas e validadas ferramentas de rastreio para esta perturbação, com o intuito de serem usadas na prática clínica diária (De & Wand, 2015).

Atualmente existem vários instrumentos validados de apoio ao diagnóstico do delirium, que foram adequados consoantes a tipologia de doentes envolvidos (Leonard et al., 2014). Assim, já existem mais de 30 instrumentos desenvolvidos e testados para a avaliação do delirium (Adamis, Sharma, Whelan, & MacDonald, 2010; C. L. Wong, Holroyd-Leduc, Simel, & Straus, 2010). No entanto, em ambiente hospitalar, o tempo muitas vezes é escasso e há uma necessidade de obter respostas rapidamente, pelo que este tipo de síndrome pode passar despercebida a muitos profissionais de saúde. Daí, ter surgido a necessidade de investigar ferramentas que possam permitir a elaboração de um diagnóstico de forma mais rápida e precisa que o habitual. E, uma vez que já se possuíam dados recolhidos entre 2014 e 2016 num hospital português, surgiu a ideia de desenvolver uma aplicação web que determina o risco de desenvolvimento de delirium de um paciente no contexto hospitalar.

Introduction:

Over the past years, average life expectancy has increased, however, alongside with this ageing population, there has been a significant increase in chronic diseases (DuGoff et al., 2014). These characterized by producing high degrees of disability and often responsible for pressures on the health system. The biological ageing to which the individual is subjected has a variable evolution, and goes through a natural degradation where numerous changes can occur in the normal functioning of the body, highlighting here the deterioration that occurs at the cognitive level, which is a very common change in older people. Small changes in the individual caused by the change of disease status, medication, hospitalization, a recently presented physical deficit, or a combination of these can promote states of confusion and disorientation. Delirium is a common manifestation of severe acute neuropsychiatric dysfunction, very prevalent in the hospital setting, which due to its variability of clinical presentation is often underdiagnosed and neglected. Delirium can affect people of all ages, but predominantly affects hospitalized older adults and is associated not only with increased morbidity and mortality, but also with increased length of stay, as well as deterioration in the individual's physical and mental status. The ability to assess delirium is an essential component in the patient assessment strategy in order to prevent or treat the occurrence of this disorder. Studies indicate that early diagnosis and an appropriate approach are associated with a reduction in morbidity and mortality rates associated with delirium (Inouye, Westendorp, & Saczynski, 2014; Mittal et al., 2011)

In recent years, several clinical assessment tools for delirium have been developed, which has represented an important methodological advance in the study and diagnosis of this disorder. Since delirium can easily go unnoticed by health care professionals, especially in ICU and ER patients, it is important to use screening tools that allow for an earlier detection of this disorder. Not only for the possible improvement in the quality of life of patients, but also to contain costs related to patient treatment. In view of this need, screening tools for this disorder have been developed and validated, with the purpose of being used in daily clinical practice (De & Wand, 2015).

Currently, there are several validated instruments to support the diagnosis of delirium that have been adapted according of patients involved (Leonard et al., 2014). Consequently, there are already more than 30 instruments developed and tested for the assessment of delirium

(Adamis, Sharma, Whelan, & MacDonald, 2010; Wong, Holroyd-Leduc, Simel, & Straus, 2010)

(Adamis, Sharma, Whelan, & MacDonald, 2010; C. L. Wong, Holroyd-Leduc, Simel, & Straus, 2010). However, in a hospital setting where time is often short and answers must be obtained quickly, it is likely that this type of syndrome goes unnoticed by many medical professionals. Hence, the need has arisen to investigate tools to make diagnosis more quickly and accurately than usual. And, since we already had data collected between 2014 and 2016 in a Portuguese hospital regarding this topic, it was decided to develop a web application that would determine the risk of a patient developing delirium in a hospital setting using machine learning algorithms.

﻿The objective is to develop and validate a predictive risk model to detect delirium using patient data.

﻿Data from electronic health records for patients hospitalized from the ED between 2014, and 2016, were extracted.

Ferramentas de diagnóstico

CAM RASS,

Machine Learning

O ML é uma área de investigação da ciência da computação que utiliza conceitos de IA e métodos estatísticos para desenvolver algoritmos que aprendem e fazem previsões sobre os dados. Este campo da IA explora o estudo e a construção de algoritmos, que permitem aprender com dados, identificar padrões em enormes quantidades de dados e tomar decisões. A maior utilidade e impacto do conhecimento extraído a partir de dados e eventos históricos é a previsão de eventos e alterações similares no futuro (Murphy, 2012). Apesar de não ser nova, esta técnica tem vindo a ganhar importância nos últimos anos e é agora utilizada numa grande variedade de aplicações. Com o rápido desenvolvimento da IA, o ML e o reconhecimento inteligente têm sido cada vez mais aplicados às necessidades da vida humana (Xia, Wang, Yan, Dong, & Wang, 2019). Um estudo realizado em 2020 por Vellido, afirma que a conjetura atual do desenvolvimento tecnológico desencadeou a ideia que a utilização de ML seria o caminho a seguir para resolver problemas relacionados com a saúde, além de ser uma mais-valia para a melhoria da qualidade dos serviços de saúde (Vellido, 2020). Do mesmo modo, Kareemi et al. (2021) destacam o potencial do ML implementado nos cuidados de saúde, ao promover uma melhoria na qualidade da medicina e ao permitir acelerar o ritmo de evolução de técnicas complexas de diagnóstico e terapêuticas.

Para além disto, foram realizadas investigações que analisaram a utilização de modelos de ML em diversas áreas da saúde, e concluíram que estes modelos de ML parecem ter melhor desempenho de diagnóstico e prognóstico em comparação com as ferramentas tradicionalmente utilizadas em contexto hospitalar (Jauk et al., 2020; Kareemi et al., 2021; Stewart, Sprivulis, & Dwivedi, 2018; Vellido, 2020).

Em 2018, foi realizado um estudo que avaliou a predição de delirium usando o algoritmo RF. Para tal, previamente foi executada uma recolha de dados, que implicou a realização do rastreio de delirium através do CAM e também a recolha dos dados de saúde eletrónicos de 64038 pacientes. Estes dados foram divididos aleatoriamente em 80% para treino e 20% para teste e aplicados ao algoritmo RF. Este modelo de previsão produziu uma área abaixo da curva ROC de 0,909, o que demonstrou que este algoritmo possui um grau elevado de precisão e potencial para fornecer um modelo preditivo útil na prática clínica (Corradi et al., 2018). Já em 2021, foi publicado um estudo de coorte retrospetivo que desenvolveu e validou algoritmos de ML para a deteção do delirium. Para a execução deste estudo foram recolhidos dados durante 5 anos e para a realização do rastreio foram utilizados o método Delirium Observation Screening Scale (DOSS) para os doentes internados e Confusion Assessment Method for the Intensive Care Unit (CAM-ICU) para doentes ventilados. Também foram recolhidos dados acerca do histórico médico, medicamentos administrados, medições fisiológicas e resultados laboratoriais. Os algoritmos estudados incluíram RL, Árvore de Decisão (AD), RF, Gradient Boosting Machine (GBM), Gaussian Naïve Bayes (GNB), Support Vector Machine (SVM), e K Nearest Neighbor (KNN). Com este estudo, foi possível concluir que os algoritmos de RF, GBM e RL apresentaram a melhor capacidade de previsão com o valor da área abaixo da curva ROC de 0,85 a 0,86. Tendo demonstrado, que o uso de algoritmos de ML para a identificação de delirium pode ser uma boa abordagem na prática clínica, na medida em que podem permitir identificar casos que passariam despercebidos (Lee, Mueller, Nick Street, & M. Carnahan, 2021). Neste sentido, é importante frisar que a identificação precoce de doentes com risco de desenvolver delirium pode facilitar a prevenção desta perturbação e assim melhorar a qualidade de vida dos pacientes. Neste seguimento, foram desenvolvidos modelos preditivos para a deteção do delirium, que demonstraram ser uma vantagem na prática clinica diária (Van Den Boogaard et al., 2012). O PREdiction of DELIRium for Intensive Care patients (PRE-DELIRIC), foi um modelo de previsão do delirium criado em 2012 para uso na medicina de cuidados intensivos. Este modelo prevê o desenvolvimento de delirium ao longo do internamento, mediante 10 preditores (idade, grupo diagnóstico, coma, admissão urgente, administração de morfina, ureia, infeção, sedação, acidose metabólica, pontuação Acute Physiology and Chronic Health Evaluation-II (APACHE-II) avaliáveis 24 horas após a admissão do doente (Liang et al., 2020; Van Den Boogaard et al., 2012). Segundo Liang et al. (2020), o PRE-DELIRIC tem um elevado valor preditivo e é sugerido que este modelo seja adotado em unidades de cuidados intensivos (UCI) para a deteção do delirium em doentes de alto risco, pois contribui para uma melhor gestão de recursos assim como uma melhoria na vida dos pacientes.

Em 2015, foi validado outro modelo para deteção precoce do delirium para cuidados intensivos, denominado por Early PREdiction of DELIRium for Intensive Care patient (E-PRE-DELIRIC). Este modelo é constituído por nove preditores: idade, histórico de alterações cognitivas, histórico de abuso de álcool, níveis de ureia no sangue, grupo diagnóstico, admissão urgente, tensão arterial média, administração de corticosteroides e insuficiência respiratória. Este estudo surgiu como necessidade de colmatar a lacuna do modelo anterior ter a limitação de exigir preditores obtidos durante as primeiras 24 h de admissão na UCI. Pelo que, o modelo E-PRE-DELIRIC utiliza os dados disponíveis na admissão à UCI para prever o desenvolvimento do delirium durante o tempo de internamento do paciente (Wassenaar et al., 2015).

**Related Work**

The rapid development of artificial intelligence has increasingly allowed the application of ML to the needs of human life (Xia, Wang, Yan, Dong, & Wang, 2019). A study conducted in 2020 by Vellido, indicates that the current conjecture of technological development has triggered the idea that the use of ML would be the way forward to solve health-related problems, as well as be an asset to improve the quality of health services (Vellido, 2020). Similarly, Kareemi et al (2021) highlight the potential of ML implementation in healthcare by promoting an improvement in the quality of medicine and allowing to accelerate the pace of evolution of complex diagnostic techniques (Kareemi, Vaillancourt, Rosenberg, Fournier, & Yadav, 2021)

The PREdiction of DELIRium for Intensive Care patients (PRE-DELIRIC), was a delirium prediction model created in 2012 for use in adult intensive care patients. The model was developed using data from 1613 consecutive intensive care patients in one hospital and was temporally validated using 549 patients from the same hospital. This model predicts the development of delirium throughout the hospital stay by 10 risk factors (﻿age, Acute Physiology and Chronic Health Evaluation-II (APACHE-II) score, admission group, coma, infection, metabolic acidosis, use of sedatives and morphine, urea concentration, and urgent admission) assessable 24 hours after patient admission (Liang et al., 2020; Van Den Boogaard et al., 2012). ﻿The model produced an AUC-ROC of 0.87 (95% confidence interval 0.85 to 0.89) and 0.86 after bootstrapping. Temporal and external validation resulted in an AUC-ROC of 0.89 (0.86 to 0.92) and 0.84 (0.82 to 0.87), respectively (Van Den Boogaard et al., 2012). According to Liang et al. (2020), the PRE-DELIRIC has a high predictive value and suggested that this model be adopted in intensive care units to detect delirium in high-risk patients, as it contributes to improve the management of resources as well as an improvement in patients' lives (Liang et al., 2020).

In 2015, an alternative model for early detection of delirium in intensive care was validated, called Early PREdiction of DELIRium for Intensive Care patient (E-PRE-DELIRIC). The E-PRE-DELIRIC model uses the data available at intensive care unit (ICU) admission to predict the development of delirium during the patient's hospital stay. Which is composed by nine predictors: age, history of cognitive impairment, history of alcohol abuse, blood urea nitrogen, admission category, urgent admission, mean arterial blood pressure, use of corticosteroids, and respiratory failure. This study emerged as a need to fill the gap in the previous model with the limitation of requiring predictors from the first 24 h of ICU admission. In total were included data from 2914 patients. The AUROC obtained was 0.76 (95 % confidence interval 0.73 to 0.77) in the development dataset and 0.75 (0.71 to 0.79) in the validation dataset. And obtained an AUCROC of 0.70 (0.67 to 0.74) in the case of delirium that developed until 2 days and 0.81 (0.78 to 0.84), for delirium that developed after 6 days. (A. Wassenaar et al., 2015)

In order to understand which of the two models would be better prepared for clinical use, the study "Delirium prediction in the intensive care unit: comparison of two delirium prediction models" was conducted in the year 2017. This study concluded that the PRE-DELIRIC model predicts delirium better, however, ICU physicians rated the user convenience of using E-PRE-DELIRIC as higher than to PRE-DELIRIC (Annelies Wassenaar et al., 2018).

In 2018, a study was conducted that evaluated the prediction of delirium using the RF algorithm. For this purpose, a data collection was previously performed, which involved screening for delirium using CAM, and collecting the electronic health data from 64038 patients. ﻿The data used included demographic data, comorbidities, medications, procedures, and physiological measures. These data were randomly divided into 80% for training and 20% for testing and applied to the RF algorithm. This predictive model produced an AUC-ROC of 0.909, which demonstrated that this algorithm is highly accurate and has the potential to provide ﻿a clinically useful predictive model (Corradi, Thompson, Mather, Waszynski, & Dicks, 2018)

As early as 2021, a retrospective cohort study was published that developed and validated ML algorithms for the detection of delirium. To conduct this study, data were collected over 5 years period and the Delirium Observation Screening Scale (DOSS) was applied to inpatients and the Confusion Assessment Method for the Intensive Care Unit (CAM-ICU) for ventilated patients. The information collected included medical history, medications administered, physiological measurements, and laboratory results. The algorithms investigated included RL, Decision Tree, RF, Gradient Boosting Machine, Gaussian Naïve Bayes, Support Vector Machine, and K Nearest Neighbour. This study concluded that ﻿random forest, gradient-boosted machine, and logistic regression models demonstrated the best predictive ability with respective AUCs of 0.85 to 0.86 (Lee, Mueller, Nick Street, & M. Carnahan, 2021)

Having shown that using ML algorithms to identify delirium may be a good approach in clinical practice, as they make it possible to recognize cases that would otherwise go unnoticed (Lee, Mueller, Nick Street, & M. Carnahan, 2021). In this sense, it is important to emphasize that early identification of patients at risk of developing delirium may facilitate the prevention of this disorder and therefore enhance the quality of life of patients. Following this, predictive models for detecting delirium have been developed and have proven to be an advantage in daily clinical practice (Van Den Boogaard et al., 2012).

**Material and Methods**

**Data**

This ﻿study used data from a Portuguese hospital, extracted between 2014 and 2016. ﻿The study population comprised patients admitted at emergency department, ﻿patients were aged between 18 and 100 years old. ﻿The outcome measure of this study was a positive delirium diagnosis determined by RASS. ﻿The data available included patient demographics, medical histories, admission category, urgent admission, physiological measurements, medications administered, and lab results. At total 26 features was used.

﻿**Outcome**

The outcome measure of this study was a positive delirium diagnosis within one day of hospitalization determined by a positive DOSS or CAM assessment. Twenty-four hours was chosen as the cut-off length for study inclusion to optimize the model’s ability to identify delirium cases most related to factors observed around the time of the ED visit. A positive outcome was defined as one or more positive screenings within 24 hours even if a prior assessment was negative.

**Predictors**

We collected data on patient demographics, medical histories, physiological measurements, medications administered, and lab results. The following variables were used to generate and test a model: age at the time of hospitalization, sex, gender, history of stroke, dementia, severe illness defined by meeting two or more Systemic Inflammatory Response Syndrome (SIRS) criteria, transient ischemic attack (TIA), diagnosis of intracranial hemorrhage in the ED, tachypnea, and visual or hearing impairment. The physiological variables collected at the time of ED evaluation were heart rate, respiratory rate, Body Mass Index (BMI), and temperature. Medications ordered were obtained with a drug flag for opioids and benzodiazepines. We defined the anticholinergic variable as receipt of drugs classified as level 2 or 3 on an updated version of the Anticholinergic Drug Scale (Supplemental table 1 displays anticholinergics received by the sample).13,14 Although there is not an established Activity of Daily Living (ADL) assessment in the ED at this institution, nursing staff in inpatient units recorded a Barthel index, a measurement of the degree of assistance required by a patient determined by 10 variables describing ADL and mobility.15 The Barthel index was included as a continuous variable where a higher Barthel index is ﻿indicative of a higher level of independence. ADL was an important predictor for delirium in the literature but the Barthel index is not available at the time of the ED visit, so we examined the models without the Barthel index as the primary analysis and without as the secondary analysis.1

﻿**Analysis**

We reported summary statistics for the population by positive delirium screening. Means and Standard Deviation (SD) summarized continuous variables; frequency counts and proportions summarized categorical variables.

Missing values were imputed using KNN-imputation which has been shown to outperform other widely used imputation methods.17 Possible outliers were identified with the IQR Extreme Value analysis. To prevent the loss of information about variability in the study, clinical reasoning was used to determine if an outlier reflects the study population. Variance inflation factors were observed to detect multicollinearity, and continuous variables were checked for linearity by examining plots of the continuous independent variables versus the logit of the outcome.

We compared the predictive performance of five machine learning models using the Python Machine Learning library Scikit-Learn.18 The algorithms included Logistic Regression (LR), Decision Tree (DT), Random Forest (RF), Gradient Boosting Machine (GBM), and Gaussian Naïve Bayes (GNB), Support Vector Machine (SVM), and K Nearest Neighbor (KNN) with an intention to identify an interpretable model. Cross-validation was implemented for both hyperparameter tuning and model evaluation with AUC as the evaluation metric. This re- sampling method was selected over repeated sub-sampling to prevent any loss of information about the positive class by ensuring every observation appears in both the training and test data.

To avoid an optimistic bias that can result from using the same cross-validation procedure for both hyperparameter tuning and model evaluation, nested cross-validation was employed. In nested cross-validation, k-fold cross-validation for hyperparameter tuning is nested inside the k- fold cross-validation for model evaluation. Using tenfold nested cross-validation, the data was randomly divided into 10 equally-sized subsets. Out of the 10 sets, 9 were used to train the classifier, and the 10th was used for testing. The training set was further partitioned into 5 folds for an inner cross-validation grid search to optimize hyperparameters. This process was repeated until each of the 10 subsets had served as the test set. Similar to a regular cross-validation procedure, evaluation metrics are obtained by averaging the test set scores of the 10 runs. By conducting model selection independently in each trial of the model fitting procedure, the risk of overfitting during hyperparameter tuning is reduced. The final models were selected using a 10- fold cross-validation grid search on all available data.

**Results**

**Discussion:**

﻿This proof-of-concept study demonstrates the value of machine learning applied to clinical data sets. Generally applicable and accurate prediction of patients at high risk of developing delirium would be of great value in identifying underlying medical conditions and modifiable risk factors.

**Conclusion:**

Introduction:

Increased life expectancy is an indicator of improved quality of life, but it is also associated with an increase in chronic diseases. Due to the multiple physiological changes that characterise the elderly, it is expected that this age group is particularly vulnerable to the adverse effects of hospitalisation. Small changes in the individual caused by the change of disease state may promote states of confusion and disorientation. Delirium is a common manifestation of severe acute neuropsychiatric dysfunction, prevalent in hospital environment, which due to its variability of clinical presentation is often underdiagnosed and neglected.

Delirium can affect people of all ages, but it predominantly affects hospitalised older adults and it is associated with increased morbidity and mortality.

Disease, medications, a newly presented physical deficit, or a combination of these can

Devido às múltiplas alterações fisiológicas que caracterizam os idosos, espera-se que este grupo etário seja particularmente vulnerável aos efeitos adversos da hospitalização. Pequenas mudanças no indivíduo causadas pela mudança de estado da doença podem promover estados de confusão e desorientação. O delírio é uma manifestação comum de grave disfunção neuropsiquiátrica aguda, prevalente em ambiente hospitalar, que devido à sua variabilidade de apresentação clínica é frequentemente subdiagnosticada e negligenciada.

O delirium pode afectar pessoas de todas as idades, mas afecta predominantemente adultos idosos hospitalizados e está associado a uma morbilidade e mortalidade acrescidas.

O delirium é uma síndrome neuropsiquiátrica grave caracterizada por um distúrbio de atenção ou de consciência~~~cite{ref\_apa}.

A implementação deste tipo de mecanismos irá ajudar a desenvolver um sistema de saúde inteligente.

# Materiais e Métodos:

**Dados**

Algoritmos de ML

Avaliação

**Data**

**﻿This ﻿study used data from a Portuguese hospital, extracted between 2014 and 2016. ﻿The study population comprised patients admitted at emergency department, ﻿patients were aged between 18 and 100 years old. ﻿The outcome measure of this study was a positive delirium diagnosis determined by RASS. ﻿The data available was patient demographics, medical histories, physiological measurements, medications administered, and lab results. At total 26 features was used.**

**Data preparation**

Data preparation consists of raw data analysis techniques to produce quality data mainly including data collecting, data cleaning, data transformation and data reduction (Zhang, Zhang, & Yang, 2003).

The first step in this dataset was to collect the missing data through the hospital's computer system. During collection, it was noted that information on 77 patients was no longer available. As a result, these data set rows were excluded. The next step was to analyse the variables, which included the elimination of columns with repeated information and the exclusion of columns with a single value (zero variation). Thus, after collecting missing data, excluding irrelevant variables and deleting rows where information was unavailable, the resulting database comprises information from 434 individuals and 54 variables.

Dimensionality reduction can be done by exploiting the redundancy of the input data and finding a smaller set of new variables, each of which is a combination of the input variables, containing basically the same information as the input variables (Sorzano, Vargas, & Montano, 2014).

The next step was to apply dimensionality reduction, wherein the redundancy of input data was explored. A small set of new variables was created, each being a combination of input variables with essentially the same information. This dimensional reduction consisted of clustering drugs variables and allowed a better organization of the data, as well as a reduction of the dimensionality without loss of relevant information. In this sense, research was conducted in order to group the drugs by their respective pharmacological group. As a result of this approach, there was a decrease from 54 to 28 independent variables.

The final step involved data transformation. Since ML algorithms at their core operate on numerical data, it was necessary to execute a data transformation to satisfy this requirement of the algorithms. This means that there was a need for categorical variable encoding techniques in this database. In addition, a normalization of the data was also performed because the database consists of 13 numerical variables which comprise different measurement scales and are therefore difficult to compare. After all these data transformations, they are ready to be submitted to the ML models.

Feature Selection

﻿Most of the time, the dataset is made up of a large number of variables, some of which may not be relevant or redundant to the classification model. For this reason, prior to constructing the model, it is important to select the best set of independent variables to be included in the predictive model. For this purpose, statistical and exploratory techniques were used to select and delete variables that contribute less to the model, namely the feature selection. The main objective of variable selection is to find a subset of variables that best correlates with the response variable, without eliminating relevant information, allowing a reduction in computational costs and an increase in the predictive power of the classifier. There are three types of algorithms for feature selection: filter methods, wrapper methods, embedded methods. The feature selection techniques used in this project were the wrapper and embedded methods (Kumar, 2014).

﻿In the wrapper approach the learning algorithm is used to determine the optimal feature subset. Different combinations of subsets are defined, and their performance is evaluated using a classification algorithm. The classification model is first constructed using a subset of samples for training, and then the model is evaluated by the rest of the samples to test the model. The resulting classifier is evaluated according to evaluation metrics such as predictive accuracy, precision, recall, f-1 score, AUC, among others. The subset with the highest classification outcomes is the subset used to construct the final model. This method is quite suitable for the selection of relevant variables, but implies a high computation cost and can be prone to overfitting. (Cherrington, Thabtah, Lu, & Xu, 2019; Suppers, van Gool, & Wessels, 2018)

The main examples of this method are the forward selection, backward elimination, both and recursive feature elimination (RFE) strategy.

The forward strategy begins with no variables in the model, then adds variables to the model one by one. At each step, each variable excluded from the model is tested in order to be included in the model. In each subsequent iteration, the most significant variable is added first.

When the newly added variable does not improve model performance, the method ends.

In the backward model, on the other hand, the procedure is initiated with all the predictors to be included in the model and in each iteration the least significant variable that allows for an improvement in model performance is eliminated. This process is repeated until no improvement in model performance is observed. The two-way model combines the two techniques mentioned above, and can be combined so that at each step, the procedure selects the best attribute and removes the worst one from among the remaining attributes. It is considered less greedy than the two previous procedures, since it reconsiders the addition of predictors in the model that was removed and vice-versa. (Chowdhury & Turin, 2020)

Finally, the RFE, select features recursively considering the sets of variables getting smaller and smaller. First, the estimator is trained on the initial set of variables and the importance of each variable is obtained through a specific attribute. Then, the less important variables are pruned from the current set of variables. This procedure is repeated recursively on the pruned set until the desired number of features to be selected is reached. (Pedregosa et al., 2012) The Sequential Feature Selector (SFS), Recursive Feature Elimination (RFE), Recursive Feature Elimination with cross validation (RFECV) and Select From Model (SFM) methods were used to search for the best set of variables, and the RF and RL algorithms were used as classifiers.

Select From Model is a meta-transformer used with estimator that has coef\_ or feature\_importances\_ attribute after fitting. In this problem, was used a random forest classifier and a logistic regression classifier followed by Select From Model for choosing features using features’ importance. Those features are selected which has coef\_ or feature\_importances values greater than given threshold value. The threshold value may be mean of the importance, median of importance, a float value or none.

Machine learning

Random Forest

Random Forest, as the name suggests, is a tree-based ensemble with each tree depending on a collection of random variables. Random Forests can be used for either a categorical response variable or a continuous response. Similarly, the predictor variables can be either categorical or continuous. These trees partition the predictor space using a sequence of binary partitions (“splits”) on individual variables. The “root” node of the tree comprises the entire predictor space. The nodes that are not split are called “terminal nodes” and form the final partition of the predictor space. Each nonterminal node splits into two descendant nodes, one on the left and one on the right, according to the value of one of the predictor variables. For a continuous predictor variable, a split is determined by a split point; points for which the predictor is smaller than the split point go to the left, the rest go to the right.

The RF classifier has incorporated the importance value of each variable into the feature\_importances\_ attribute, also known as the Gini index. Starting with the SFM function, it selects the variables whose value is higher than the selected threshold. Each decision tree calculates the importance of each variable based on its capacity to increase leaf purity. The greater the increase in leaf purity, more important the variable is. This procedure is performed in each tree, then the average of all trees generated is calculated and finally the value is normalized, so the sum of the variables scores is 1.

An alternative method used was the RFE with cross-validation. Cross-validation is a technique for evaluating ML models by training several ML models on available subsets of the input data and evaluating them in the complementary data subset. The primary idea of RFECV is to select the best set of variables using cross-validation. First, the estimator is trained on the initial set and the importance of each variable is obtained with the feature\_importances\_ attribute. Then, the less significant variables are removed, this procedure is repeated recursively on the pruned set until there are no more variables to exclude or the desired number of variables to select is reached.

Finally, sequential search technique was used, whose main objective is to improve the subset of selected variables through iterations that check if adding or removing variables to the subset tends to improve its performance. The SFS function was used, this function contains two configurable parameters to change the configuration of the forward parameter between False or True and the floating parameter also between False or True. Different combinations of the values can be used to switch between backward, forward and bidirectional methods.

Logistic Regression

LR is a statistical technique designed to generate observations based on a set of observations, a model that allows the prediction of values taken by a categorical variable, based on one or several continuous and/or binary independent variables (Hosmer, Lemeshow, & Sturdivant, 2013) This technique uses similar general principles used in linear regression, the difference is in the response variable which in LR is binary (dichotomic) whereas in the linear regression model it is continuous (Hosmer et al., 2013).

According to Stoltzfus (2011), regression techniques are versatile when applied to medical research because of their ability to predict outcomes and control variables. It also argues that RL is an effective and powerful way to analyse the impact of a group of independent variables on a binary outcome, quantifying the contribution of each variable (Stoltzfus, 2011). (Stoltzfus, 2011)

The same techniques mentioned in the RF model were used for feature selection using the logistic regression classification algorithm. LR classifier has incorporated the coefficient value of each variable into the coef\_ attribute. In general, the methods used to find the coefficients of the logistic function follow an iterative process of selection of a candidate variable and calculation of the logarithm of the probability. This process is repeated until convergence is achieved and the maximum likelihood is found.

Model evaluation

Classification models should be evaluated before being adopted in a real context, because if the classifier is poorly calibrated, it may mislead healthcare professionals and consequently cause harm to the target population. Therefore, in order to minimize this type of occurrence, it is imperative to evaluate the quality of the predictions obtained. For this evaluation, the model was submitted to evaluation metrics of the predictive model.

Recall may be defined as the probability that the model correctly classifies a person with the delirium syndrome, given that the individual carries the syndrome. This measure is also known as true positive rate, in other words, this measure assesses the ability of the test to detect delirium when it is indeed present, and its value can be estimated using the following formula (J. Han et al., 2012):

ROC CURVE

One way to evaluate the ability of a diagnostic test to discriminate between two populations is by ROC analysis (Fawcett, 2006). This analysis is based on signal detection theory and was developed during World War II, where it was used to analyse radar images (Lloyd & Appel, 1976). The science of signal detection theory was later extended to other scientific domains, including diagnostic medicine (Lusted, 1971). Being a commonly used tool in medical diagnostics due to its discriminative capacity.

In general, the ROC curve can show the performance of a ML model for binary classification. It is a two-dimensional graphical representation, plotting sensitivity (true positive rate) on the y-axis against 1-specificity (false positive rate) on the x-axis. Whether or not the test can be affirmed as having the ability to discriminate between individuals with and without delirium is directly linked to a measure of ROC curve accuracy, called the area under the ROC curve (AUC-ROC). Using this measure, it was possible to transform the ROC performance into a scalar value, which allowed to evaluate the discriminant ability of the ROC curve. A closer curve in the upper left corner presents a higher discriminating capacity and the AUC can reach the maximum value of 1, meaning perfect discrimination. On the other hand, if the curve approaches the diagonal, the discriminant capability of the model is null.

In summary, the ROC analysis provides important information about diagnostic test performance, and the closer the curve approaches the upper left corner, the greater discriminatory capability of the test.

PRECISION-RECALL

This measure is useful to evaluate the quality of the classifier results for cases involving unbalanced data sets, which is an alternative to the ROC curve. The Precision-Recall (P-R) curve indicates the balance between precision and recall for different thresholds. There is a great difference between the visual representation of between the ROC curve and the P-R curve, because the objective of the ROC space is in the upper left corner while the objective of the P-R space is in the upper right corner. Thus, a high area under the P-R curve represents both high recall and high precision, where high precision is associated with a low false positive rate and high recall associated with low false negative rate.

Embedded Approach is done with a specific learning algorithm that performs feature selection in the process of training.

The analysis of the characteristics present in the dataset allows the discovery patterns and trends that can provide valuable information. The purpose of this analysis is to extract, contextualize and organize the information, so as to generate new hypotheses or models that allow the problem to be solved (Kuhn & Johnson, 2020)

**ML algorithms**

**﻿We compared the predictive performance of two machine learning models using the Python Machine Learning library Scikit-Learn. The algorithms included Logistic Regression (LR) and Random Forest (RF) ﻿with an intention to identify an interpretable model. ﻿Cross-validation was implemented for both hyperparameter tuning and model evaluation with AUC as the evaluation metric.**

**﻿This resampling method selected used was to prevent any loss of information about the positive class by ensuring every observation appears in both the training and test data. ﻿To avoid an optimistic bias that can result from using the same cross-validation procedure for both hyperparameter tuning and model evaluation, nested cross-validation was employed. ﻿In nested cross-validation, k-fold cross-validation for hyperparameter tuning is nested inside the k- fold cross-validation for model evaluation. Using tenfold nested cross-validation, the data was randomly divided into 10 equally-sized subsets.**

**Predictive modeling, machine learning projects, such as classic cation and regression, always Involve some form of data preparation. The space c data preparation required for a dataset Depends on the species case of the data, such as the variable types, as well as the algorithms that Will be used to model them that may impose expectations or requirements on the data.**

# Results:

﻿The Area Under Curve (AUC), sensitivity, specificity for the RF ﻿were 0.76-0.77, 0.90, 0.33-0.35, respectively. ﻿Our study reported the importance of variables, such as age a.

﻿These findings highlighted several modifiable variables that we can approach in the ED and hospital setting. ﻿The comparison of the machine learning-driven models added a significant impact on how to identify patients who have a high risk of developing delirium while minimizing bias.

# Conclusão:

Reconhecer, diagnosticar, gerir e prevenir o delírio nestas circunstâncias inovadoras e adversas tem vindo com desafios únicos.

Conclusion:

﻿This study demonstrated the use of machine learning algorithms to identify the combination of variables that are predictive of delirium.

Recognising, diagnosing, managing and preventing delirium in these novel and adverse circumstances has come with unique challenges.

﻿Since delirium is missed in the ED and our prediction model will help to identify a high-risk group, which fits with using the model to identify prevalent and impending delirium.

﻿The discovery of a predictive model that clinicians can use as a clinical decision aid could lead to improved detection of delirium and identification of a high-risk group. This contribution is significant because the findings will introduce a clinical decision aid that either clinicians use actively or receive passively from machine learning algorithms, overcoming the limitation of misdiagnosis or under diagnosis by clinical gestalt alone to detect delirium.

﻿The model can predict delirium for the complete stay in the intensive care within 24 hours of admission. We can now identify patients who have a high risk of developing delirium during their stay at healthcare setting. This will facilitate targeted initiation of preventive measures. Our study shows that the use of this model is significantly better than the predictions of the attending caregivers, and it should therefore be used daily in intensive care practice. A version of the web application model can be accessed at <https://share.streamlit.io/natawild/deliriumapp/main/AppDelirium.py>.

This article is about the importance of

Increased life expectancy is an indicator of improved quality of life, but it is also associated with an increase in chronic diseases. Due to the multiple physiological changes that characterise the elderly, it is

expected that this age group is particularly vulnerable to the adverse effects of hospitalisation. Small changes in the individual caused by the change of disease state may promote states of confusion and disorientation. Delirium is a common manifestation of severe acute neuropsychiatric dysfunction, prevalent in hospital environment, which due to its variability of clinical presentation is often underdiagnosed and neglected.

Delirium can affect people of all ages, but it predominantly affects hospitalised older adults and it is associated with increased morbidity and mortality.

Delirium is a severe neuropsychiatric syndrome characterized by a disturbance of attention or awareness~\cite{ref\_apa}.

The implementation of this kind of mechanisms will help in developing a smart healthcare system.

Adamis, D., Sharma, N., Whelan, P. J. P., & MacDonald, A. J. D. (2010). Delirium scales: A review of current evidence. *Aging and Mental Health*, *14*(5), 543–555. https://doi.org/10.1080/13607860903421011

Cherrington, M., Thabtah, F., Lu, J., & Xu, Q. (2019). Feature Selection: Filter Methods Performance Challenges. In *2019 International Conference on Computer and Information Sciences (ICCIS)* (pp. 1–4). IEEE. https://doi.org/10.1109/ICCISci.2019.8716478

Chowdhury, M. Z. I., & Turin, T. C. (2020). Variable selection strategies and its importance in clinical prediction modelling. *Family Medicine and Community Health*, *8*(1), e000262. https://doi.org/10.1136/fmch-2019-000262

Corradi, J. P., Thompson, S., Mather, J. F., Waszynski, C. M., & Dicks, R. S. (2018). Prediction of Incident Delirium Using a Random Forest classifier. *Journal of Medical Systems*, *42*(12). https://doi.org/10.1007/s10916-018-1109-0

De, J., & Wand, A. P. F. (2015). Delirium screening: A systematic review of delirium screening tools in hospitalized patients. *Gerontologist*, *55*(6), 1079–1099. https://doi.org/10.1093/geront/gnv100

DuGoff, E. H., Canudas-Romo, V., Buttorff, C., Leff, B., & Anderson, G. F. (2014). Multiple Chronic Conditions and Life Expectancy. *Medical Care*, *52*(8), 688–694. https://doi.org/10.1097/MLR.0000000000000166

Fawcett, T. (2006). An introduction to ROC analysis. *Pattern Recognition Letters*, *27*(8), 861–874. https://doi.org/10.1016/j.patrec.2005.10.010

Hosmer, D. W., Lemeshow, S., & Sturdivant, R. X. (2013). *Applied Logistic Regression.* *Wiley Series in Probability and Statistics*.

Inouye, S. K., Westendorp, R. G. J., & Saczynski, J. S. (2014). Delirium in elderly people. *The Lancet*, *383*(9920), 911–922. https://doi.org/10.1016/S0140-6736(13)60688-1

Kareemi, H., Vaillancourt, C., Rosenberg, H., Fournier, K., & Yadav, K. (2021). Machine Learning Versus Usual Care for Diagnostic and Prognostic Prediction in the Emergency Department: A Systematic Review. *Academic Emergency Medicine*, *28*(2), 184–196. https://doi.org/10.1111/acem.14190

Kumar, V. (2014). Feature Selection: A literature Review. *The Smart Computing Review*, *4*(3). https://doi.org/10.6029/smartcr.2014.03.007

Lee, S., Mueller, B., Nick Street, W., & M. Carnahan, R. (2021). Machine learning algorithm to predict delirium from emergency department data. *BMJ*. https://doi.org/https://doi.org/10.1101/2021.02.19.21251956

Liang, S., Chau, J. P. C., Lo, S. H. S., Bai, L., Yao, L., & Choi, K. C. (2020). Validation of PREdiction of DELIRium in ICu patients (PRE-DELIRIC) among patients in intensive care units: A retrospective cohort study. *Nursing in Critical Care*, (August), 1–7. https://doi.org/10.1111/nicc.12550

Lloyd, M. A., & Appel, J. B. (1976). Signal Detection Theory and the Psychophysics of Pain: An Introduction and Review. *Psychosomatic Medicine*, *38*(2), 79–94. https://doi.org/10.1097/00006842-197603000-00002

Lusted, L. B. (1971). Decision-Making Studies in Patient Management. *New England Journal of Medicine*, *284*(8), 416–424. https://doi.org/10.1056/NEJM197102252840805

Mittal, V., Muralee, S., Williamson, D., McEnerney, N., Thomas, J., Cash, M., & Tampi, R. R. (2011). Delirium in the elderly: A comprehensive review. *American Journal of Alzheimer’s Disease and Other Dementias*, *26*(2), 97–109. https://doi.org/10.1177/1533317510397331

Pedregosa, F., Varoquaux, G., Gramfort, A., Michel, V., Thirion, B., Grisel, O., … Duchesnay, É. (2012). Scikit-learn: Machine Learning in Python. *Journal of Machine Learning Research*, *12*(May 2014), 2825–2830. Retrieved from http://arxiv.org/abs/1201.0490

Sorzano, C. O. S., Vargas, J., & Montano, A. P. (2014). A survey of dimensionality reduction techniques, 1–35. Retrieved from http://arxiv.org/abs/1403.2877

Stoltzfus, J. C. (2011). Logistic regression: A brief primer. *Academic Emergency Medicine*, *18*(10), 1099–1104. https://doi.org/10.1111/j.1553-2712.2011.01185.x

Suppers, A., van Gool, A., & Wessels, H. (2018). Integrated Chemometrics and Statistics to Drive Successful Proteomics Biomarker Discovery. *Proteomes*, *6*(2), 20. https://doi.org/10.3390/proteomes6020020

Van Den Boogaard, M., Pickkers, P., Slooter, A. J. C., Kuiper, M. A., Spronk, P. E., Van Der Voort, P. H. J., … Schoonhoven, L. (2012). Development and validation of PRE-DELIRIC (PREdiction of DELIRium in ICu patients) delirium prediction model for intensive care patients: Observational multicentre study. *BMJ (Online)*, *344*(7845), 17. https://doi.org/10.1136/bmj.e420

Vellido, A. (2020). The importance of interpretability and visualization in machine learning for applications in medicine and health care. *Neural Computing and Applications*, *32*(24), 18069–18083. https://doi.org/10.1007/s00521-019-04051-w

Wassenaar, A., van den Boogaard, M., van Achterberg, T., Slooter, A. J. C., Kuiper, M. A., Hoogendoorn, M. E., … Pickkers, P. (2015). Multinational development and validation of an early prediction model for delirium in ICU patients. *Intensive Care Medicine*, *41*(6), 1048–1056. https://doi.org/10.1007/s00134-015-3777-2

Wassenaar, Annelies, Schoonhoven, L., Devlin, J. W., van Haren, F. M. P., Slooter, A. J. C., Jorens, P. G., … van den Boogaard, M. (2018). Delirium prediction in the intensive care unit: comparison of two delirium prediction models. *Critical Care*, *22*(1), 114. https://doi.org/10.1186/s13054-018-2037-6

Wong, C. L., Holroyd-Leduc, J., Simel, D. L., & Straus, S. E. (2010). Does this patient have delirium?: value of bedside instruments. *Jama*, *304*(7), 779–786.

Xia, H., Wang, C., Yan, L., Dong, X., & Wang, Y. (2019). Machine Learning Based Medicine Distribution System. In *Proceedings of the 2019 10th IEEE International Conference on Intelligent Data Acquisition and Advanced Computing Systems: Technology and Applications, IDAACS 2019* (Vol. 2, pp. 912–915). IEEE. https://doi.org/10.1109/IDAACS.2019.8924236

Zhang, S., Zhang, C., & Yang, Q. (2003). Data preparation for data mining. *Applied Artificial Intelligence*, *17*(5–6), 375–381. https://doi.org/10.1080/713827180

If you do not modify the content of a field, the default value will be applied. Please see the values entered in this form in the table below.

Delirium is a common but serious condition that is under recognized and associated with poor outcomes. However, delirium can be prevented and treated if it is diagnosed in time. It is therefore essential that all hospital staff be aware of the possibility of delirium developing, and that prompt assessment and appropriate management are ensured. This web application has been designed to support healthcare staff and alert them to the development of delirium in patients admitted to a hospital environment.