Sistema preditivo automatizado de saída em UTI: uma inovação

Hudson F. Golino, MSc., Ph.D (Cand.)

Sistema preditivo automatizado de saída em UTI: uma inovação tecnológica

Conteúdo

Conteúdo 2
Resumo 3
Dados 3
Métodos Estatísticos 5
Objetivos da Pesquisa ..... **Erro! Indicador não definido.**
Resultados 6
1) Análises Descritivas 6
2) Análises Preditivas 12
Modelos inovadores para realizar predição de saída hospitalar em UTI 13
Resultados da Análise Preditiva 18
Discussão 28
Referências 30
Principais Códigos Utilizados ..... **Erro! Indicador não definido.**

Lista de Figuras

Figura 1 Frequência de Dias de Ventilação Mecânica Por Saída Hospitalar entre Julho de 2012 e Dezembro de 2013. 6
Figura 2 Distribuição dos Destinos Hospitalares mês a mês, entre Julho e Dezembro. 7
Figura 3 Frequência de Saídas Hospitalares por Destino entre Julho e Dezembro nos anos de 2012 e 2013. . 7
Figura 4 Diagnóstico Primário por Ano, entre Julho e Dezembro. 7
Figura 5 Diagnóstico primário discriminado mês a mês entre Julho e Dezembro, dos anos de 2012 e 2013. . 8
Figura 6 Saídas Hospitalares discriminada mês a mês, entre Julho e Dezembro, nos anos de 2012 e 2013. ... 9
Figura 7 Distribuição das categorias de idade discretizada mês a mês, entre Julho e Dezembro. 10
Figura 8 Distrição das taxas de mortalidade entre Julho e Dezembro, nos anos de 2012 e 2013. 11
Figura 9 Modelo de árvore gerado. 18
Figura 10 Erros por classe no modelo de *Bagging*. 20
Figura 11 Distribuição dos errros por classe no modelo de Random Forest 22
Figura 12 Distribuição dos erros de predição do modelo de *boosting* por iteração em cada conjunto de dados 25
Figura 13 Grafo indireto acíclico forçado da matrix de proximidade 29
Figura 14 Grafo indireto acíclico forçado da matrix de proximidade 30

Resumo

A literatura aponta que o desenvolvimento de sistemas para estimar o risco de óbito de pacientes por meio de métodos objetivos é crucial para o avanço das ciências clínicas (Knaus et. al., 1991). Nos últimos anos tem ocorrido um avanço significativo nas ciências de dados, especialmente com o desenvolvimento de métodos estatísticos, matemáticos e computacionais para realizar predições em tempo real (Hastie, Tibshirani & Friedman, 2009). Esse avanço na ciência de dados possibilita a construção de ferramentas automatizadas para predição de óbito em unidades intensivas de tratamento, que pode incrementar o poder preditivo das "acute measures" usualmente empregadas nos centros de saúde. O objetivo desse trabalho é mostrar os resultados iniciais de um sistema de predição que usa um dos modelos mais inovadores da estatística computacional para predizer o risco de óbito de pacientes internados em UTI, por meio de dados clínicos usualmente monitorados pela equipe médica. Esse modelo inovador recebe o nome de "Random Forest", e consiste em um algorítimo de predição que constrói uma assembléia de árvores de predição, realizado operações em subamostras aleatórias dos casos e das variáveis preditivas, levando à modelos com alta acurácia e estabilidade (Hastie, Tibshirani & Friedman, 2009). O presente estudo utilizou uma amostra 439 pacientes internados em uma UTI de um hospital particular. A equipe médica registrava informações clínicas dos pacientes assim que esses eram admitidos na UTI tais como idade, sexo, pontuação na escala Glasgow, uso de ventilação mecânica invasiva, dentre outras variáveis clínicas em um total de 59 variáveis. A amostra foi dividida em duas subamostras aleatórias (treino e teste) para validação cruzada. O modelo apresentou sensibilidade de 100% e especificidade de 87,50% na subamostra de treino, e sensibilidade de 81,81% e especificidade de 75% na subamostra de teste. Esse modelo pode ser utilizado para a predição automatizada do risco de óbito em UTI's

Dados

· Versão do Banco de Dados: Dados oriundos da UTI, com informações de 439 pacientes, internados entre Julho de 2012 e Dezembro de 2013.

· Critérios de inclusão/exclusão de variáveis/pacientes: Variáveis com variância zero foram eliminadas das análises. Não houve exclusão de pacientes da análise, independentemente do número total de dados faltantes no banco de dados.

· Informações sobre o banco de dados:

- Dimensionalidade inicial: 439 pacientes/ 95 variáveis
- Dimensionalidade após a exclusão das variáveis: 439 pacientes/ 59 variáveis
- Variáveis, níveis, tipo e missingness (após exclusão das variáveis):

Tabela 1.

Nome das variáveis, níveis, tipo, total de dados faltantes (#Missing), e se foi utilizada na predição.

Variáveis	Níveis	Tipo	# Missing	Utilizada na predição
Registro		double	0	Não
Peso_kg_		integer	0	Não
IMC		double	0	Não
Data_Entrada_Hosp	287	integer	7	Não
Hora_Entrada_Hosp	51	integer	0	Não
Data_Entrada_UTI	284	integer	0	Não
Hora_Entrada_UTI	88	integer	3	Não
Dia_Entrada_UTI		integer	0	Não
Mes_Entrada_UTI		integer	0	Não
Ano_Entrada_UTI		integer	0	Não
Dia_Semana_entrada_UTI	7	integer	0	Sim
Tempo_ingresso_UTI		integer	7	Não
Delay_Entrada_UTI_min		integer	46	Sim
Data_Ultima_Alta	44	integer	392	Não
Tipo_Reinternacao	2	integer	392	Não
Diag_Primario	3	integer	1	Sim
Grupo_Etario	19	integer	0	Sim
Glasgow		integer	0	Sim
Motivo_Internacao_Apache	78	integer	0	Não
Insuf..organica.ou.imunodeprimido	2	integer	0	Sim
T_Max		double	0	Sim
FC_Max		integer	0	Sim
FR_Max		integer	0	Sim
PAm_Max		integer	0	Sim
pH_Max		double	0	Sim
FiO2_Medida1		integer	0	Sim
PaO2_Medida1		integer	0	Sim
PaCO2_Medida1		integer	0	Sim
Na_Max		integer	0	Sim
K_Max		double	0	Sim
Ht_Max		double	0	Sim
Leucocitos_Max		integer	311	Sim
Ureia_Med1		integer	0	Sim
Glicose_Med1		integer	214	Sim
Cirrose	2	integer	0	Sim
Imunossupressao	2	integer	0	Sim
Linfoma	2	integer	0	Sim
Score_APACHE_III		integer	0	Sim
Risco_APACHE_III_Hosp_		double	0	Não
Risco_obito		double	0	Não

Horario_Saida	47	integer	0	Não
Destino	6	integer	0	Não
CID10_Principal_Saida	45	integer	0	Não
CID10_3_Saida_	28	integer	354	Não
CID10_5_Saida_	8	integer	428	Não
Cateter_Vesical		integer	0	Sim
Dreno_de_Torax		integer	0	Sim
Intubacao_Nasotraqueal		integer	0	Sim
Marcapasso_provisorio		integer	0	Sim
Monitorizacao_da_Pressao_Intracraniana		integer	0	Sim
Metodos_Dialiticos		integer	0	Sim
Nutricao_Parenteral		integer	0	Sim
Puncao_Venosa_Profunda__jugular_		integer	0	Sim
Puncao_Venosa_Profunda__subclavia_		integer	0	Sim
Ventilacao_Mecânica_Invasiva		integer	0	Sim
Perda_de_Puncao_Venosa_Profunda		integer	0	Sim
Pneumotorax_por_Puncao_Venosa		integer	0	Sim
Pneumonia_Associada_à_Ventilacao		integer	0	Sim
Saida_Hospitalar	2	integer	8	Sim

Métodos Estatísticos

Todas as análises empregadas no presente relatório foram elaboradas utilizando-se o software aberto de código livre R para computação estatística (R Core Team, 2013). As análises são divididas em duas partes: descritiva e preditiva. Na análise descritiva, o objetivo foi verificar a distribuição de algumas variáveis de acordo com o ano, semestre ou tipo de saída hospitalar, assim como verificar a taxa de mortalidade, prevalências e outros indicadores relevantes. Já na análise preditiva, modelos de *learning trees* foram utilizados para poder predizer a saída hospitalar categórica (alta/óbito), visando a maximização da acurácia, da sensibilidade e da especificidade. Os detalhes de cada uma das técnicas, assim como os procedimentos realizados serão descritos na seção específica de *análise preditiva.*

Resultados

1) Análises Descritivas

Dos 439 pacientes internados na UTI do Hospital entre Julho de 2012 e Dezembro de 2013, 91% tiveram alta da unidade. De todos os pacientes internados, 3,8% tiveram pneumonia associada à ventilação, apenas um teve pneumotórax por punção venosa (0,002%) e três tiveram perda de punção venosa profunda (0,006%). Em termos de ventilação mecânica invasiva, 80% não demandaram sua utilização, e dos que demandaram 82,02% a utilizaram por apenas um dia, e 17,97% por dois dias ou mais (ver figura 1 abaixo).

Figura 1 Frequência de Dias de Ventilação Mecânica Por Saída Hospitalar entre Julho de 2012 e Dezembro de 2013.

Apenas 1% dos internados demandou nutrição parenteral, 4% demandaram métodos dialíticos, 2,72% tiveram monitoramento da pressão intracraniana, 0,9% utilizaram marcapasso provisório, 0,2% intubação nasotraqueal, 6% utilizaram dreno de tórax e 35% fizeram uso de cateter vesical. O destino principal da saída do UTI foi a enfermaria (81%), seguida por óbito na UTI (9%), residência sem assistência (7%), remoção para outro hospital (1%), residência com assistência (1%) e transferência para outra UTI (1%). A distribuição dos destinos (e o respectivo tipo de saída) entre Julho e Dezembro dos anos de 2012 e 2013 estão representados na figura 2 abaixo, e a figura 3 apresenta a distribuição dos destinos mês a mês. As figuras subsequentes apresentam a distribuição ou frequência de algumas das variáveis registradas na UTI. No entanto, essas figuras não serão detalhadas em forma de texto, para evitar repetição de informação.

Figura 2 Frequência de Saídas Hospitalares por Destino entre Julho e Dezembro nos anos de 2012 e 2013.

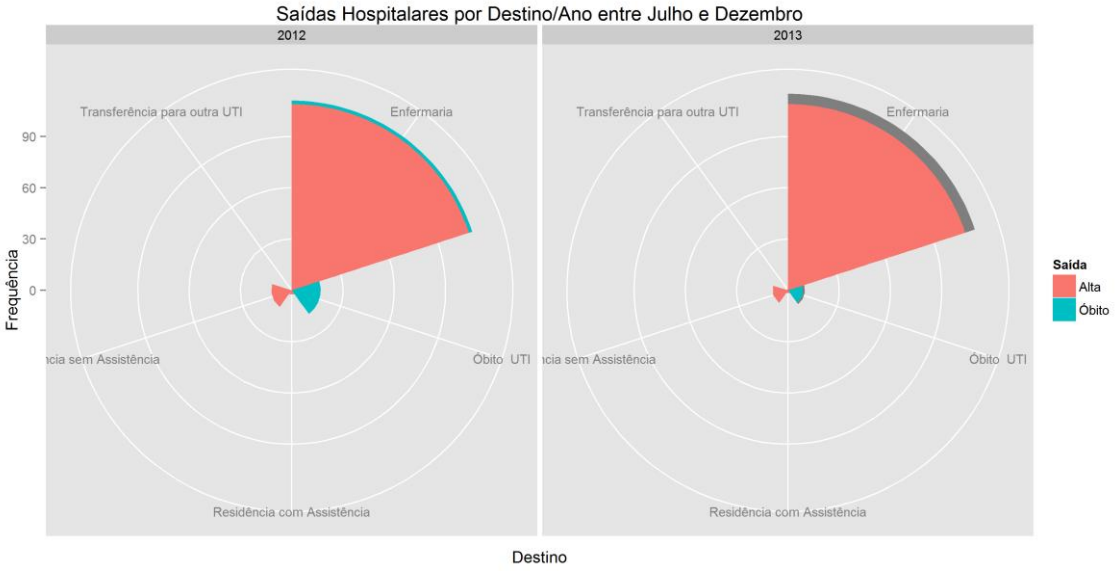


Figura 3 Distribuição dos Destinos Hospitalares mês a mês, entre Julho e Dezembro.

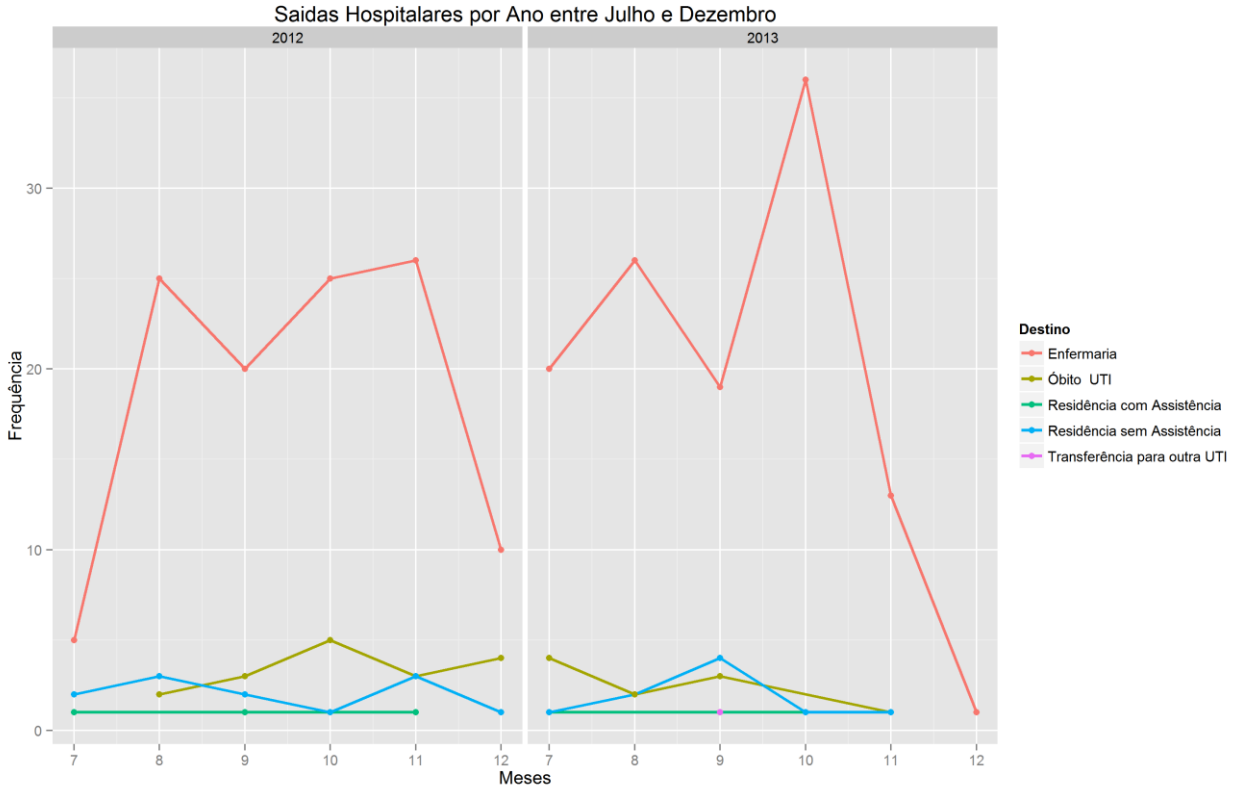


Figura 4 Diagnóstico Primário por Ano, entre Julho e Dezembro.

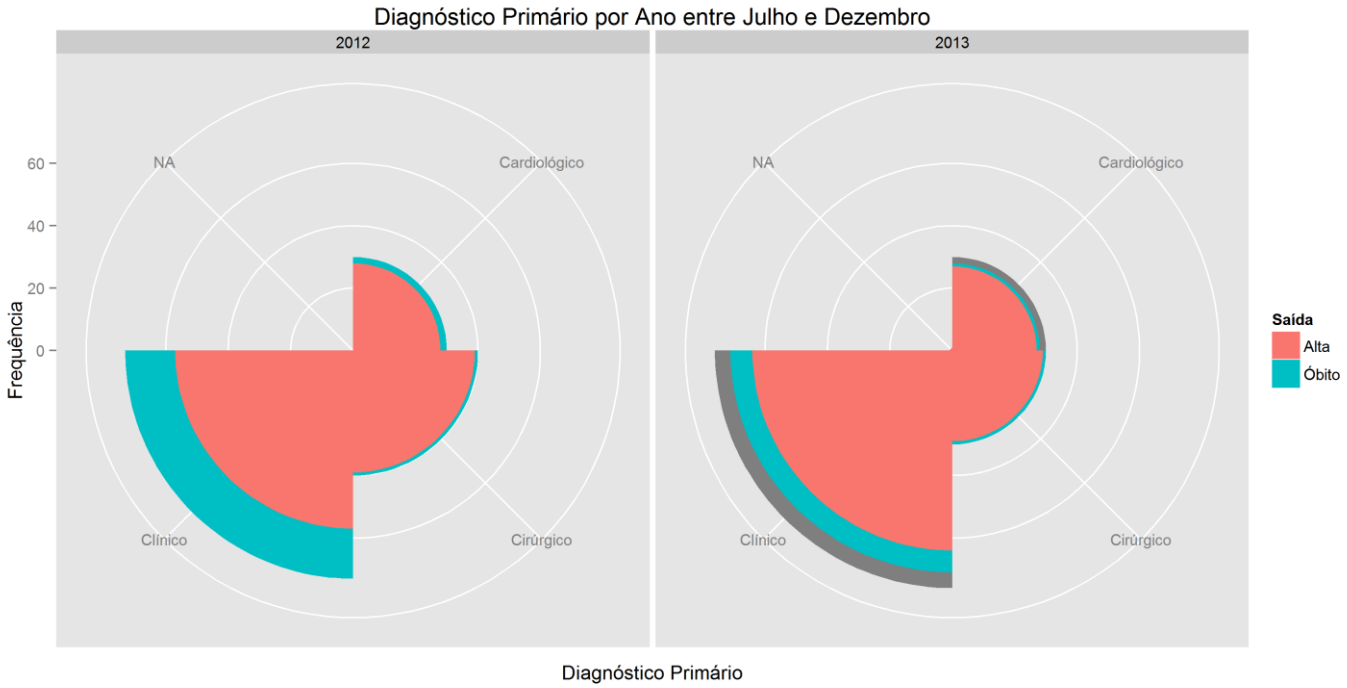


Figura 5 Diagnóstico primário discriminado mês a mês entre Julho e Dezembro, dos anos de 2012 e 2013.

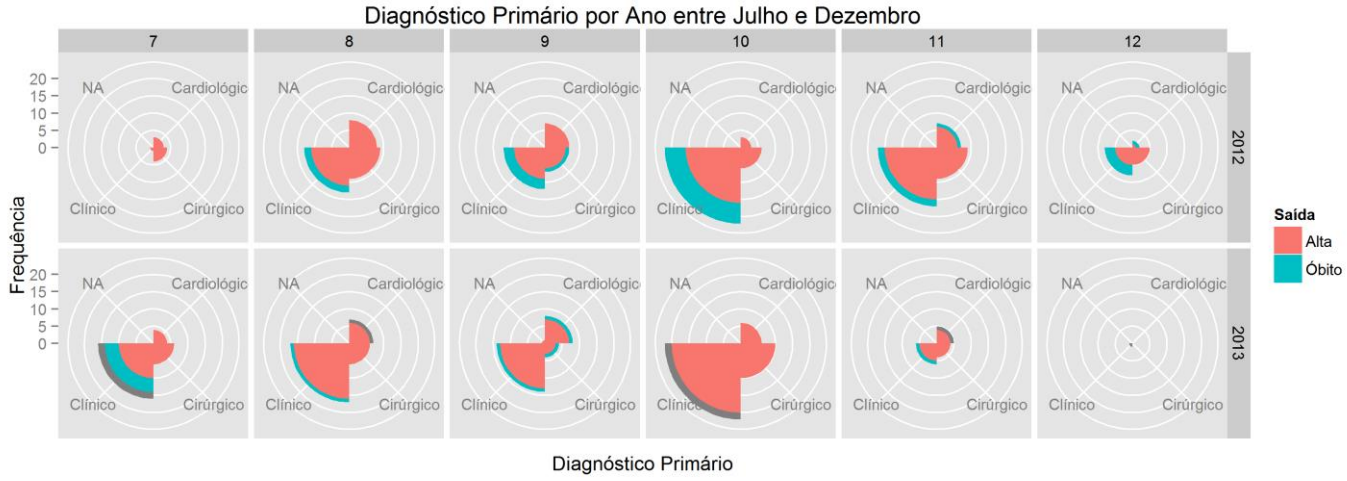


Figura 6 Saídas Hospitalares discriminada mês a mês, entre Julho e Dezembro, nos anos de 2012 e 2013.

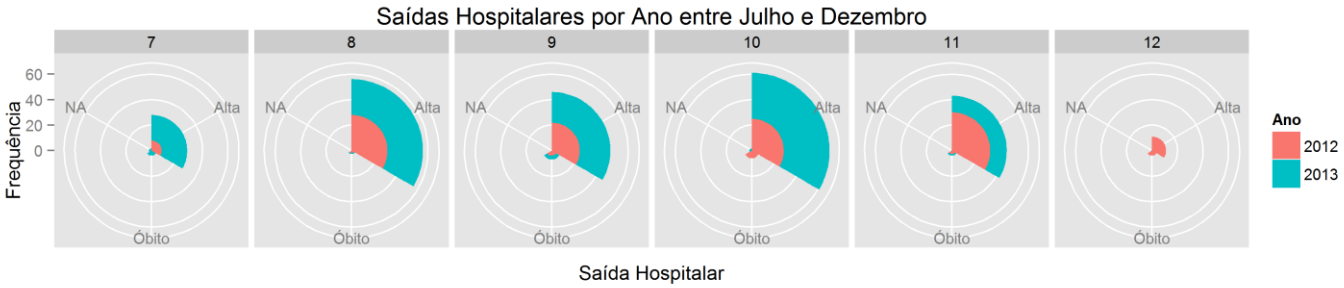
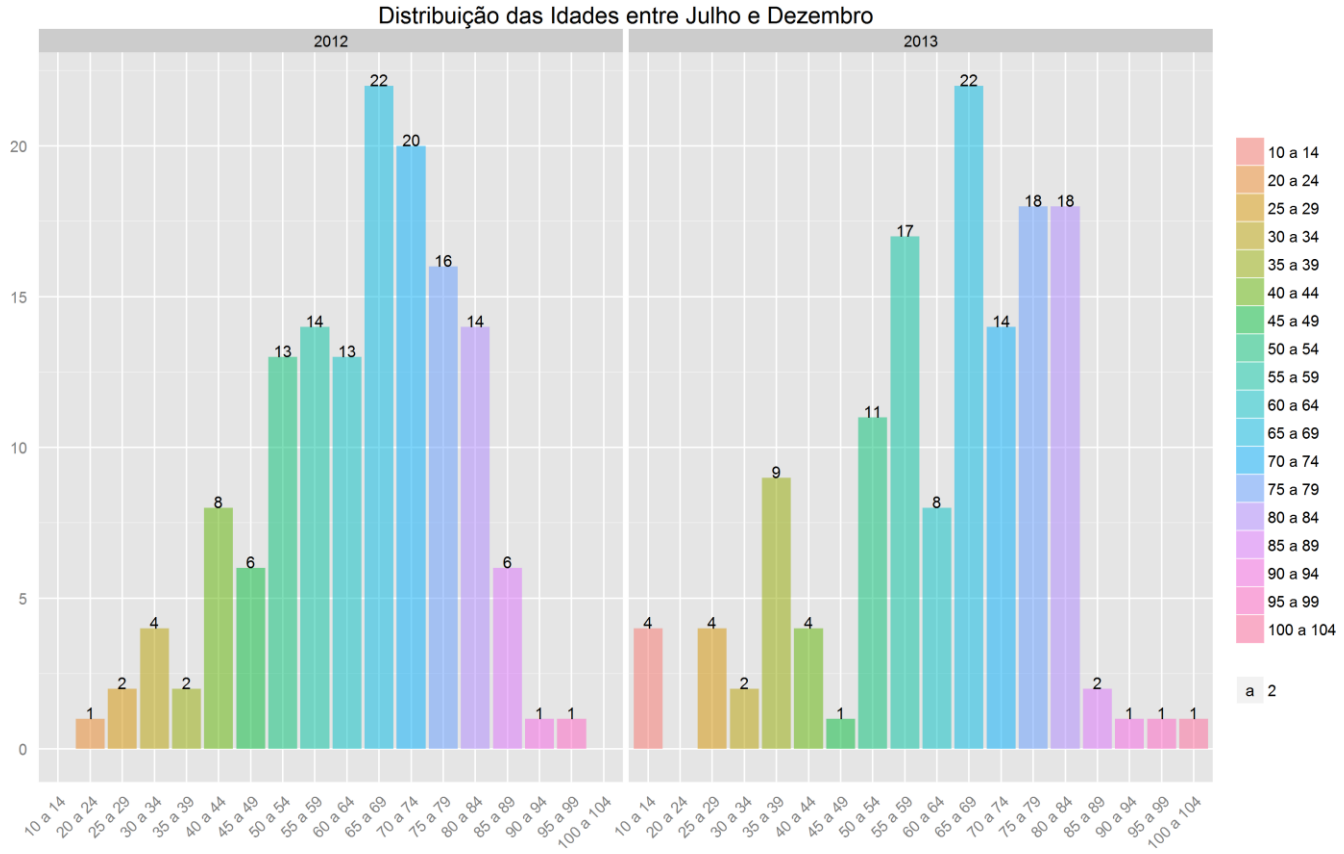
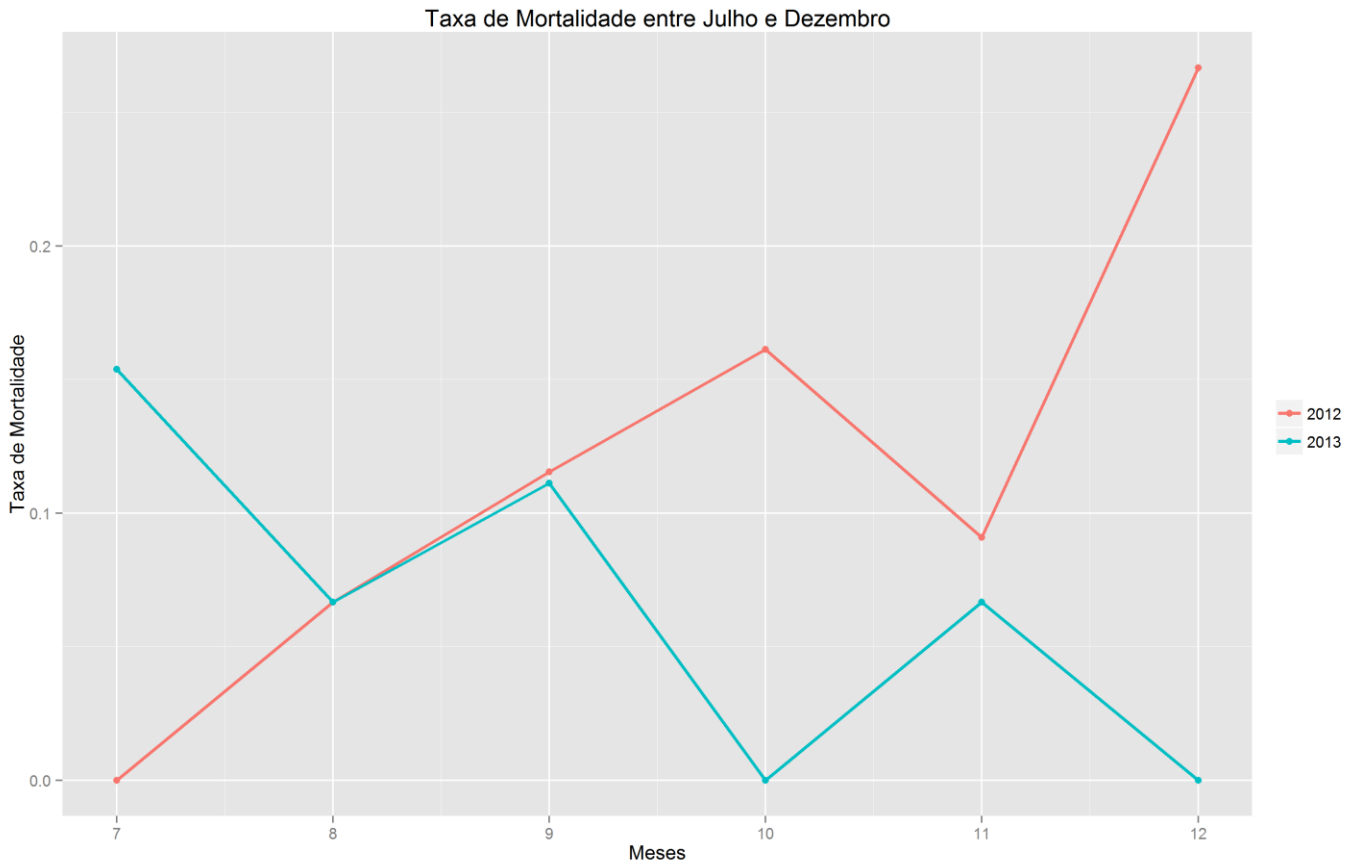


Figura 7 Distribuição das categorias de idade discretizada mês a mês, entre Julho e Dezembro.



Por último, a figura 8 abaixo apresenta a distribuição da taxa de mortalidade na UTI, entre os meses de Julho e Dezembro, nos anos de 2012 e 2013. É preciso levar em consideração que, apesar de ficar evidente uma tendência de aumento da taxa de mortalidade no ano de 2012 e um declínio no ano 2013, é necessário cautela na interpretação do resultado. É possível que ambas as tendências sejam baseadas na utilização do sistema de coleta de dados.

Figura 8 Distrição das taxas de mortalidade entre Julho e Dezembro, nos anos de 2012 e 2013.



2) Análises Preditivas

Os métodos usualmente empregados para verificar a relação entre variáveis médicas e a saída hospitalar em UTI são correlações, regressão linear e logística, análise de variância (ANOVA), análise multivariada da variância (MANOVA), dentre outras. A correlação não é utilizada no procedimento de predição, mas proporciona informações a respeito da direção e da força de relação entre as variáveis empregadas e as saídas hospitalares. Apesar de ser um método útil, a correlação não é adequada para verificar se uma variável é boa ou má preditora de outra variável. Se duas variáveis apresentam coeficiente de correlação pequeno e não estatisticamente significativo, não necessariamente implicará que uma variável é ruim na predição de outra variável.

Tabela 2

Técnicas usuais para verificar a relação entre variáveis médicas e saída hospitalar: pressupostos básicos.

	Main Assumptions							
Technique	Distribuição	Relação entre as variáveis	Homocedasticidade?	Sensível à *outliers*?	Independência?	Sensível à colinearidade?	Demanda uma razão amostra/preditores grande?	Sensível à dados faltantes?
Correlação	Normal Bivariada	Linear	Sim	Sim	NA	NA	NA	Sim
Regressão Linear Simples	Normal	Linear	Sim	Sim	Preditores independentes	NA	Sim	Sim
Regressão Linear Múltipla	Normal	Linear	Sim	Sim	Preditores independentes/ Erros independentes	Sim	Sim	Sim
ANOVA	Normal	Linear	Sim	Sim	Preditores independentes	Sim	Sim	Sim
MANOVA	Normal	Linear	Sim	Sim	Preditores independentes	Sim	Sim	Sim
Regressão Logística	As probabilidades condicionais verdadeiras são uma função logística das variáveis independentes	As variáveis independents não são uma combinação linear uma das outras	Não	Sim	Preditores independentes	NA	Sim	Sim

Já as outras técnicas são diretamente utilizadas na predição de saída hospitalar, sendo que as regressões linear e logística são as mais populares. Além dos pressupostos restritos, as técnicas usualmente empregadas não possuem, no geral, acurácia elevada se comparada com modelos mais recentes. Neste relatório, serão apresentadas quatro técnicas de predição do campo de *Machine Learning* genéricamente chamadas de *tree-based models*: *single learning trees, bagging, Random Forest* e *boosting*. Essas técnicas serão aplicadas na predição da saída hospitalar do UTI do Hospital, utilizando variáveis coletadas pela equipe médica nos pacientes internados entre Julho de 2012 e Dezembro de 2013. As técnicas de árvore não fazem qualquer tipo de pressuposição em relação à normalidade, linearidade, homocedasticidade, colinearidade ou independência (Geurts, Irrthum & Wehenkel, 2009). Não demandam uma razão grande entre amostra e preditores e são mais adequadas que as técnicas usuais para verificar efeitos de interação entre os preditores. Adicionalmente, essas técnicas podem levar à melhoras consideráveis na acurácia da predição da saída hospitalar, uma vez que os *tree-based models* são o que há de estado-da-arte em termos de acurácia de predição (Flach, 2012; Geurts et al., 2009).

Modelos inovadores para realizar predição de saída hospitalar em UTI

O campo denominado de *Machine Learning* é uma nova área da ciência composta por uma ampla gama de métodos computacionais e estatísticos que são utilizados para extrair modelos à partir de um sistema de relações entre observações ou medidas (Hastie et al., 2009; Geurts et al., 2009). A extração de modelos pode ser empregada para resolver diferentes tipos de problemas, como predições, inferências e descobrimento de novas relações não verificadas anteriormente (Flach, 2012; Hastie et al., 2009).

As técnicas de *Machine Learning* são divididas em duas áreas principais, que desempenham diferentes tipos de tarefas: aprendizagem não-supervisionada e supervisionda. No campo da aprendizagem não-supervisionada, o objetivo é descobrir, detectar ou aprender quais são as relações, estruturas, tendências ou padrões nos dados. Existe um vetor com d componentes de observações ou medidas de características (preditores), $\mathfrak{X} = \mathfrak{F}_1 \times \mathfrak{F}_2 \times \mathfrak{F}_3 \times \ldots \times \mathfrak{F}_d$, mas não há respostas associadas (Flach, 2012; James, Witten, Hastie, & Tibshirani, 2013). As características $\mathfrak{F}$ podem ser de quaisquer natureza: nominal, ordinal, intervalar ou de razão.

No campo da aprendizagem supervisionada, por sua vez, para cada observação (ou medida) dos preditores x_i, $i = 1, \ldots, n$, há uma resposta associada, ou variável de saída y_i. O vetor x_i pertence ao espaço de características $\mathfrak{X}$, $x_i \in \mathfrak{X}$, e o vetor y_i pertence ao espaço de saída $\mathfrak{Y}$, $y_i \in \mathfrak{Y}$. A tarefa na aprendizagem supervisionada, pode ser de dois tipos: regressão ou classificação. A regressão é utilizada quando a variável de saída tem natureza intervalar ou de razão, enquanto a classificação é utilizada quando a variável de saída é de natureza categórica. Quando a tarefa é do tipo de *classificação* (por exemplo, classificando as pessoas em grupos como óbito vs. alta hospitalar), o objetivo é construer uma função (l) que mapeie o espaço de características no espaço de saída compost por um conjunto pequeno e finite de classes $\mathcal{C} = \{\mathcal{C}_1, \mathcal{C}_2, \ldots, \mathcal{C}_k\}$,

de tal forma que $l: \mathfrak{X} \rightarrow \mathcal{C}$. Nesse caso, o espaço de saída *é* o conjunto finite de classes: $\mathfrak{F} \equiv \mathcal{C}$. Em suma, no problema de classificação uma variável de saída categótica (p.ex óbito vs alta hospitalar) é predito por um conjuento de características (ou preditores, ou variáveis independentes). No caso da tarefa de regressão, o valor de uma variável de saída de natureza intervalar ou de razão (por exemplo o valor da pressão arterial sistólica) é predito por um conjunto de características específicas (por exemplo IMC, idade, etc). O presente relatório irá centrar-se na tarefa de classificação, uma vez que o objetivo é construir um Sistema de predição de saída do UTI para ser implementado no sistema assistencialista SERVO.

Dentre os métodos de classificação do *Machine Learning*, os chamados *tree-based models*, do campo de aprendizagem supervisionada, são de especial interesse no campo de pesquisas médicas, uma vez que são úteis para: 1) descobrir quais variáveis, ou combinação de variáveis, melhor predizem uma variável de saída (p.ex. saída hospitalar); 2) identificar quais valores de corte, para cada uma das variáveis independentes, predizem maximamente um determinado resultado; e 3) estudar os efeitos de interação entre as variáveis independentes que levam às predições mais acuradas.

As árvores de classificação particionam o espaço de características em várias regiões R_m distintas e mutuamente exclusivas. Cada região é ajustada com um modelo específico que desempenha a função de classificação atribuindo uma das classes $\mathcal{C}_k$ à um espaço particular. A classe é atribuída à região R_m do espaço de características identificando a classe majoritária naquela região. De forma à se chegar em uma solução que melhor separa o espaço de características em nodos (ou regiões) mais puros, a técnica de particionamentos binários recursivos é utilizada. Um nodo é considerado puro quando 100% dos casos são de uma mesma classe (por exemplo: óbito). Um nodo com 90% de pessoas que entraram em óbito e com 10% de pessoas que tiveram alta da UTI é considerado mais puro que um nodo com 50% de cada classe. A técnica de particionamento binário recursive funciona da seguinte maneira. O espaço de características é dividido em duas regiões utilizando-se um valor de corte de uma variável preditora (x_i) de forma que gere a configuração mais pura possível. Cada região gerada por meio desse primeiro particionamento é, então, modelado de acordo com a sua classe majoritária. Os nodos gerados no particionamento um são, por sua vez, divididos em outros nodos, usando a variável preditiva que gere o melhor ajuste dentre todas as pertencentes ao espaço de características. Esse processo de particionamento continua até que o espaço de características atinja a configuração mais pura possível, com R_m regiões ou nodos classificados com uma classe $\mathcal{C}_k$ distinta. As *learning trees* possuem dois parâmetros básicos de ajuste (para ajustes mais refinados verificar Breiman, Friedman, Olshen & Stone, 1984): 1) o número de variáveis utilizadas na predição $n(\mathfrak{X})$, e 2) a complexidade da árvore, que é o número possível de nodos terminais $\alpha|T|$. Nodos terminais são aqueles nos quais a classificação é definitivamente realizada. Se mais de um preditor é dado, então a seleção de qual variável será utilizada para realizar o particionamento de cada região será determinada por meio daquela que proporcione uma partição com configuração mais pura. É importante apontar que, em uma árvore de classificação, o primeiro particionamento indica a variável mais importante na predição. Finalmente, Leek (2013) sintetiza como um algorítmo de árvore funciona: 1) particiona as observações em

grupos de forma iterativa; 2) particiona os dados onde as variáveis são maximamente preditivas e 3) maximiza a homogeneidade em cada grupo.

A qualidade das predições realizadas utilizando-se *learning trees* pode ser verificada utilizando a taxa de erro de classificação e o desvio residual médio (Hastie et al., 2009). De forma a se calcular ambos os índices, primeiramente é necessário calcular a proporção de uma classe $\mathcal{C}_k$ em um nodo m. Como citado anteriormente, a classe que será atribuída à uma região ou nodo particular, será aquela com a maior proporção nesse nodo. Matematicamente, a proporção de uma classe $\mathcal{C}_k$ em um nodo m da região R_m, com N_m pessoas será:

$$\hat{p}_{m\mathcal{C}_k} = \frac{1}{N_m} \sum_{x_i \in R_m} I(y_i = \mathcal{C}_k)$$

A função de classificação que atribuirá a classe $\mathcal{C}_k$ ao nodo m é: $max_{\mathcal{C}_k} \hat{p}_{m\mathcal{C}_k}$. Já a taxa de erro de classificação é simplesmente a proporção de observações que não pertencem à classe $\mathcal{C}_k$ na região m:

$$\frac{1}{N_m} \sum_{x_i \in R_m} I(y_i \neq \mathcal{C}_k) = 1 - \hat{p}_{m\mathcal{C}_k}$$

O desvio médio residual, por sua vez, é obtido por meio da seguinte formula:

$$-2 \sum_m \sum_{\mathcal{C}_k} n_{m\mathcal{C}_k} \log \hat{p}_{m\mathcal{C}_k} \Big/ n - |T|$$

onde $n_{m\mathcal{C}_k}$ representa o número de pessoas (ou casos/observações) da classe $\mathcal{C}_k$ na região m, enquanto n é o tamanho da amostra e $|T|$ é o número de nodos terminais (James et al., 2013).

O desvio é preferível ao erro de classificação, uma vez que é mais sensível à pureza dos nodos na classificação. Por exemplo, vamos supor que duas árvores (A e B), possuam 800 observações cada, com pessoas que tiveram alta ou foram à óbito (50% de cada classe). A árvore A possui dois nodos, sendo A_1 com 300 pessoas que tiveram alta e 100 pessoas que vieram à óbito, e A_2 com 100 pessoas que tiveram alta e 300 que entraram em óbito. A árvore B também possui dois nodos: B_1 com 200 pessoas que tiveram alta e 400 óbitos, e B_2 com 200 altas e nenhum óbito. A taxa de erro de classificação da árvore A e da árvore B são iguais (0,25). No entanto, a árvore B produziu nodos mais puros, uma vez que o nodo B_2 é inteiramente compost por pessoas que tiveram alta. Portanto a árvore B possui menor desvio médio que a árvore A. É possível, também, calcular um pseudo- R^2 para o modelo de árvore pode ser calculado à partir do desvio:

$$\text{Pseudo-}R^2 = 1 - \left({Desvio}/{Desvio\ Nulo} \right).$$

Geurts, Irrthum e Wehenkel (2009) argumentam que as *learning trees* estão entre os algorítmos mais populares do campo de *Machine Learning*, devido à três principais características: interpretabilidade, flexibilidade e facilidade de uso. Interpretabilidade significa que o modelo construído para mapear o espaço de características ao espaço de saída é fácil de compreender, uma vez que é um conjunto de regras do tipo

se-então. James, Witten, Hastie e Tibshirani (2013) apontam que os modelos de árvore são mais fáceis de explicar que os modelos clássicos de predição, uma vez que os primeiros se assemelham mais à forma como o ser humano pensa e toma decições. Já a flexibilidade significa que as técnicas de árvores são aplicáveis à uma ampla gama de problemas, lidam com diferentes tipos de variáveis (incluindo variáveis nominais, ordinais, intervalares e de razão), são técnicas não-paramétricas e não possuem nenhum pressuposto sobre normalidade, linearidade ou independência (Geurts, et al., 2009). Além disso, as técnicas de árvore são sensíveis ao impacto da adição de novas variáveis no modelo preditivo, sendo especialmente relevante para o estudo da validade incremental. Essas técnicas também verificam quais variáveis, ou combinações delas, melhor predizem um dado resultado, assim como calculam quais pontos de corte predizem maximamente esse resultado. Por ultimo, a facilidade de uso das técnicas diz respeito à simplicidade computacional das mesmas.

Apesar das qualidades das *learning trees* apontadas anteriormente, as técnicas sofrem de duas limitações interrelacionadas. A primeira é conhecida como o problema do *overfitting* (ou hiperajuste). Uma vez que o espaço de características é ligado ao espaço de saída por meio de particionamentos binários recursivos, os modelos de árvore podem aprender muito sobre os dados, modelando-os de tal forma que se tornam dependents da amostra. Ao se tornarem dependentes das amostras, no sentido de que o particionamento fica muito apropriado para o conjunto de dados em mãos, os modelos irão ter desempenhos preditivos ruins em novos bancos de dados. O Segundo problema é exatamente a consequência do hiperajuste, e é conhecido como o problema da variância. O erro de predição no conjunto de dados de treino, que é o conjunto de variáveis preditivas e de resposta utilizados para gerar uma árvore de classificação pela primeira vez, pode ser muito diferente do erro de predição em um novo conjunto de dados (conjunto teste). Na presença de hiperajuste ao conjunto de dados de treino, os erros de predição apresentarão uma grande diferença entre os conjuntos de treino e de teste. Ademais, as árvores de classificação não possuem a mesma acurácia preditiva dos outros métodos de *Machine Learning* (James et al., 2013). De forma à previnir o hiperajuste, o problema da variância e também para aumentar a acurácia de predição nas árvores de classificação, uma estratégia chamada de *técnicas de assembléia* pode ser empregada.

As técnicas de assembléia são simplesmente a junção de várias árvores para realizar uma tarefa de classificação baseada na predição realizada por cada árvore individualmente. Há três principais técnicas de assembléias para os modelos de árvores: *bagging*, *Random Forest* e *boosting*. As primeiras duas técnicas aumentam a acurácia de predição e diminuem a variância entre os conjuntos treino e teste, assim como evitam o hiperajuste. Já a técnica de *boosting* apenas aumenta a acurácia, uma vez que pode levar ao hiperajuste (James, Witten, Hastie and Tibshirani, 2013).

O nome *bagging* (Breiman, 2001b) é a abreviação em Inglês de *bootstrap aggregating*, e é um procedimento geral para reduzir a variância em árvores de classificação (Flach, 2012, Hastie et al., 2009; James et al., 2013). O procedimento gera um número B_j de *bootstraps* à partir do conjunto de dados de treino, gerando uma árvore que atribui uma classe $\mathcal{C}_k$ à regiões R_m do espaço de características para cada j.

Por último, a classe k de cada região m é registrada para cada árvore B, e a predição é feita por meio de voto majoritário (Hastie et al., 2009; James et al., 2013). O voto majoritário é simplesmente a classe que mais aparece na região *m* considerando-se todas as B árvores. Como as *bagged trees* não utilizam todos os casos (apenas uma subamostra, usualmente 2/3 do conjunto de dados), os remanescentes (conhecidos como *out-of-bag* ou OOB) sõa utilizados para verificar a acurácia da predição. O erro *out-of-bag* pode ser computado como "uma estimativa válida do erro de teste do modelo, uma vez que a resposta para cada observação é predita usando apenas árvores que não foram ajustadas utilizando essas observações" (p. 323, James et al., 2013). As *bagged trees* possuim dois parâmetros básicos de ajuste: 1) o número de variáveis utilizadas na predição $n(\mathfrak{X})$, que no caso das *bagged trees* é o número total de preditores do espaço de características, e 2) o tamanho *j* do conjunto de *bootstraps B*, que é igual ao número de árvores geradas.

A segunda técnica de assembléia é denominada *Random Forest* (Breiman, 2001a). Essa técnica difere-se da anterior uma vez que utiliza uma subamostra aleatória com reposição (*n*) do conjunto original de dados (*N*) para gerar as árvores, assim como seleciona uma subamostra $\mathfrak{X}_m$ do espaço de características $\mathfrak{X}$ em cada nodo, de forma que o número de variáveis selecionadas é menor que o número total de elementos do espaço de características: $n(\mathfrak{X}_m) < n(\mathfrak{X})$. Como aponta Breiman (2001a), o valor de $n(\mathfrak{X}_m)$ é mantido constante durante todo o procedimento de cômputo da floresta, e é usualmente definido como $\sqrt{n(\mathfrak{X})}$. Por meio da geração de sub amostras aleatórias da amostra original de casos e de variáveis, a técnica de *Random Forest* aprimora o método de *bagging* uma vez que descorrelaciona as árvores geradas (Hastie et al., 2009). Esse procedimento também descorrelaciona os erros cometidos por cada árvore, gerando predições mais acuradas.

E porque a descorrelação é importante? James et al. (2013) criam um cenário que torna essa característica clara. Vamos seguir o argumento dos autores, que é bastante interessante. Imagine que temos um preditor bastante forte no nosso espaço de características, junto com outros preditores moderadamente fortes. No procedimento de *bagging* o preditor forte irá estar presente na primeira partição na maioria das árvores, uma vez que é a variável que melhor separa as classes $\mathcal{C}_k$. Por consequência, as *bagged trees* serão bastante semelhantes umas às outras, com a mesma variável na primeira partição, tornando as predições altamente correlacionadas. Consequentemente, os erros serão altamente correlacionados também. Isso não irá levar à uma diminuição na variância se comparado com uma árvore de classificação única. Já a técnica de *Random Forest*, força cada particionamento a considerer apenas uma subamostra das variáveis, abrindo chance para que as outras variáveis menos fortes façam o trabalho delas. O preditor forte será deixado de fora em uma série de situações, tornando as árvores bastante diferentes umas das outras. Como resultado, as árvores apresentarão menor variância no erro de classificação e no erro OOB, levando à uma predição mais confiável. A técnica de *Random Forest* possui dois parâmetros de ajuste básicos: 1) o tamanho da subamostra de variáveis utilizadas em cada particionamento, $n(\mathfrak{X}_m)$, que obrigatoriamente tem que ser menor que o número total de variáveis existentes ($n(\mathfrak{X}_m) < n(\mathfrak{X})$), sendo geralmente computado como $\sqrt{n(\mathfrak{X})}$; 2) o tamanho *j* do conjunto *B*, que é igual ao número de árvores geradas.

A última técnica a ser apresentada no presente relatório é o *boosting* (Freund & Schapire, 1997). Esse é um método geral adapatativo de predição, e não uma técnica de assembléia tradicional como as anteriores. Nela, cada árvore é construída baseada na árvore anterior, de forma à aumentar o poder de predição. O método de *boosting* aprende a partir dos erros das árvores anteriores, podendo levar ao hiperajuste se o número de árvores geradas for grande. A técnica de *boosting* tem três parâmetros principais de ajuste: 1) o tamanho j do conjunto B, que é igual ao número de árvores geradas; 2) o parâmetro de redução de erros λ (*shrinkage*), que é a taxa de aprendizagem de uma árvore para a outra; e 3) a complexidade da árvore, que é o número possível de nodos terminais $d = \alpha|T|$. James, Witten, Hastie e Tibshirani (2013) apontam que λ é usualmente ajustado como 0,01 ou 0.001, e que quão menor for λ, maior deverá ser o número de árvores (B), para que se consiga predições acuradas.

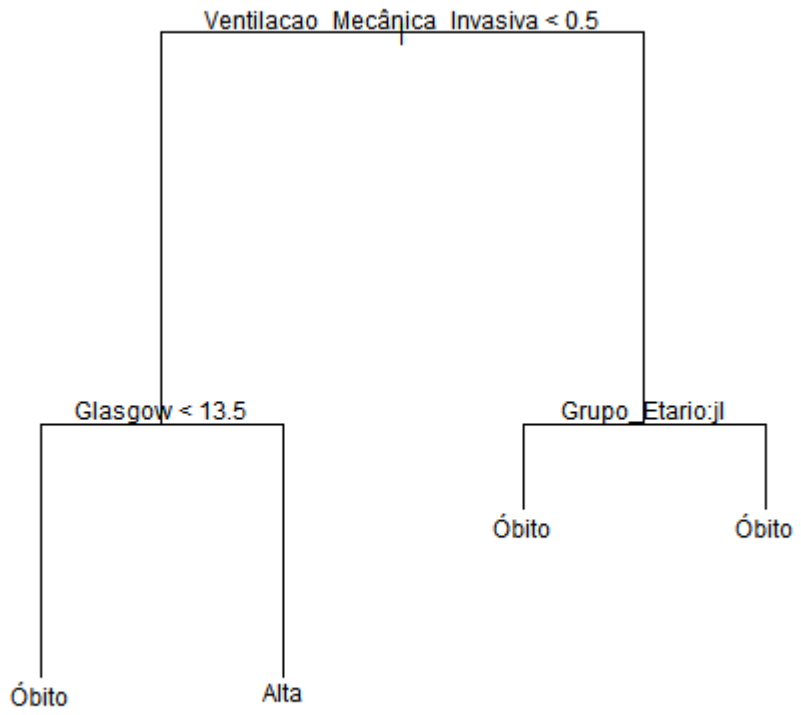
As técnicas de *Machine Learning* apresentadas nesse relatório podem ser úteis para descobrir quais variáveis médicas, ou quais combinações complexas das mesmas, melhor predizem a saída hospitalar em UTI. As técnicas de árvores possuem uma série de vantagens sobre as técnicas mais tradicionais, uma vez que não fazem nenhuma pressuposição acerca da normalidade, independência ou linearidade das variáveis. Além disso, são técnicas não-paramétricas, lidam com diferentes tipos de variáveis, aplicam-se à diferentes tipos de problemas, lidam de forma adequada com valores faltantes (*missings*), e quando combinadas com técnicas de assembléias proporcionam resultados estado-da-arte em termos de acurácia de predição (Geurts et al., 2009). Uma das limitações dos métodos apresentados acima diz respeito ao balanceamento das classes. Os métodos de árvore necessitam de classes relativamente do mesmo tamanho, uma vez que se uma das classes for muito mais frequênte que as outras, os algoritmos irão previlegiar a classe dominante, enviesando os resultados.

Uma vez que apenas 9% dos internados na UTI do Hospital entre Julho de 2012 e Dezembro de 2013, é necessário fazer uma seleção aleatória dos casos que tiveram alta hospitalar. Essa estratégia visa a balancear as classes, e evitar o enviesamento nas predições. Portanto, as análises abaixo empregaram uma subamostra com 86 pacientes, sendo metade no grupo que teve alta hospitalar e a outra metade no grupo que veio à óbito.

Resultados da Análise Preditiva

O primeiro modelo gerado foi o modelo de *single tree*. O resultado no conjunto de dados de treino apontou uma taxa erro de classificação de 0,08, e um desvio médio residual de 0,32. O pseudo-R^2 do modelo apontou que 78,10% da variância foi corretamente explicada pelo modelo. As variáveis selecionadas que melhor particionavam o espaço de características foram a ventilação mecânica invasiva, o escore da escala de Glasgow e os grupos etários maior que 50 anos. A sensibilidade do modelo foi de 100% e a especificidade de 85,18%. O modelo de *single tree* gerado pode ser graficamente visualizado na figura 9 abaixo.

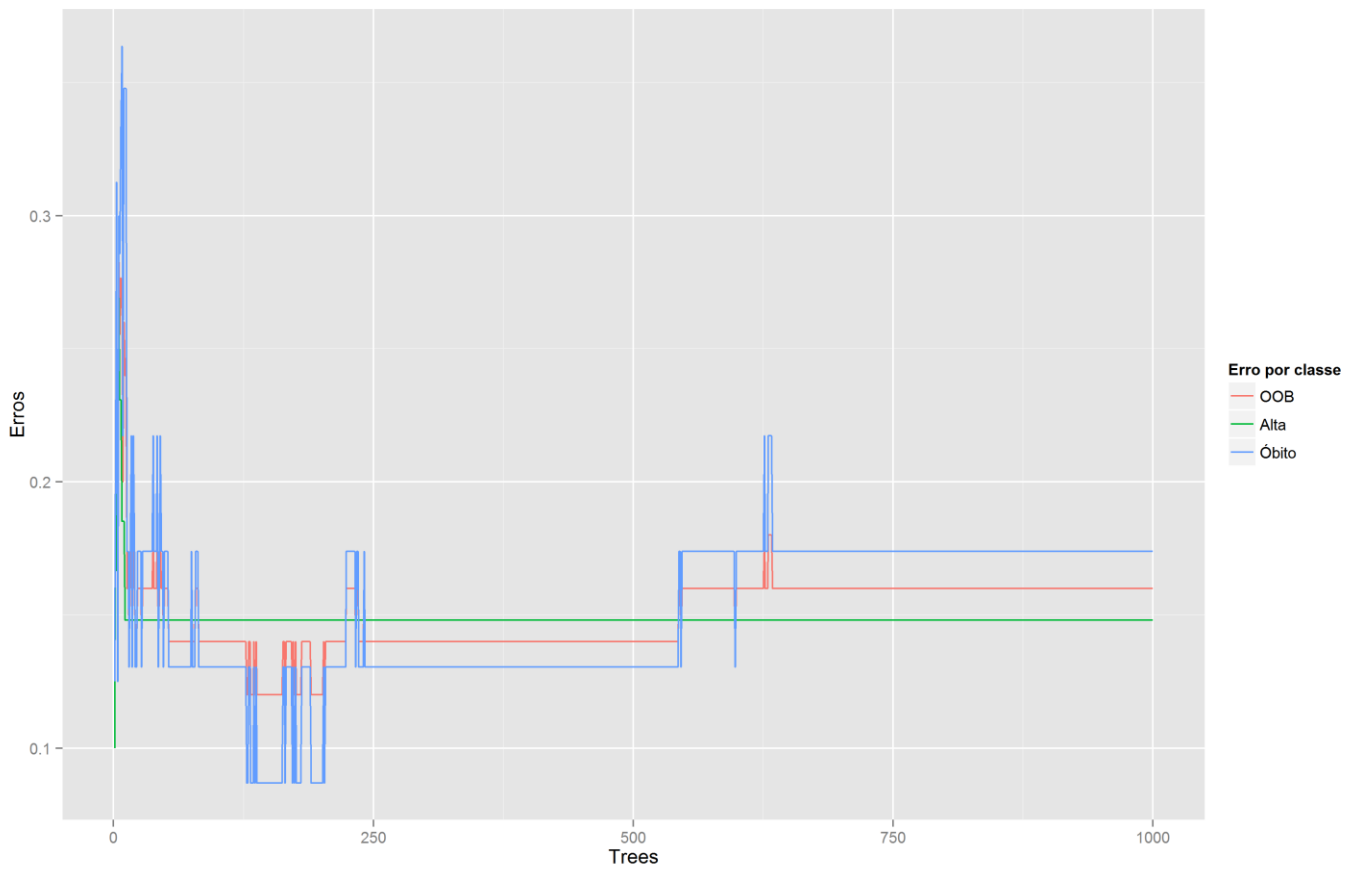
Figura 9 Modelo de árvore gerado.



Esse modelo foi testado na amostra de validação, e obteve uma acurácia geral de 88,83%, com sensibilidade de 100% e especificidade de 75%. Já no grupo de dados de teste, o modelo apresentou uma acurácia geral de 78,95%, com sensibilidade de 90,90% e especificidade de 62,50%. A diferença do ajuste entre o conjunto de treino e o de validação foi de 3,17%, e entre o primeiro e o grupo de teste foi de 13,05%.

O modelo de *bagging*, por sua vez, apresentou uma taxa de erro no conjunto *out-of-bag* de 16%, representando uma acurácia de predição de aproximadamente 84%. O modelo apresentou, ainda, sensibilidade de 82,61% e especificidade de 85,19%. A distribuição do erro *OOB* (vermelho), do erro de classificação do grupo com alta hospitalar (verde) e do grupo de óbito (azul) está apresentada na figura 10 abaixo.

Figura 10 Erros por classe no modelo de *Bagging*.



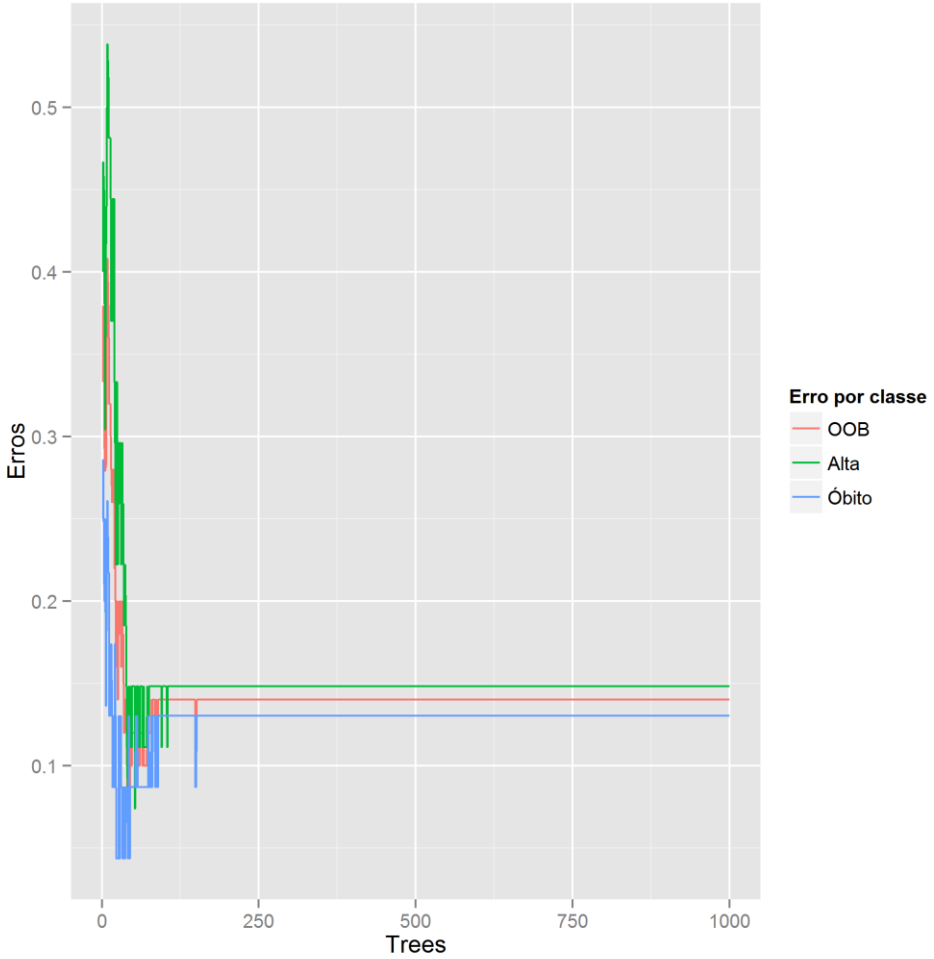
A aplicação do modelo de *bagging* no conjunto de dados de validação apontou uma acurácia total 94,12%, com sensibilidade de 100% e especificidade de 87,50%. Já no conjunto de dados de teste, o modelo de *bagging* apresentou acurácia total de 73,69%, sensibilidade de 81,81% e especificidade de 62,50%. A diferença do ajuste entre o conjunto de treino e o de validação foi de 10,12%, e entre o primeiro e o grupo de teste foi de 10,31%. Em termos de quais variáveis apresentaram maior impacto na predição da saída na UTI, o índice de Gini fornece informações importantes. As quatro variáveis mais importantes na predição, no modelo de *bagging*, foram: ventilação mecânica invasiva, o escore da escala de Glasgow, o grupo etário e a uréia.

Tabela 3
Importância das variáveis na predição Segundo o modelo de Bagging.

Variável	Alta	Óbito	Decaimento Médio da Acurácia	Decaimento Médio do índice Gini
Ventilação Mecânica Invasiva	0,19	0,11	0,15	6,10
Glasgow	0,16	0,06	0,11	3,96
Grupo Etário	-0,01	-0,03	-0,01	1,78
Uréia Med1	0,02	0,01	0,02	1,31
Punção Venosa Profunda subclavia	0,03	0,02	0,03	0,93
FiO2 Medida1	0,01	0,00	0,00	0,39
FR Max	0,00	-0,01	-0,01	0,33
PAm Max	0,00	0,00	0,00	0,23
Dreno de Tórax	0,00	0,00	0,00	0,18
FC Max	0,00	0,00	0,00	0,18
Catéter Vesical	0,00	0,00	0,00	0,12
Punção Venosa Profunda jugular	0,00	0,00	0,00	0,08
Cirrose	0,00	0,00	0,00	0,06
PaO2 Medida1	0,00	0,00	0,00	0,03
Nutrição Parenteral	0,00	0,00	0,00	0,03
Marcapasso provisório	0,00	0,00	0,00	0,01
Pneumonia Associada à Ventilação	0,00	0,00	0,00	0,00
Metodos Dialíticos	0,00	0,00	0,00	0,00
Insuf orgânica ou imunodeprimido	0,00	0,00	0,00	0,00
Imunossupressão	0,00	0,00	0,00	0,00
Monitoramento da Pressão Intracraniana	0,00	0,00	0,00	0,00
Perda de Punção Venosa Profunda	0,00	0,00	0,00	0,00

O modelo de *Random Forest* apontou uma taxa de erro no conjunto *out-of-bag* de 14%, representando uma acurácia de predição de aproximadamente 86%. O modelo apresentou, ainda, sensibilidade de 85,19% e especificidade de 86,96%. A distribuição do erro *OOB* (vermelho), do erro de classificação do grupo com alta hospitalar (verde) e do grupo de óbito (azul) está apresentada na figura 11 abaixo. A aplicação do modelo de *Random Forest* no conjunto de dados de validação apontou uma acurácia total 94,12%, com sensibilidade de 100% e especificidade de 87,50%. Já no conjunto de dados de teste, o modelo de *Random Forest* apresentou acurácia total de 78,95%, sensibilidade de 81,81% e especificidade de 75%. A diferença do ajuste entre o conjunto de treino e o de validação foi de 8,12%, e entre o primeiro e o grupo de teste foi de 7,05%.

Figura 11 Distribuição dos errros por classe no modelo de Random Forest



Em termos de quais variáveis apresentaram maior impacto na predição da saída na UTI, a tabela abaixo apresenta um sumário. As quatro variáveis mais importantes na predição foram: o escore na escala de Glasgow, a ventilação mecânica invasive, a uréia e a punção venosa profunda subclavia.

Tabela 4

Importância das variáveis na predição Segundo o modelo de Random Forest.

Variável	Alta	Óbito	Decaimento Médio da Acurácia	Decaimento Médio do índice Gini
Glasgow	0,11	0,05	0,08	2,94
Ventilação Mecânica Invasiva	0,08	0,06	0,07	2,63
Uréia Med1	0,05	0,03	0,04	1,87
Punção Venosa Profunda subclavia	0,07	0,04	0,05	1,83
Grupo Etário	0,00	-0,01	-0,01	1,77
FiO2 Medida1	0,04	0,02	0,03	1,15
Catéter Vesical	-0,01	0,02	0,01	0,65
Pam Max	0,00	0,00	0,00	0,49
FC Max	0,00	0,00	0,00	0,39
FR Max	0,00	0,00	0,00	0,30
Punção Venosa Profunda jugular	0,00	0,00	0,00	0,27
Dreno de Tórax	0,00	0,00	0,00	0,18
Cirrose	0,00	0,00	0,00	0,17
Métodos Dialíticos	0,00	0,00	0,00	0,08
PaO2 Medida1	0,00	0,00	0,00	0,05
Nutrição Parenteral	0,00	0,00	0,00	0,04
Marcapasso provisório	0,00	0,00	0,00	0,03
Insuf. Orgânica ou imunodeprimido	0,00	0,00	0,00	0,03
Pneumonia Associada à Ventilação	0,00	0,00	0,00	0,03
Imunossupressão	0,00	0,00	0,00	0,00
Monitoramento da Pressão Intracraniana	0,00	0,00	0,00	0,00
Perda de Punção Venosa Profunda	0,00	0,00	0,00	0,00

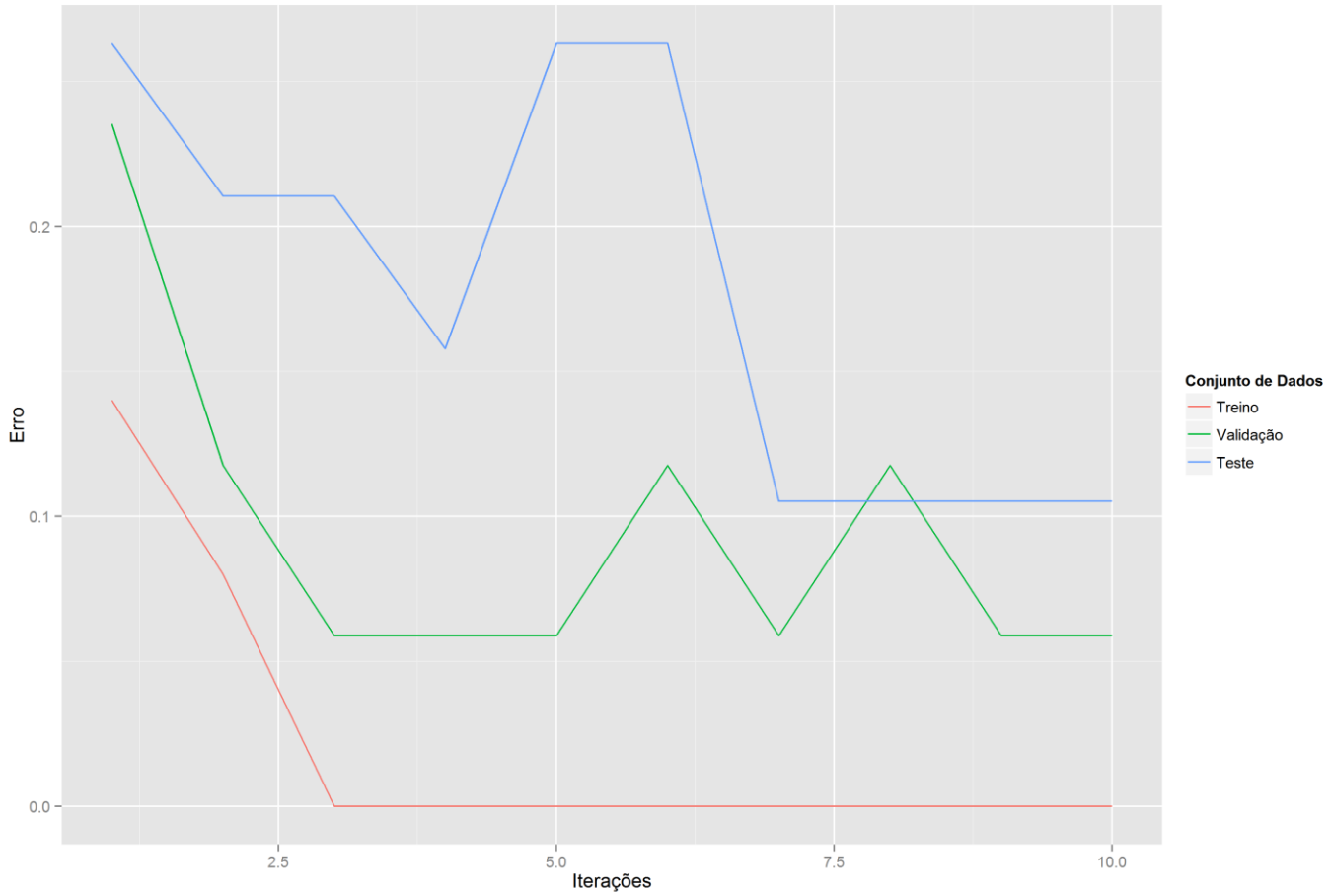
O modelo de *Boosting* apontou uma taxa acurácia de predição 100%, com sensibilidade de 100% e especificidade de 100%. O desempenho do modelo no conjunto de dados de validação apontou uma acurácia total de 94,12%, com sensibilidade de 90% e especificidade de 100%. Já no conjunto de dados de teste, o modelo apresentou uma acurácia total de 89,48%, com sensibilidade de 90,90% e especificidade de 87,50%. A diferença do ajuste entre o conjunto de treino e o de validação foi de 5,88%, e entre o primeiro e o grupo de teste foi de 10,52%. A tabela abaixo apresenta o ganho no índice Gini, para cada variável utilizada na predição. As mais importantes foram o grupo etário, o escore da escala de Glasgow, a uréia, punção venosa profunda subclavia e a ventilação mecânica invasive.

Tabela 5
Importância das variávies Segundo o modelo de Boosting

Variáveis	Ganho no indice gini
Grupo Etário	26,16
Glasgow	24,29
Uréia Med1	14,58
Punção Venosa Profunda subclavia	12,23
Ventilação Mecânica Invasiva	7,45
Cirrose	6,42
Dreno de Tórax	2,71
PAm Max	1,88
FR Max	1,45
Cateter Vesical	1,16
Punção Venosa Profunda jugular	0,84
FC Max	0,84
FiO2 Medida1	0,00
Imunossupressão	0,00
Insuf. Orgânica ou imunodeprimido	0,00
Marcapasso provisório	0,00
Métodos Dialíticos	0,00
Monitoramento da Pressão Intracraniana	0,00
Nutrição Parenteral	0,00
PaO2 Medida1	0,00
Perda de Punção Venosa Profunda	0,00
Pneumonia Associada à Ventilação	0,00

A figura 12 abaixo mostra os erros de predição em cada iteração do modelo de boost, para os conjuntos de dados de treino, validação e teste.

Figura 12 Distribuição dos erros de predição do modelo de *boosting* por iteração em cada conjunto de dados



De forma à verificar se há diferenças estatisticamente significativas entre as estimativas de acurácia, sensibilidade e especificidade em cada modelo para cada conjunto de dados, foi utilizado o procedimento de Marascuilo (1966). Esse procedimento aponta se há diferenças estatisticamente significativas entre todos os pares de proporções, por meio do cômputo de um valor de teste e de um "*critical range*" para todas as comparações pareadas. Se o valor do teste exceder o valor do *critical range*, então a diferença entre as proporções é considerada significativa no nível de 0,05. Uma explicação mais aprofundada do procedimento pode ser verificada no website do NIST/Semantech: (http://www.itl.nist.gov/div898/handbook/prc/section4/prc474.htm). A tabela 6 abaixo apresenta uma síntese dos valores da acurácia total, sensibilidade e especificidade para cada modelo, em cada conjunto de dados, assim como suas as diferenças pareadas. Já a tabela 7 apresenta o valor de teste do procedimento de Marascuilo, o *critical range* e se a diferença é estatisticamente significativa, em todas as comparações realizadas.

Tabela 6

Síntese da performance dos modelos

	Treino			Validação			Teste			Diferença 1 (Treino-Validação)			Diferença 2 (Treino-Teste)			Diferença 3 (Validação-Teste)			Diferença Média		
	Acurácia Total	Sensibilidade	Especificidade	Acurácia Total	Sensibilidade	Especificidade	Acurácia Total	Sensibilidade	Especificidade	Acurácia Total	Sensibilidade	Especificidade	Acurácia Total	Sensibilidade	Especificidade	Acurácia Total	Sensibilidade	Especificidade	Acurácia Total	Sensibilidade	Especificidade
Learning Tree	0,920	1,000	0,852	0,888	1,000	0,750	0,789	0,909	0,625	0,032	0,000	0,102	0,131	0,091	0,227	0,099	0,091	0,125	0,087	0,061	0,151
Bagging	0,840	0,826	0,851	0,941	1,000	0,875	0,737	0,818	0,625	-0,101	-0,174	-0,024	0,103	0,008	0,226	0,204	0,182	0,250	0,069	0,005	0,151
Random Forest	0,860	0,851	0,869	0,941	1,000	0,875	0,789	0,818	0,750	-0,081	-0,149	-0,006	0,071	0,033	0,119	0,152	0,182	0,125	0,047	0,022	0,079
Boosting	1,000	1,000	1,000	0,941	0,900	1,000	0,894	0,909	0,875	0,059	0,100	0,000	0,106	0,091	0,125	0,047	-0,009	0,125	0,071	0,061	0,083

Tabela 7
Significância das comparações pareadas.

	Significância da Diferença 1: Acurácia Total			Significância da Diferença 1: Sensibilidade			Significância da Diferença 1: Especificidade			Significância da Diferença 2: Acurácia Total			Significância da Diferença 2: Sensibilidade			Significância da Diferença 2: Especificidade			Significância da Diferença 3: Acurácia Total			Significância da Diferença 3: Sensibilidade			Significância da Diferença 3: Especificidade			Significância da Diferença Média: Acurácia Total			Significância da Diferença Média: Sensibilidade			Significância da Diferença Média: Especificidade		
	Valor	Critical Range	A diferença é significativa?	Valor	Critical Range	A diferença é significativa?	Valor	Critical Range	A diferença é significativa?	Valor	Critical Range	A diferença é significativa?	Valor	Critical Range	A diferença é significativa?	Valor	Critical Range	A diferença é significativa?	Valor	Critical Range	A diferença é significativa?	Valor	Critical Range	A diferença é significativa?	Valor	Critical Range	A diferença é significativa?	Valor	Critical Range	A diferença é significativa?	Valor	Critical Range	A diferença é significativa?	Valor	Critical Range	A diferença é significativa?
Learning Tree - Bagging	0,07	0,06	Sim	0,17	0,07	Sim	0,08	0,06	Sim	0,03	0,08	Não	0,08	0,05	Sim	0,00	0,11	Não	0,11	0,09	Sim	0,09	0,09	Sim	0,13	0,10	Sim	0,02	0,07	Não	0,06	0,04	Sim	0,00	0,09	Não
Learning Tree - Random Forest	0,05	0,06	Não	0,15	0,06	Sim	0,10	0,06	Sim	0,06	0,08	Não	0,06	0,06	Não	0,11	0,09	Sim	0,05	0,08	Não	0,09	0,09	Sim	0,00	0,08	Não	0,04	0,06	Não	0,04	0,05	Não	0,07	0,08	Não
Learning Tree - Boosting	0,03	0,05	Não	0,10	0,05	Sim	0,10	0,05	Sim	0,03	0,08	Não	0,00	0,07	Não	0,10	0,10	Sim	0,05	0,07	Não	0,08	0,05	Sim	0,00	0,08	Não	0,02	0,07	Não	0,00	0,06	Não	0,07	0,08	Não
Bagging - Random Forest	0,02	0,07	Não	0,03	0,09	Não	0,02	0,03	Não	0,03	0,07	Não	0,03	0,04	Não	0,11	0,09	Sim	0,05	0,10	Não	0,00	0,10	Não	0,13	0,10	Sim	0,02	0,06	Não	0,02	0,03	Não	0,07	0,08	Não
Bagging - Boosting	0,04	0,07	Não	0,07	0,09	Não	0,02	0,03	Não	0,00	0,08	Não	0,08	0,05	Sim	0,10	0,10	Sim	0,16	0,08	Sim	0,17	0,07	Sim	0,13	0,10	Sim	0,00	0,06	Não	0,06	0,04	Sim	0,07	0,08	Não
Random Forest - Boosting	0,02	0,06	Não	0,05	0,08	Não	0,01	0,01	Não	0,04	0,07	Não	0,06	0,06	Não	0,01	0,08	Não	0,11	0,07	Sim	0,17	0,07	Sim	0,00	0,08	Não	0,02	0,06	Não	0,04	0,05	Não	0,00	0,07	Não

Discussão

O presente relatório teve como objetivos realizar análises descritivas e preditivas utilizando os dados coletados na UTI de um hospital particular, entre Julho de 2012 e Dezembro de 2013. A análise descritiva verificou que a incidência de pneumonia associada à ventilação foi de 3,8% enquanto a incidência de pneumotórax por punção venosa foi de 0,002% e de perda de punção venosa profunda foi de 0,006%. Ademais, dos pacientes que demandaram a utilização de ventilação mecânica invasive, cerca de 82% utilizaram-na por apenas um dia. A maior parte dos pacientes que deixaram a UTI foi para a enfermaria (81%), e a maioria dos internados tinham diagnóstico primário clínico. A faixa etária principal de internamento na UTI é a que vai de 65 à 69 anos (ver figura 7). A taxa de óbito média entre Julho e Dezembro de 2012 foi de 13,38%, e a taxa média no mesmo período no ano de 2013 foi de 7,29%.

Do ponto de vista das análises preditivas, o objetivo era construir um modelo de predição de saída da UTI (alta vs. óbito), utilizando os dados coletados pela equipe médica no período investigado. Para maximizar a capacidade preditiva do modelo, técnicas de *Machine Learning* foram utilizadas. Dentre os principais benefícios dessas técnicas encontra-se a capacidade de lidar com quaisquer tipo de variáveis, a sua não-linearidade, e a ausência de pressuposições acerca da normalidade, linearidade, colinearidade e independência dos preditores. Além disso, as técnicas de *Machine Learning* são consideradas estado-da-arte em termos de acurácia. Como a taxa média de óbito foi de aproximadamente 9%, o conjunto de dados encontrava-se bastante desbalanceado em termos das classes-alvo (alta vs. óbito). Para que as estimativas não se enviesassem à classe dominante (alta), uma seleção aleatória de aproximadamente 43 casos com alta da UTI foi realizado, de forma que a subamostra de predição contivesse 86 casos, metade em cada classe. Essa subamostra foi ainda dividida em três conjuntos (treino, validação e teste), com proporções de 60%, 20% e 20%, para a realização de validação cruzada dos modelos gerados. Os modelos de *bagging*, *Random Forest* e *boosting* foram os que apresentaram maior acurácia total no conjunto de dados de validação (94,10% cada). Já no conjunto de dados de teste, o modelo de *boosting* apresentou a acurácia mais elevada (89,40%). Em termos de sensibilidade, os modelos de *single tree*, *bagging* e *Random Forest* empataram no conjunto de dados de validação, com classificação perfeita de óbito (100%), enquanto no conjunto de dados de teste foram os modelos de *single tree* e *boosting* os que tiveram melhor desempenho (90,90%). Já em termos da especificidade, o modelo de *boosting* foi superior tanto no conjunto de dados de validação (100%) quanto no de teste (87,50%). A variabilidade entre os conjuntos de dados para cada modelo aponta diferentes níveis de hiperajuste. A média da diferença entre os conjuntos de dados para a acurácia total não é significativamente diferente para todas as comparações pareadas realizadas (ver tabela 7). Isso significa que todos os modelos apresentaram um nível de hiperajuste relativamente semelhante em termos de acurácia total da predição. No entanto, o modelo de *bagging* é o que apresenta a menor diferença média em sensibilidade, em relação aos três conjuntos de dados, sendo significativamente diferente dos modelos de *single tree* e *boosting*. Isso significa que o modelo de *bagging* é o mais estável na predição correta de óbito no ambiente da UTI. No entanto, esse modelo não apresenta diferença média significativa em relação ao modelo de *Random Forest*. Em termos da especificidade, não houve diferenças significativas em nível de 95% de significância nas diferenças médias entre os modelos. À partir desses resultados, é possível apontar os modelos de *Bagging* e *Random Forest* como os melhores modelos para predizer a saída do UTI. A acurácia desses modelos variou entre 84% e 94%, com a sensibilidade variando entre 82% e 100% e com a especificidade variando entre 62% e 87%.

Apesar de não ser o foco do presente relatório, foi possível realizar algumas análises inferenciais à partir dos modelos de *Machine Learning*. Para o modelo de *Random Forest*, as quatro variáveis que mais impactam na saída da UTI são o escore na escala de Glasgow, os dias de utilização da ventilação mecânica invasiva, o nível de uréia e se a quantidade de dias que a pessoa fez uso de punção venosa profunda subclávia. Já o modelo de *bagging* adiciona o grupo etário no conjunto de variáveis que mais impactam a

saída da UTI. Por último, as figuras 13 e 14 abaixo apresentam a qualidade das predições dos modelos de *Random Forest* e *bagging*. Como cada um dos modelos estima mil árvores para realizar as predições, em cada uma dessas árvores as pessoas são alocadas em nodos específicos com predições específicas. Todas as vezes que uma pessoa x_1 fica em um mesmo nodo que uma pessoa x_2, elas recebem um ponto de proximidade. Após o cômputo das mil árvores, é possível obter uma matriz de proximidade para cada par de pessoas da amostra. Essa matriz de proximidade é a média do número de vezes que duas pessoas estiveram em um mesmo nodo de classificação. Ela pode ser interpretada como sendo uma *matriz de correlação* das classificações (alta ou óbito) atribuída à cada um dos pacientes utilizados na predição. Se cada um dos pacientes for representado por um elemento em uma rede, e a ligação entre os pacientes for os valores dos mesmos na matriz de proximidade, então é possível elaborar um *grafo acíclico indireto.* Quanto mais claro for a separação entre as classes (alta vs. óbito), melhor é a predição gerada pelos modelos. As figuras 13 e 14 abaixo apresentam os grafos acíclicos indiretos obtidos por meio da matriz de proximidade das classificações, distribuídos espacialmente de acordo com o *peso* do escore de proximidade. Quanto maior o escore de proximidade entre dois pacientes, mais próximos espacialmente eles estarão representados no grafo. Essa aproximação espacial é obtida por meio do algoritmo de Fruchterman–Reingold (1991)

Figura 13 Grafo indireto acíclico forçado da matrix de proximidade das classificações do *Random Forest.*

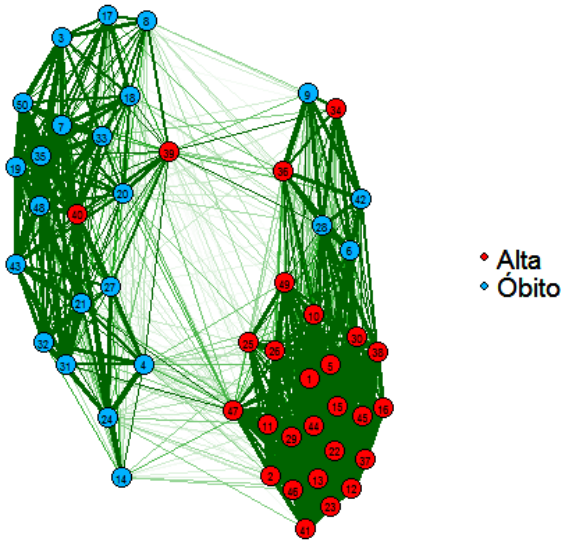
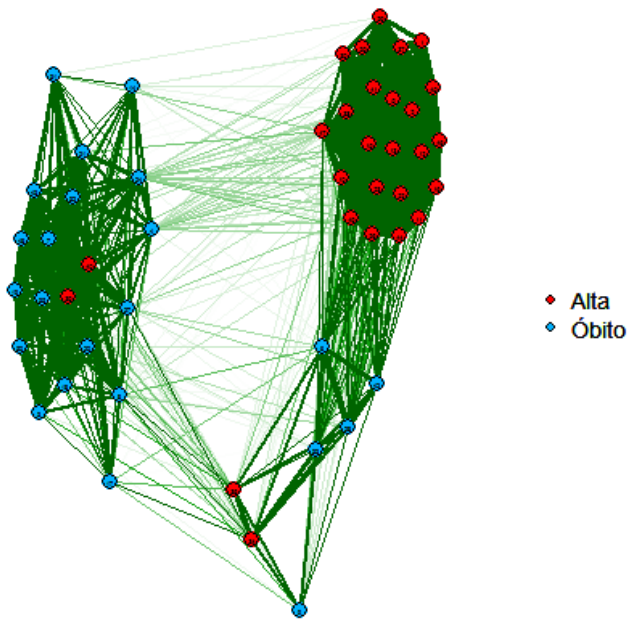


Figura 14 Grafo indireto acíclico forçado da matrix de proximidade das classificações do modelo de *Bagging*.



Referências

Breiman, L. (2001a). Random Forests. *Machine Learning, 1*(45), 5-32. DOI 10.1023/A:1010933404324

Breiman, L. (2001b). Bagging Predictors. *Machine Learning 24* (2), 23–140.

Breiman, L., Friedman, J., Olshen, R., & Stone, C. (1984). *Classification and Regression Trees*. New York: Chapman & Hall.

Edelsbrunner, P. & Schneider, M. (2013). Modelling for Prediction vs. Modelling for Understanding: Commentary on Musso et al. (2013). *Frontline Learning Research, 2*, 99-101.

Flach, P. (2012). *Machine Learning: The art and science of algorithms that make sense of data*. Cambridge: Cambridge University Press.

Freund, Y. & Schapire, R.E. (1997). A Decision-Theoretic Generalization of On-Line Learning and an Application to Boosting. *Journal of Computer and System Sciences, 55*(1), 119-139.

Fruchterman, T. & Reingold, E. (1991). Graph drawing by force-directed placement. Software - Pract. Exp. 21, 1129?1164

Geurts, P., Irrthum, A., & Wehenkel, L. (2009). Supervised learning with decision tree-based methods in computational and systems biology. *Molecular Biosystems, 5*(12), 1593-1605.

Hastie, T., Tibshirani, R., & Friedman, J. (2009). *The Elements of Statistical Learning: Data Mining, Inference and Prediction* (2nd Edition). New York, NY: Springer.

James, G., Witten, D., Hastie, T., & Tibshirani, R. (2013). An introduction to Statistical Learning with applications in R. New York, NY: Springer.

Leek, J.T. (2013). *Predicting with trees*. Coursera's Data Analysis Class Material. Available at https://github.com/jtleek/dataanalysis.

Liaw, A. & Wiener, M. (2012). randomForest: Breiman and Cutler's random forests for classification and regression. R package version 4.6-7. URL

Marascuilo, L.A. (1966). Large-sample multiple comparisons. *Psychological Bulletin, 65*(5), 280-290. doi: 10.1037/h0023189

R Core Team (2013). R: A language and environment for statistical computing. R Foundation for Statistical Computing, Vienna, Austria. URL http://www.R-project.org/.

Ripley, B.D. (2013). Package 'tree': Classification and regression trees. R package version 1.0-33. URL: http://cran.r-project.org/web/packages/tree.