UNIVERSIDADE FEDERAL DO RIO GRANDE DO NORTE	
FICHA DE EXPECTATIVA DE RESPOSTA DA PROVA ESCRITA	
Edital nº:	**008/2018**
Carreira:	( ) MAGISTÉRIO SUPERIOR (X) MAGISTÉRIO EBTT
Unidade Acadêmica:	**INSTITUTO METRÓPOLE DIGITAL**
Área de Conhecimento:	**INTELIGÊNCIA COMPUTACIONAL**

GABARITO DAS QUESTÕES DE MÚLTIPLA ESCOLHA									
1	**C**	5	**A**	9	**B**	13	**A**	17	**A**
2	**A**	6	**B**	10	**D**	14	**D**	18	**B**
3	**C**	7	**D**	11	**A**	15	**C**	19	**C**
4	**D**	8	**A**	12	**D**	16	**B**	20	**C**

CRITÉRIOS DE AVALIAÇÃO PARA TODAS AS QUESTÕES DISCURSIVAS

- Clareza e propriedade no uso da linguagem;
- Coerência e coesão textual;
- Domínio dos conteúdos, evidenciando a compreensão dos temas objeto da prova;
- Domínio e precisão no uso de conceitos;
- Coerência no desenvolvimento das ideias e capacidade argumentativa.

QUESTÃO 1: valor (0,00 a 3,40 pts)

Com o avanço nos estudos das ciências da vida, em particular a biologia molecular contribuindo cada vez mais para geração de dados genômicos, ao longo das últimas décadas surgiram diversas bases de dados biológicas. Seja com o propósito de auxiliar no desenvolvimento da indústria, na agricultura, no tratamento de doenças, ou na investigação científica, essas bases de dados vêm se consolidando como uma valiosa fonte de informação. Tais bases de dados, além de englobar um grande volume de informação, são conhecidas por apresentarem múltiplos formatos e por explorarem diversos atributos em variados níveis de informação. Esse aumento horizontal das bases de dados leva a um problema grande para os algoritmos de Mineração de dados ou Inteligência Computacional, que é o aumento da dimensionalidade. Além de elevar o custo computacional com o aumento de atributos, leva a outros problemas, por exemplo: autocorrelação, alta variância, entre tantos outros.

Com relação à aplicação de métodos de inteligência computacional em análise de dados com alta dimensionalidade, responda:

a) Em uma etapa de pré-processamento de dados é comum a aplicação de métodos de redução de dimensionalidade. Justifique a necessidade da aplicação de um método de redução de dimensionalidade e descreva métodos que possam ser aplicados neste caso.

Resposta: O candidato deve dissertar sobre métodos de redução de dimensionalidade ressaltando: custo computacional, maldição da dimensionalidade, métodos baseados em seleção, métodos baseados em extração. Como justificativa o candidato deve indicar que a redução de dimensionalidade pode acelerar o treinamento do modelo ou mesmo permitir encontrar relações que seriam inviáveis no tratamento do dado cru. Alternativamente o candidato pode relatar a capacidade de achar relações entre diferentes atributos utilizando estes métodos e a possibilidade de visualização da diferença entre as variáveis.

b) Descreva a aplicação de um método de inteligência computacional que permita implementar um sistema de modelo preditivo a partir dos dados coletados de uma base de dados com alta dimensionalidade, considerando etapa de treinamento do modelo e aplicação.

Resposta: O candidato deve descrever em alto nível a implementação de um sistema de análise preditivo, discutindo desde a modelagem dos dados, a escolha do algoritmo, a separação dos conjuntos de treino e teste e a avaliação do modelo obtido. Para cada um desses pontos o candidato deve discutir diferentes opções e justificar suas escolhas.

QUESTÃO 2: valor (0,00 a 3,30 pts)

Com o advento da tecnologia e a difusão da utilização de técnicas de Inteligência Computacional nas mais diferentes áreas de aplicação, tem-se buscado sistemas cada vez mais eficientes e eficazes. Contudo, nem sempre é possível obter um único modelo com tais características. Neste contexto, a combinação de diferentes modelos de classificação tem surgido como uma alternativa possível para solucionar este problema. Diferentes estratégias de construção de comitês de classificadores têm sido propostas na literatura, tanto para comitês homogêneos quanto heterogêneos.

a) Descreva o funcionamento das estratégias de construção de comitês de classificadores *Bagging*, *Boosting* e *Stacking*, citando suas principais características, vantagens e desvantagens.

Resposta: O candidato deve descrever o algoritmo das estratégias Bagging, Boosting e Stacking, as características/propriedades de cada estratégia, suas vantagens e desvantagens. Como características/propriedades o candidato deve mencionar as técnicas de amostragem para formação do conjunto de treinamento para geração de cada componente do comitê, se são gerados comitês homogêneos ou heterogêneos, métodos de combinação das saídas dos componentes do comitê etc.

b) Defina em que situações o uso de comitês de classificadores é mais indicado, apresentando os atributos que devem ser considerados para avaliação.

Resposta: O candidato deve abordar as circunstâncias que motivam a utilização de comitês de classificadores, mencionando as melhorias que podem ser obtidas através de sua utilização. O candidato deve ressaltar a importância da diversidade em comitês de classificadores, mencionando os benefícios que podem trazer para a acurácia dos comitês, mencionando a relação entre diversidade e acurácia. Além disso, o candidato deve mencionar que dependendo do grau de diversidade do comitê, ao invés de melhorar o desempenho

geral, pode deteriorá-lo. O candidato deve mencionar que a avaliação da aplicabilidade de comitês perpassa por uma comparação com o desempenho de classificadores individuais. Ademais, o candidato pode mencionar e/ou descrever as medidas de diversidade existentes na literatura.

QUESTÃO 3: valor (0,00 a 3,30 pts)

Técnicas de processamento de linguagem natural (PLN) têm sido usadas tanto no contexto de dados estáticos como em streaming. Um exemplo real de aplicação de PLN a streaming é a análise de dados provenientes de mídias sociais. Considere uma aplicação que vise analisar postagens em uma rede social de mensagens de redes curtas (ex. Twitter).

a) Descreva as formas de representação de textos como dados passíveis de processamento pelos algoritmos tradicionais de aprendizado de máquina e quais os pré-processamentos essenciais para tratar tais dados.

Resposta: O candidato deve inicialmente descrever formas de representação de textos, como as baseadas em frequência (bag-of-words, IDF e TF-IDF), n-gramas e word embeddings. Com relação ao pré-processamento, o candidato deve falar sobre as abordagens mais comumente utilizadas, como a remoção de stopwords, pontuações, sufixos, e diacríticos.

b) Considere agora que sua aplicação deverá ser executada sobre um streaming de dados (coleta em tempo real). Defina uma estratégia para que sua abordagem descrita na questão anterior escale adequadamente para este contexto.

Resposta: O candidato deverá demonstrar consciência de que este tipo de aplicação é afetado tanto pela escala dos dados quanto pela sua dinamicidade. Quanto a escala, deve demonstrar consciência de que é impraticável armazenar todo o histórico de dados visto, sendo necessário o uso de sketches. Sobre a dinamicidade dos dados, deve demonstrar consciência de que modelos podem precisar ser retreinados em funções de novos dados, bem como do custo computacional acarretado em razão disto. Para concluir, o candidato deve demonstrar conhecimento de como algoritmos podem fazer uso de janelas para aprender de forma dinâmica.

Natal/RN, 23 de setembro de 2018.

Jorge Estefano Santana De Souza
PRESIDENTE

Leonardo César Teonácio Bezerra
1o EXAMINADOR

Araken de Medeiros Santos
2o EXAMINADOR

HORA DE AFIXAÇÃO DA ATA NO QUADRO DE AVISOS
13:25h