Instituto Serzedello Corrêa
Escola Superior do Tribunal de Contas da União

CURSO DE ESPECIALIZAÇÃO EM

ANÁLISE DE DADOS PARA O CONTROLE

Projeto Pedagógico
2018

Instituto Serzedello Corrêa
Escola Superior do Tribunal de Contas da União

SUMÁRIO

Instituto Serzedello Corrêa
Escola Superior do Tribunal de Contas da União

1. IDENTIFICAÇÃO

Nome do curso	**Especialização em Análise de Dados para o Controle**
Área do conhecimento	1.03.00.00-7 – Ciência da Computação
Modalidade	Presencial
Instituição promotora	Instituto Serzedello Corrêa (ISC) do Tribunal de Contas da União (TCU)
Número de vagas	30 vagas

2. JUSTIFICATIVA

As tecnologias da informação e da comunicação (TIC) se tornaram em poucas décadas essenciais e incontornáveis, seja para as corporações, o governo e o próprio cidadão. De um papel inicialmente baseado na automação de processos e registro de transações, as TIC evoluíram para uma imensa plataforma de serviços, interconexão, armazenamento e distribuição de informações digitais.

Enquanto organizações que lidam essencialmente com informação e conhecimento, a questão talvez mais relevante diante dessa revolução para as entidades de fiscalização seja como se apropriar desse universo de tecnologias e conteúdo informacional posto ao nosso alcance para dele retirar ferramental e conhecimento útil à prática do controle.

Dentro dessa perspectiva, esse curso busca especializar o auditor na compreensão e uso de uma gama de teorias, metodologias e ferramentas tecnológicas voltadas à manipulação e análise de grandes bases de dados. Teremos como pano de fundo a Ciência de Dados, um campo interdisciplinar que lida com processos e sistemas destinados a extrair conhecimento ou discernimento de dados em diversos formatos, estruturados ou não, combinando técnicas e teorias

provenientes de vários campos das áreas da matemática, estatística, ciência da informação e ciência da computação.

Tendo sempre um enfoque prático e aplicado, o curso cobrirá cada etapa do macroprocesso de descoberta de conhecimento trabalhando casos e exemplos em escala real e lidando com temas e bases de dados afetos à realidade cotidiana do auditor e da administração pública.

3. ENTIDADE CERTIFICADORA

3.1. Institucional

Criado por lei federal (art. 88 da Lei nº 8.443, de 16/7/1992), o Instituto Serzedello Corrêa (ISC) foi concebido pelo ideal de propiciar condições para o desenvolvimento humano e organizacional do Tribunal, provendo a Secretaria do TCU de pessoas qualificadas para o exercício das atividades de controle externo da Administração Pública.

As competências do Instituto estão definidas na Resolução-TCU nº 284, de 30 de dezembro de 2016, que define a estrutura, as competências e a distribuição das funções de confiança das unidades da Secretaria do Tribunal de Contas da União; na Resolução-TCU nº 212, de 25 de junho de 2008, que dispõe sobre o desenvolvimento de ações de educação no âmbito do TCU; em seu Regimento Interno; e nos regulamentos específicos de cada segmento educacional, inclusive, no da pós-graduação. A Resolução-TCU nº 212/08 estabelece para o Programa de Pós-Graduação os seguintes objetivos:

> I – promover a pesquisa científica e a geração de conhecimento em nível avançado em áreas de interesse do TCU, com vistas a melhorar a eficiência, a eficácia e a efetividade das ações realizadas pelo Tribunal no cumprimento de sua missão institucional;
>
> II – aprimorar a qualificação e a especialização dos servidores do TCU e ampliar o corpo docente do Instituto Serzedello Corrêa, com vistas à promoção de futuros projetos de pós-graduação de interesse institucional; e
>
> III – criar as condições necessárias à preservação de uma cultura organizacional comprometida com a inovação e com a permanente adequação das competências dos servidores aos objetivos do Tribunal.

Em 14 de fevereiro de 2017, o Ministério da Educação publicou o despacho do Ministro em que homologa o Parecer nº 657/2016, da Câmara de Educação Superior, do Conselho Nacional de Educação, favorável ao credenciamento da Escola de Governo Instituto Serzadello Corrêa e unidades vinculadas, para oferta de pós-graduação lato sensu, em regime presencial e a distância, observando-se o prazo de 8 (oito) anos.

4. OBJETIVOS

4.1. Objetivo Geral

Especializar servidores na extração de conhecimento útil à atividade de controle e auditoria da administração pública a partir de grandes e diversas bases de dados eletrônicos por meio do uso de metodologias e ferramentas tecnológicas de análise, mineração e visualização de dados e informação.

4.2. Objetivos Específicos

São objetivos específicos do curso:

a) Sensibilizar o auditor quanto ao potencial do uso de dados para a melhoria das atividades inerentes ao controle;
b) Treinar o auditor no uso de ferramentas e recursos informacionais disponíveis no TCU;
c) Apresentar as principais técnicas de análise de dados mostrando quando e como aplica-las a casos práticos;
d) Preparar o auditor para as diversas etapas do processo de descoberta de conhecimento dando enfoque na criação de produtos de uso continuado.

5. PÚBLICO-ALVO

O curso destina-se a profissionais que tenham familiaridade com dados armazenados em bases tabulares, preferencialmente bancos de dados, que tenham noções de programação e que atendam aos requisitos:

a) Ser servidor ativo do Tribunal de Contas da União (TCU) ou do Ministério da Transparência e Controladoria-Geral da União (CGU);

b) No caso de servidor do TCU, ocupar o cargo de Auditor Federal de Controle Externo (AUFC) ou o cargo de Técnico Federal de Controle Externo (TEFC) e ter anuência expressa do dirigente da unidade de lotação;

c) ter conhecimento de inglês suficiente para leitura e interpretação de textos;

d) ter familiaridade com dados armazenados em bases tabulares, preferencialmente bancos de dados;

e) ter noções de programação.

Serão ofertadas 30 vagas, assim distribuídas:

1) Duas vagas para servidores da CGU.

2) Vinte e oito vagas para servidores do TCU, distribuídas para os candidatos classificados conforme ordem decrescente da pontuação obtida, respeitando-se os seguintes critérios:

    a) Reserva para cada Unidade organizacional que tenha candidato participante no presente processo seletivo de um quantitativo de até 7% (sete por cento) do total de sua lotação autorizada disponível, de acordo com as diretrizes estabelecidas pela Portaria-TCU nº 61, de 23 de fevereiro de 2018, arredondando-se as frações para o primeiro número inteiro imediatamente superior.

    b) Atendido o critério previsto no subitem anterior e havendo vagas remanescentes, estas serão preenchidas independentemente da unidade de lotação e conforme a ordem de classificação geral dos candidatos.

O ingresso no curso ocorrerá por meio de processo seletivo regido por edital próprio e obedecerá aos seguintes critérios:

    a) Entrega de toda a documentação prevista no edital de abertura;

a) Aprovação de um pré-projeto de pesquisa, consistindo de uma proposta preliminar, que poderá sofrer aprimoramentos e modificações ao longo do curso, mas na qual o candidato deverá: (i) demonstrar conhecimento da temática abordada pelo curso ao identificar um tema de interesse; (ii) delimitar um problema ou objeto de pesquisa; (iii) elaborar a justificativa do estudo, evidenciando sua vinculação aos objetivos estratégicos ou às necessidades de desenvolvimento institucional do TCU, ou ainda a questões de caráter inovador que podem ser estudadas; (iv) descrever a metodologia mais adequada ao estudo; (v) enunciar os objetivos e resultados esperados (vi) apontar possiblidades de aplicação e incorporação dos conhecimentos a processos de trabalho ou ao ambiente organizacional e o possível alcance e resultados da disseminação do conhecimento adquirido ou produzido com a pesquisa.

b) Avaliação do currículo do candidato, descrevendo sua experiência profissional. Será dada especial atenção nessa avaliação à experiência do candidato no uso de dados tabulares e linguagens de programação.

c) Avaliação da capacidade de elaboração de texto de caráter acadêmico-científico, a ser aferida a partir da análise do pré-projeto de pesquisa;

d) Conhecimento de inglês suficiente para leitura e interpretação de textos;

O cronograma do processo seletivo, requisitos e critérios de aceitação e participação de candidatos com o perfil mais adequado para o curso, bem como o procedimento para a matrícula, serão fixados em edital.

6. PERFIL DO EGRESSO

O egresso do curso de pós-graduação em análise de dados retornará às suas atividades laborais com uma maior capacidade de contribuir para a melhoria do controle da gestão pública ao utilizar, de forma mais ampla e eficaz, o universo de dados e informações relacionadas ao dia-a-dia da administração pública. Essa contribuição será exercida profissionalmente na área de atuação de cada egresso, mesmo que não atue diretamente no Controle Externo. Seja fiscalizando obras públicas e programas de governo, analisando tomadas e/ou prestações de contas, participando de

auditorias em órgãos públicos ou empresas estatais, controlando a regulação e/ou os processos de desestatização, promovendo o controle social e a gestão pública responsável, o estudante egresso do Instituto deverá ser capaz de, no exercício de suas atividades laborais, promover e proteger o interesse público explorando de forma mais eficiente dados, informações e recursos tecnológicos à sua disposição.

A partir desse processo de desenvolvimento técnico e profissional, buscar-se-á incentivar o servidor egresso a continuar sua busca por autodesenvolvimento, por meio de uma educação por toda a vida, em benefício da Administração Pública e da sociedade brasileira.

7. CONCEPÇÃO DO PROGRAMA

O curso de Especialização em Análise de Dados é uma ação de formação continuada que pretende aprimorar o conhecimento instrumental e teórico dos servidores envolvidos nessa área de atuação. Trata-se de um processo formativo em serviço, com base na educação formal em pós-graduação *lato sensu*, com o objetivo maior de gerar conhecimentos que possibilitem a realização de auditorias com alto nível de especialização.

A estrutura do curso é composta de 15 (quinze) disciplinas, perfazendo um total de 360 (trezentos e sessenta) horas/aula mais um trabalho de conclusão na forma de monografia.

8. COORDENAÇÃO

O curso será coordenado pelo Serviço Pós-Graduação (Sepos), com o apoio do Serviço de Ações Educacionais Presenciais (Sedup).

9. CARGA HORÁRIA

A carga horária do curso será de 300 horas de aulas presenciais, acrescidas de 60 horas de atividades na modalidade ensino a distância. Será proposta uma disciplina de nivelamento precedendo o início do primeiro período letivo, correspondente a 28 horas aula, das quais 20 horas na modalidade ensino a distância. No total serão 360 horas/aula. O curso terminará por um trabalho de conclusão orientado, estimado em 74 horas de dedicação. Ao final estima-se que o curso demande em torno de 434 horas de dedicação distribuídos em no máximo quinze meses.

10. PERÍODO E PERIODICIDADE

O curso será ofertado a partir de agosto de 2018, conforme o seguinte cronograma geral:

Período Letivo	**Início (datas prováveis)**	**Término (datas prováveis)**
Nivelamento	25/06/2018	03/08/2018
1º período	06/08/2018	26/11/2018
2º período	04/02/2019	24/06/2019
TCC	25/06/2019	31/10/2019

As aulas presenciais serão ministradas semanalmente às segundas-feiras das 8h às 12hrs e das 14h às 18hrs, totalizando 8 (oito) horas semanais. Aulas presenciais poderão ser seguidas de uma atividade extraclasse, na modalidade ensino a distância, que deverá ser desenvolvida antes do encontro seguinte. Se necessário, poderá haver a realização de atividades fora desses horários ou a antecipação de aulas para outros dias da semana quando as segundas-feiras não forem dias úteis. O cronograma detalhado das aulas será definido e comunicado aos alunos oportunamente.

11. CORPO DOCENTE

O corpo docente do curso de Especialização em Auditoria do Setor Público será constituído, principalmente, por servidores pertencentes ao quadro do TCU que possuem titulação, experiência pedagógica e conhecimento profissional na área do curso. Também poderão compor o corpo docente professores convidados de outras instituições que, por sua qualificação, tenham a possibilidade de complementar a formação oferecida pelos docentes internos, obedecendo aos parâmetros estabelecidos pelo MEC e à legislação vigente.

12. METODOLOGIA

As atividades pedagógicas serão desenvolvidas segundo uma abordagem que privilegia a associação entre teoria e prática, por meio de aulas expositivas, discussões e trabalhos em grupo, estudos de

casos, leitura crítica de textos, debates em sala de aula, seminários e palestras com profissionais e professores convidados, além de outras julgadas pertinentes pelos professores e alunos.

Ressalte-se que, na realização dessas atividades didáticas, o objetivo maior é estabelecer não apenas o trânsito entre trabalho e educação, como é comum e desejado em um processo de formação em serviço, mas também uma rede de interconexões entre os saberes e as práticas que fundamentam o campo de estudo.

As estratégias de ensino deverão ser cuidadosamente selecionadas e planejadas, de modo a propiciar situações que:

- Busquem fortalecer a integração entre teoria e prática, valorizando a experiência prévia do aluno.
- Viabilizem posicionamentos críticos.
- Proponham problemas e questões, como pontos de partida para discussões.
- Definam a relevância de um problema por sua capacidade de propiciar o saber pensar, não se reduzindo, assim, à aplicação mecânica de fórmulas feitas.
- Provoquem a necessidade de busca de informação.
- Enfatizem a manipulação do conhecimento, não a sua aquisição.
- Otimizem a argumentação e a contra argumentação para a comprovação de pontos de vista, seja durante aulas expositivas ou por meio do uso de técnicas de ensino grupal ativas, que privilegiem o debate em torno dos temas do curso.
- Dissolvam receitas prontas, criando oportunidades para tentativas e erros.
- Desmistifiquem o erro, desencadeando a preocupação com a provisoriedade do conhecimento, a necessidade de formulação de argumentações mais sólidas.

- Tratem o conhecimento como um processo, tendo em vista que ele deve ser retomado, superado e transformado em novos conhecimentos.

- Contribuam para a implementação de um processo de aprendizagem emancipatório, permitindo a abertura de espaços para a construção do próprio conhecimento na área de abrangência do curso proposto.

- Adotem estilo de ensino com treinamento interativo, galgado em combinação de abordagens que fomentam conhecimento e compartilhamento de experiências, que despertem a atenção, estimulem e mantenham o interesse e o envolvimento da turma, com programas interativos, além dos recursos audiovisuais pertinentes, palestras, chats entre outras atividades.

- Adaptem-se ao perfil do aluno, dos diferentes níveis de ganhos, bem como ao grau de dificuldade identificado durante o processo de ensino-aprendizagem.

A adoção de métodos de ensino-aprendizagem com base nesses critérios contrapõe-se a uma postura preocupada em repassar conhecimentos a serem apenas copiados e reproduzidos, desafiando os alunos a participarem ativamente das aulas, evitando uma atitude de meros espectadores, e fomentando sua capacidade de problematizar e buscar respostas próprias, calcadas em argumentos sólidos.

Estudos de casos serão realizados permitindo que os alunos participem ativamente do processo de construção do conhecimento.

Com exceção dos dois únicos componentes curriculares de cunho teórico, a saber “O processo de descoberta de conhecimento” e “Inovação e Metodologia Científica”, todos os demais utilizarão uma plataforma tecnológica aberta única conhecida como Jupiter Notebook, onde conteúdo teórico, exemplos e exercícios são apresentados numa mesma interface ativa, permitindo ao aluno experimentar em sala de aula desde o momento em que recebe a teoria dos novos conceitos e técnicas. Nessa mesma plataforma o aluno poderá resolver exercícios na sala de aula e em atividades extraclasse, utilizando a linguagem Python e acessando dados das mais diversas fontes.

13. INTERDISCIPLINARIDADE

O foco do curso nas questões referentes à análise de dados para descoberta de conhecimento favorece a interdisciplinaridade, uma vez que requer conhecimentos de várias áreas. Espera-se que os alunos percebam essa relação e saibam localizá-la nas diferentes disciplinas do curso. Mais especificamente, o diálogo necessário entre as diversas áreas do saber será incentivado na elaboração do TCC, que, mesmo quando situado em um aspecto específico a ser abordado, não poderá deixar de fazer referência ao conjunto das disciplinas do curso.

14. TRABALHO DE CONCLUSÃO DO CURSO (TCC)

No segundo semestre das disciplinas presenciais, o discente deverá elaborar um trabalho de conclusão de curso cuja defesa ocorrerá diante de uma banca de professores nas últimas semanas do curso. O TCC deverá ser elaborado individualmente. Para a elaboração do TCC será designado um professor orientador. O TCC seguirá as normas de padronização editadas pela Associação Brasileira de Normas e Técnicas (ABNT).

15. ATIVIDADES COMPLEMENTARES

As disciplinas presenciais de maior complexidade serão complementadas por atividades laboratoriais, desenvolvidas na modalidade de ensino a distância, com o acompanhamento de tutores. Independentemente da programação dessas disciplinas laboratoriais, outras atividades complementares poderão ser desenvolvidas, como a participação em eventos e visitas, devendo ocorrer de acordo com a disponibilidade dos alunos e o interesse das instituições participantes.

16. TECNOLOGIA

Além das comunicações e do atendimento alternativo, que serão feitos via *e-mail*, os alunos serão inscritos em uma comunidade virtual de aprendizagem, que receberá o nome do curso, para facilitar a interação tanto entre alunos e professores quanto entre os próprios alunos. Essa comunidade será parte da plataforma de ensino a distância do ISC.

A estrutura do ISC também comporta um sistema para registro e acompanhamento dos cursos (ISCnet). Neste sistema serão lançadas as informações do curso e das disciplinas, as matrículas, as

notas de cada disciplina, a frequência dos discentes, a avaliação do TCC. Também permitirá a emissão do certificado de curso, com os elementos necessários para sua validade, segundo a Resolução CNE/CES nº 1, de 8/6/2007.

17. LOCAL DE REALIZAÇÃO DO CURSO

O curso será realizado nas instalações do Instituto Serzedello Corrêa (ISC), que está localizado no Setor de Clubes Esportivos Sul (SCES), Trecho 3, Polo 8, Lote 3 CEP 70200-003 – Brasília-DF.

Trata-se de complexo arquitetônico com área construída total de 24.552 m² formado por duas edificações. O prédio principal possui 3 andares. No térreo há um Anfiteatro com capacidade para 45 pessoas e duas oficinas para atividades diversificadas com capacidade para 64 participantes. No primeiro pavimento localizam-se 9 salas de aula para um total de 360 alunos, sendo que 3 funcionam também como laboratórios de informática, uma das salas conta com recursos de gravação e transmissão de vídeo, que é integrada com uma sala de conferência, e outra possui espaço duplo com capacidade para até 80 participantes. No segundo piso localiza-se a área administrativa do Instituto.

O segundo prédio abriga o complexo cultural do TCU, composto por Museu, Espaço Cultural, área educativa do Centro Cultural, com capacidade para 60 alunos, e auditório, com capacidade 484 pessoas.

Entre os prédios há uma área central de convívio, composta por duas praças onde podem ser realizadas atividades culturais diversificadas. O conjunto também possui uma biblioteca com espaço coletivo e salas individuais para estudo, espaço de convivência para os alunos, com computadores e acesso à internet por wi-fi, e espaços reservados para um restaurante, uma lanchonete e um café. Há ainda a disponibilidade total de 443 vagas de estacionamento, distribuídas áreas de garagem cobertas e descobertas.

18. ACERVO

18.1. Bibliografia básica do curso.

A bibliografia básica do curso poderá ser constituída de material digital ou de livros e/ou periódicos que podem ser adquiridos pela Biblioteca Ministro Ruben Rosa para acesso aos alunos do curso, ou disponibilizados na comunidade virtual de aprendizagem quando se tratar de conteúdo digital de livre divulgação.

Cada disciplina do curso contará também outras fontes bibliográficas que deverão ser indicadas como leitura recomendada para os alunos no decorrer das aulas. Essa bibliografia deverá ser constituída de preferencialmente de material digital, a ser disponibilizado na comunidade virtual de aprendizagem. Será evitada a reprodução física do material de leitura. Todo material disponibilizado deverá obedecer à legislação de direitos autorais. Livros e/ou periódicos de conteúdo integral que compõem a bibliografia básica das disciplinas podem ser adquiridos pela Biblioteca Ministro Ruben Rosa quando indicados previamente pelos docentes.

18.2. Bibliografia complementar

A bibliografia do curso será ampliada por bibliografia complementar fornecida pelos docentes de cada disciplina. O acervo poderá estar eventualmente disponível na Biblioteca Ministro Ruben Rosa.

19. SISTEMAS DE AVALIAÇÃO E FREQUÊNCIA MÍNIMA

19.1. Frequência mínima

A frequência mínima exigida para a aprovação é de 50% de presença por disciplina e de 75% do total de disciplinas.

19.2. Sistema de Avaliação por disciplina

O aproveitamento acadêmico poderá ser medido por meio de provas, seminários, trabalhos e participação, individuais ou em grupo. O rendimento escolar será aferido por disciplina, abrangendo sempre os aspectos de assiduidade e aprendizagem que será apurada por pontos cumulativos, em

uma escala de 0 (zero) a 10 (dez). Será considerado aprovado o aluno que alcançar rendimento acadêmico mínimo igual ou superior a 6 (seis).

19.3. Avaliação do TCC

A avaliação do trabalho de conclusão do curso será realizada em dois momentos. No início do segundo semestre os alunos deverão defender seu projeto de TCC diante de uma banca unificada de professores, que por banca constituída por, pelo menos, dois professores, sendo um deles o orientador. Os alunos receberão da banca examinadora as menções “aprovado”, “aprovado com restrições”, ou “não aprovado”. Será considerado “aprovado” o aluno que receber essa menção de todos os membros da banca. Será considerado “aprovado com restrições” o discente que receber essa menção de pelo menos um dos membros da banca. Será considerado “não aprovado” aquele aluno que receber esse conceito de todos os membros da banca. No caso de aprovação com restrições, as modificações sugeridas deverão ser efetuadas pelo aluno no prazo máximo de 60 (sessenta) dias e apresentadas ao orientador, que atestará a validade das modificações realizadas, a fim de que possa ser considerado aprovado em caráter definitivo.

20. CERTIFICAÇÃO

Ao discente que obedecer às exigências do item anterior (frequência mínima de 50% por disciplina e 75% no conjunto de disciplinas, nota mínima de 6 por disciplina e de 7 no TCC) será conferido o grau de Especialista.

21. CONTEÚDO PROGRAMÁTICO

21.1. Disciplinas e carga horária

Disciplinas	Carga horária presencial	EaD
As Ferramentas do Cientista de Dados	8h	20h
O processo de descoberta de conhecimento	12h	-

Metodologia Científica	12h	-
Obtendo e preparando dados	32h	-
Inferência estatística	16h	-
Análise exploratória de dados	24h	-
BI e Visualização de dados	24h	-
Modelos de regressão	16h	-
Técnicas de Mineração de Dados	48h	-
Análise de Políticas Públicas a partir de dados oficiais	24h	-
Aprendizagem de máquina	40h	-
Análise de dados espaciais e georeferenciados	16h	
Tópicos especiais em Análise de Dados	16h	-
Implementando produtos baseados em dados	12h	-
Atividades Laboratoriais I	-	18h
Atividades Laboratoriais II	-	22h
Total disciplinas	**304h**	**60h**
Trabalho de conclusão de curso		**74h**

Total geral	**438h**

21.2. Ementas das disciplinas

As Ferramentas do Cientista de Dados (28h)

Objetivos: Nesta disciplina o aluno receberá uma introdução às ferramentas e técnicas mais importantes da "caixa de ferramenta do cientista de dados". Dois pilares principais serão abordados: os bancos de dados relacionais e a linguagem Python, uma das ferramentas de programação mais populares na mineração de dados. Embora esses dois pilares não esgotem o universo de tecnologias úteis à análise de dados, eles são com certeza as ferramentas do dia-a-dia do cientista de dados e, por isso, serão também a base ferramental desse curso. Trata-se também de um módulo de nivelamento. Profissionais com experiência em programação e no uso de bancos de dados talvez não necessitem do conteúdo que será ministrado, mas aqueles com experiência esporádica ou insuficiente nesses temas poderão adquirir a base necessária para atravessar os módulos seguintes. Profissionais sem nenhuma experiência nesses temas precisarão de uma dedicação especial a esse módulo, buscando inclusive ajuda extra na literatura especializada se sentirem dificuldade em acompanhar o conteúdo das aulas.

Estrutura da disciplina:

Parte I –Bancos de Dados Relacionais e a linguagem SQL.

Parte III – Introdução à programação em Python utilizando o Jupiter Notebook.

Modalidade: hibrida – presencial e ensino a distância

O processo de descoberta de conhecimento (12h)

Objetivos: Nesta disciplina introdutória apresentaremos a ciência de dados, seus conceitos básicos, objetivos, possibilidades e limitações. Apresentaremos ainda o macroprocesso de descoberta de conhecimento e as principais metodologias para mineração de dados.

Estrutura da disciplina:

Parte I: Conceituação de descoberta de conhecimento. Perfil e atitude do cientista de dados. Ética e limites no uso de dados. Objetivos, possibilidades e áreas de aplicação da análise de dados. A importância e o potencial da análise de dados na esfera pública.

Parte II: O processo de referência CRISP-DM. As fases, tarefas genéricas e tarefas específicas. O foco no negócio. A busca pelas questões a serem respondidas.

Parte III: Os principais paradigmas da mineração de dados. Estado da arte da ciência de dados. Avanços recentes. Novas tecnologias. Perspectivas para o futuro.

Parte IV: Apresentação de casos de sucesso: classificação, regressão, clusterização, regras de associação, mineração de textos, classificação de imagens.

Modalidade: presencial.

Metodologia Científica (12h)

Objetivos: Conhecer e correlacionar os fundamentos, os métodos e as técnicas de análise presentes na produção do conhecimento científico. Compreender as diversas fases de elaboração e desenvolvimento de pesquisas e trabalhos acadêmicos. Elaborar e desenvolver pesquisas e trabalhos científicos obedecendo às orientações e normas vigentes nas Instituições de Ensino e Pesquisa no Brasil e na Associação Brasileira de Normas Técnicas.

Estrutura da disciplina:
Parte I: Fundamentos da Metodologia Científica. Fases de elaboração e desenvolvimento de pesquisas e trabalhos acadêmicos. Métodos e técnicas de pesquisa.

Parte II: A Comunicação Científica. A organização de texto científico. Normas para Elaboração de Trabalhos Acadêmicos (Normas ABNT).

Modalidade: presencial.

Obtendo e preparando dados (32h)

Ementa: Para analisar dados é preciso antes obtê-los. Esta disciplina tratará das principais formas de obtenção de dados úteis à descoberta de conhecimento. Serão abordadas técnicas de extração, transformação e carga de dados provenientes de diversas fontes de dados, tais como bancos de dados relacionais, internet e APIs. Um enfoque particular será dado às principais fontes de dados governamentais e à infraestrutura de concentração de dados criada pelo TCU, o LabContas. Serão abordadas ainda as principais técnicas de limpeza e preparação de dados. Por meio delas é possível obter dados de melhor qualidade e melhor adequados às etapas subsequentes de exploração e análise.

Estrutura da disciplina:

Parte I: Busca das fontes de dados relevantes. Dados brutos e dados processados. Grau de credibilidade. Dado aberto, sensível e sigiloso. Anonimização. Limites e riscos do uso da análise de dados: indício versus prova.

Parte II: Engenharia de dados. Processo de ETL. Amostragem. Qualidade do dado. Limpeza.

Parte III: Dados estruturados: Bancos de dados relacionais. Consultas SQL. Views. Flat files. LabContas.

Parte IV: Dados semi-estruturados: Arquivos XML, JSON e HTML. Parsing. Expressões regulares. Bancos NoSQL.

Parte V: Dados não estruturados: Dados textuais. Conversão de documentos. Indexação. Codificações de caracteres. Técnicas de Pré-processamento de Textos. OCR. Reconhecimento de fala.

Parte VI: Extração de dados da web (Web scraping). Noções de protocolo HTTP e APIs. Acesso a fontes de dados governamentais.

Parte VII: Criação e aplicação de questionários. Utilização do LimeSurvey.

Modalidade: presencial

Inferência estatística (16h)

Ementa: O objetivo dessa disciplina é apresentar técnicas de proposição de afirmações sobre uma população a partir de evidências fornecidas em uma amostra. A população é supostamente bem maior que o conjunto de dados observados contidos na amostra. Essa generalização inclui o teste de hipóteses e a derivação de proposições a respeito da população, acompanhadas de uma medida de precisão. A inferência estatística difere da estatística descritiva pois esta última se preocupa somente com as propriedades do dado observado enquanto a primeira levanta proposições que dizem respeito à população como um todo. As duas principais escolas de inferência são a inferência frequentista (clássica) e a inferência bayesiana. Existe uma grande variedade de meios para executar a inferência, incluindo a modelagem estatística, estratégias orientadas a dados e o uso explícito de randomização.

Estrutura da disciplina:

Parte I: Fundamentos da Probabilidade. Variáveis aleatórias. Independência. Expectância. Probabilidade condicional. Funções de densidade. Teorema de Bayes.

Parte II: Variância. Média. Desvio padrão. Distribuição Binomial. Distribuição Normal. Distribuição de Poisson.

Parte III: Intervalos de confiança. Teste de hipóteses. Hipótese nula. P-values.

Modalidade: presencial

Análise exploratória de dados (24h)

Ementa: Esta disciplina trata das principais técnicas exploratórias que permitem a descrição e a sumarização dos dados brutos. Estas técnicas são tipicamente utilizadas antes da concepção de modelos preditivos ou inferenciais e servem também de base para modelos estatísticos mais complexos. As técnicas exploratórias são importantes para o teste, aprimoramento e até descarte de hipóteses preliminares que o analista tem a respeito do negócio objeto do estudo. Serão

apresentadas ainda as principais técnicas estatísticas descritivas multivariadas utilizadas na visualização de dados de alta dimensionalidade.

Estrutura da disciplina:

Parte I: Estatística descritiva. Sumarização. Média e desvio padrão. Outliers. Boxplot.

Parte II: Avaliação da relevância das variáveis para fins analíticos. Redução da dimensionalidade. Análise Fatorial. Criação de novas variáveis. Análise de Componentes Principais.

Parte III: Clusterização. K-Means. HCA (Hierarchical Cluster Analysis).

Parte IV: Técnicas de visualização de dados de alta dimensionalidade.

Modalidade: presencial

Modelos de regressão (16h)

Ementa: Esta disciplina trata da análise por regressão, mínimos quadrados e inferência utilizando modelos de regressão. Casos especiais de modelos de regressão, como o ANOVA, ANCOVA e a regressão logística serão tratados também. A disciplina abordará técnicas recentes de seleção de modelos e novos usos dos modelos de regressão.

Estrutura da disciplina:

Parte I: Mínimos quadrados e regressão linear

Parte II: Regressão multivariável e polinomial. Interpretação dos coeficientes. Residuais. Inferência com regressão. Predição. Seleção de modelos.

Parte III: Regressão logística e regressão de Poisson.

Parte IV: Técnicas de aprendizagem de máquina para regressão.

Modalidade: presencial

Business Intelligence e Visualização de dados (24h)

Ementa: Tão importante quanto a descoberta de conhecimentos em grandes bases de dados é a habilidade de explorar e comunicar os resultados e conclusões obtidas. Dados e informações em formato bruto, na forma de tabelas e listas, são de difícil interpretação pelo usuário final, principalmente quando a semântica desses resultados envolve regras de negócio e formatos muito específicos. Diversas tecnologias permitem a construção de modelos de dados voltados à consulta, geração de relatórios ad hoc e interfaces interativas por meio das quais o consumidor do conhecimento gerado pela análise de dados pode dele se apropriar e interpreta-lo dentro do seu próprio contexto de negócio e de atuação. Nessa disciplina abordaremos algumas dessas tecnologias assim como alguns princípios da preparação dos dados para visualização e da construção dessas interfaces.

Parte I: Preparando dados para visualização e comunicação de resultados: agregações, desnormalizações, modelos OLAP.

Parte II: Princípios e boas práticas da visualização de dados.

Parte III: Self-service BI. Construindo painéis e dashboards interativos.

Modalidade: presencial

Técnicas de Mineração de Dados (48h)

Ementa: Esta disciplina e a de Aprendizagem de Máquina cobrem os principais componentes da construção e aplicação de modelos preditivos. A aprendizagem de máquina está revolucionando diversos domínios do conhecimento ao permitir a elaboração de modelos preditivos a partir de quantidades cada vez maiores de dados e dimensões. Esses modelos permitem a compreensão de fenômenos complexos, que são identificáveis somente a partir da análise de grandes e abrangentes séries de eventos. Serão apresentados conceitos básicos, tais como conjuntos de treinamento, validação e de teste, underfitting, overfitting e indicadores de qualidade de um modelo preditivo. Serão abordados os principais modelos e algoritmos de aprendizagem de máquina, tais como as

árvores de decisão, Naive Bayes, Random Forests, Support Vector Machines e Redes Neurais. As disciplinas cobrirão todo o processo de construção de modelos preditivos, incluindo a coleta de dados, criação de novas variáveis (engenharia de requisitos), preparação e treinamento dos modelos e avaliação.

Estrutura da disciplina:

Parte I: Predição, erros e validação cruzada. Matriz de confusão. Underfitting. Overfitting. Comparação de modelos.

Parte II: Seleção do paradigma adequado ao problema tratado. Maldição da dimensionalidade. Seleção de variáveis preditivas. Entropia.

Parte III: Classificação. Árvores de decisão. Random Forests. Bagging. Boosting.

Parte IV: Naive Bayes.

Parte V: Regras de associação. Algoritmo Apriori.

Parte VI: Redes Neurais. Princípios, conceitos e modelos.

Parte VII: Support Vector Machines.

Parte VIII: Ensembles. Máquinas de comitê.

Modalidade: presencial

Análise de Políticas Públicas a partir de dados oficiais (24h)

Ementa: Nesta disciplina será apresentada uma visão holística dos elementos introduzidos nos módulos anteriores de forma pragmática, aplicando-os à análise de políticas públicas de forma ampla e aprofundada. Serão abordados os diversos aspectos envolvidos nesse tipo de análise, tal como a identificação e preparação das bases de dados relevantes, principais bases de dados abertos disponíveis, seleção ou construção de indicadores, aplicação das técnicas adequadas a cada aspecto da análise, construção de modelos preditivos e construção de artefatos de visualização dos

resultados. Serão selecionadas até três políticas públicas relevantes e os alunos, divididos em equipes, realizarão um trabalho prático dirigido seguindo as etapas do macroprocesso de descoberta de conhecimento.

Parte I: Identificando as dimensões relevantes da política pública. Selecionando bases de dados. Criando indicadores.

Parte II: Selecionando questões de análise pela perspectiva do auditor. Escolhendo os paradigmas de análise.

Parte III: Criando modelos preditivos.

Parte IV: Construindo artefatos de consulta e visualização dos resultados.

Modalidade: presencial

Aprendizagem de máquina (40h)

Ementa: Dando continuidade ao tema da mineração de dados esta disciplina abordará mais alguns modelos e conceitos relevantes com foco em técnicas de aprendizagem de máquina.

Estrutura da disciplina:

Parte I: Redes Neurais profundas: Deep Learning.

Parte II: Mineração de textos. Busca e recuperação.

Parte III: Análise de Dados de Redes Sociais. Processamento de Linguagem Natural. Sumarização. Análise de Sentimentos.

Parte IV: Mineração em dados sequenciais. Hidden Markov Models. Conditional Random Fields. Maximum Entropy Markov Models. Recurrent Neural Networks.

Modalidade: presencial

Análise de dados espaciais e georreferenciados (16h)

Ementa: O geoprocessamento permite a combinação de informações cartográficas (mapas, cartas topográficas e plantas) e de sensoriamento remoto a quaisquer outras bases de informações às que se possa associar coordenadas geográficas de latitude e longitude, obtidas explicitamente por meio sistemas de localização (GPS) ou derivadas por meio de processos de geocodificação. Essa combinação adiciona uma poderosa dimensão aos dados tabulares convencionais, permitindo diversas modalidades de análise nas mais diversas aplicações. Indo desde a simples inspeção visual de imagens de sensoriamento remoto em intervalos de tempo escolhidos, que permite a verificação de fatos, medições e comparação entre dados declarados e situação in loco, passando pela classificação automática de padrões simples, tais como ocupação do solo, biomas, ocupação urbana, detecção de mudanças, e indo até sofisticados métodos de análise e busca de padrões de interesse, o geoprocessamento pode ser uma poderosa ferramenta no apoio às ações de controle. Nessa disciplina serão apresentados os principais conceitos e métodos associados ao geoprocessamento, exemplos de aplicação no Controle e noções de uso dos recursos e ferramentas disponíveis.

Estrutura da disciplina:

Parte I: Explorando a dimensão geográfica de dados georeferenciados. Utilizando ferramentas de visualização espacial: Google Earth. Tipos básicos de arquivos geo: vetorial e raster. Sistemas de coordenadas.

Parte II: Geocodificação de endereços e entidades georeferenciadas.

Parte III: Noções de cartografia. Criação de mapas temáticos com sobreposição de camadas.

Parte IV: Obtendo e manipulando imagens de sensoriamento remoto. Provedores de imagens de satélite. Bandas espectrais. Criação de mosaicos. Utilizando imagens de drones.

Parte V: Combinando dados tabulares e dados georeferenciados: junções espaciais.

Modalidade: presencial

Tópicos Especiais em Análise de Dados (16h)

Ementa: Nesta disciplina serão apresentados alguns conceitos disruptivos que estão revolucionando a análise de dados, permitindo que limitações tecnológicas e conceituais sejam vencidas e que quantidades cada vez maiores de dados, dimensões e temas sejam manipuláveis simultaneamente à busca de novos conhecimentos. A disciplina terá a forma de seminários, proferidos por especialistas convidados. Os temas a seguir são sugestões, que podem ser alterados segundo o interesse do grupo e disponibilidade dos especialistas.

Parte I: Noções de Big Data. Hadoop, MapReduce e NoSQL.

Parte II: Cloud Computing. Análise de dados utilizando a computação em nuvem (Machine Learning APIs).

Parte III: Geoprocessamento utilizando a computação em nuvem.

Modalidade: presencial

Implementando produtos baseados em dados (16h)

Ementa: Toda análise baseada em dados bem-sucedida pode se tornar num produto disponibilizado para consumo de um público selecionado, o que exige que ele seja continuamente atualizado de forma automatizada e que possa evoluir. Esta disciplina trata dos diversos aspectos relacionados à criação de produtos baseados em dados. Será dada ênfase na construção de produtos de dados úteis à esfera pública e no seu compartilhamento entre instituições.

Estrutura da disciplina:

Parte I: A avaliação de modelos produzidos pela análise de dados e a integração ao negócio e aos processos de trabalho institucionais. As questões de mineração foram respondidas? Como monitorar a qualidade dos modelos no tempo? Quando um novo ciclo de descoberta de conhecimento é necessário?

Parte II: Automatização de procedimentos de ETL. Criando e mantendo um repositório de dados estável e atualizado. Integrando dados de fontes e instituições diferentes.

Parte III: Desenvolvendo produtos de dados em Python. Plano de manutenção e evolução de produtos baseados em dados. Compartilhamento de código.

Modalidade: presencial

Atividades laboratoriais I (18h)

Ementa: Essa disciplina agrega as atividades práticas extraclasse que serão demandas pelos módulos presenciais do primeiro semestre. Tratam-se de análises feitas em torno de casos selecionados, onde alguma técnica específica será utilizada na manipulação de dados relevantes com acompanhamento de tutores. Essas atividades buscam adicionar uma dimensão prática e aplicada ao curso de especialização, permitindo ao estudante a aquisição de habilidades, no uso das metodologias e tecnologias estudadas, que serão úteis na sua atividade profissional enquanto egresso.

Modalidade: ensino a distância

Atividades laboratoriais II (26h)

Ementa: Essa disciplina agrega as atividades práticas extraclasse que serão demandas pelos módulos presenciais do segundo semestre.

Modalidade: ensino a distância.