

UNIVERSIDADE FEDERAL DE JUIZ DE FORA  
INSTITUTO DE CIÊNCIAS EXATAS  
BACHARELADO EM CIÊNCIA DA COMPUTAÇÃO

# Aplicação de técnicas de aprendizado de máquina como apoio para a construção de um sistema de recomendação de eventos

**Hygor Xavier Araújo**

JUIZ DE FORA  
NOVEMBRO, 2017

# Aplicação de técnicas de aprendizado de máquina como apoio para a construção de um sistema de recomendação de eventos

HYGOR XAVIER ARAÚJO

Universidade Federal de Juiz de Fora  
Instituto de Ciências Exatas  
Departamento de Ciência da Computação  
Bacharelado em Ciência da Computação

Orientador: Saulo Moraes Villela  
Coorientador: Jairo Francisco de Souza

JUIZ DE FORA  
NOVEMBRO, 2017

APLICAÇÃO DE TÉCNICAS DE APRENDIZADO DE MÁQUINA  
COMO APOIO PARA A CONSTRUÇÃO DE UM SISTEMA DE  
RECOMENDAÇÃO DE EVENTOS

Hygor Xavier Araújo

MONOGRAFIA SUBMETIDA AO CORPO DOCENTE DO INSTITUTO DE CIÊNCIAS  
EXATAS DA UNIVERSIDADE FEDERAL DE JUIZ DE FORA, COMO PARTE INTE-  
GRANTE DOS REQUISITOS NECESSÁRIOS PARA A OBTENÇÃO DO GRAU DE  
BACHAREL EM CIÊNCIA DA COMPUTAÇÃO.

Aprovada por:

Saulo Moraes Villela  
Doutor em Engenharia de Sistemas e Computação

Jairo Francisco de Souza  
Doutor em Informática

Heder Soares Bernardino  
Doutor em Modelagem Computacional

Victor Ströele de Andrade Menezes  
Doutor em Engenharia de Sistemas e Computação

JUIZ DE FORA  
29 DE NOVEMBRO, 2017

*Aos meus pais, familiares e amigos.*

## Resumo

Atualmente existe uma grande facilidade em divulgar eventos sociais na internet. Por causa disso, existe um grande volume de dados disponível que dificulta a uma pessoa interessada em encontrar eventos achar aqueles que são realmente de seu interesse. Desta forma, sistemas de recomendação surgem como uma possível solução para este problema ao lidar com a sobrecarga de informação e apresentar ao usuário informações que mais provavelmente lhe interessam. Este trabalho faz um estudo de uma aplicação de uma técnica híbrida para recomendação de eventos a usuários de um sistema.

**Palavras-chave:** sistemas de recomendação, sistemas móveis, aprendizado de máquina.

## Abstract

Currently there is a great facility in publicizing social events on the internet. As a consequence, there is a great amount of data available that makes it difficult for a person interested in finding events to find the ones that are really of interest to them. In this way, recommendation systems emerge as a possible solution to this problem when dealing with information overload and presenting the user with information that is most likely to interest him. This work presents a study of an application of a hybrid technique for recommending events to users of a system.

**Keywords:** recommendation systems, mobile systems, machine learning.

## Agradecimentos

Aos meus pais pelo amor, sustento e apoio incondicional. A minha irmã, por toda ajuda e por ser um exemplo de dedicação. A minha companheira e amiga Ana Paula, por todo seu apoio e presença fundamental.

A todos os meus familiares, amigos e colegas, pelo encorajamento e apoio.

Aos meus orientadores, Saulo e Jairo, por todo seu apoio, dedicação e paciência para que este trabalho fosse realizado.

Aos professores do Departamento de Ciência da Computação pelos seus ensinamentos e aos funcionários do curso, que durante esses anos, contribuíram de algum modo para o nosso enriquecimento pessoal e profissional.

# Conteúdo

<b>Lista de Figuras</b>	<b>7</b>
<b>Lista de Tabelas</b>	<b>8</b>
<b>Lista de Abreviações</b>	<b>9</b>
<b>1 Introdução</b>	<b>10</b>
1.1 Apresentação do tema . . . . .	10
1.2 Problema . . . . .	10
1.3 Justificativa . . . . .	11
1.4 Objetivos . . . . .	12
1.5 Estrutura do trabalho . . . . .	12
<b>2 Fundamentação teórica</b>	<b>14</b>
2.1 Sistemas de recomendação . . . . .	14
2.1.1 Recomendação baseada em conteúdo . . . . .	14
2.1.2 Recomendação colaborativa . . . . .	15
2.1.3 Recomendação móvel e baseada em contexto . . . . .	15
2.1.4 Abordagem híbrida de recomendação . . . . .	16
2.2 Aprendizagem de máquina . . . . .	16
2.2.1 <i>K-Means</i> . . . . .	17
2.2.2 <i>K-Nearest-Neighbors</i> . . . . .	18
2.2.3 <i>Nearest Centroid</i> . . . . .	18
2.2.4 <i>Support Vector Machines</i> . . . . .	19
2.2.5 <i>K-Fold Cross-validation</i> . . . . .	20
<b>3 Sistema de divulgação de eventos</b>	<b>21</b>
3.1 Componentes . . . . .	21
3.2 Requisitos . . . . .	22
3.2.1 Funcionais . . . . .	22
3.2.2 Não funcionais . . . . .	24
3.3 Casos de uso . . . . .	25
3.4 Base de dados . . . . .	28
3.5 Protótipo . . . . .	29
3.5.1 Protótipo web . . . . .	30
3.5.2 Protótipo Android . . . . .	35
<b>4 Abordagem desenvolvida</b>	<b>38</b>
4.1 Dados utilizados . . . . .	38
4.1.1 Geração de dados . . . . .	38
4.2 Agrupamento . . . . .	40
4.3 Avaliação da classificação . . . . .	43
4.4 Seletor para recomendação de eventos . . . . .	45
<b>5 Análise de resultados</b>	<b>47</b>



<b>6</b>	<b>Considerações finais</b>	<b>49</b>
6.1	Trabalhos futuros . . . . .	49
	<b>Bibliografia</b>	<b>51</b>
<b>A</b>	<b>Representação dos clusters de eventos e participantes</b>	<b>52</b>
A.1	Eventos . . . . .	52
A.2	Participantes . . . . .	52
<b>B</b>	<b>Representação dos dados utilizados para teste</b>	<b>68</b>
<b>C</b>	<b>Diagrama ER do banco de dados</b>	<b>71</b>

## Lista de Figuras

3.1	Relação entre componentes do sistema . . . . .	22
3.2	Formulário para criação de uma conta de promotor de eventos no sistema web. . . . .	30
3.3	Tela para acesso do usuário promotor de eventos ao sistema. . . . .	31
3.4	Lista de eventos previamente cadastrados pelo usuário. . . . .	31
3.5	Visualização e edição de dados de um eventos. . . . .	32
3.6	Passo 1 do formulário para criação de um novo evento. . . . .	33
3.7	Passo 2 do formulário para criação de um novo evento. . . . .	33
3.8	Passo 3 do formulário para criação de um novo evento. . . . .	34
3.9	Passo 4 do formulário para criação de um novo evento. . . . .	34
3.10	Tela de acesso ao sistema para o usuário participante e menu de navegação interno . . . . .	35
3.11	Cadastro de usuário participante no sistema . . . . .	35
3.12	Exibição do perfil do usuário com seus dados e também o relacionamento com outros participantes . . . . .	36
3.13	Busca por eventos, exibição de detalhes de um evento e avaliação do mesmo	36
3.14	Tela de busca por participantes . . . . .	37
4.1	Acurácia dos classificadores de eventos . . . . .	44
4.2	Acurácia dos classificadores de participantes . . . . .	44
4.3	Exemplos de matrizes de confusão do K-NN, NC e SVM para eventos . . .	45
4.4	Exemplo de matrizes de confusão do K-NN, NC e SVM para participantes	45
A.1	Boxplots de features para cada cluster de eventos . . . . .	53
A.2	Representação do cluster 0 de eventos . . . . .	54
A.3	Representação do cluster 1 de eventos . . . . .	55
A.4	Representação do cluster 2 de eventos . . . . .	56
A.5	Representação do cluster 3 de eventos . . . . .	57
A.6	Representação do cluster 4 de eventos . . . . .	58
A.7	Representação do cluster 5 de eventos . . . . .	59
A.8	Representação do cluster 6 de eventos . . . . .	60
A.9	Representação do cluster 7 de eventos . . . . .	61
A.10	Boxplots de features para cada cluster de participantes . . . . .	62
A.11	Representação do cluster 0 de participantes . . . . .	63
A.12	Representação do cluster 1 de participantes . . . . .	64
A.13	Representação do cluster 2 de participantes . . . . .	65
A.14	Representação do cluster 3 de participantes . . . . .	66
A.15	Representação do cluster 5 de participantes . . . . .	67
C.1	Diagrama Entidade Relacionamento do banco de dados desenvolvido para o sistema . . . . .	71
C.2	Recorte do diagrama com foco na tabela evento. . . . .	72
C.3	Recorte do diagrama com foco em tabelas associadas a tabela evento. . . .	73
C.4	Recorte do diagrama com foco na tabela participante. . . . .	74

## Lista de Tabelas

3.1	Requisitos funcionais do sistema web para o promotor de eventos . . . . .	23
3.2	Requisitos funcionais do sistema . . . . .	24
3.3	Requisitos não funcionais do sistema . . . . .	25
4.1	Features de eventos . . . . .	40
4.2	Features de participantes . . . . .	41
4.3	Pesos para cada interação, utilizado na clusterização do participante . . . . .	42
4.4	Exemplo de separação dos dados de eventos em conjunto de treinamento e validação estratificados . . . . .	43
4.5	Exemplo de separação dos dados de participantes em conjunto de treinamento e validação estratificados . . . . .	43
4.6	Exemplo de grafo gerado. <i>Clusters</i> de participantes como linhas e de eventos como colunas. . . . .	46
4.7	Exemplo de grafo gerado com pesos normalizados no intervalo $[-1, 1]$ . <i>Clusters</i> de participantes como linhas e de eventos como colunas. . . . .	46
5.1	Distâncias entre centroides de clusters de eventos . . . . .	47
5.2	Distâncias entre centroides de clusters de participantes . . . . .	48
5.3	Eventos com classificações divergentes . . . . .	48
5.4	Participantes com classificações divergentes . . . . .	48
A.1	Número de eventos em cada cluster . . . . .	52
A.2	Número de participantes em cada cluster . . . . .	52
B.1	Dados de teste de eventos já classificados . . . . .	68
B.2	Dados de teste de participantes já classificados - parte 1 . . . . .	69
B.3	Dados de teste de participantes já classificados - parte 2 . . . . .	70

## Lista de Abreviações

K-NN *K-Nearest-Neighbors*

NC *Neares Centroid*

SVM *Support Vector Machine*

# 1 Introdução

## 1.1 Apresentação do tema

A internet se tornou um grande e importante canal de comunicação de eventos sociais. Não somente em canais de notícias e entretenimento, a divulgação de eventos ocorre também em redes sociais que não tem o mesmo como objetivo, como Facebook<sup>1</sup>, e em outras com o objetivo específico dessa divulgação, como Meetup<sup>2</sup>. Além disso, também houve o surgimento de plataformas para divulgação e venda de ingressos para eventos, como Ingresso Certo<sup>3</sup> e Sympla<sup>4</sup>. Estes sistemas permitem a exposição de informações sobre diversos eventos a um grande número de pessoas. A facilidade de divulgação destes eventos através da internet, no entanto, gera um grande volume de informação disponível para os usuários que, com uma grande variedade de eventos para escolha, possui uma difícil tarefa de encontrar eventos que sejam de seu interesse.

Sistemas de recomendação surgem como uma possível solução para este problema, permitindo que seja oferecido ao usuário eventos que possam ter uma maior relevância para ele. Através da aplicação de técnicas que permitem utilizar o perfil do usuário e seu contexto é possível oferecer ao mesmo recomendações personalizadas de eventos que sejam mais compatíveis com seu gosto.

## 1.2 Problema

Apesar da internet possibilitar uma grande divulgação de informações referentes a eventos, se tornou difícil, para o usuário, a tarefa de selecionar quais são de seu interesse devido ao grande volume de eventos disponíveis para sua escolha. A utilização de sistemas de recomendação no auxílio desta tarefa acontece de forma natural considerando sua aplicação em diversas outras áreas, como, para a recomendação de livros, filmes, músicas,

---

<sup>1</sup><http://facebook.com>

<sup>2</sup><http://meetup.com>

<sup>3</sup><http://ingressocerto.com>

<sup>4</sup><http://sympla.com.br>

notícias, artigos e produtos em geral.

Em sistemas de recomendação existe o problema da partida fria, em que o sistema não é capaz de realizar recomendações pela falta de dados relacionados a um usuário ou item. Para a recomendação de eventos já que estes são geralmente divulgados por um curto tempo e não possuem informação histórica de participação, por acontecer em um momento futuro a sua divulgação (MACEDO; MARINHO; SANTOS, 2015), ou seja, não existem dados de avaliação dos usuários sobre o item a ser recomendado ampliando o problema da partida fria.

### 1.3 Justificativa

Como já mencionado, atualmente existem diversas plataformas que têm como seu produto eventos, sejam eles profissionais, esportivos, artísticos ou para lazer. Estas plataformas possuem como objetivo a divulgação dos eventos para atrair participantes ou também a realização de vendas de ingressos para participação nos mesmos. Pelo grande volume de eventos cadastrados e a diversidade dos mesmos, surge a necessidade de direcionar seu *marketing* de um evento para os usuários certos, aqueles que possuem uma maior chance de terem interesse pelo evento.

Considerando esta necessidade, tem-se como uma possível solução o uso de um sistema de recomendação, já que estes mecanismos têm como objetivo oferecer itens que sejam de maior interesse a um usuário. Tomando os itens recomendados como eventos e os usuários como possíveis participantes dos mesmos, tem-se então que categorizar os usuários, a partir de seus dados disponíveis, de forma que seja possível identificar quais eventos na plataforma são de seu interesse.

Uma forma de identificação do usuário seria através de dados coletados através de seus dispositivos móveis (*smartphone*), que com o grande crescimento da utilização por todo o mundo, sua grande presença no dia a dia das pessoas e capacidade de captar informações referentes ao seu usuário (localização, agenda, clima etc.), torna possível o uso de informações referentes ao contexto de cada um na construção de sistemas inteligentes e personalizados.

Desta forma, esse trabalho propõe a criação de uma aplicação para dispositi-

vos móveis para divulgação de eventos que utilize um mecanismo de recomendação dos mesmos, tendo como objetivo explorar e propor técnicas de aprendizado de máquina e mineração de dados no desenvolvimento de um sistema que ofereça aos usuários recomendações que melhor se enquadrem em seus interesses.

As técnicas escolhidas para utilização foram o *K-Nearest-Neighbors*, *Nearest Centroid*, *Support Vector Machine*, *K-Means* e *Stratified K-Fold Cross Validation*. Com base nos experimentos realizados em (MACEDO; MARINHO, 2014) escolheu-se utilizar inicialmente o KNN e como referência de comparação entre métodos baseados em distância o NC. O *Stratified K-Fold Cross Validation* foi utilizado por causa do desbalanceamento das classes na base de dados. Já o *K-Means* e o SVM foram escolhidos pela sua disponibilidade no pacote de ferramentas utilizado *SciPy* (JONES et al., 2001–2017).

## 1.4 Objetivos

Esse trabalho tem como objetivo demonstrar a aplicação de técnicas de aprendizado de máquina para auxiliar o processo de recomendação em um sistema de divulgação de eventos. Além disso, realizar uma análise das técnicas utilizadas e avaliar os resultados obtidos.

## 1.5 Estrutura do trabalho

Esse trabalho é composto por seis capítulos. Seguido deste capítulo introdutório, o segundo capítulo apresenta a fundamentação teórica utilizada para o desenvolvimento deste trabalho, quais as áreas relacionadas e técnicas utilizadas.

No terceiro capítulo é apresentado o sistema de divulgação de eventos que foi idealizado e planejado. Aqui também estão presentes toda a documentação desenvolvida para o sistema, além de uma descrição da base de dados e um protótipo do mesmo.

No quarto capítulo é descrito como foi desenvolvida toda a parte relacionada às recomendações do sistema e logo a seguir, no capítulo cinco, são apresentadas as análises e resultados obtidos.

Ao final, no capítulo seis é feita uma conclusão do trabalho apresentando algumas

---

considerações finais e indicações de trabalhos futuros para a evolução do trabalho.



## 2 Fundamentação teórica

Esse capítulo apresenta a base de conceitos, métodos e técnicas utilizados para o desenvolvimento do trabalho.

### 2.1 Sistemas de recomendação

Sistemas de recomendação vem se tornando uma parte intrínseca do nosso dia a dia, desde recomendações de livros, músicas e notícias, já se tornou comum, e até mesmo esperado, encontrar recomendações de itens em quase qualquer site ou sistema acessado.

Este tipo de ferramenta oferece ao usuário de um sistema, recomendações que se adequam ao seu interesse. Para descobrir qual o gosto pessoal de um usuário existem abordagens explícitas, através de questionários, por exemplo, e implícitas, onde o perfil do usuário é aprendido através de suas interações com o sistema ao passar do tempo (ADOMAVICIUS; TUZHILIN, 2005). É uma área multidisciplinar, que envolve diversos campos como inteligência artificial, interação humano-computador, mineração de dados, recuperação da informação, entre outros. Possui diversas aplicações práticas, principalmente no auxílio do usuário ao lidar com sobrecarga de informação e ao fornecer recomendações personalizadas. A seguir são definidas classificações utilizadas na literatura para sistemas de recomendação. Estas classificações são definidas com base na abordagem utilizada para a geração das recomendações pelo sistema.

#### 2.1.1 Recomendação baseada em conteúdo

Nesse tipo de abordagem de recomendação o usuário irá receber recomendações de itens similares a outros que ele tenha demonstrado preferência no passado. Para avaliar se um novo item deve ser recomendado a um usuário é feita uma estimativa da possível avaliação do usuário sobre o item baseado na similaridade entre o novo item e todos os itens já avaliados pelo usuário.

Existem problemas conhecidos ao se utilizar esse tipo de abordagem para realizar

recomendações. Um destes problemas está relacionado à escolha das características que definem o item que será recomendado. Uma escolha ruim pode, por exemplo, fazer com que dois itens sejam considerados similares (ou iguais), levando a erros no processo de recomendação. Outra limitação bem conhecida desse tipo de recomendação é o problema do novo usuário no sistema, que possuiria poucas avaliações (ou nenhuma) dos itens do sistema, tornando difícil a avaliação de suas preferências pelo mecanismo de recomendação, este problema é conhecido como problema da partida fria.

### 2.1.2 Recomendação colaborativa

Diferente do método de recomendação baseada em conteúdo, no método de recomendação colaborativa o usuário irá receber recomendações de acordo com a avaliação de outros usuários, que possuam preferências similares às suas, sobre itens ainda não avaliados por este.

Assim como o método de recomendação baseada em conteúdo, a recomendação colaborativa também sofre do mesmo problema do novo usuário onde o sistema ainda não possui informação suficiente para aprender as preferências do usuário. Além do problema do novo usuário, esse método também é afetado pelo problema do novo item, pela sua dependência de avaliações dos usuários sobre os itens faz com que um item sem avaliações não possa ser recomendado.

### 2.1.3 Recomendação móvel e baseada em contexto

O uso de dispositivos móveis pelos usuários permite explorar um conhecimento adicional sobre o contexto do usuário (local, tempo, clima, etc.), permitindo utilizar estes dados para fornecer serviços melhores e mais personalizáveis (RICCI, 2011).

Grande parte das abordagens existentes de recomendações focam em quais são os itens mais relevantes para o usuário, considerando apenas o espaço Usuário x Item, sem levar em consideração seu contexto (tempo, localização, clima, companhia de outras pessoas, agenda pessoal). A importância deste tipo de informação para a geração de melhores recomendações já é reconhecida (ADOMAVICIUS; TUZHILIN, 2011).

Aplicando esse conceito para melhoria dos sistemas tradicionais (Usuário x Item),

iniciou-se a pesquisa em sistemas de recomendação sensíveis ao contexto, que oferece recomendações ao usuário considerando o espaço Usuário x Item x Contexto.

Um problema deste método é como utilizar o alto volume de dados proveniente de dispositivos móveis de forma a acrescentar informações relevantes que podem ser utilizadas para a recomendação.

### 2.1.4 Abordagem híbrida de recomendação

Esse método combina duas ou mais abordagens de recomendação. O objetivo principal deste método é prover uma forma de que um método possa solucionar as limitações de outro.

Adomavicius e Tuzhilin (2005) definem algumas formas de combinar métodos baseados em conteúdo e colaborativos que podem ser resumidos de forma geral em: implementar os dois métodos separadamente e combinar suas predições; ou incorporar características de um método em outro e, por último, construir um modelo unificado que incorpora características de ambos os métodos.

## 2.2 Aprendizagem de máquina

Para a realização das recomendações em um sistema de recomendação diversas técnicas podem ser empregadas, algumas comumente utilizadas são: TF-IDF, clusterização, classificadores, árvore de decisão, redes neurais artificiais, regressão, modelos probabilísticos, vizinho mais próximo, entre outros (ADOMAVICIUS; TUZHILIN, 2005). Muitas destas técnicas estão relacionadas a aprendizagem de máquina, uma sub-área da Inteligência Artificial com o objetivo de permitir que uma máquina seja capaz de aprender e se adaptar de acordo com os dados disponíveis para ela.

Dentro da área de aprendizado de máquina existem várias formas de aprendizado, duas delas que são relevantes para este trabalho são o aprendizado supervisionado e o não supervisionado. Nos métodos de aprendizado supervisionado é utilizado um conjunto de dados chamado conjunto de treinamento que possui todos os seus dados classificados com rótulos que se referem a classes relacionadas ao problema. Estes dados rotulados são então

utilizados para o treinamento de um modelo capaz de prever um rótulo de um novo dado. No aprendizado não supervisionado também é utilizado um conjunto de treinamento, no entanto, os dados não possuem um rótulo e sua distribuição no espaço do problema é que é utilizada para realizar a construção do modelo e fazer previsões sobre dados novos.

Atualmente, a aplicação de técnicas de aprendizado de máquina tem sido algo muito visado pelo mercado devido ao seu grande impacto ao trazer “inteligência” para um sistema e oferecer aos seus usuários uma experiência mais personalizada e pessoal. Além disso, também permite a descoberta de novos conhecimentos que não são tão evidentes quando observados apenas de forma superficial. Com isso, houve um aumento nas pesquisas nesta área e em áreas relacionadas, tanto na academia quanto na indústria.

A seguir são descritas as técnicas de aprendizado de máquina utilizadas nesse trabalho.

### 2.2.1 *K-Means*

O algoritmo *K-means* é um método utilizado em aprendizado não supervisionado que particiona um conjunto de dados em agrupamentos de forma que amostras semelhantes pertençam ao mesmo agrupamento. Este particionamento é atingido quando o algoritmo tenta iterativamente encontrar partições que minimizem o erro quadrático (minimizar variação dentro de um agrupamento) para o número de agrupamentos especificados (FACELI K., 2011).

A entrada do algoritmo é um conjunto de dados  $X$  e o número  $k$  de agrupamentos, a saída é a partição de  $X$  em  $k$  grupos. No começo os centroides são escolhidos aleatoriamente para cada um dos  $k$  grupos, então enquanto houver alterações na associação de uma amostra a um cluster é feito o cálculo da distância de cada amostra a cada centroide, fazendo a associação da amostra com o centroide mais próximo e, por fim, recalculando o centroide do cluster antes de verificar se houve alguma alteração. Em sua formulação padrão a distância Euclidiana é utilizada.

Algumas limitações deste algoritmo são sua sensibilidade a definição inicial dos centroides, que além de ter que definir sua quantidade previamente também pode fazer o algoritmo convergir para uma solução de ótimo local, e a sua medida de distância que,

por exemplo, ao utilizar a distância Euclidiana pode limitar as amostras utilizadas a pertencerem a este espaço.

### 2.2.2 *K-Nearest-Neighbors*

O *K-Nearest-Neighbors* ou K-NN, é um método utilizado em aprendizado supervisionado para classificação e regressão que utiliza a distância entre os dados para realizar uma predição, assumindo uma hipótese de que dados similares estarão organizados em uma mesma região do espaço de entrada e dados não similares localizam-se a distâncias maiores (FACELI K., 2011). Para o cálculo da distância é usualmente utilizado a distância euclidiana, dada pela equação 2.1. Na equação,  $\mathbf{x}_i$  e  $\mathbf{x}_j$  são vetores em um espaço de dimensão  $d$ .

$$d(\mathbf{x}^i, \mathbf{x}^j) = \sum_{l=1}^d \sqrt{(\mathbf{x}_l^i - \mathbf{x}_l^j)^2} \quad (2.1)$$

O algoritmo K-NN recebe como entrada o número  $k$  de vizinhos a serem usados e um conjunto de dados com rótulos já conhecidos e, na fase de treinamento, memoriza estes dados. Para classificar um novo dado sem rótulo é calculada a distância entre a nova amostra e todos os dados do conjunto de treinamento são então selecionadas as  $k$  amostras mais próximas e cada uma vota com seu próprio rótulo qual o rótulo da nova amostra.

Este método possui alguns problemas semelhantes ao *K-Means*, como a determinação do valor de  $k$  de forma manual e a influência da distância no seu desempenho. Algumas limitações relacionadas a distância são, por exemplo, a pressuposição de que os atributos são numéricos, no entanto, em problemas reais existem muitos atributos categóricos, e a escala utilizada por cada atributo, que pode influenciar a medida de distância.

### 2.2.3 *Nearest Centroid*

O *Nearest Centroid* é um método utilizado em aprendizado supervisionado para classificar novas amostras com base no centroide mais próximo dos dados de treinamento utilizados.

O algoritmo recebe como entrada um conjunto de dados com rótulos conhecidos, para cada rótulo (classe) é encontrado o centroide da mesma através do cálculo da média das amostras de cada classe. Para prever a classe de uma nova amostra é feito o cálculo da distância (usualmente euclidiana) da nova amostra para os centroides encontrados durante a fase de treinamento do algoritmo, a amostra então recebe o rótulo do centroide mais próximo a ela.

### 2.2.4 *Support Vector Machines*

*Support Vector Machines* é um modelo de aprendizado supervisionado utilizado para classificação e regressão. O objetivo do algoritmo é encontrar uma hipótese  $h$  que separe os dados de treinamento com a margem máxima, sendo a margem definida pelos vetores de suporte. Para o caso linear binário de classificação, pode-se definir a hipótese como em (FACELI K., 2011):

$$h(\mathbf{x}) = \mathbf{w} \cdot \mathbf{x} + b \quad (2.2)$$

Onde  $\mathbf{x}$  é um vetor que representa um dado no espaço,  $\mathbf{w}$  é o vetor normal ao hiperplano e  $b$  é a distância do hiperplano a origem. E usar uma função sinal para obter o rótulo de uma amostra, que para o caso binário podem ser definidos como positivo (+1) e negativo (-1).

$$\text{sign}(h(\mathbf{x})) = \begin{cases} +1 & \text{se } \mathbf{w} \cdot \mathbf{x} + b > 0 \\ -1 & \text{se } \mathbf{w} \cdot \mathbf{x} + b < 0 \end{cases} \quad (2.3)$$

Com algumas manipulações algébricas é possível concluir que o aumento da margem está relacionado a minimização da norma do vetor  $\mathbf{w}$  e o problema pode ser definido na forma (FACELI K., 2011):

$$\begin{aligned} \min_{\mathbf{w}, b} & \frac{1}{2} \|\mathbf{w}\|^2 \\ \text{s.a. } & y_i(\mathbf{w} \cdot \mathbf{x} + b) - 1 \geq 0 \quad \forall i = 1, \dots, n \end{aligned} \quad (2.4)$$

Para o caso multiclasse há dois tipos de abordagens: o um contra todos e o um contra um. No um contra todos é feito o treinamento de um classificador para cada classe,

considerando a classe em que está sendo treinado como a positiva e todas as outras como negativa. No um contra um o treinamento é feito para cada par de classes, sendo mais custoso computacionalmente.

### 2.2.5 *K-Fold Cross-validation*

Para avaliação de modelos gerados por algoritmos de aprendizagem é comum a utilização de um conjunto de dados (com rótulos conhecidos) que não foi utilizado durante a fase de treinamento para avaliação do desempenho do modelo (FACELI K., 2011).

Um dos métodos utilizados para essa avaliação é o *K-Fold Cross-validation*. Neste método o conjunto de dados é dividido em  $k$  subconjuntos de tamanhos iguais. A cada iteração um destes subconjuntos é separado como dados de teste e os outros  $k - 1$  são utilizados para o treinamento. Com a criação do modelo o subconjunto de teste é então utilizado para medir a acurácia do mesmo, fazendo a alternância de forma circular do subconjunto de testes e repetindo este processo  $k$  vezes, é feito o cálculo da média da acurácia do modelo.

Uma variação deste método é chamada de *Stratified K-Fold Cross-validation*. Nesta variação a divisão dos dados em subconjuntos é feita de forma a manter a proporção das classes em cada subconjunto. Por exemplo, se  $k = 3$ , um conjunto de dados pertencentes as classes  $y = \{0, 1\}$ , sendo 40% da classe 0 e 60% da classe 1, cada um dos 3 subconjunto irá possuir também 40% dos dados na classe 0 e 60% na classe 1. Este tipo de avaliação é importante quando há um desbalanceamento das classes na base de dados.

## 3 Sistema de divulgação de eventos

Para a elaboração desse trabalho foi realizado um planejamento para o desenvolvimento de um sistema de divulgação de eventos. Foi determinado que devem existir dois tipos de usuários no sistema, um usuário do tipo promotor de eventos e outro usuário do tipo participante de eventos. Além da divulgação, o sistema também possui a capacidade de realizar recomendações para os usuários, que são os participantes de eventos. A seguir são descritas as documentações criadas durante a análise para o desenvolvimento do sistema.

### 3.1 Componentes

A figura 3.1 mostra um esboço da relação entre os componentes do sistema, foi proposta a utilização de um modelo cliente-servidor, onde a comunicação entre as aplicações cliente e o servidor é feita através de um *WebService RESTful*. Uma outra possibilidade, além do que é exibido na figura, seria separar o *WebService* e o banco de dados em servidores distintos.

- **Cliente Web:** utilizado pelo usuário promotor de eventos. O usuário promotor será capaz, através desta aplicação web, de criar e gerenciar seus eventos que ficarão disponíveis para os usuários participantes. A aplicação web não chegou a ser implementada, no entanto, como exemplo de tecnologia que poderia ser utilizada para sua construção podemos citar a biblioteca ReactJS (REACTJS..., 2017).
- **Cliente Mobile:** utilizado pelo usuário participante de eventos. Inicialmente planejado para plataforma Android (ANDROID..., 2017), permitirá um participante visualizar os eventos criados pelos promotores, avaliar e receber recomendações de eventos.
- ***WebService RESTful*:** responsável pela comunicação entre as aplicações cliente e o servidor. Define a interface utilizada para manipulação e acesso aos dados do sistema utilizando o padrão REST (*Representational State Transfer*). As requisições



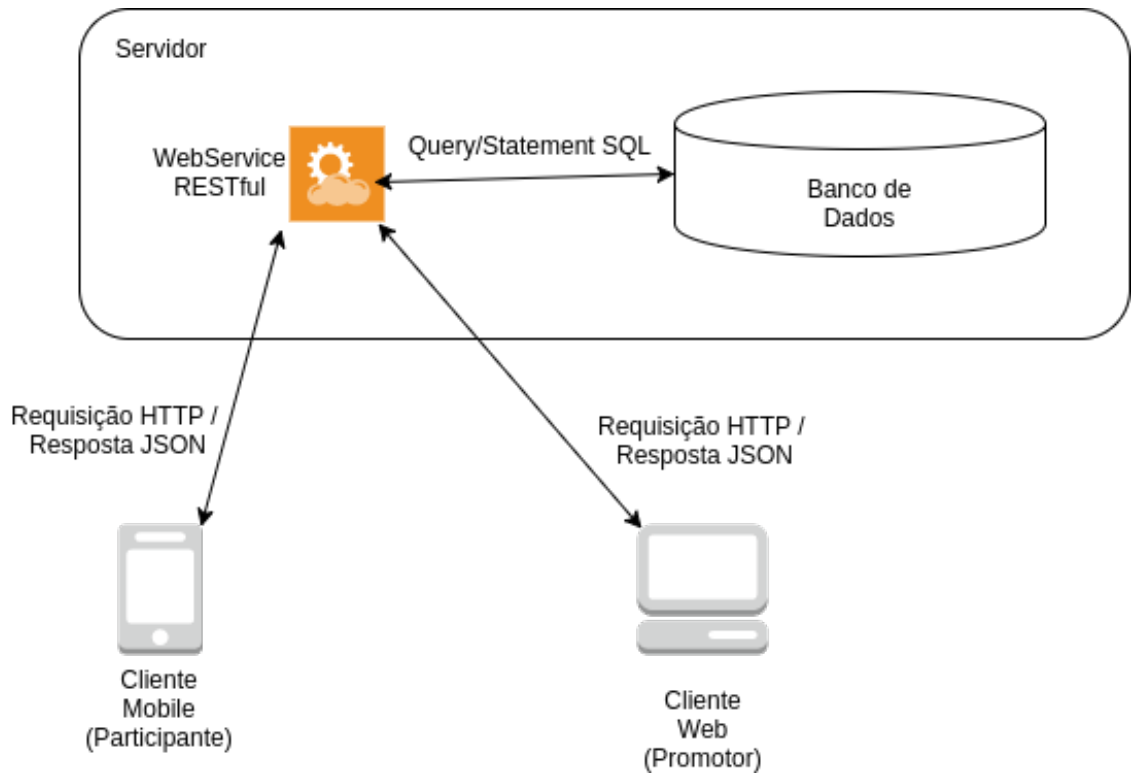


Figura 3.1: Relação entre componentes do sistema

e respostas são feitas utilizando o protocolo HTTP e os recursos são transferidos no formato JSON (*JavaScript Object Notation*). Este *WebService* foi implementado utilizando o *framework SpringBoot* (SPRINGBOOT..., 2017).

- Banco de dados: utilizado para persistência dos dados de todo sistema, será discutido com mais detalhes posteriormente.

## 3.2 Requisitos

Após realizada uma análise sobre as necessidades do projeto, foi elaborado um documento de requisitos do mesmo para descrever, de maneira concisa e clara, as características e funcionalidades esperadas do sistema.

### 3.2.1 Funcionais

Um requisito funcional é utilizado para definir uma função de um sistema ou parte dele. Ele é o conjunto de entradas, o comportamento e a saída desejada da funcionalidade.

<b>Sistema Web</b>	
RF01	Como um promotor de eventos, eu gostaria de realizar o cadastro dos meus eventos.
RF02	Como um promotor de eventos, eu espero que o sistema persista os dados dos eventos previamente cadastrados.
RF03	Como um promotor de eventos, eu gostaria de visualizar uma lista contendo todos os eventos previamente cadastrados por mim.
RF04	Como um promotor de eventos, eu gostaria de editar os dados de um evento previamente cadastrado por mim.
RF05	Como um promotor de eventos, eu gostaria de alterar o <i>status</i> de um evento, informando que o mesmo já foi realizado ou cancelado.
RF06	Como um promotor de eventos, eu gostaria de realizar login no sistema para que eu possa realizar o gerenciamento dos meus eventos.
RF07	Como um promotor de eventos, eu gostaria que existam filtros na lista de eventos que me permitam filtrar por: <i>status</i> , data, classificação e nome.
RF08	Como um promotor de eventos, eu gostaria de poder me cadastrar na plataforma e ter acesso para que eu possa divulgar os meus eventos.
RF09	Como um promotor de eventos, eu gostaria de visualizar os detalhes de um determinado evento.

Tabela 3.1: Requisitos funcionais do sistema web para o promotor de eventos

<b>Sistema Android</b>	
RF01	Como um participante de eventos, eu gostaria de poder me cadastrar no sistema para ter acesso a informações de eventos.
RF02	Como um participante de eventos, eu gostaria de ter acesso a uma lista de eventos pertencentes a diversos promotores de eventos.
RF03	Como um participante de eventos, eu gostaria de poder editar as informações inseridas no ato do meu cadastro.
RF04	Como um participante de eventos, eu gostaria de declarar interesse em um determinado evento.
RF05	Como um participante de eventos, eu gostaria de vincular outros participantes a minha rede de contatos.
RF06	Como um participante de eventos, eu gostaria que existam filtros na lista de eventos que me permitam filtrar por: <i>status</i> , data, classificação e nome.
RF07	Como um participante de eventos, eu gostaria que o sistema me recomendasse eventos baseando-se nas minhas informações de cadastro e também nas informações de eventos que eu tenha declarado interesse.
RF08	Como um participante de eventos, eu gostaria de poder fazer login no sistema para acessar informações sobre eventos.
RF09	Como um participante de eventos, eu gostaria de visualizar uma lista contendo meus contatos.
RF10	Como um participante de eventos, eu gostaria de remover um participante da minha lista de contatos.
RF11	Como um participante de eventos, eu gostaria de realizar uma busca por novos contatos.
RF12	Como um participante de eventos, eu gostaria de visualizar detalhes de um evento.

Tabela 3.2: Requisitos funcionais do sistema

### 3.2.2 Não funcionais

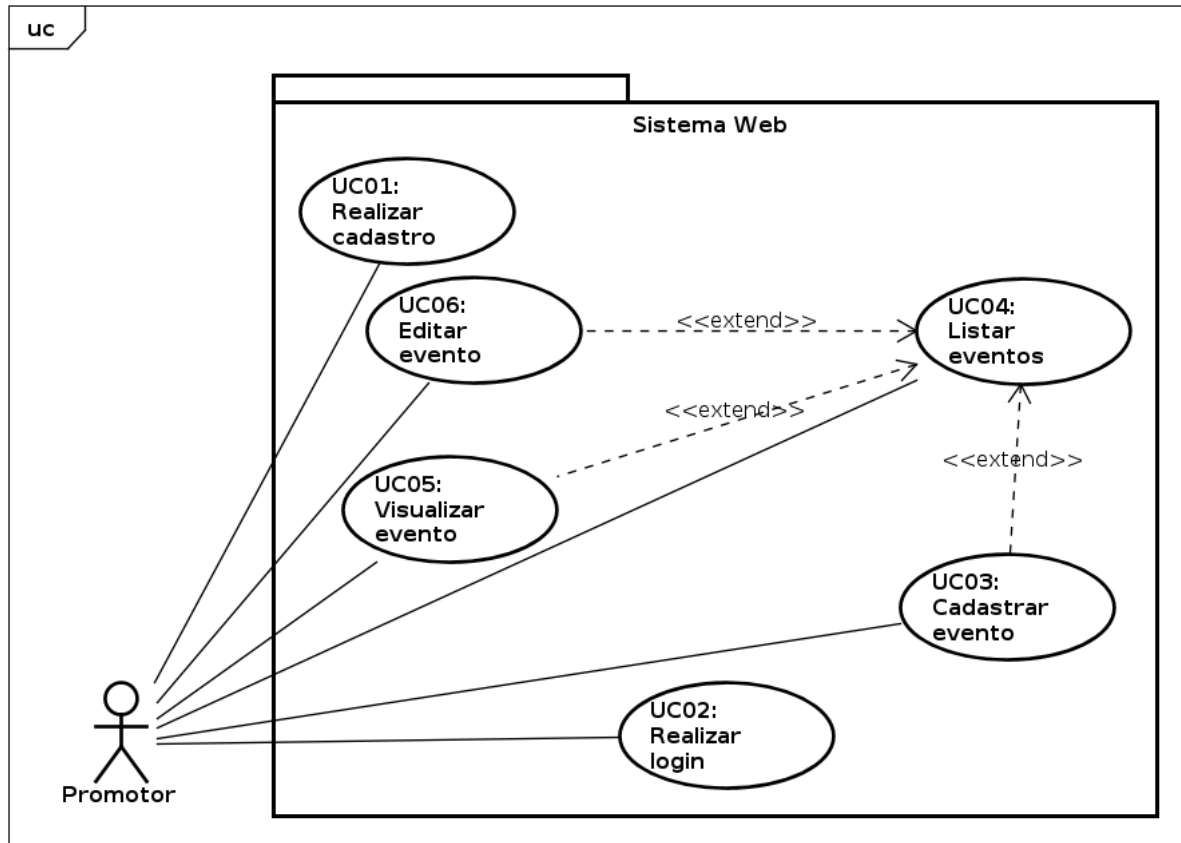
Requisitos não funcionais são relacionados ao uso da aplicação em termos de desempenho, usabilidade, confiabilidade, disponibilidade, segurança e tecnologias envolvidas. Muitas vezes, os requisitos não funcionais acabam gerando (ou coibindo) requisitos funcionais.

RNF01	A comunicação entre aplicativo e servidor deverá ser realizada através de uma API RESTful.
RNF02	Deverão ser aplicados padrões de projetos que sejam pertinentes de forma a melhorar o desempenho do sistema.
RNF03	Para o projeto de interfaces do sistema, deve ser seguido as orientações do <i>material design</i> de forma a melhorar a experiência do usuário enquanto utiliza o sistema.
RNF04	O sistema deve ser confiável, deixando transparente para o usuário, que ao executar uma operação, a mesma foi concluída com sucesso.
RNF05	O sistema deve assegurar que os dados dos seus usuários não serão acessados por partes não autorizadas.
RNF06	O sistema deve estar disponível para os usuários ao menos 99,99% do tempo.

Tabela 3.3: Requisitos não funcionais do sistema

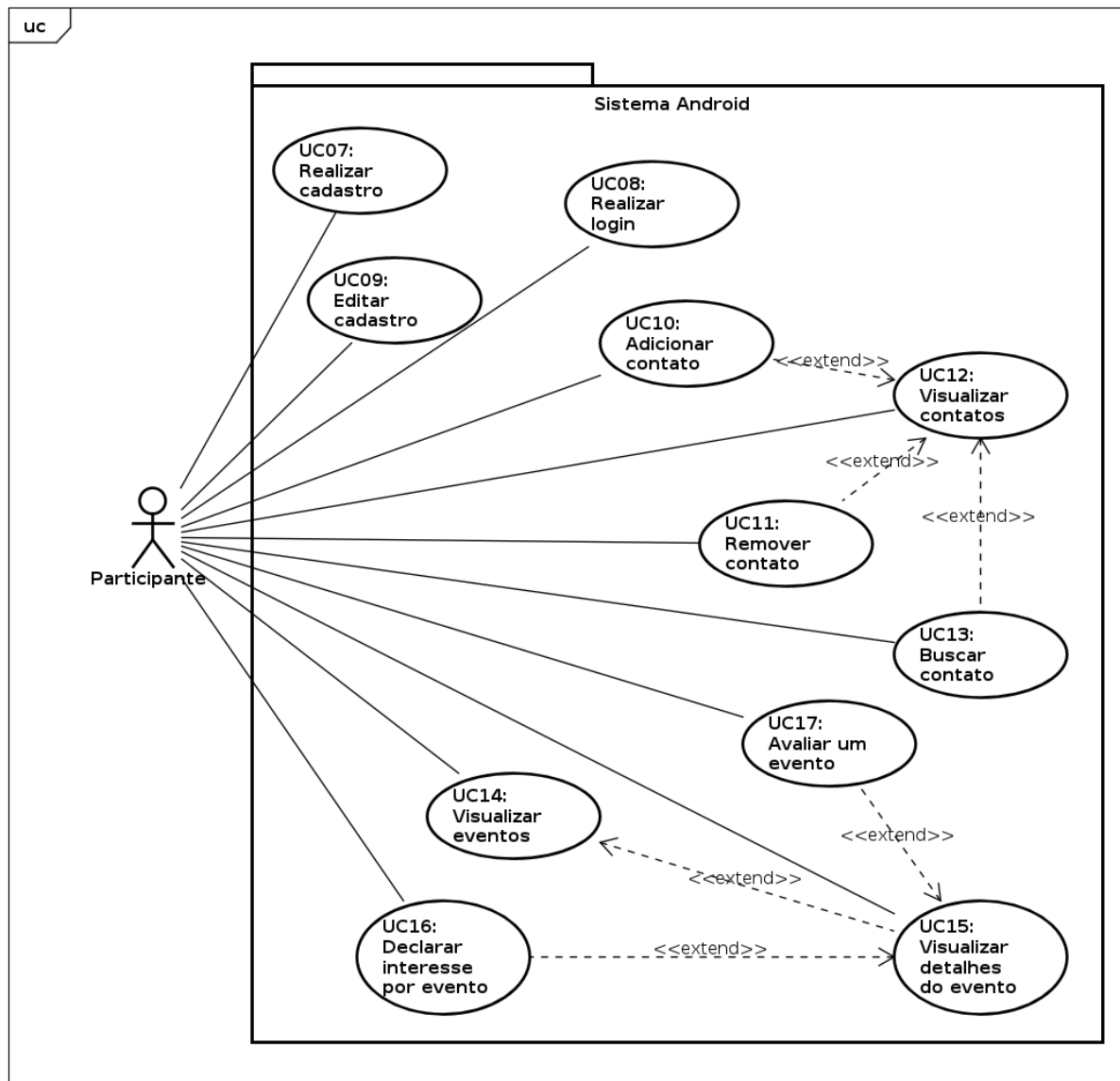
### 3.3 Casos de uso

Após a definição dos requisitos funcionais e não funcionais foi elaborado um diagrama dos casos de uso do sistema. A seguir é apresentado o diagrama e as definições dos casos de uso.



### 1. Web

- UC01 - Realizar cadastro: Um usuário será capaz de acessar uma página para realizar seu cadastro no sistema informando seus dados.
- UC02 - Realizar login: Um usuário será capaz de acessar uma página onde informará suas informações de login (email e senha) para acessar o sistema.
- UC03 - Cadastrar evento: Ao acessar o sistema, um usuário será capaz de cadastrar um novo evento através de um formulário.
- UC04 - Listar eventos: Ao acessar o sistema, um usuário será capaz de visualizar seus eventos previamente cadastrados em uma lista. Nesta mesma visualização o usuário será capaz de filtrar os eventos listados.
- UC05 - Visualizar evento: Ao acessar o sistema e visualizar seus eventos, um usuário será capaz de selecionar um evento e visualizar mais detalhes sobre o mesmo.
- UC06 - Editar evento: Ao acessar o sistema e visualizar seus eventos, um usuário será capaz de editar os dados de um evento selecionado.



## 2. Android

- UC07 - Realizar cadastro: Um usuário será capaz de realizar seu cadastro no sistema ao completar um formulário informando seus dados.
- UC08 - Realizar login: Um usuário será capaz de acessar o sistema ao informar seus dados de login (email e senha).
- UC09 - Editar cadastro: Ao acessar o sistema, um usuário poderá alterar algumas de suas informações informadas durante o cadastro.
- UC10 - Adicionar contato: Ao visualizar um outro usuário do sistema, um usuário será capaz de adicionar este usuário em seus contatos.

- UC11 - Remover contato: Ao visualizar seus contatos, um usuário será capaz de remover qualquer um destes.
- UC12 - Visualizar contatos: Ao acessar o sistema, um usuário será capaz de visualizar uma lista de seus contatos previamente adicionados.
- UC13 - Buscar contato: Ao acessar o sistema, um usuário será capaz de realizar uma busca por outros usuários utilizando o nome destes usuários. O sistema deverá então listar os usuários que correspondem a busca realizada.
- UC14 - Visualizar eventos: Ao acessar o sistema, um usuário será capaz de ver uma lista dos eventos cadastrados no sistema.
- UC15 - Visualizar detalhes do evento: Ao selecionar um evento, um usuário será capaz de ver mais detalhes sobre o evento selecionado.
- UC16 - Declarar interesse por evento: Ao visualizar os detalhes de um evento, um usuário será capaz de declarar qual seu interesse pelo evento.
- UC17 - Avaliar um evento: Ao visualizar os detalhes de um evento e ao ter informado que compareceu no evento, um usuário será capaz de avaliar o evento em uma pontuação de 1 a 5.

## 3.4 Base de dados

Para a persistência dos dados do sistema foi criado um banco de dados relacional com PostgreSQL. Inicialmente foi criado um modelo para o banco que foi incrementado de acordo com a evolução do projeto.

Um promotor de eventos precisa de divulgar seus eventos de forma a facilitar para os possíveis participantes dos eventos encontrar as informações que precisam. Para isso ele se cadastra no sistema informando alguns de seus dados, como: nome, e-mail e senha para acesso ao sistema e um documento (CPF ou CNPJ) para sua identificação.

Após acessar o sistema ele inicia o cadastramento dos seus eventos. No cadastro de um evento ele pode informar dados como o título do evento, uma descrição, a censura, quais as datas e horários para início e fim do evento, um breve texto descrevendo a atração do evento, informações de contato (telefone, email, site), qual o local do evento, definir *tags*

para um evento e cadastrar informações referentes aos tipos de ingressos do evento com informações do preço, sexo, se é promocional para estudante e, caso houver, a quais setores do evento este ingresso dá acesso. Além disso, o promotor deve classificar seu evento de acordo com uma classificação previamente definida no sistema. Esta classificação é dividida em três níveis: categoria, estilo e gênero. A categoria é o nível mais superior e mais amplo da classificação. Cada categoria possui um conjunto de estilos que, por sua vez, possuem gêneros. Os eventos possuem também um atributo de status, indicando, por exemplo, se ainda não foi iniciado, se já finalizou ou se foi cancelado pelo promotor.

Por outro lado, um participante de eventos deseja encontrar eventos de seu interesse, para isso ele se cadastra no sistema informando seus dados pessoais, como: nome, sobrenome, sexo, data de nascimento, estado civil, CEP do seu local de moradia, informações sobre o seu nível educacional e área de estudo e também um e-mail e senha para acesso. Após acessar o sistema e visualizar os eventos disponíveis um participante pode ter "interações" com um evento, por exemplo, indicar se está interessado ou não em um evento, se já foi em um evento passado e neste caso classificar o evento que participou. Além de poder buscar e interagir com eventos, um participante também pode seguir ou ser seguido por outros participantes, gerando uma relação entre participantes distintos.

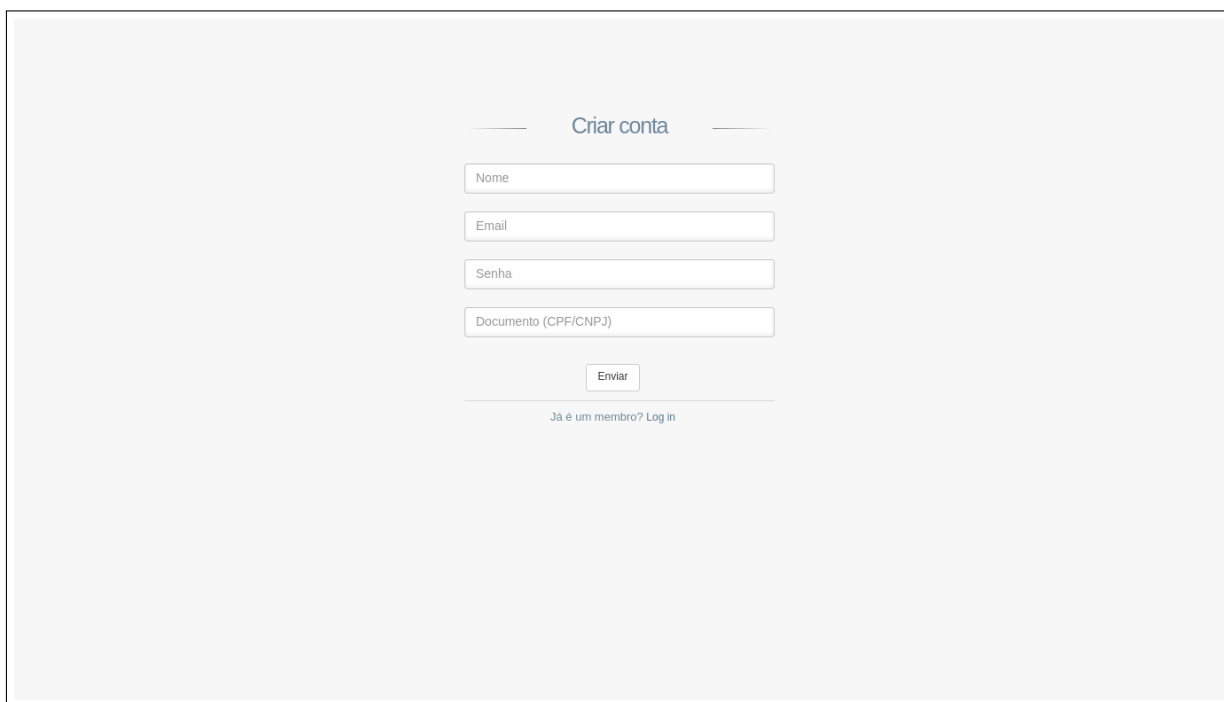
Em um capítulo posterior será feita a descrição de como a base foi populada. O diagrama de entidade relacionamento do banco de dados está disponível no apêndice C.

## 3.5 Protótipo

Nessa seção são mostradas algumas telas do protótipo do sistema, criado a partir das especificações definidas anteriormente na documentação. É importante ressaltar que o protótipo proposto a seguir visa somente definir um MVP (*minimum viable product*) do sistema, onde, por exemplo, em alguns formulários são requisitadas apenas informações necessárias para o funcionamento básico da divulgação e recomendação de eventos que foi desenvolvida nesse trabalho.



### 3.5.1 Protótipo web



Um formulário web centralizado para a criação de uma conta. O formulário é contido dentro de um retângulo cinza claro com uma borda preta. No topo, o texto "Criar conta" está centralizado e sublinhado por duas linhas horizontais. Abaixo dele, há quatro campos de entrada de texto empilhados verticalmente, cada um com uma borda cinza e um ícone de lupa no canto superior direito. Os campos são rotulados "Nome", "Email", "Senha" e "Documento (CPF/CNPJ)". Abaixo dos campos, há um botão "Enviar" com uma borda cinza. Na base do formulário, o texto "Já é um membro? Log in" está centralizado.

Figura 3.2: Formulário para criação de uma conta de promotor de eventos no sistema web.

A Figura 3.2 mostra a primeira tela do protótipo, onde o usuário promotor de eventos pode fazer seu cadastro no sistema. Esta tela mostra a funcionalidade descrita no caso de uso UC01.

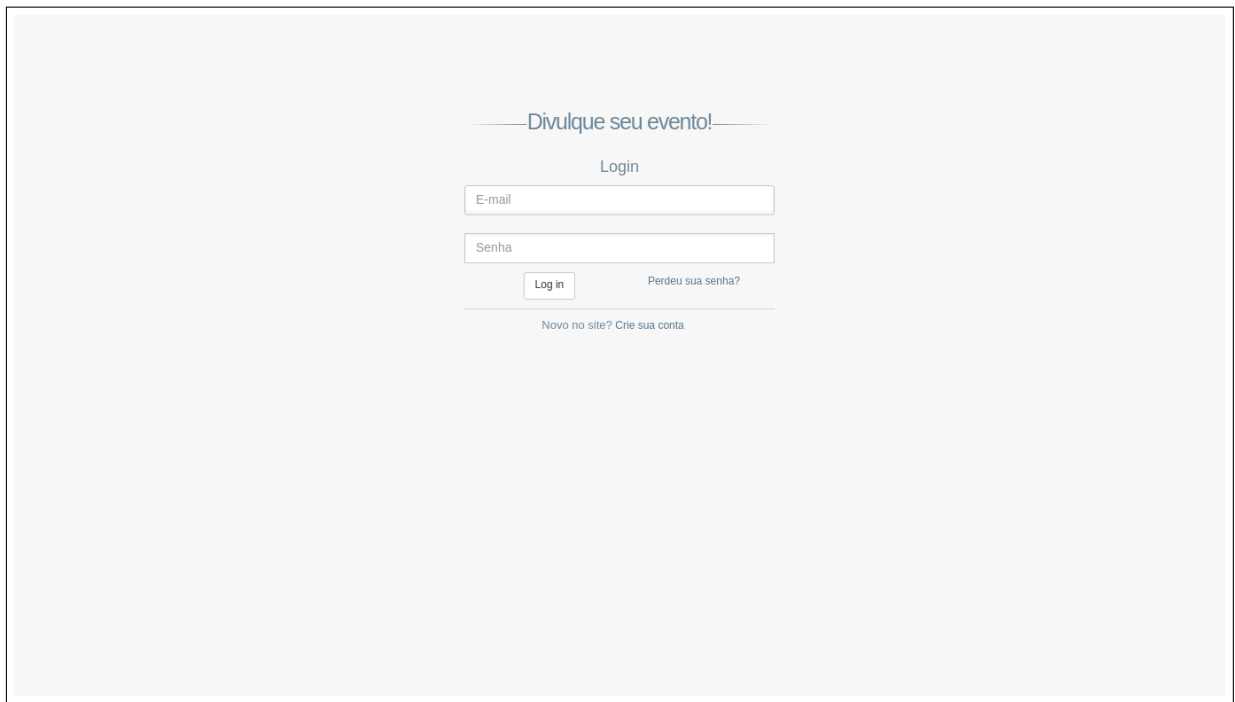


Figura 3.3: Tela para acesso do usuário promotor de eventos ao sistema.

A Figura 3.3 mostra a tela do protótipo, onde o usuário promotor de eventos pode fazer o acesso no sistema. Esta tela demonstra a funcionalidade descrita no caso de uso UC02.

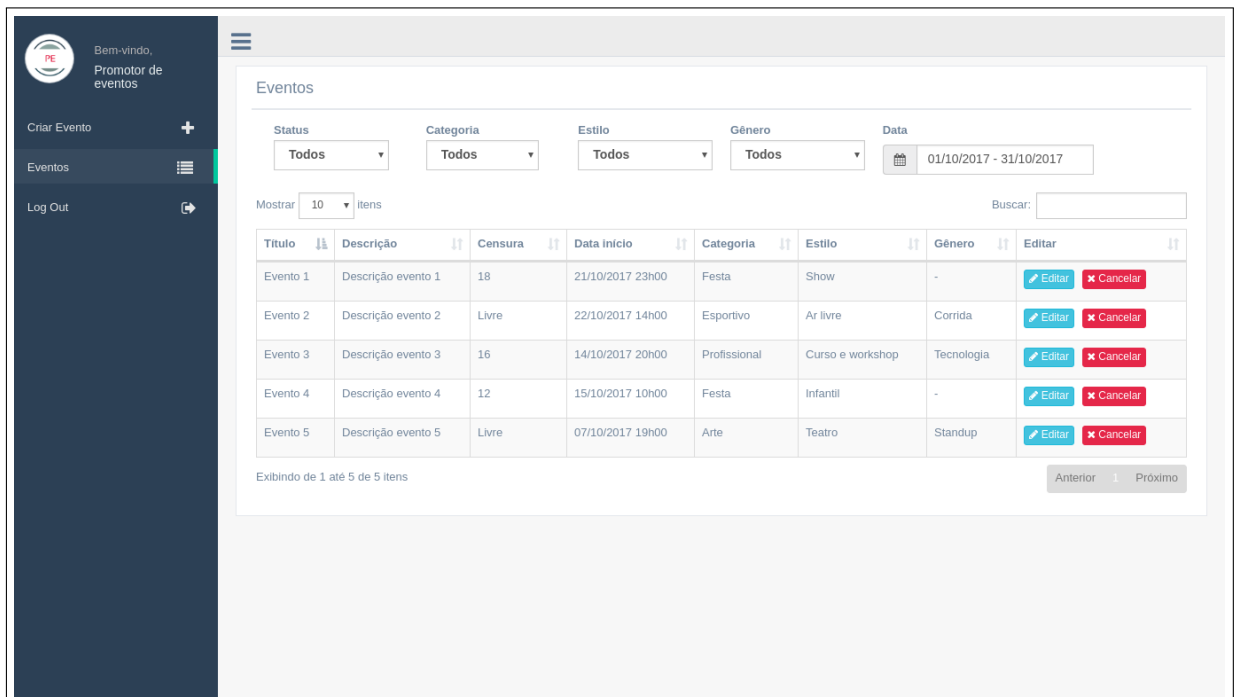


Figura 3.4: Lista de eventos previamente cadastrados pelo usuário.

Na Figura 3.4 é exibida uma lista de eventos cadastrados pelo usuário e os filtros

que podem ser aplicados sobre eles. Esta tela demonstra a funcionalidade descrita no caso de uso UC04.

The screenshot shows a web application interface for editing an event. On the left is a dark sidebar with a logo and the text 'Bem-vindo, Promotor de eventos'. Below the logo are three menu items: 'Criar Evento' with a plus icon, 'Eventos' with a list icon, and 'Log Out' with a right-pointing arrow icon. The main content area is titled 'Editar evento' and contains the following form fields:

- Título \***: Text input with value 'Evento 1'
- Descrição \***: Text area with value 'Descrição Evento 1'
- Censura \***: Radio button group with options 'Livre', '10', '12', '14', '16', and '18'. '18' is selected.
- Data de inicio \***: Text input with value '21/10/2017 23h00'
- Data de fim \***: Text input with value '22/10/2017 6h00'
- Local \***: Text input with value 'Local 1 - Rua Primeira, 01, 10000-001, Primeira Cidade - PE'
- Categoria**: Dropdown menu with value 'Festa'
- Estilo**: Dropdown menu with value 'Show'
- Gênero**: Dropdown menu with value 'Nenhum'

At the bottom of the form are two buttons: a green 'Salvar' button and a red 'Cancelar' button.

Figura 3.5: Visualização e edição de dados de um eventos.

Na Figura 3.5 são exibidos os dados de um evento de forma que os mesmos possam ser alterados pelo usuário. Esta tela demonstra a funcionalidade descrita nos casos de uso UC05 e UC06.

Bem-vindo,  
Promotor de eventos

Criar Evento +

Eventos

Log Out

### Criar evento

1 Passo 1 Descrição

2 Passo 2 Data

3 Passo 3 Local

4 Passo 4 Classificação

Título \*

Descrição \*

Censura \*

Figura 3.6: Passo 1 do formulário para criação de um novo evento.

As Figuras 3.6, 3.7, 3.8 e 3.9 mostram o formulário de criação de um novo evento no sistema. Estas telas demonstram a funcionalidade descrita no caso de uso UC03.

Bem-vindo,  
Promotor de eventos

Criar Evento +

Eventos

Log Out

### Criar evento

1 Passo 1 Descrição

2 Passo 2 Data

3 Passo 3 Local

4 Passo 4 Classificação

Data de início \* < novembro 2017 >

Do	2ª	3ª	4ª	5ª	6ª	Sá
29	30	31	1	2	3	4
5	6	7	8	9	10	11
12	13	14	15	16	17	18
19	20	21	22	23	24	25
26	27	28	29	30	1	2
3	4	5	6	7	8	9

16 : 25

Data de fim \* < novembro 2017 >

Do	2ª	3ª	4ª	5ª	6ª	Sá
29	30	31	1	2	3	4
5	6	7	8	9	10	11
12	13	14	15	16	17	18
19	20	21	22	23	24	25
26	27	28	29	30	1	2
3	4	5	6	7	8	9

16 : 25

Figura 3.7: Passo 2 do formulário para criação de um novo evento.

Bem-vindo,  
Promotor de eventos

Criar Evento +

Eventos

Log Out

### Criar evento

1 Passo 1 Descrição

2 Passo 2 Data

3 Passo 3 Local

4 Passo 4 Classificação

Informe um endereço \*

Ou crie um novo:

Logradouro \*

Número \*  CEP \*

Bairro \*  Complemento

Cidade \*  Estado \*

Anterior Próximo Finalizar

Figura 3.8: Passo 3 do formulário para criação de um novo evento.

Bem-vindo,  
Promotor de eventos

Criar Evento +

Eventos

Log Out

### Criar evento

1 Passo 1 Descrição

2 Passo 2 Data

3 Passo 3 Local

4 Passo 4 Classificação

Categoria \*

Estilo \*

Gênero \*

Anterior Próximo Finalizar

Figura 3.9: Passo 4 do formulário para criação de um novo evento.

### 3.5.2 Protótipo Android

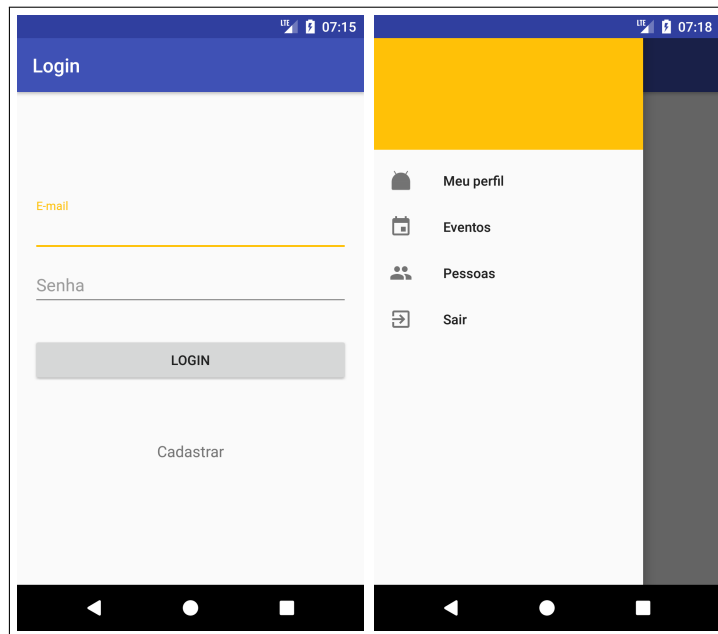


Figura 3.10: Tela de acesso ao sistema para o usuário participante e menu de navegação interno

A Figura 3.10 mostra a tela de login do usuário participante e também um menu para navegação no aplicativo após o login. Esta tela demonstra a funcionalidade descrita no caso de uso UC08.

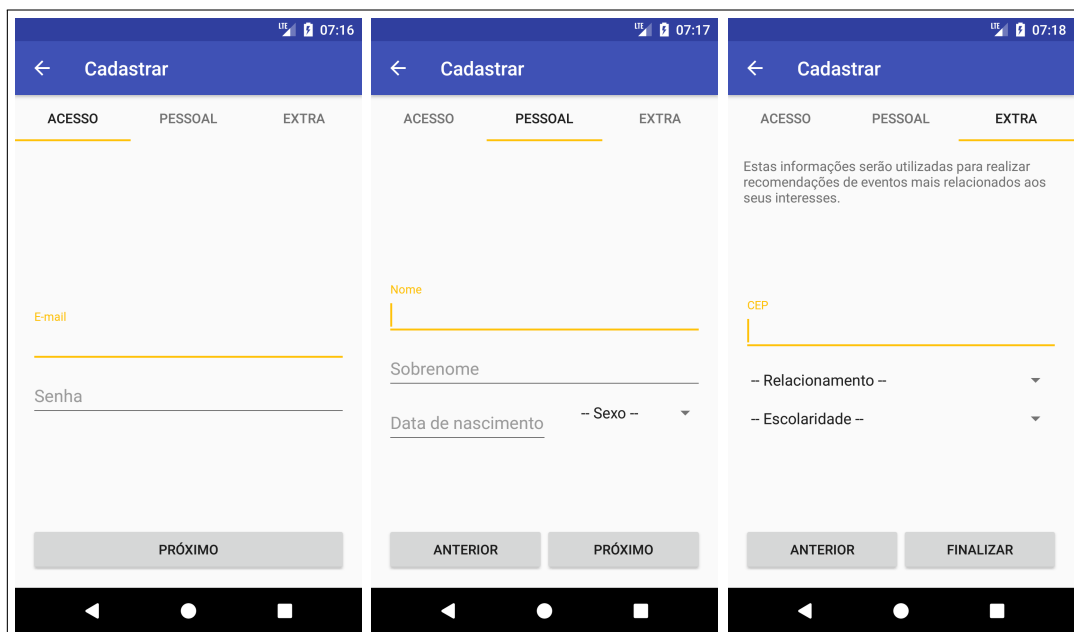


Figura 3.11: Cadastro de usuário participante no sistema

A Figura 3.11 mostra os passos para o cadastro de um novo usuário no sistema

como participante. Estas telas demonstram a funcionalidade descrita no caso de uso UC07.

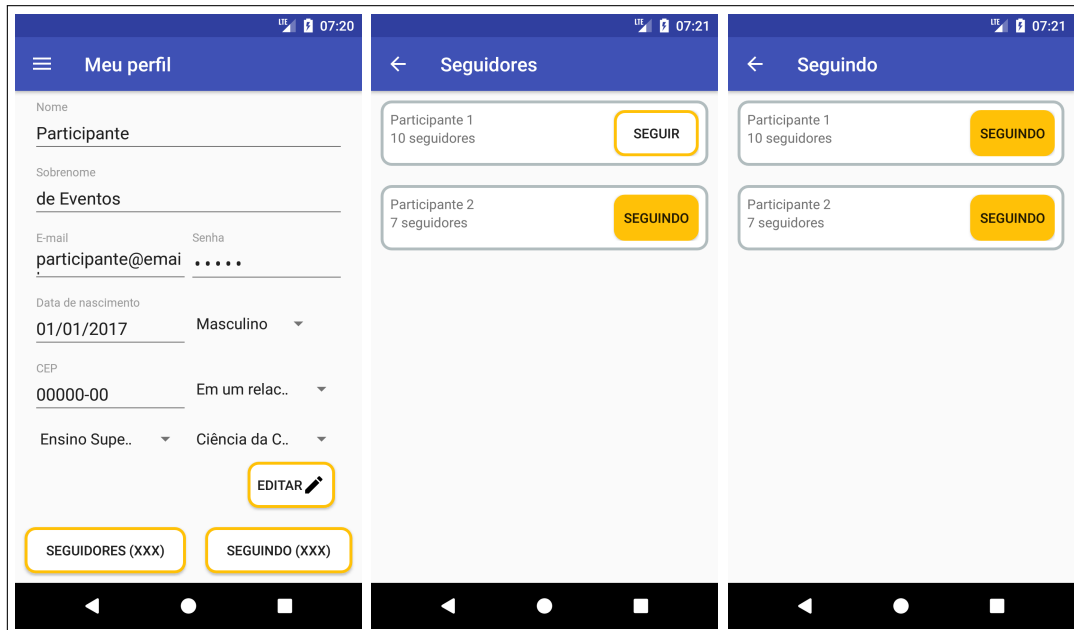


Figura 3.12: Exibição do perfil do usuário com seus dados e também o relacionamento com outros participantes

A Figura 3.12 mostra o perfil de um usuário com seus dados e também o seu relacionamento com outros participantes (sendo seguido ou seguindo outros). Estas telas demonstram as funcionalidades descritas nos casos de uso UC09, UC11 e UC12.

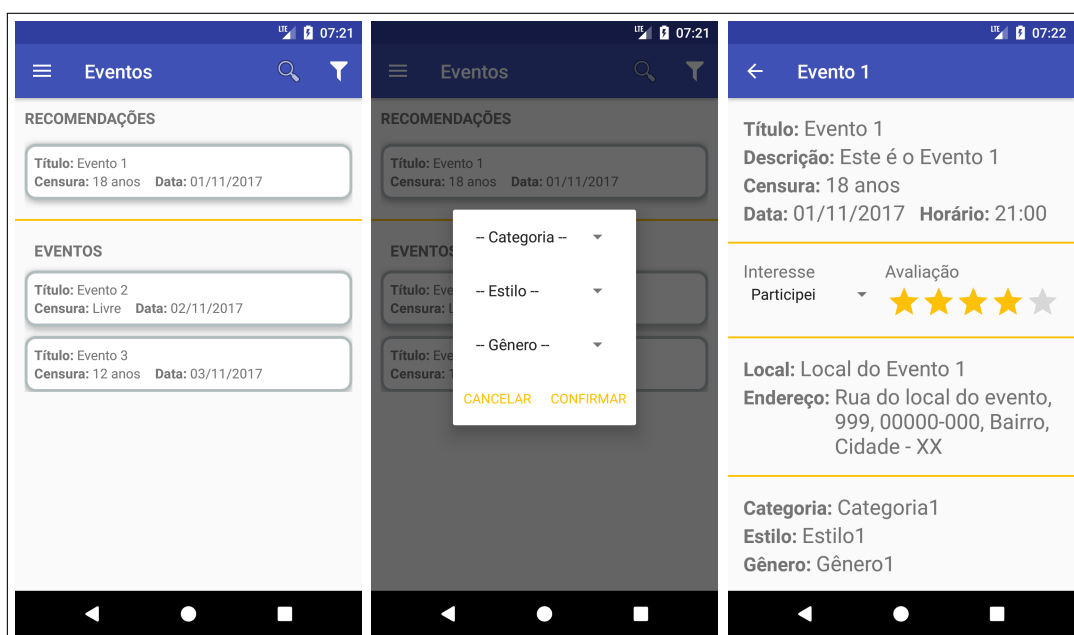


Figura 3.13: Busca por eventos, exibição de detalhes de um evento e avaliação do mesmo

A Figura 3.13 mostra as telas onde o usuário pode buscar e receber recomendações de eventos. Também é feita a exibição de detalhes de um evento onde o participante pode definir seu interesse e avaliar o mesmo caso tenha participado do evento. Estas telas demonstram as funcionalidades descritas nos casos de uso UC14, UC15, UC16 e UC17.



Figura 3.14: Tela de busca por participantes

A Figura 3.14 mostra a tela onde um usuário pode buscar e seguir outros usuários. Esta tela demonstra as funcionalidades descritas nos casos de uso UC10 e UC13.



## 4 Abordagem desenvolvida

Nesse capítulo é descrito como as técnicas mencionadas anteriormente foram utilizadas para auxiliar na tarefa de recomendação de eventos.

### 4.1 Dados utilizados

Para iniciar o processo de desenvolvimento de um método de recomendação é necessário ter uma base de dados populada com dados sobre os eventos, participantes e com algum tipo de interação entre estes. Como não havia uma base disponível com estes tipos de dados e também não seria de fácil obtenção dados reais, foi necessária a criação de uma base de dados artificial.

A seguir é descrito o procedimento utilizado para geração dos dados necessários para a avaliação do método de recomendação.

#### 4.1.1 Geração de dados

Para a geração de dados foram escritos *scripts* na linguagem Python (PYTHON, 2017) que fazem inserções diretamente no banco de dados criado em PostgreSQL (POSTGRESQL, 2017).

Antes de iniciar a geração de dados algumas informações que seriam necessárias foram inseridas no banco para facilitar este processo, por exemplo, informações relacionadas às localizações dos eventos e as possíveis classificações de acordo com as categorias, estilos e gêneros previamente gerados, e informações relacionadas a escolaridade do participante e seu estado civil.

Primeiramente foram gerados os dados de eventos, com base na estrutura das tabelas utilizadas para persistir estes dados foram escolhidos os campos da mesma que poderiam se tornar atributos para o modelo de aprendizado a ser gerado. Desta forma, após realizar algumas transformações dos campos originais da tabela foram escolhidas os seguintes atributos: dia da semana, hora de início, se é feriado, qual a idade de censura,

o CEP da localização, a categoria, estilo e gênero do evento.

Com a escolha dos atributos foram então definidas configurações para o *script* de geração de eventos. Para executar o *script* é necessário configurar o número de eventos a serem gerados, qual o período do dia o evento se inicia (manhã de 7-12h, tarde de 13-18h e noite de 19-24h), se ocorre nos dias de semana ou no fim de semana, qual a censura do evento e a qual categoria devem pertencer. A partir destas configurações iniciais e outras fixas (mês e ano do evento, combinações de categoria, estilo e gênero, etc) foram gerados 300 eventos para serem utilizados na fase de treinamento do modelo de recomendação.

De forma similar o *script* para geração de participantes também possui algumas configurações, como a quantidade de participantes que devem ser gerados, qual o intervalo de idades os participantes devem estar, dados os estados civis em que um participante pode ter (solteiro(a), em um relacionamento sério, casado(a), divorciado(a), viúvo(a)) é informado um vetor com as probabilidades de cada um, da mesma forma também é informado um vetor com as probabilidades dos níveis de escolaridade (fundamental, médio e superior) e foi mantida uma proporção de 47% de homens em relação as mulheres. Assim, foram gerados 1000 participantes para serem também utilizados no conjunto de treinamento dos algoritmos utilizados.

Por fim, foram criadas as interações entre participantes e eventos, um participante poderia definir seu interesse pelo evento das seguintes formas: interessado, participei e vou (todas com um valor de 1), desinteressado, não participei e não vou (todas com um valor de -1) e uma avaliação com valor de 1 a 5 caso tenha participado do evento. Para cada participante foram geradas pelo menos 15 interações com eventos selecionados de forma randômica caso alguma interação fosse de participação no evento, era gerada uma extra do tipo avaliação com um valor randômico de 1 a 5.

Após a geração dos dados de eventos e participantes foi verificado a existência de amostras duplicadas na base através do cálculo da distância euclidiana entre amostras e foram encontradas menos de 10% de duplicação de participantes e menos de 1% de duplicação de eventos. Estas duplicações foram alteradas manualmente até que não houvesse mais repetições na base.

## 4.2 Agrupamento

Como os dados de eventos e participantes não possuíam um rótulo conhecido que auxiliasse na determinação de suas preferências, foi inicialmente feita uma clusterização de ambos como forma de identificar participantes e eventos semelhantes.

Para a análise dos dados e o desenvolvimento das recomendações foi utilizado o ecossistema SciPy baseado em Python que possui pacotes de *software* de código aberto para matemática, ciência e engenharia (JONES et al., 2001–2017). Com destaque para os pacotes scikit-learn, pandas, numpy e matplotlib.

### Eventos

Features	Valor
dia_semana	[0, 6]
hora_inicio	[1, 24]
feriado	{0, 1}
censura	{0, 10, 12, 14, 16, 18}
cep	Qualquer CEP válido, ex.: 36010000, 36036330, etc
categoria	[1, 6]
estilo	[1, 20]
genero	[1, 48]

Tabela 4.1: Features de eventos

Como um pré-processamento dos dados antes de utilizar o algoritmo *K-Means* para gerar o agrupamento dos eventos foi feita a normalização dos dados para tratar problemas com a diferente escala das *features*, como no caso do CEP que chega a ter valores  $10^7$  maiores que outras *features*. Todos os valores foram normalizados no intervalo de 0 a 1 através do seguinte cálculo:

$$z_i = \frac{x_i - \min(x)}{\max(x) - \min(x)} \quad (4.1)$$

Onde  $x_i$  é o valor atual da *feature*  $x$  em uma amostra  $i$ , a função  $\min(x)$  retorna o valor mínimo de  $x$  entre todas as amostras e  $\max(x)$  retorna o valor máximo,  $z_i$  é o valor normalizado.

Para definir o número de *clusters* de eventos a serem utilizados foi medido o coeficiente de silhueta variando de 2 a 10 *clusters*. O melhor resultado obtido para o

coeficiente foi de 8 *clusters*, com um valor de 0,2966, que foi o valor utilizado.

A configuração utilizada para o algoritmo foi então de 8 *clusters*, com 10 execuções com diferentes sementes para inicialização dos centroides, fazendo no máximo 300 iterações a cada execução e uma estratégia de escolha dos clusters chamada de *k-means++* que seleciona centroides distantes entre si.

O resultado da execução do algoritmo de agrupamento pode ser observado no apêndice A.

## Participantes

Features	Valor
cep	Qualquer CEP válido, ex.: 36010000, 36036330, etc
sexo	{0, 1}
estado_civil	{1, 2, 3, 4, 5}
nivel_escolaridade	{1, 2, 3}
grande_area	[1, 9]
area	[1, 99]
idade	[18, 50]
arte	[0, 1]
profissional	[0, 1]
festa	[0, 1]
esportivo	[0, 1]
gastronômico	[0, 1]
vestuário	[0, 1]

Tabela 4.2: Features de participantes

Além dos campos pertencentes ao participante que se tornaram *features* também são inclusas as categorias de eventos, desta forma o gosto do participante faz parte da sua representação. Os valores para estas *features* é dado pelas interações dos participantes com os eventos fazendo a soma dos pesos definidos na Tabela 4.3 de acordo com o tipo de interação feita e a categoria do evento em que o participante demonstrou interesse. Após a realização do somatório dos pesos eles são normalizados de forma que os valores nas categorias estejam no intervalo de  $[-1, 1]$ , isto é feito dividindo todos os valores pelo maior valor em módulo.

A interação de “Participei” foi considerada como zero pois, na geração de dados, sempre que um participante tinha uma interação de “Participei” era criado uma avaliação do evento, portanto esta interação foi considerada irrelevante. Para a interação de “Não

<b>Interação</b>	<b>Peso</b>	<b>Avaliação</b>	<b>Peso</b>
Interessado	1	1	-4
Desinteressado	-1	2	-2
Participei	0	3	0
Não participei	0	4	2
Vou	2	5	4
Não vou	-2		

Tabela 4.3: Pesos para cada interação, utilizado na clusterização do participante

participei” seu valor também é zero pois não é possível definir se o motivo da não participação foi por falta de interesse ou não, não sendo possível assim definir um valor positivo ou negativo para essa interação.

As *features* de sexo, grande área e área, apesar de serem exibidas como forma de caracterizar os clusters, não foram utilizadas no processo de clusterização e de classificação. A escolha pela não utilização delas se deve por um viés causado no modelo que sempre dividia os *clusters* em dois quando elas estavam presentes. Isto ocorre, no caso do sexo, por estar bem balanceado a quantidade de homens e mulheres e então causar uma divisão no espaço por esta característica. No caso da grande área e área a divisão ocorre porque os níveis de escolaridade fundamental e médio não possuem estas características (valor 0 na tabela), aumentando, assim, a distância entre estes e o nível superior.

De forma similar aos eventos, também foi realizada a normalização dos dados dos participantes no intervalo de  $[0, 1]$ . Para a determinação do número de *clusters*, no entanto, apesar de também ter sido feito o cálculo do coeficiente de silhueta, este demonstrou um melhor resultado para números pequenos de *clusters* (2 e 3). Para evitar que houvesse um enviesamento do modelo, o número de clusters de participantes foi escolhido como 5.

A configuração do algoritmo foi similar a utilizada para os eventos, exceto pelo número de *clusters* que foi 5.

O resultado da execução do algoritmo de agrupamento pode ser observado no apêndice A.

## 4.3 Avaliação da classificação

Para realizar a classificação de novos itens (eventos e participantes) nos *clusters* encontrados foram avaliados três modelos: K-NN, *Nearest Centroid* e o SVM.

O algoritmo K-NN foi utilizado com  $k = 5$ . O SVM foi utilizado com fator de penalidade  $C = 1,0$  e a estratégia utilizada para o problema multiclasse é o um contra um. A distância utilizada pelos algoritmos é a euclidiana.

Para avaliação destes modelos foi utilizada a técnica *Stratified K-Fold* com  $k = 10$ .

Cluster	0	1	2	3	4	5	6	7
<b>Nº de itens total</b>	45	42	42	52	44	39	13	23
<b>Nº de itens para treinamento</b>	40	38	38	47	39	35	12	21
<b>Nº de itens para validação</b>	5	4	4	5	5	4	1	2

Tabela 4.4: Exemplo de separação dos dados de eventos em conjunto de treinamento e validação estratificados

Cluster	0	1	2	3	4
<b>Nº de itens total</b>	224	171	222	168	215
<b>Nº de itens para treinamento</b>	202	154	200	151	193
<b>Nº de itens para validação</b>	22	17	22	17	22

Tabela 4.5: Exemplo de separação dos dados de participantes em conjunto de treinamento e validação estratificados

Os seguintes valores de acurácia e desvio padrão foram obtidos como resultado da validação cruzada.

- Eventos:
  - K-NN: acurácia média 0,966 e desvio padrão 0,046
  - NC: acurácia média 1,00 e desvio padrão 0,00
  - SVM: acurácia média 0,959 e desvio padrão 0,044
- Participantes:
  - K-NN: acurácia média 0,852 e desvio padrão 0,030
  - NC: acurácia média 0,979 e desvio padrão 0,015
  - SVM: acurácia média 0,933 e desvio padrão 0,013

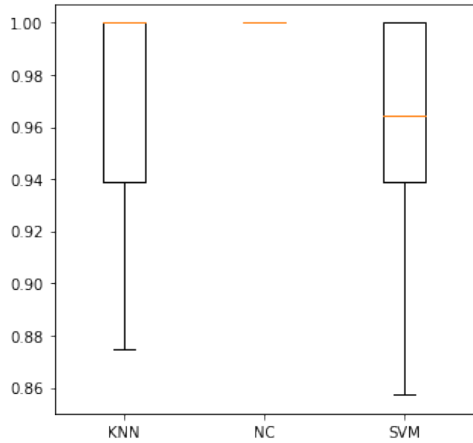


Figura 4.1: Acurácia dos classificadores de eventos

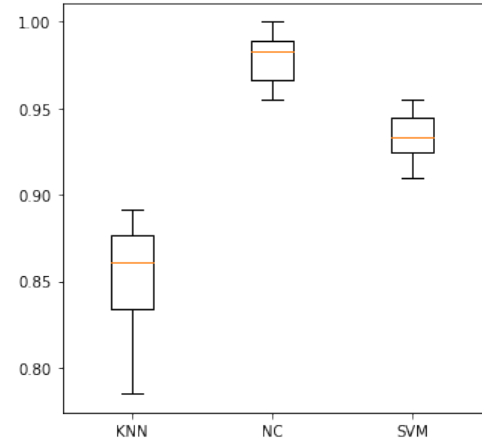


Figura 4.2: Acurácia dos classificadores de participantes

Pode-se perceber que tanto para a classificação de eventos quanto de participantes o *Nearest Centroid* obteve melhores resultados. Este resultado pode ser explicado pela estratégia utilizada pelo NC, como o algoritmo busca o centroide de cada classe e as classes existentes no conjunto de treinamento são os agrupamentos encontrados pelo *K-Means*, o centroide encontrado pelo NC é muito próximo do que foi encontrado pelo *K-Means*, levando a essa melhor acurácia nas predições. Pela análise dos gráficos de *boxplot* pode-se ver também que o NC teve um desempenho mais estável do que os outros modelos.

Para o caso do K-NN e do SVM o seu desempenho não tão bom quanto o do NC pode ser explicado a partir das medições obtidas do coeficiente de silhueta dos agrupamentos. Como os valores da silhueta foram pequenos (positivos e próximos a 0) quer dizer que os agrupamentos encontrados estão mais próximos uns dos outros e as amostras são esparsas no espaço, o que prejudica estes dois modelos. No K-NN, devido a proximidade dos agrupamentos, provavelmente estão sendo utilizados vizinhos que pertencem a mais de um agrupamento e no SVM tanto pela proximidade quanto pela esparsidade dos dados não está sendo possível encontrar um hiperplano que separe completamente os dados.

Pelas matrizes de confusão do participante quando utilizando o K-NN e SVM pode-se notar que ocorrem erros de classificação de um mesmo rótulo para diversos outros, isto pode ser mais uma indicação da proximidade entre os agrupamentos e, como mencionado anteriormente, afeta estes dois métodos de aprendizado.

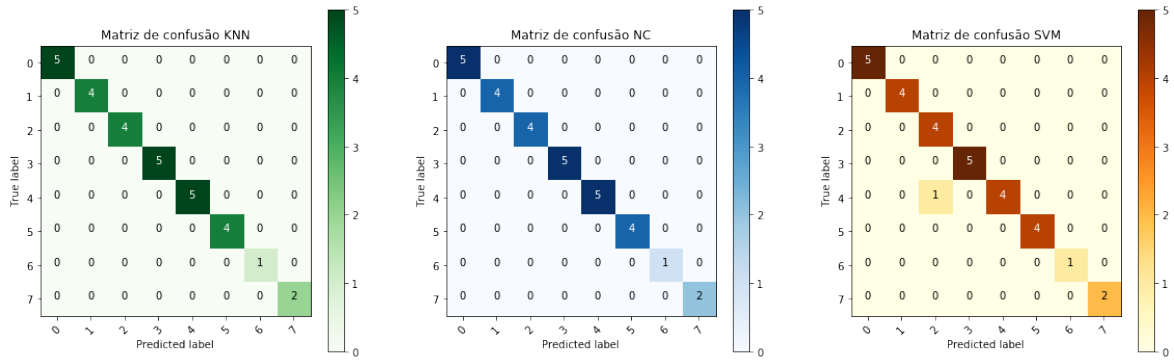


Figura 4.3: Exemplos de matrizes de confusão do K-NN, NC e SVM para eventos

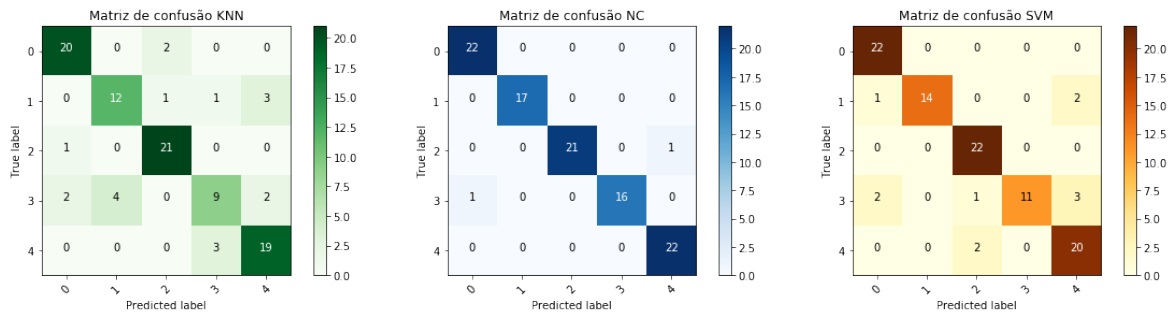


Figura 4.4: Exemplo de matrizes de confusão do K-NN, NC e SVM para participantes

## 4.4 Seletor para recomendação de eventos

Para auxiliar a recomendação de eventos foi utilizada uma simplificação do método definido em (SAWANT, 2013), construindo um grafo bipartido a partir de vértices que representam os *clusters* encontrados de eventos e participantes. Pode-se identificar os dois conjuntos de vértices do grafo como o conjunto  $E$  de *clusters* de eventos, o conjunto  $P$  de *clusters* de participantes e  $A$  representando as arestas do grafo que conectam vértices de um conjunto no outro, assim tem-se o grafo  $G = (E, P, A)$ . Desta forma podemos definir um método de selecionar eventos que podem ser recomendados a participantes de um *cluster*, diminuindo o número de eventos que devem ser analisados para a recomendação de acordo com o *cluster* ao qual o participantes pertence.

Dados dois vértices  $p \in P$  e  $e \in E$ , durante a construção do grafo, uma nova aresta  $a = (p, e)$  é criada caso exista um participante em  $p$  que possua algum tipo de interação com um evento em  $e$ . O peso dado a aresta segue o mesmo padrão determinado na Tabela 4.3. Caso a aresta já exista, o valor é somado ao seu peso atual.

Com o grafo finalizado é possível, então, realizar uma seleção de eventos a serem recomendados de acordo com os *clusters* de eventos e os *clusters* de participantes. Para



isto basta selecionar na linha do grafo (representado com uma matriz) referente ao *cluster* de participante desejado qual o maior valor encontrado e verificar em qual coluna do mesmo este se encontra, o que indica qual o *cluster* de evento recomendado.

	0	1	2	3	4	5	6	7
0	59	37	-24	57	42	-12	-11	84
1	9	-44	14	-5	-25	-18	83	-38
2	33	-28	-12	-11	-36	-17	-36	47
3	22	2	-11	5	8	-11	24	14
4	-38	42	-40	10	-8	39	-9	-7

Tabela 4.6: Exemplo de grafo gerado. *Clusters* de participantes como linhas e de eventos como colunas.

A partir do grafo apresentado na Tabela 4.6, pode-se listar a recomendação para cada cluster:

- *Cluster* 0 de participante é recomendado o *cluster* 7 de eventos
- *Cluster* 1 de participante é recomendado o *cluster* 6 de eventos
- *Cluster* 2 de participante é recomendado o *cluster* 7 de eventos
- *Cluster* 3 de participante é recomendado o *cluster* 6 de eventos
- *Cluster* 4 de participante é recomendado o *cluster* 1 de eventos

Desta forma, no entanto, estamos relacionando os participantes a um único agrupamento de eventos que pode ser uma opção muito limitada. Uma outra opção seria normalizar os pesos do grafo 4.6 no intervalo  $[-1, 1]$  usando o maior valor absoluto (como feito anteriormente nas *features* do participante) e recomendar, por exemplo, eventos de *clusters* que possuem um peso acima de 0,70 como indicado no grafo 4.7.

	<b>0</b>	<b>1</b>	<b>2</b>	<b>3</b>	<b>4</b>	<b>5</b>	<b>6</b>	<b>7</b>
<b>0</b>	<b>0.7023</b>	0.4404	-0.2857	0.6785	0.5000	-0.1428	-0.1309	<b>1.000</b>
<b>1</b>	0.1084	-0.5301	0.1686	-0.0602	-0.3012	-0.2168	<b>1.0000</b>	-0.4578
<b>2</b>	<b>0.7021</b>	-0.5957	-0.2553	-0.2340	-0.7659	-0.3617	-0.7659	<b>1.0000</b>
<b>3</b>	<b>0.9166</b>	0.0833	-0.4583	0.2083	0.3333	-0.4583	<b>1.0000</b>	0.5833
<b>4</b>	-0.9047	<b>1.0000</b>	-0.9523	0.2380	-0.1904	<b>0.9285</b>	-0.2142	-0.1666

Tabela 4.7: Exemplo de grafo gerado com pesos normalizados no intervalo  $[-1, 1]$ . *Clusters* de participantes como linhas e de eventos como colunas.

## 5 Análise de resultados

Para analisar os resultados obtidos com as recomendações foram criados 100 novos participantes e 30 novos eventos (10% da quantidade em cada um dos conjuntos de treinamento gerados) utilizando os mesmos *scripts* mencionados anteriormente.

Após utilizar os modelos treinados do K-NN, NC e SVM para prever a classe destes novos dados, notou-se que entre os 30 eventos gerados 4 possuíam uma divergência na classe escolhida e entre os 100 participantes 12 também apresentaram uma divergência de acordo com o modelo utilizado. Considerando-se a classificação dada pelo NC como a mais provável de estar correta, dado os resultados de acurácia apresentados no capítulo anterior, e comparando as outras duas viu-se que as divergências ocorreram de forma em que uma ou ambas discordavam do NC, houve uma divergência maior entre as classificações dadas pelo K-NN.

No caso das diferenças nas classificações de participantes houve muitas relacionadas aos agrupamentos 3 e 4. Ao verificar a Tabela 5.2, pode-se perceber a proximidade entre estes agrupamentos, o que poderia explicar a divergência gerada pelo K-NN e o SVM.

	<b>0</b>	<b>1</b>	<b>2</b>	<b>3</b>	<b>4</b>	<b>5</b>	<b>6</b>	<b>7</b>
<b>0</b>	0.0	1.1371	0.8454	0.9104	<b>0.7327</b>	1.0898	1.5949	1.0790
<b>1</b>	1.1371	0.0	1.3515	<b>0.7281</b>	0.8966	1.0486	1.3575	1.3637
<b>2</b>	<b>0.8454</b>	1.3515	0.0	1.3566	0.8880	1.1198	1.5813	1.1968
<b>3</b>	0.9104	<b>0.7281</b>	1.3566	0.0	1.1666	1.0750	1.4733	1.3322
<b>4</b>	<b>0.7327</b>	0.8966	0.8880	1.1666	0.0	1.0053	1.3923	1.0767
<b>5</b>	1.0898	1.0486	1.1198	1.0750	<b>1.0053</b>	0.0	1.0554	1.3576
<b>6</b>	1.5949	1.3575	1.5813	1.4733	1.3923	1.0554	0.0	<b>0.9961</b>
<b>7</b>	1.0790	1.3637	1.1968	1.3322	1.0767	1.3576	<b>0.9961</b>	0.0

Tabela 5.1: Distâncias entre centroides de clusters de eventos

Na utilização dos modelos construídos para classificar os 30 eventos gerados para teste foi identificado que em 4 destes eventos houve uma divergência entre as classificações encontradas, como indica a tabela 5. Fazendo-se uma comparação manual entre os dados destes eventos e os *clusters* existentes pode-se ver que o K-NN conseguiu acertar a clas-

	<b>0</b>	<b>1</b>	<b>2</b>	<b>3</b>	<b>4</b>
<b>0</b>	0.0	0.7491	<b>0.5809</b>	0.7840	0.9185
<b>1</b>	0.7491	0.0	0.8890	<b>0.6246</b>	0.6300
<b>2</b>	<b>0.5809</b>	0.8890	0.0	0.9493	0.7259
<b>3</b>	0.7840	<b>0.6246</b>	0.9493	0.0	0.6326
<b>4</b>	0.9185	<b>0.6300</b>	0.7259	0.6326	0.0

Tabela 5.2: Distâncias entre centroides de clusters de participantes

sificação de dois eventos, o NC também acertou dois e o SVM acertou os quatro. Desta forma, apesar da melhor performance do NC na fase de treinamento e validação pode-se ver que isto não é o suficiente para defini-lo como a melhor opção a ser utilizada e que um método como o SVM talvez proporcione uma melhor generalização do que o NC.

dia_semana	hora_inicio	feriado	censura	cep	categoria	estilo	genero	knn_cluster	nc_cluster	svm_cluster
4	23	1	14	36025290	5	17	0	7	3	3
6	23	0	16	36033007	3	10	0	0	3	0
0	22	0	12	36036220	3	7	0	3	4	3
0	18	0	16	36047362	2	6	19	3	4	4

Tabela 5.3: Eventos com classificações divergentes

De forma similar também foi identificado que entre os 100 participantes gerados para teste 12 tiveram classificações divergentes, como indicado na tabela 5. Através da análise manual dos dados destes participantes e dos *clusters* pode-se ver que o K-NN teria acertado a classificação de sete participantes, o NC de seis e o SVM de cinco. Novamente vê-se que o NC não demonstrou um desempenho tão diferente dos outros modelos.

cep	sexo	estado_civil	nivel_escolaridade	grande_area	area	idade	vestuário	esportivo	gastronômico	festa	arte	profissional	knn_cluster	nc_cluster	svm_cluster
36039080	1	2	3	1	2	37	0.8	1.0	-0.4	0.1	-0.7	3	4	4	
36047362	1	1	3	7	70	37	-1.0	1.0	-0.125	0.75	0.625	0.0	4	3	3
36039080	1	2	2	0	0	44	-0.888	-0.222	0.555	-0.888	-1.0	-0.111	4	4	1
36025110	0	2	3	4	38	39	0.0	-0.6	0.6	0.9	1.0	-0.9	3	4	2
36033007	0	1	3	7	69	47	0.555	-0.222	-0.444	1.0	0.111	0.444	3	4	4
36010532	1	2	3	6	52	48	0.125	0.125	-1.0	-0.375	-0.875	-0.375	4	2	2
36025290	0	1	1	0	0	32	0.2	-0.8	-0.5	-0.5	-0.6	-1.0	0	1	1
36047362	0	1	3	3	22	40	-0.6	0.8	0.4	1.0	-0.6	0.9	3	3	4
36016030	1	1	3	7	68	24	0.7	-0.6	0.6	0.6	-1.0	0.4	2	3	2
36039080	1	1	3	3	25	43	-0.285	0.571	0.857	1.0	0.714	0.857	3	4	4
36047362	0	2	3	3	25	29	-0.5	0.3	-0.8	-0.2	-1.0	1.0	4	3	3
36016030	0	1	3	3	26	35	-1.0	-0.888	0.222	0.666	-0.111	1.0	3	2	2

Tabela 5.4: Participantes com classificações divergentes

Os dados de participantes e eventos utilizados para teste podem ser encontrados no apêndice B.

## 6 Considerações finais

Observou-se que, entre os três modelos de identidade utilizados, o *Nearest Centroid* obteve melhores resultados. Isso ocorreu devido a utilização dos clusters encontrados pelo *K-Means* serem utilizados como classes pelos algoritmos de aprendizagem. Como o NC também possui um método que se utiliza do centroide de uma classe para prever uma classificação, os centroides encontrados por ele foram muito próximos aos do *K-Means*, fazendo com que suas classificações fossem mais próximas do *cluster* real e garantindo sua acurácia alta para a validação com o conjunto de treinamento.

Por não haver dados que indicam se um participante aceitou ou não uma recomendação de evento, não é possível ter uma medida do quão efetiva as recomendações feitas são e se são realmente relevantes para um participante.

A forma como os dados foram gerados também limita a análise feita sobre os dados simulados. Já que os dados foram criados de forma randômica e seguindo uma distribuição normal os *clusters* formados eram todos muito similares, o que dificulta encontrar preferências para cada agrupamento.

### 6.1 Trabalhos futuros

Para realização de trabalhos futuros que tragam contribuições para este trabalho, seria de grande interesse que fossem obtidos dados reais para a base. Apenas com esta diferença já seria possível fazer uma comparação entre os resultados obtidos com os dados simulados e dados reais e então validar se os modelos obtidos neste primeiro trabalho realmente fazem sentido para uma aplicação real. Considerando a aplicação em um contexto real seria interessante avaliar formas de lidar com o problema da partida fria em um sistema de divulgação de eventos.

A estrutura da base de dados definida possui a capacidade de armazenar mais informações do que foram utilizadas, por exemplo, informações sobre as formas de ingresso em um evento com detalhes sobre sexo, acesso a setores, valor e se é uma forma de ingresso

especial para estudantes, e também informações sobre como os participantes podem se conectar pelo sistema através de uma rede de seguidores e seguidos. É possível que utilizar estas informações em uma aplicação com dados reais traga um grande aumento de informações relevantes para o problema. Com estas informações seria possível também realizar recomendações de participantes para outros participantes.

Como melhoria ao sistema atual, ao invés de se requisitar ao participante o CEP da sua residência seria interessante coletar informações sobre a sua localização durante o dia através de seu *smartphone* e então inferir a partir destes dados locais importantes para o usuário (sua residência, trabalho, etc) e utilizar esta informação para identificar usuários semelhantes e como isto poderia afetar a recomendação. Uma melhoria na forma como são feitas as recomendações seria de, após encontrar o *cluster* de eventos que seria recomendado ao participante, utilizar informações sobre o seu gosto para filtrar quais eventos dentro daquele *cluster* são mais relevantes para o participante específico.

Outro caminho a ser explorado é o de classificar um evento de forma automática sem a utilização de categorias, estilos e gêneros definidas de forma manual no trabalho atual. Isto poderia ser feito ao se aproveitar de campos de texto como o título do evento, sua descrição e tags relacionadas ao evento para gerar classificações automáticas dos eventos cadastrados no sistema.

## Bibliografia

- ADOMAVICIUS, G.; TUZHILIN, A. Toward the next generation of recommender systems: A survey of the state-of-the-art and possible extensions. *IEEE Trans. on Knowl. and Data Eng.*, IEEE Educational Activities Department, Piscataway, NJ, USA, v. 17, n. 6, p. 734–749, jun. 2005. ISSN 1041-4347. Disponível em: [⟨https://doi.org/10.1109/TKDE.2005.99⟩](https://doi.org/10.1109/TKDE.2005.99).
- ADOMAVICIUS, G.; TUZHILIN, A. Context-aware recommender systems. In: \_\_\_\_\_. *Recommender Systems Handbook*. Boston, MA: Springer US, 2011. p. 217–253. ISBN 978-0-387-85820-3. Disponível em: [⟨https://doi.org/10.1007/978-0-387-85820-3\\_7⟩](https://doi.org/10.1007/978-0-387-85820-3_7).
- ANDROID (Plataforma). 2017. [Online; acessado em 24/11/2017]. Disponível em: [⟨https://developer.android.com/index.html⟩](https://developer.android.com/index.html).
- FACELI K., L. A. C. G. J. e. C. A. C. P. L. F. *Inteligência Artificial - Uma Abordagem de Aprendizado de Máquina*. [S.l.]: LTC, 2011. ISBN 9788521618805.
- JONES, E. et al. *SciPy: Open source scientific tools for Python*. 2001–2017. [Online; acessado em 24/11/2017]. Disponível em: [⟨http://www.scipy.org/⟩](http://www.scipy.org/).
- MACEDO, A. Q.; MARINHO, L. B. Event recommendation in event-based social networks. In: *Proceedings of the Int. Work. on Social Personalization*. [S.l.: s.n.], 2014.
- MACEDO, A. Q.; MARINHO, L. B.; SANTOS, R. L. Context-aware event recommendation in event-based social networks. In: ACM. *Proceedings of the 9th ACM Conference on Recommender Systems*. [S.l.], 2015. p. 123–130.
- POSTGRESQL. 2017. [Online; acessado em 24/11/2017]. Disponível em: [⟨https://www.postgresql.org/⟩](https://www.postgresql.org/).
- PYTHON. 2017. [Online; acessado em 24/11/2017]. Disponível em: [⟨https://www.python.org/⟩](https://www.python.org/).
- REACTJS (Biblioteca). 2017. [Online; acessado em 24/11/2017]. Disponível em: [⟨https://reactjs.org/⟩](https://reactjs.org/).
- RICCI, F. Mobile recommender systems. v. 12, p. 205–231, 01 2011.
- SAWANT, S. Collaborative filtering using weighted bipartite graph projection: A recommendation system for yelp. 2013.
- SPRINGBOOT (Framework). 2017. [Online; acessado em 24/11/2017]. Disponível em: [⟨http://projects.spring.io/spring-boot/⟩](http://projects.spring.io/spring-boot/).

# A Representação dos clusters de eventos e participantes

## A.1 Eventos

<b>Cluster</b>	0	1	2	3	4	5	6	7
<b>Nº de itens</b>	45	42	42	52	44	39	13	23

Tabela A.1: Número de eventos em cada cluster

## A.2 Participantes

<b>Cluster</b>	0	1	2	3	4
<b>Nº de itens</b>	224	171	222	168	215

Tabela A.2: Número de participantes em cada cluster

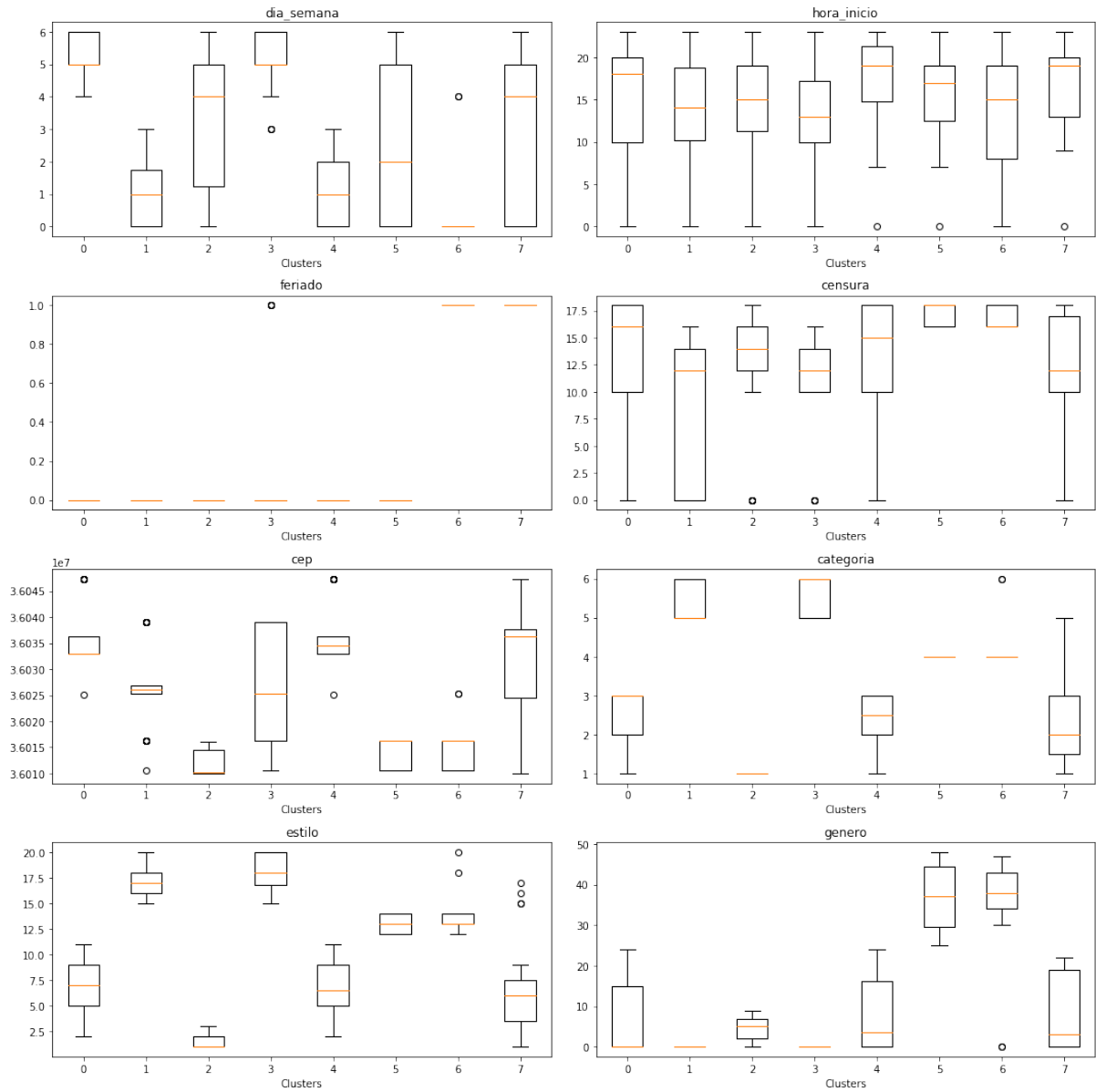


Figura A.1: Boxplots de features para cada cluster de eventos



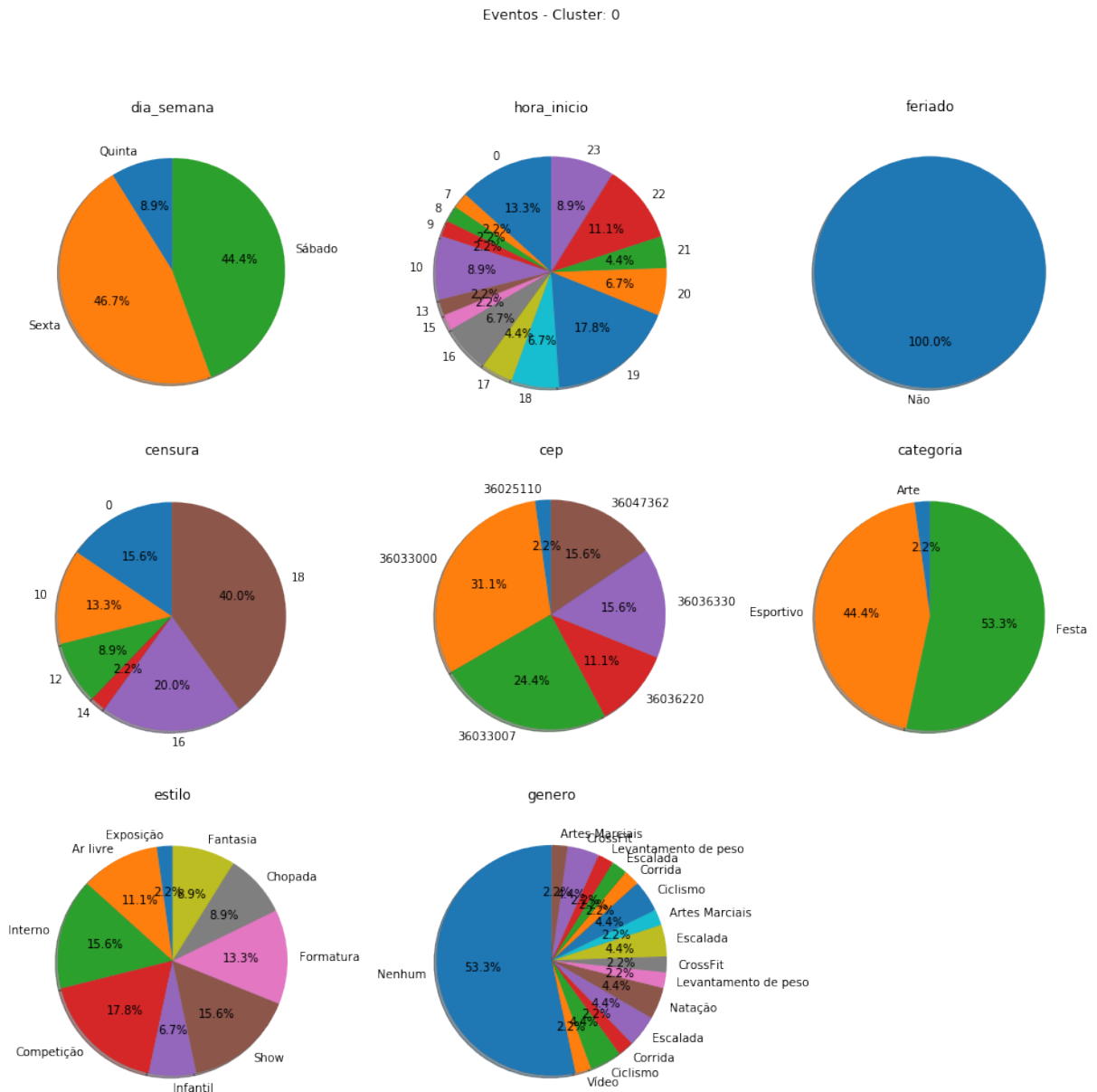


Figura A.2: Representação do cluster 0 de eventos

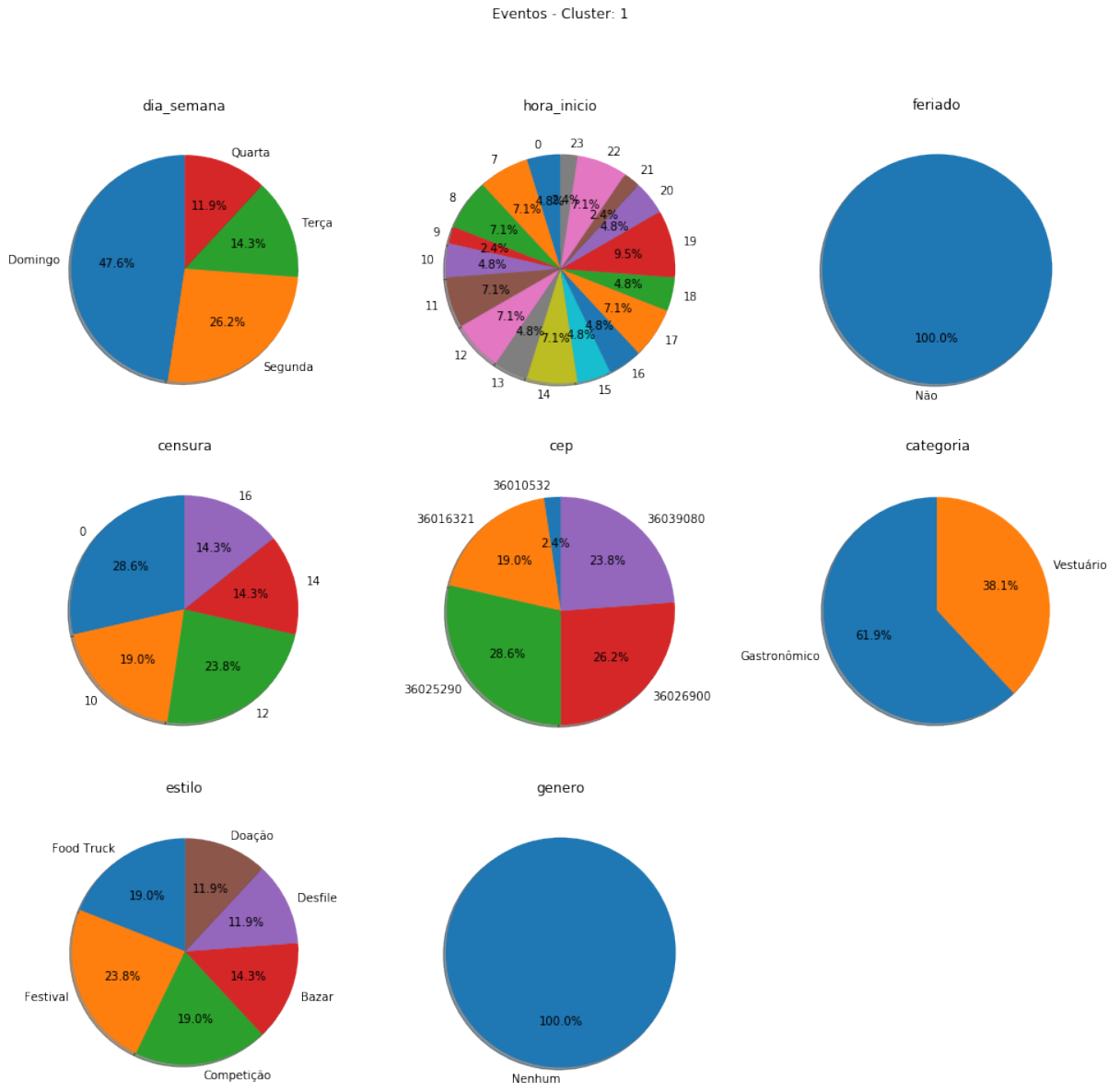


Figura A.3: Representação do cluster 1 de eventos

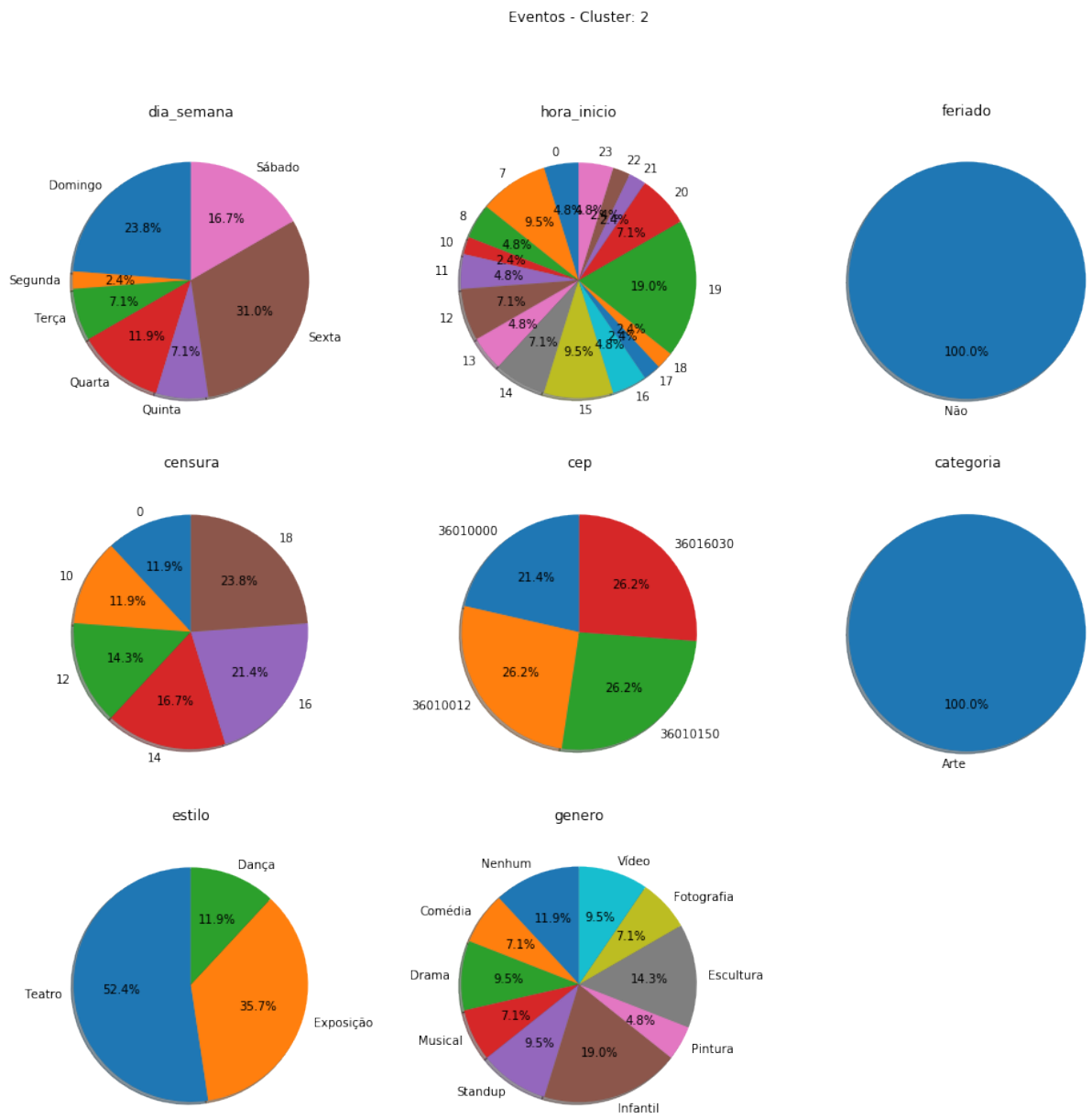


Figura A.4: Representação do cluster 2 de eventos

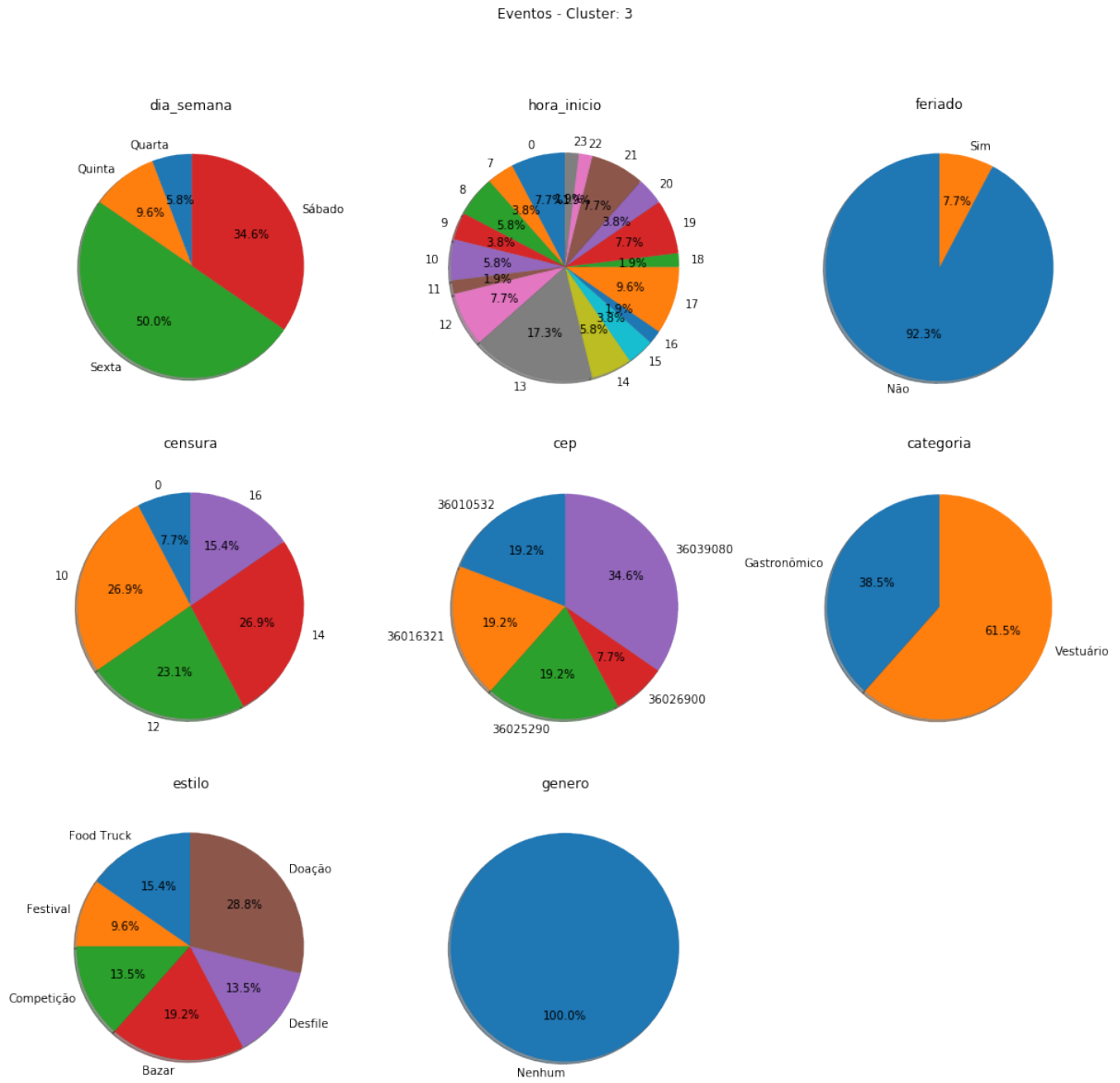


Figura A.5: Representação do cluster 3 de eventos



Figura A.6: Representação do cluster 4 de eventos

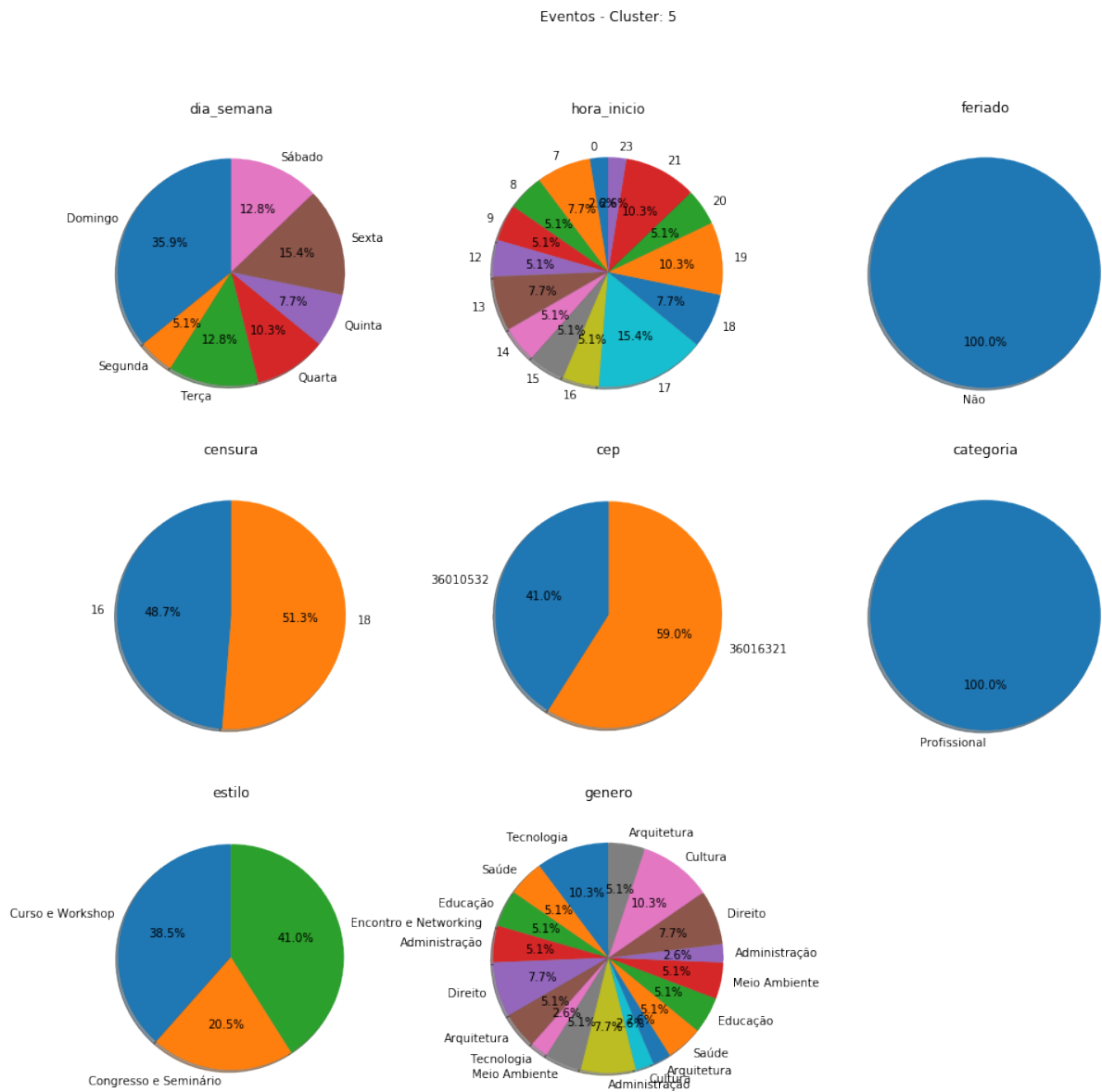


Figura A.7: Representação do cluster 5 de eventos

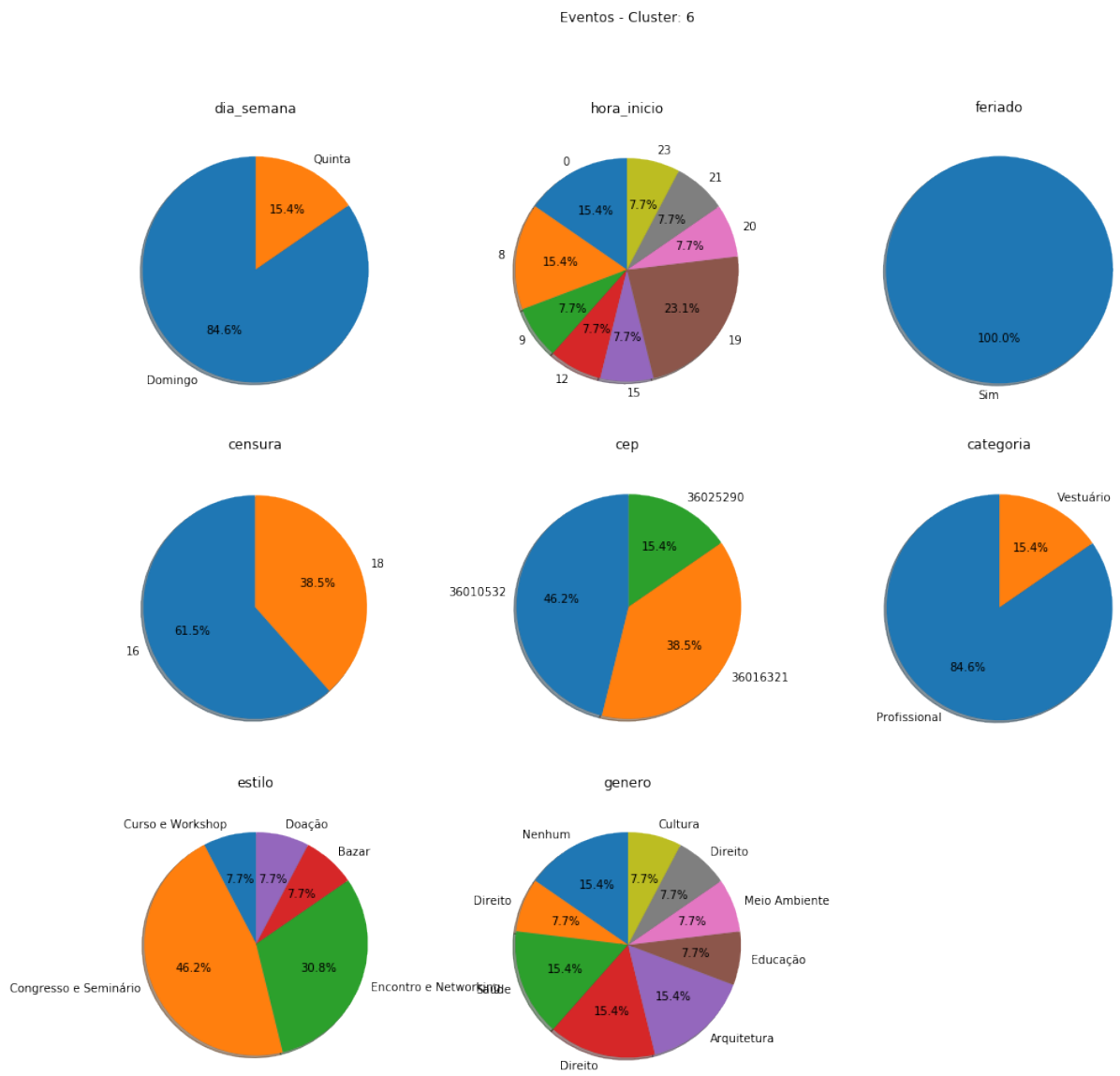


Figura A.8: Representação do cluster 6 de eventos

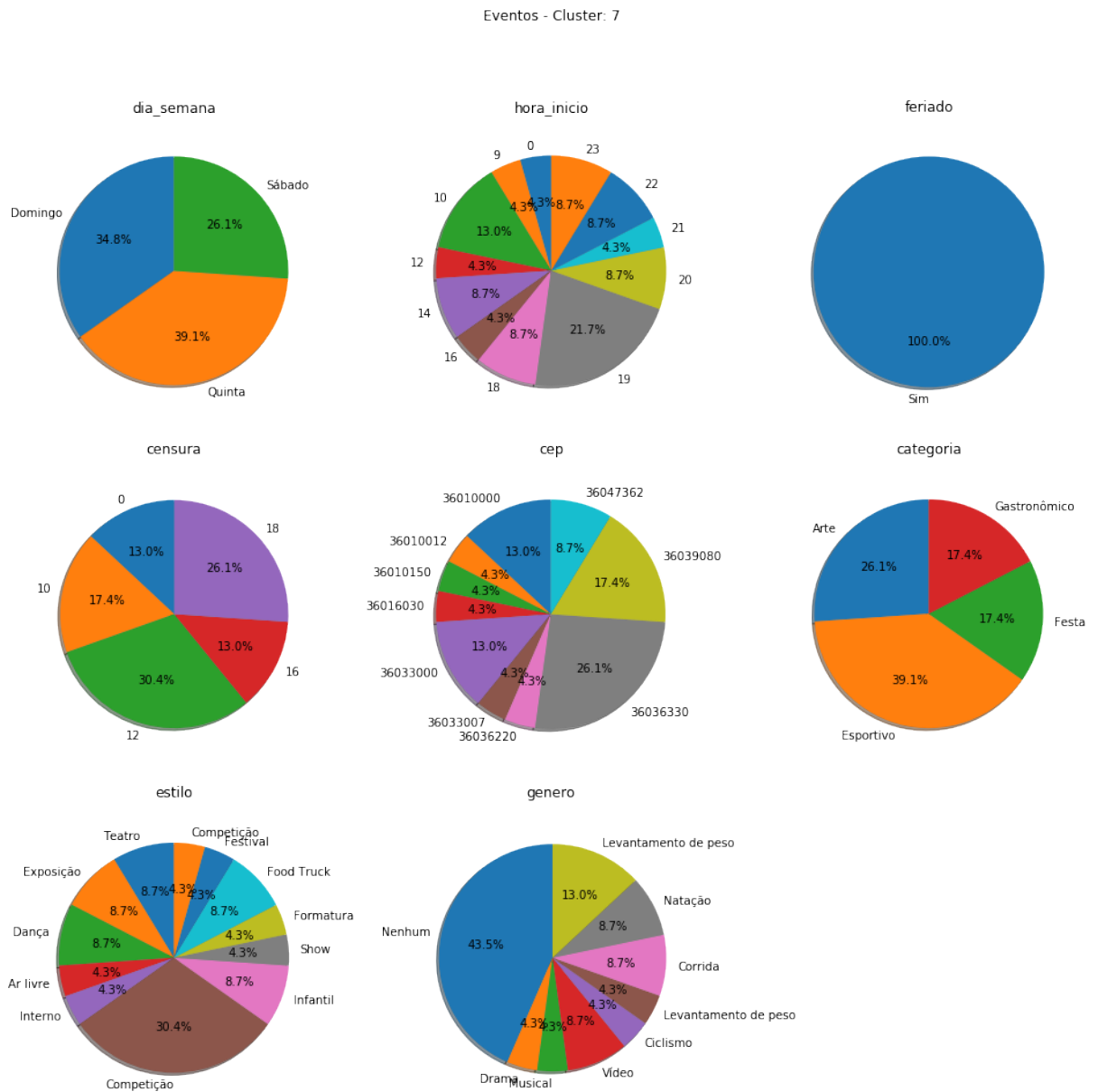


Figura A.9: Representação do cluster 7 de eventos



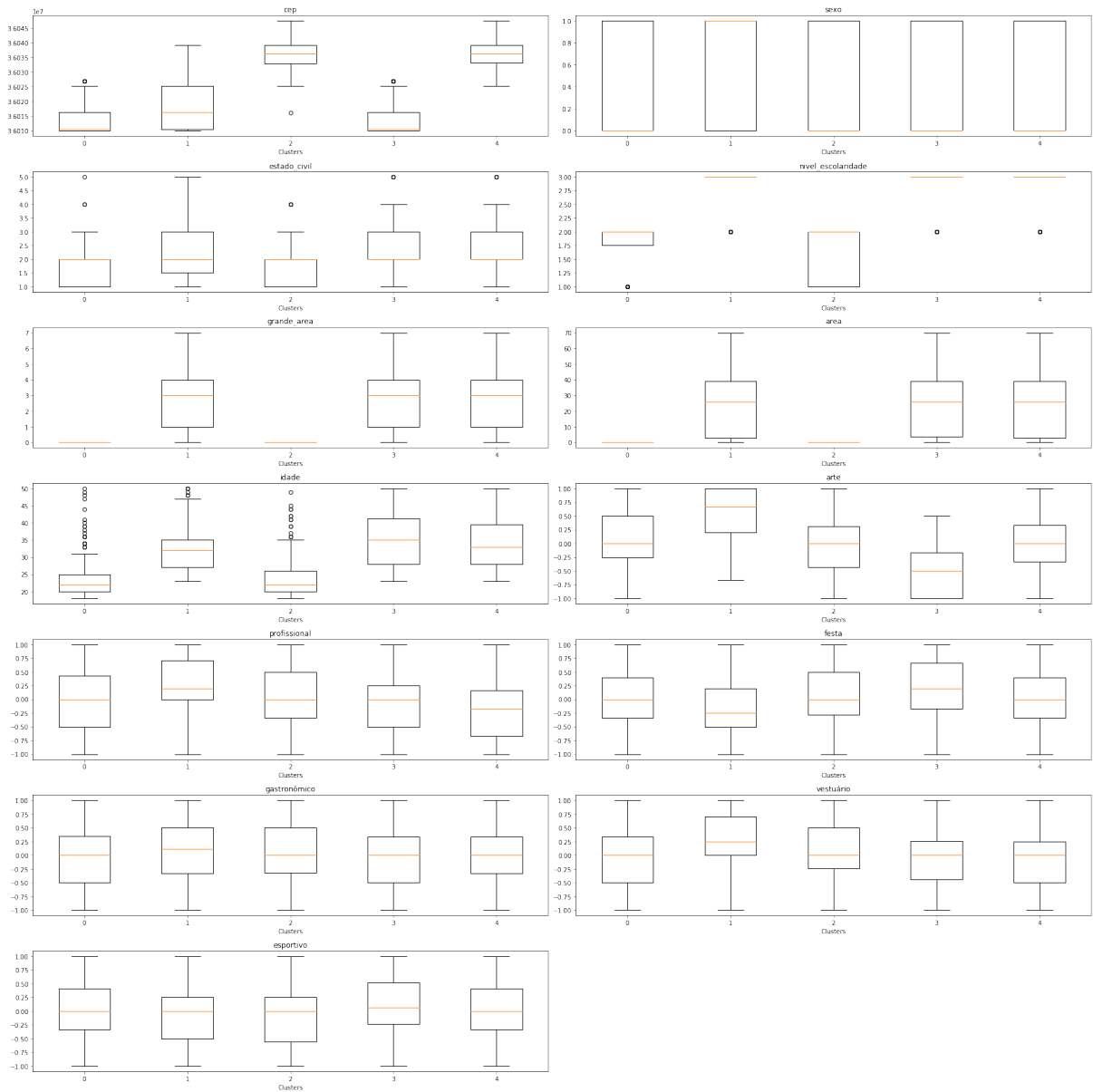


Figura A.10: Boxplots de features para cada cluster de participantes

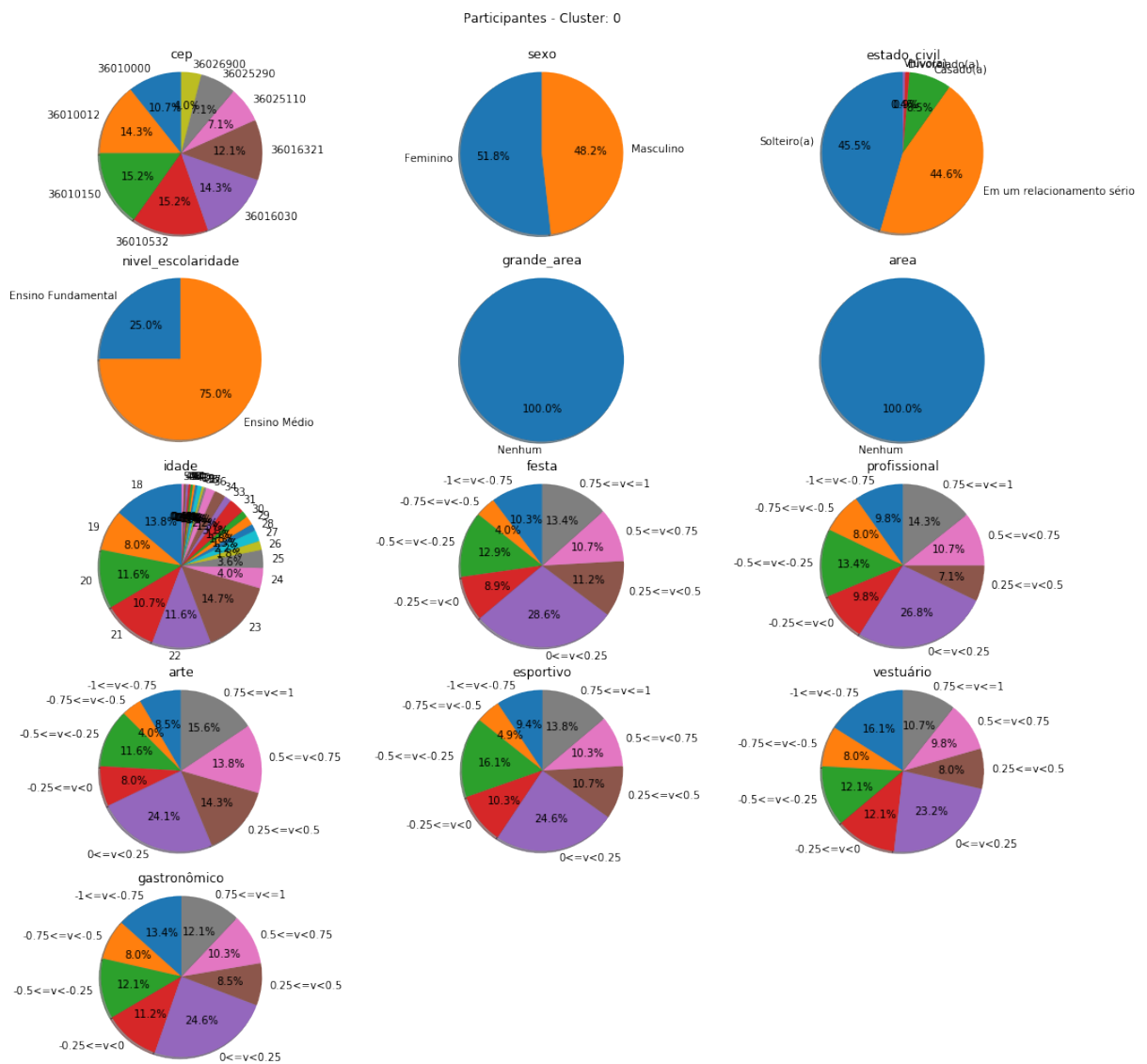


Figura A.11: Representação do cluster 0 de participantes

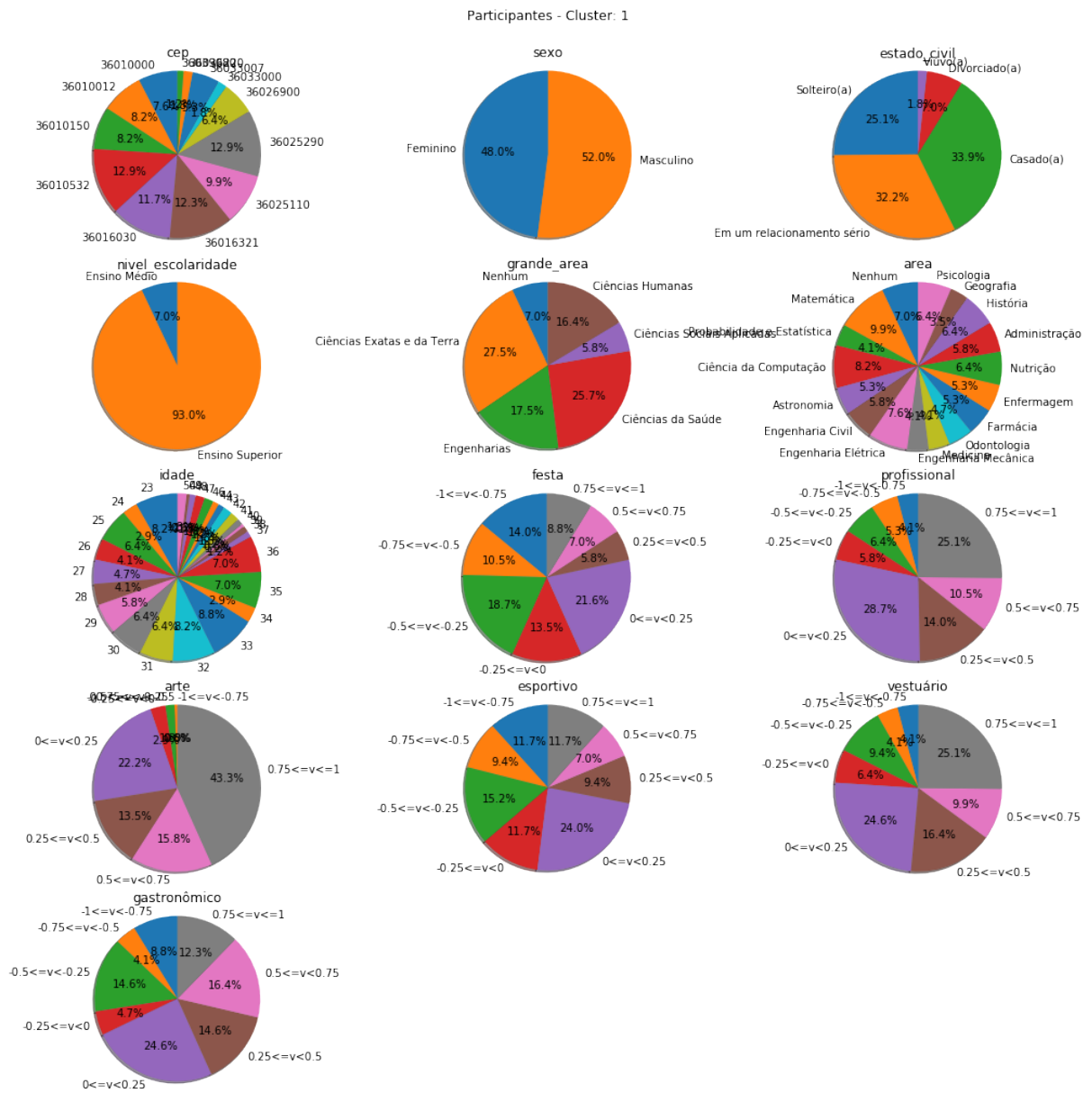


Figura A.12: Representação do cluster 1 de participantes

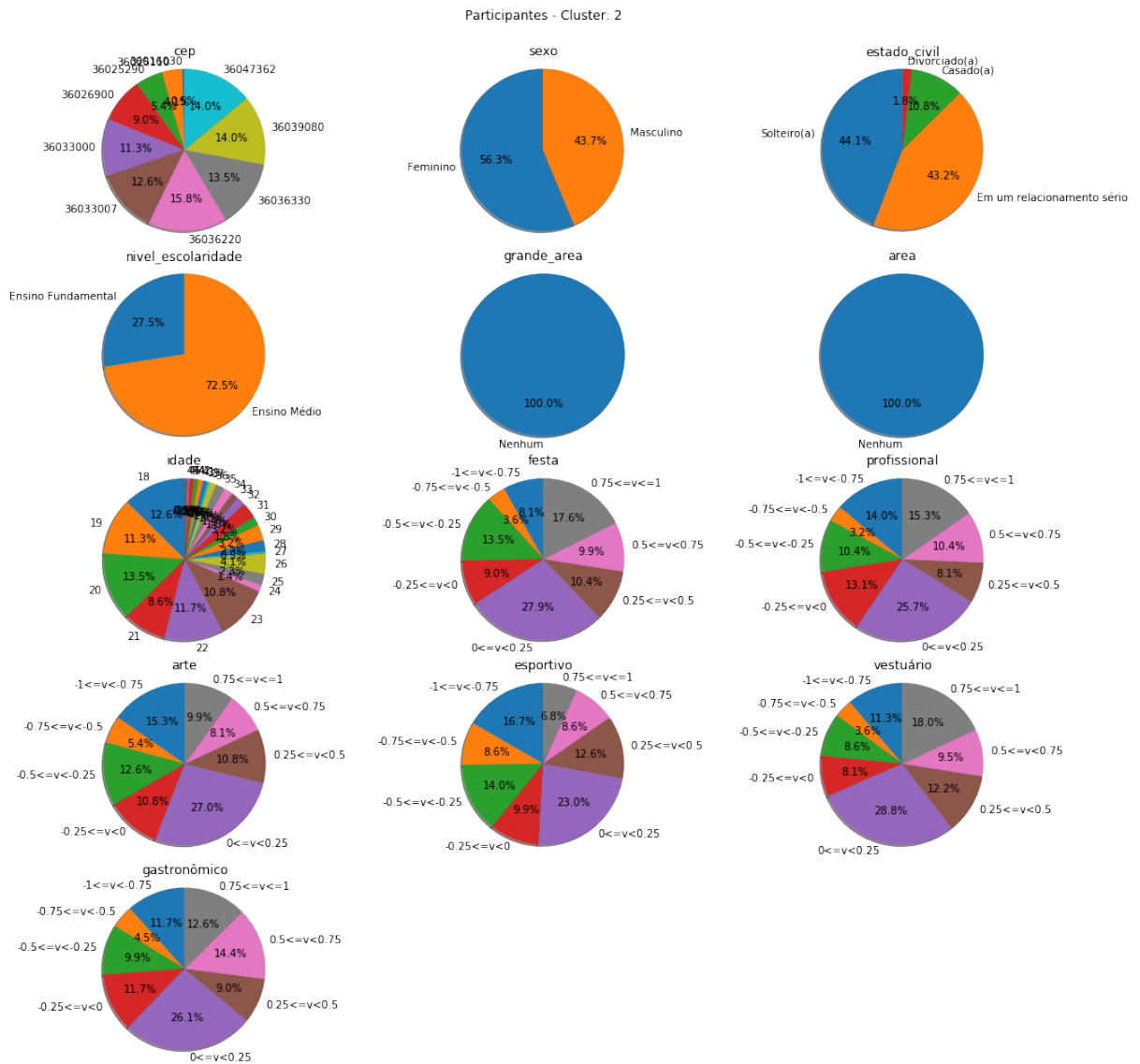


Figura A.13: Representação do cluster 2 de participantes

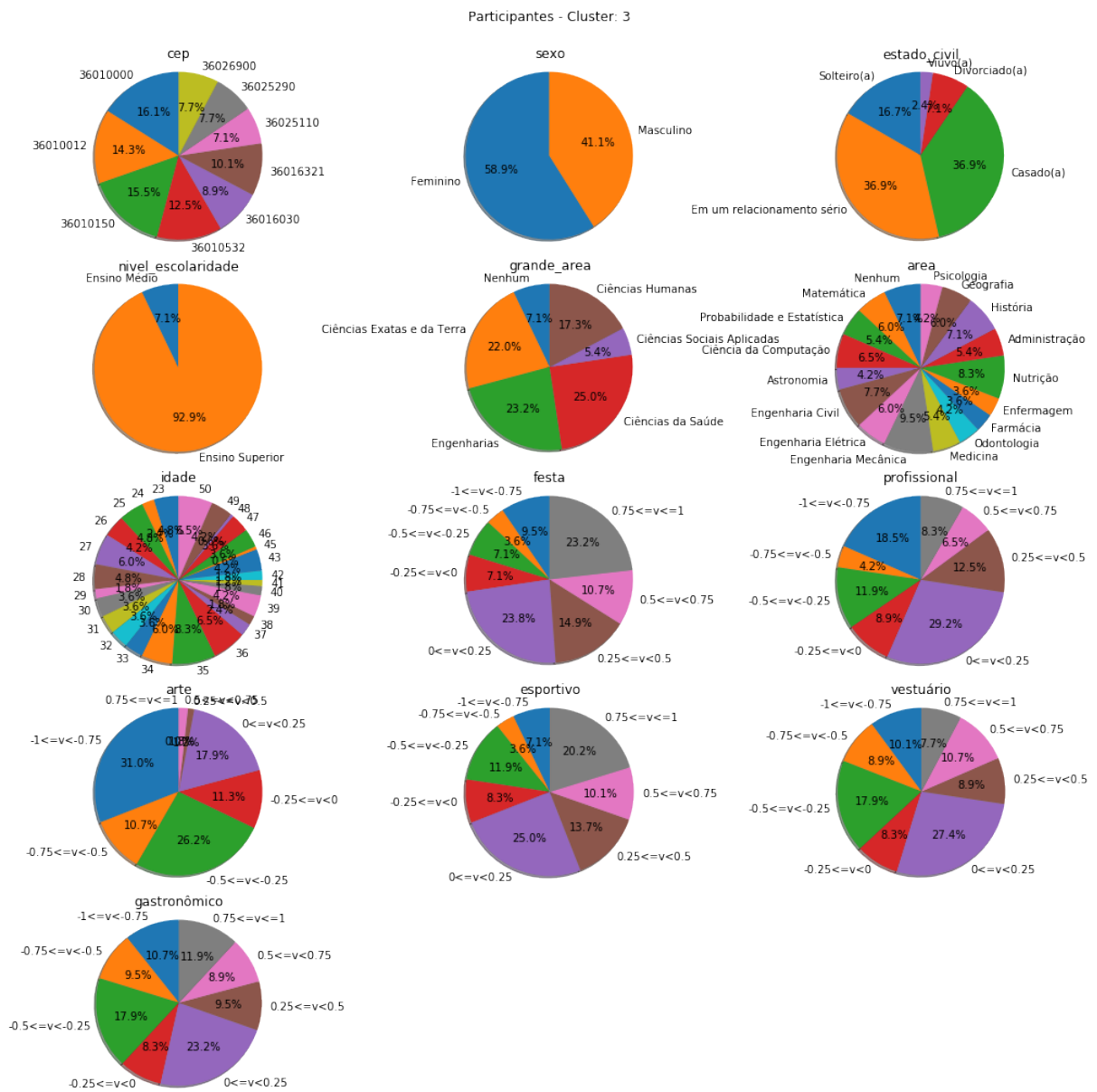


Figura A.14: Representação do cluster 3 de participantes

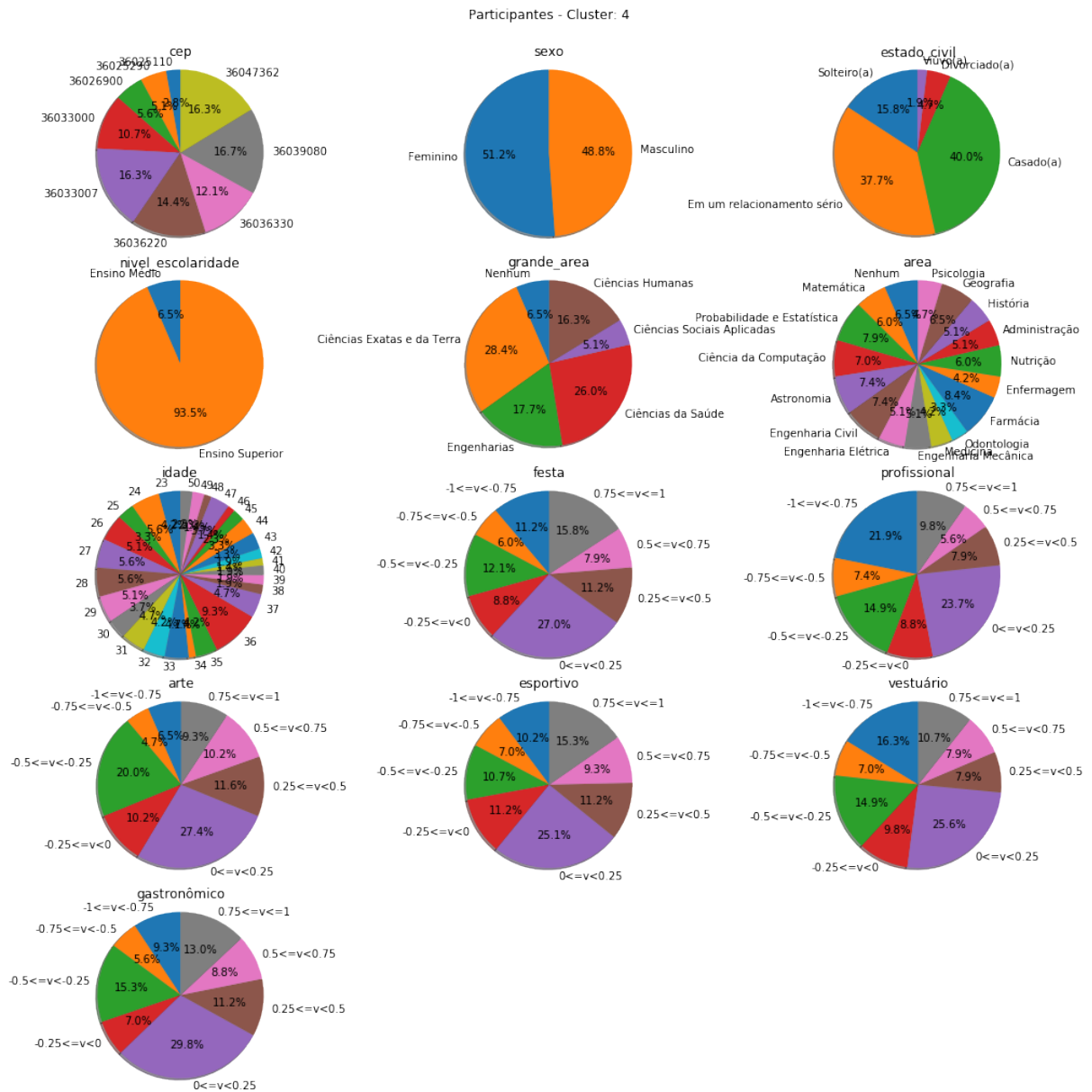


Figura A.15: Representação do cluster 5 de participantes

## B Representação dos dados utilizados para teste

	dia_semana	hora_inicio	feriado	censura	cep	categoria	estilo	genero	knn_cluster	nc_cluster	svm_cluster
0	1	0	0	18	36039080	6	20	0	3	3	3
1	3	21	0	18	36039080	6	18	0	3	3	3
2	5	21	0	16	36039080	6	20	0	3	3	3
3	5	19	0	18	36039080	6	20	0	3	3	3
4	0	20	0	18	36039080	6	19	0	3	3	3
5	6	0	1	16	36026900	5	15	0	3	3	3
6	0	23	0	16	36025290	5	17	0	3	3	3
7	5	21	0	16	36025290	5	15	0	3	3	3
8	3	23	0	0	36039080	5	15	0	3	3	3
9	4	23	1	14	36025290	5	17	0	7	3	3
10	1	20	0	16	36016321	4	12	32	5	5	5
11	1	22	0	16	36016321	4	12	25	5	5	5
12	0	19	1	16	36010532	4	13	33	6	6	6
13	6	22	0	16	36010532	4	13	39	5	5	5
14	5	20	0	18	36016321	4	14	41	5	5	5
15	6	23	0	16	36033007	3	10	0	0	3	0
16	6	22	1	18	36036220	3	9	0	7	7	7
17	0	22	0	12	36036220	3	7	0	3	4	3
18	1	23	0	14	36036220	3	11	0	3	3	3
19	3	20	0	10	36033000	3	9	0	3	3	3
20	4	14	0	12	36036330	2	6	18	0	0	0
21	1	18	0	10	36036330	2	5	13	0	0	0
22	0	13	0	16	36047362	2	6	24	4	4	4
23	6	17	0	18	36036330	2	6	23	0	0	0
24	0	18	0	16	36047362	2	6	19	3	4	4
25	0	0	0	16	36010000	1	2	8	5	5	5
26	0	21	1	16	36016030	1	1	3	7	7	7
27	0	0	1	18	36025110	1	2	9	7	7	7
28	2	22	0	10	36010150	1	1	2	0	0	0
29	2	20	0	10	36010150	1	1	3	0	0	0

Tabela B.1: Dados de teste de eventos já classificados

cep	sexo	estado_civil	nivel_escolaridade	grande_area	area	idade	gastronômico	arte	feira	profissional	esportivo	vestuário	knn_cluster	nc_cluster	svm_cluster
0	36010532	0	3	4	36	36	0.000000	-0.600000	0.200000	1.000000	-0.900000	-0.800000	5	5	5
1	36010150	1	3	4	38	34	0.777778	-0.777778	0.111111	0.888889	-1.000000	0.666667	0	0	0
2	36016321	0	1	4	37	44	-0.444444	-1.000000	-0.888889	0.333333	0.333333	-0.888889	0	0	0
3	36026900	1	3	4	36	39	-0.333333	-0.888889	0.777778	0.333333	-1.000000	0.888889	7	5	5
4	36010532	1	2	4	37	40	-0.700000	-0.300000	1.000000	1.000000	0.800000	-0.900000	5	5	5
5	36010150	1	3	3	32	37	0.888889	0.888889	0.222222	-1.000000	-0.333333	-1.000000	0	0	0
6	36026900	1	2	1	3	27	1.000000	-0.857143	-0.142857	0.428571	-0.857143	1.000000	4	4	4
7	36047362	1	3	7	68	50	-0.875000	1.000000	-0.125000	0.875000	-0.875000	-0.625000	7	7	7
8	36039080	1	2	1	2	37	0.800000	0.800000	1.000000	-0.400000	0.100000	-0.700000	4	7	4
9	36016030	1	3	1	2	31	0.777778	-0.888889	-1.000000	1.000000	1.000000	-0.444444	0	0	0
10	36016321	1	3	3	22	37	0.625000	0.875000	0.375000	0.000000	0.500000	1.000000	5	5	5
11	36010000	1	1	3	69	33	0.800000	0.200000	-0.200000	-0.600000	-1.000000	0.000000	0	0	0
12	36047362	1	1	7	70	37	-1.000000	1.000000	-0.125000	0.750000	0.625000	0.000000	7	7	7
13	36025290	0	2	3	26	28	0.444444	-0.111111	1.000000	-0.888889	-0.777778	0.666667	4	4	4
14	36026900	0	2	3	26	22	0.500000	-0.833333	0.000000	-1.000000	0.833333	0.500000	4	4	4
15	36036220	0	2	4	38	46	-1.000000	-0.888889	0.777778	0.888889	0.000000	0.000000	7	7	7
16	36016030	0	2	1	3	38	1.000000	1.000000	0.555556	0.888889	0.444444	-0.888889	0	5	5
17	36039080	1	2	0	0	44	-0.888889	-0.222222	0.555556	-0.888889	-1.000000	-0.111111	1	1	1
18	36033007	1	1	7	69	39	1.000000	0.800000	0.400000	-0.100000	0.200000	0.100000	4	4	4
19	36010150	0	1	3	22	43	-1.000000	0.400000	1.000000	-0.500000	0.900000	-0.700000	0	0	0
20	36010012	1	3	4	36	25	0.555556	-1.000000	-0.444444	-0.111111	-0.666667	-0.333333	0	0	0
21	36047362	1	1	0	29	29	0.250000	0.250000	0.625000	-1.000000	-0.625000	0.500000	8	8	8
22	36033007	0	1	1	1	49	0.285714	1.000000	-0.285714	-0.857143	0.428571	1.000000	4	7	7
23	36036330	0	1	3	35	36	-0.333333	0.500000	0.166667	0.333333	-0.500000	-1.000000	4	4	4
24	36036220	0	3	4	37	48	-0.800000	-0.500000	-0.600000	-0.300000	-0.300000	-1.000000	7	7	7
25	36016030	1	3	4	35	36	0.500000	-0.500000	-0.500000	-0.200000	-1.000000	-1.000000	5	5	5
26	36047362	0	1	4	37	48	-0.666667	0.333333	0.777778	0.111111	1.000000	0.888889	7	7	7
27	36025110	0	2	3	38	39	0.000000	-0.600000	0.600000	0.900000	1.000000	-0.900000	5	5	7
28	36033000	0	1	1	2	28	0.571429	1.000000	0.285714	0.285714	0.285714	-0.85714	4	4	4
29	36010532	1	2	3	1	28	-0.700000	0.500000	1.000000	0.200000	-0.900000	-0.400000	0	0	0
30	36010150	1	3	3	25	24	-1.000000	-0.500000	0.500000	-0.625000	0.000000	-0.875000	0	0	0
31	36016321	0	3	0	0	40	0.900000	-0.900000	-0.700000	-1.000000	0.800000	-0.600000	6	6	6
32	36025290	0	3	1	0	37	0.700000	-0.700000	0.100000	0.900000	-1.000000	-0.800000	1	1	1
33	36036220	0	1	0	0	43	-0.400000	0.900000	0.200000	-0.600000	1.000000	-0.200000	8	8	8
34	36010000	0	2	4	35	34	1.000000	1.000000	0.111111	0.555556	-0.777778	-0.777778	0	0	0
35	36010000	0	1	0	0	34	0.625000	0.500000	0.625000	0.625000	-0.125000	-1.000000	2	2	2
36	36036220	1	2	7	68	35	-0.857143	-1.000000	0.000000	-0.285714	0.714286	-0.714286	7	7	7
37	36039080	1	1	7	69	25	1.000000	-1.000000	-0.400000	0.600000	0.100000	-0.600000	4	4	4
38	36010000	0	3	4	36	36	0.700000	-1.000000	0.600000	-0.100000	1.000000	-0.800000	5	5	5
39	36036220	1	1	0	0	28	-0.333333	1.000000	0.166667	-0.833333	0.166667	-0.500000	3	3	3
40	36026900	0	2	3	22	30	0.700000	0.700000	0.600000	0.000000	1.000000	-0.300000	4	4	4
41	36010532	1	3	3	22	28	1.000000	-0.444444	-0.333333	0.222222	-0.333333	-0.444444	0	5	5
42	36016030	0	3	3	26	29	0.600000	0.500000	0.100000	-1.000000	-1.000000	-0.200000	5	5	5
43	36033007	0	1	3	69	47	0.555556	-0.222222	-0.444444	1.000000	0.111111	-0.444444	7	7	7
44	36010532	1	2	6	52	48	0.125000	0.125000	-1.000000	-0.375000	-0.875000	-0.375000	5	5	5
45	36036220	1	3	4	39	42	0.555556	-0.444444	0.000000	-0.444444	0.555556	1.000000	7	7	7
46	36036330	1	1	7	68	35	0.900000	-1.000000	1.000000	1.000000	0.000000	1.000000	4	4	4
47	36039080	1	1	3	2	36	-1.000000	-0.875000	0.250000	-0.500000	0.625000	-0.625000	4	4	4
48	36010000	0	3	3	37	25	1.000000	1.000000	0.000000	-0.800000	0.200000	0.400000	0	5	0
49	36025290	0	1	0	0	32	0.200000	-0.800000	-0.500000	-0.500000	-0.600000	-1.000000	2	8	8

Tabela B.2: Dados de teste de participantes já classificados - parte 1



cep	sexo	estado_civil	nivel_escolaridade	grande_area	area	idade	gastro_nômico	arte	festa	profissional	esportivo	vestuário	knn_cluster	nc_cluster	svm_cluster
50	36047362	0	1	2	22	40	0.800000	0.400000	1.000000	0.900000	-0.600000	0.900000	4	7	7
51	36029000	1	2	0	0	39	0.555556	0.111111	-0.666667	0.777778	0.777778	0.777778	1	1	1
52	36029290	1	1	3	22	35	0.600000	-1.000000	-1.000000	0.400000	0.000000	0.000000	4	4	4
53	36036330	0	1	1	0	23	0.300000	-0.400000	0.300000	-0.400000	-0.400000	-0.200000	8	8	8
54	36010000	0	1	0	0	28	0.500000	-0.125000	0.750000	1.000000	0.000000	0.000000	2	2	2
55	36016321	1	1	3	4	36	0.444444	-0.222222	0.777778	0.777778	-1.000000	0.111111	0	0	0
56	36047362	0	2	3	26	35	0.500000	1.000000	0.750000	0.750000	0.750000	-0.750000	4	7	7
57	36016030	1	1	3	68	24	0.700000	-0.600000	0.600000	-1.000000	0.400000	0.400000	0	0	0
58	36039080	1	1	3	35	22	0.900000	-1.000000	0.900000	-0.600000	-0.600000	-0.600000	4	4	4
59	36010150	0	2	0	32	0	0.555556	0.777778	1.000000	-1.000000	0.000000	0.000000	6	6	2
60	36016321	0	2	7	70	42	0.857143	0.857143	1.000000	0.714286	-1.000000	1.000000	5	5	5
61	36016030	1	2	3	26	38	-1.000000	-0.333333	-0.444444	-0.222222	-1.000000	-0.777778	0	5	0
62	36036330	0	1	1	0	31	0.600000	0.400000	0.700000	-1.000000	-0.800000	0.400000	8	8	8
63	36010150	0	1	3	3	32	0.000000	1.000000	0.666667	1.000000	0.666667	-0.444444	0	0	0
64	36039080	1	1	3	25	43	-0.285714	0.571429	0.857143	1.000000	0.714286	0.857143	4	7	7
65	36010012	0	1	3	35	33	-0.428571	0.857143	-0.285714	1.000000	-0.571429	0.571429	0	0	0
66	36039080	0	2	3	37	40	-0.857143	0.428571	-1.000000	0.285714	0.000000	0.000000	7	7	7
67	36036220	1	1	3	67	47	-1.000000	-0.375000	0.375000	-1.000000	-0.125000	-0.125000	7	7	7
68	36036220	1	1	3	39	30	-0.700000	-0.600000	1.000000	-0.600000	-0.200000	-0.200000	4	4	4
69	36047362	1	3	3	2	25	-0.857143	0.571429	-0.428571	-1.000000	0.285714	0.714286	4	7	7
70	36016321	1	1	1	0	26	-0.833333	0.500000	0.500000	1.000000	0.333333	0.333333	2	2	2
71	36036330	0	3	0	0	37	0.000000	-0.500000	0.000000	0.000000	-1.000000	-1.000000	1	1	1
72	36047362	0	2	3	3	31	0.100000	-1.000000	0.700000	0.400000	0.900000	0.400000	7	7	7
73	36039080	0	2	3	25	22	-1.000000	0.200000	0.200000	0.600000	0.000000	-0.100000	4	4	4
74	36010532	1	2	3	37	28	0.750000	-1.000000	-0.875000	-0.375000	-0.875000	-1.000000	0	0	0
75	36047362	1	1	4	35	26	-0.222222	-0.222222	-1.000000	-0.555556	-0.222222	-0.666667	4	4	4
76	36029110	0	3	1	2	45	0.428571	-0.285714	0.285714	1.000000	0.857143	0.857143	5	5	5
77	36016321	0	2	3	26	47	0.888889	0.555556	0.555556	0.000000	-0.222222	1.000000	5	5	5
78	36047362	0	2	3	25	29	-0.500000	0.300000	-0.800000	-0.200000	-1.000000	1.000000	7	7	4
79	36010532	1	1	3	26	44	-0.888889	-0.222222	-0.888889	1.000000	-0.333333	-0.888889	0	0	0
80	36010012	1	2	3	38	42	-0.200000	-0.700000	0.400000	0.000000	0.000000	0.700000	5	5	5
81	36010150	0	2	3	25	30	0.333333	-0.222222	1.000000	-0.666667	-1.000000	0.666667	0	0	0
82	36039080	0	1	3	70	36	-0.333333	-0.777778	-0.555556	-0.222222	-1.000000	-0.555556	4	4	4
83	36010532	1	2	3	2	44	-0.400000	-0.100000	1.000000	-0.100000	0.600000	-0.400000	5	5	5
84	36010012	1	3	3	35	41	-0.166667	0.666667	1.000000	-0.833333	-0.333333	0.166667	5	5	5
85	36010532	0	3	1	0	31	-1.000000	-0.200000	-1.000000	-1.000000	0.600000	0.600000	6	6	6
86	36010012	0	3	6	52	46	1.000000	-0.571429	0.857143	-0.428571	0.571429	0.571429	5	5	5
87	36036330	0	2	3	70	50	-0.700000	1.000000	0.300000	0.500000	0.700000	0.200000	7	7	7
88	36033007	1	2	3	35	43	1.000000	-0.666667	-0.500000	0.000000	-0.666667	-0.666667	7	7	7
89	36010532	0	1	3	25	25	1.000000	-0.600000	0.200000	0.100000	-0.900000	-0.600000	0	0	0
90	36016321	0	1	3	4	46	-1.000000	-0.800000	1.000000	-0.200000	0.300000	0.300000	0	0	0
91	36016030	0	1	3	26	35	-1.000000	-0.888889	0.222222	0.666667	-0.111111	1.000000	0	0	0
92	36047362	0	2	0	0	28	-0.333333	-0.111111	1.000000	0.444444	-0.555556	-0.555556	3	1	3
93	36029290	1	2	3	26	40	-0.200000	0.400000	-1.000000	0.400000	0.000000	0.000000	7	7	7
94	36016321	0	3	6	52	38	0.777778	0.000000	-0.333333	0.200000	-0.444444	-0.444444	5	5	5
95	36010532	1	3	3	38	48	0.100000	-0.500000	-1.000000	0.000000	-0.800000	-0.800000	5	5	5
96	36010150	0	2	3	68	36	0.500000	-0.200000	1.000000	0.600000	-0.400000	-0.300000	0	5	0
97	36047362	0	1	3	39	28	0.600000	-1.000000	-0.800000	-0.200000	-0.600000	-0.600000	4	4	4
98	36010012	0	2	3	1	4	0.200000	0.800000	-0.900000	0.200000	-1.000000	-0.100000	5	5	5
99	36033000	1	1	3	69	43	1.000000	-0.300000	0.800000	0.800000	0.600000	0.600000	4	7	7

Tabela B.3: Dados de teste de participantes já classificados - parte 2

## C Diagrama ER do banco de dados

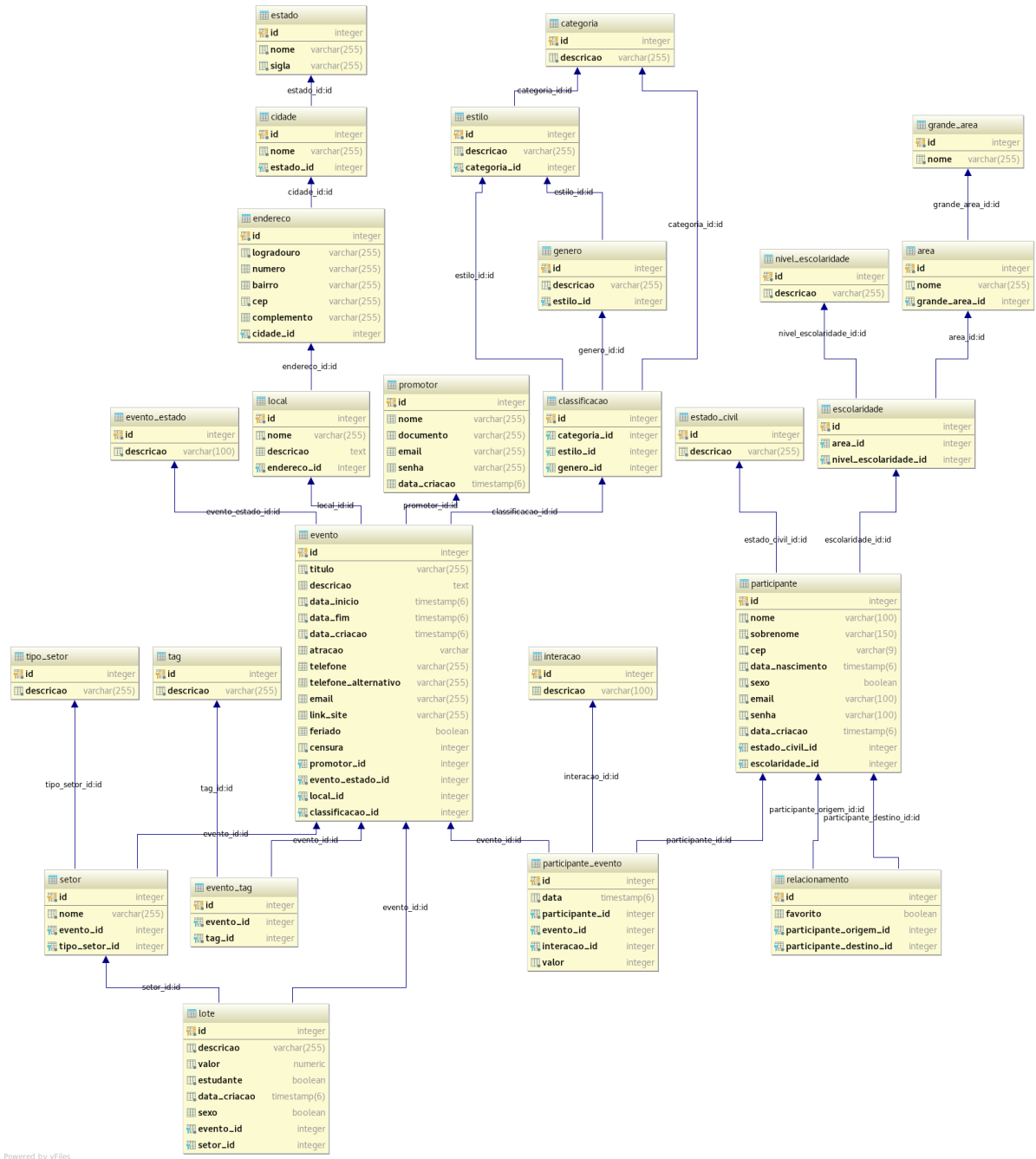
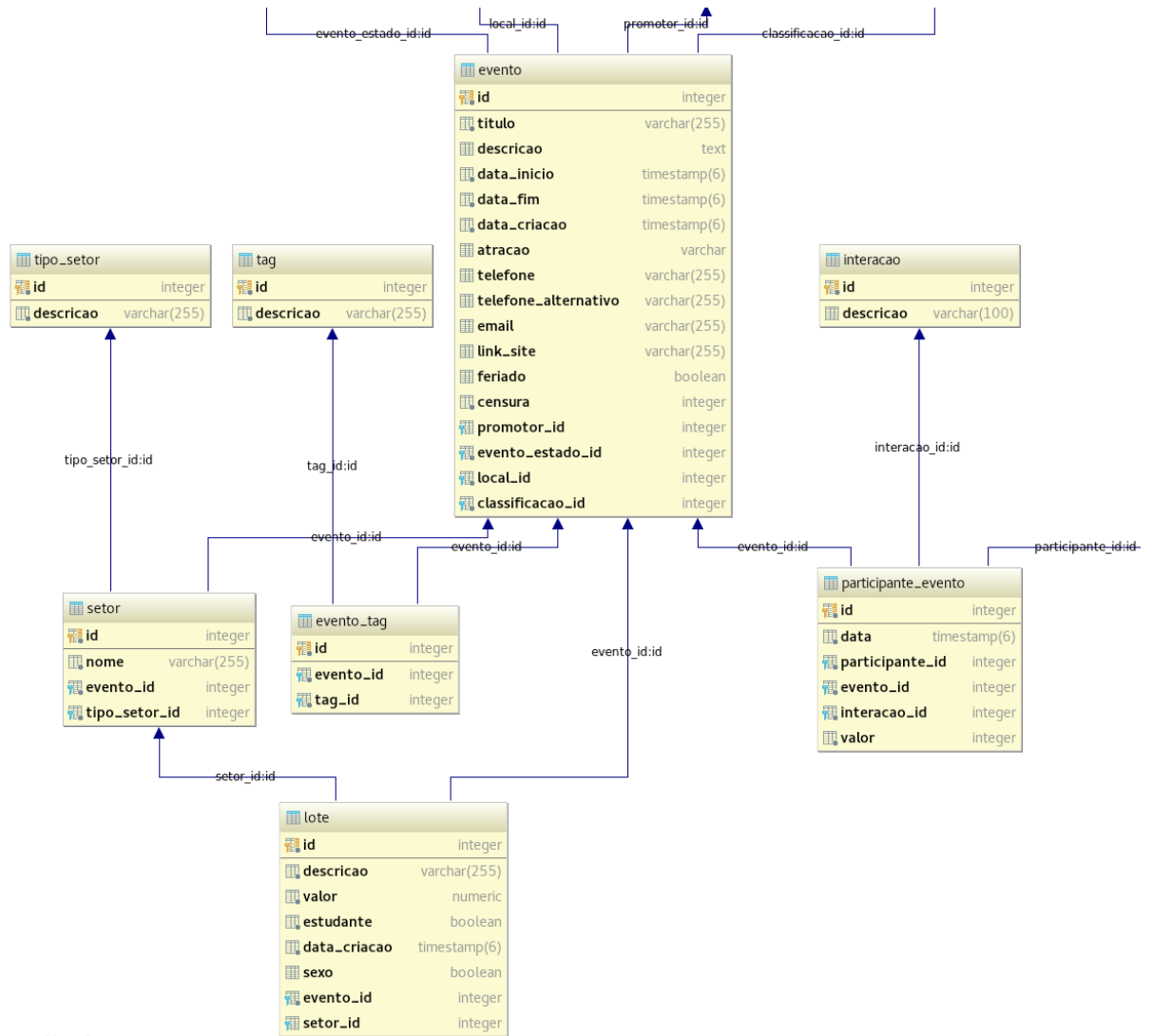


Figura C.1: Diagrama Entidade Relacionamento do banco de dados desenvolvido para o sistema

Para uma melhor visualização do diagrama do banco de dados foram feitos alguns recortes que são exibidos a seguir.



Powered by yFiles

Figura C.2: Recorte do diagrama com foco na tabela evento.

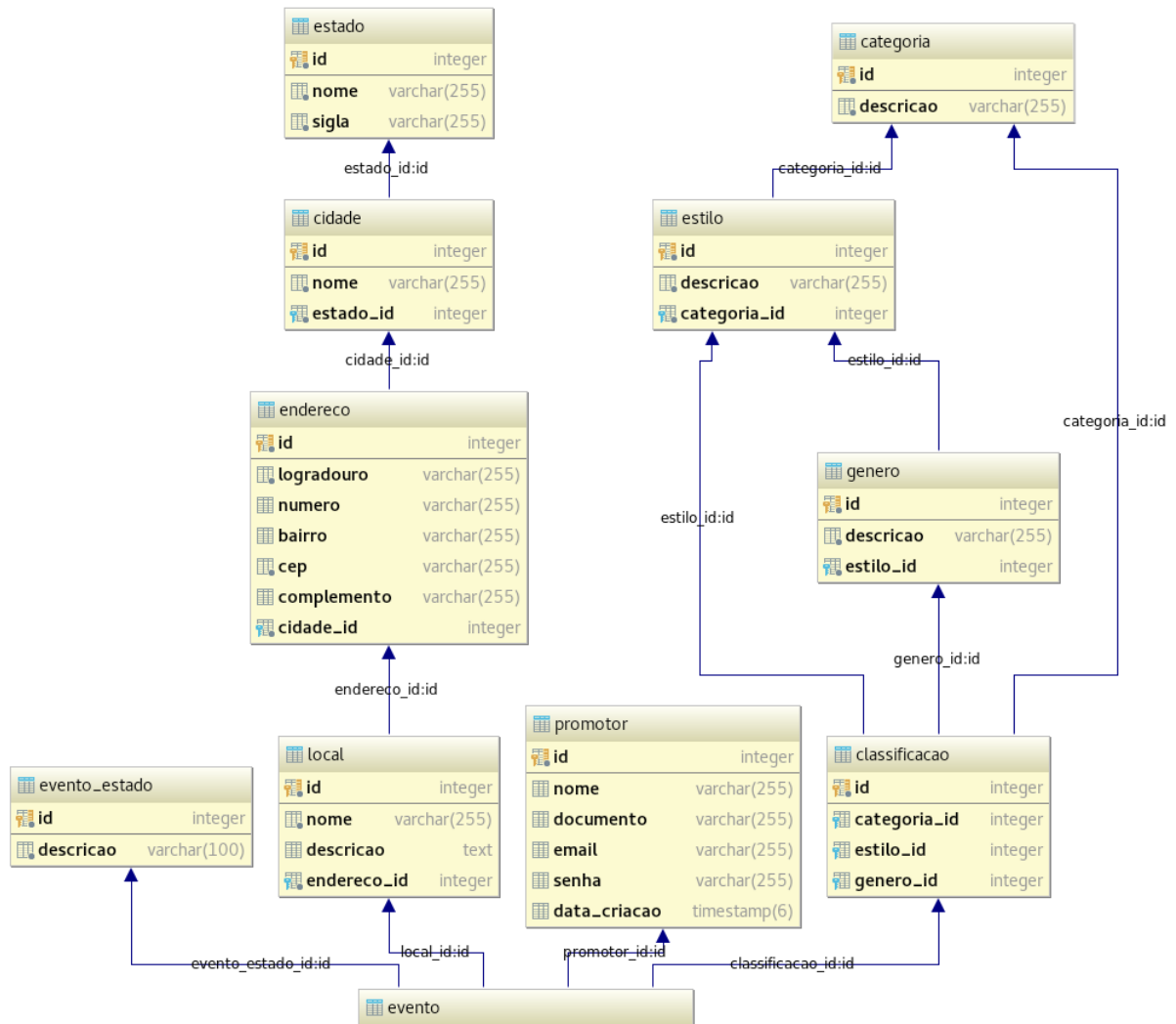


Figura C.3: Recorte do diagrama com foco em tabelas associadas a tabela evento.

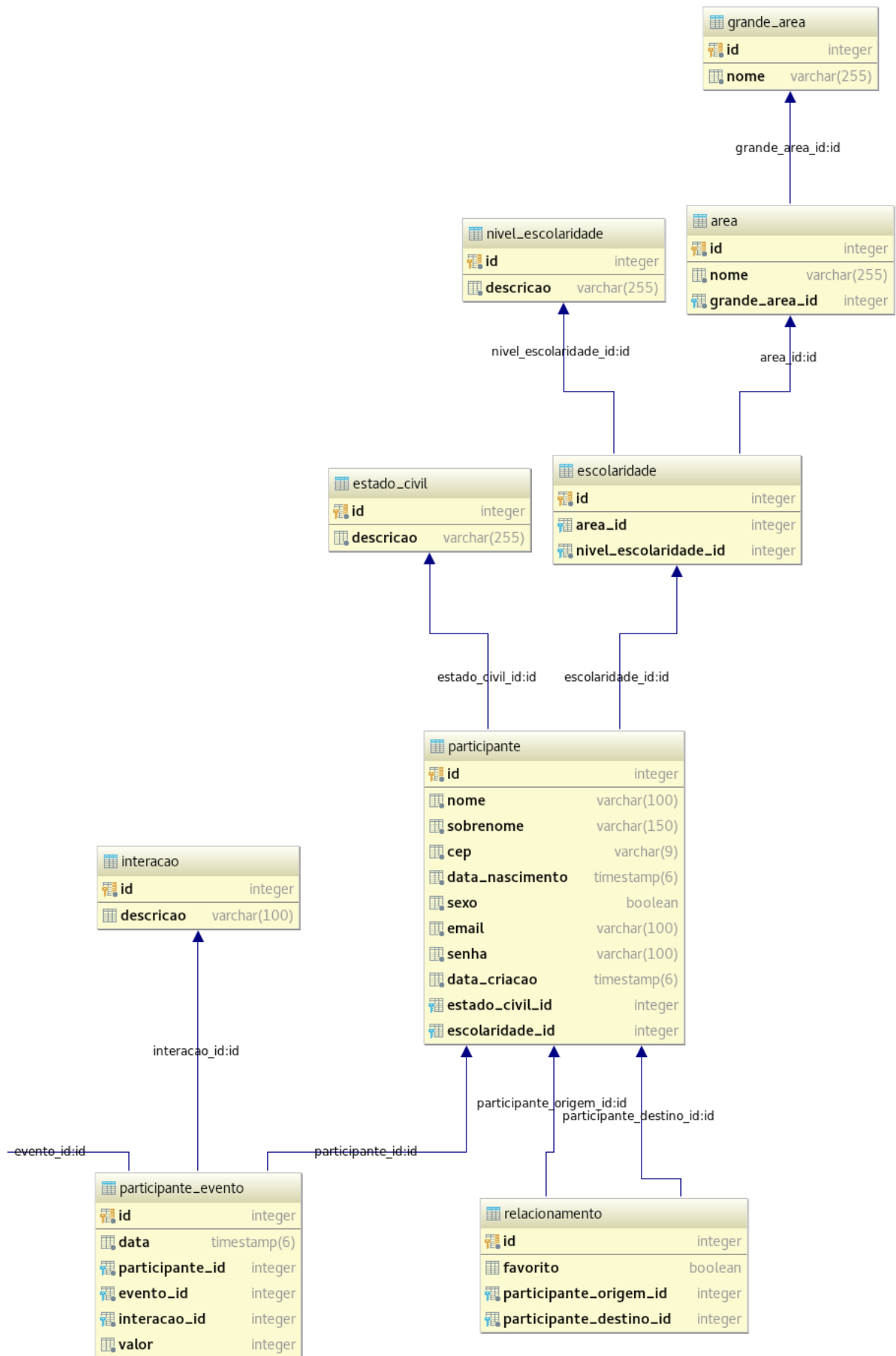


Figura C.4: Recorte do diagrama com foco na tabela participante.