Projeto de Formatura – Turmas 2017

PCS - Departamento de Engenharia de Computação e Sistemas Digitais

Engenharia de Computação

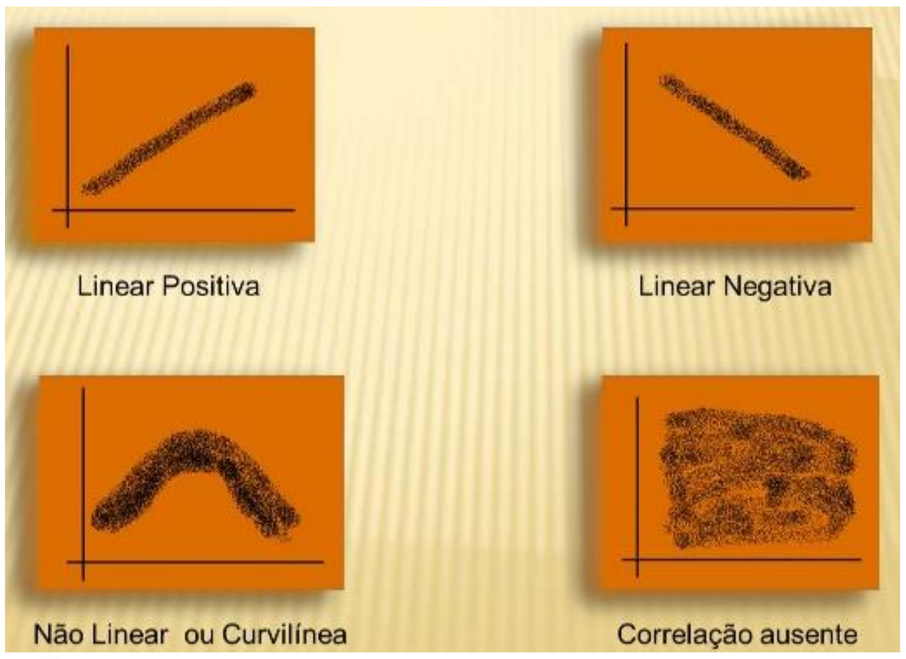
Tema:

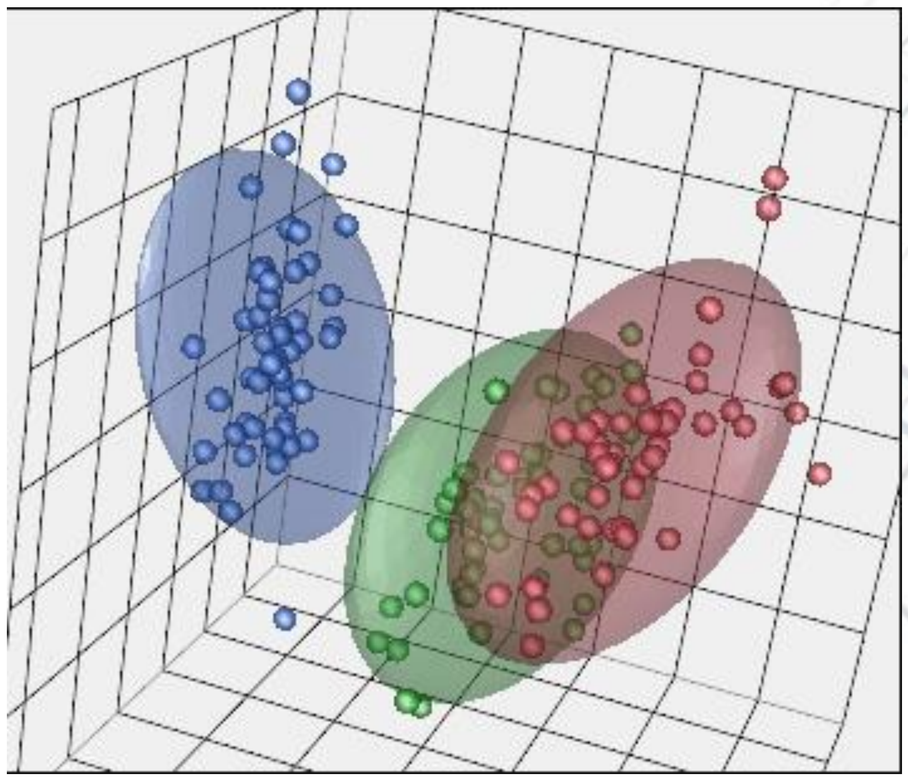
Predição de Ação e Classificação de Usuários em Jogos Mobile

O objetivo deste projeto de formatura consistiu de verificar o impacto de medidas de correlação estatística em modelos preditivos de ação de usuários em jogos mobile, construídos por meio do uso de algoritmos de aprendizado de máquina, e também de validar a construção de um modelo de classificação de usuários de acordo com seu perfil de comportamento, com base em algoritmos de clusterização de dados.

Correlação é definida como o estudo do grau de associação entre duas variáveis. O coeficiente de correlação é uma medida padronizada entre duas variáveis, cujo objetivo é indicar a força e a direção do relacionamento entre essas duas variáveis aleatórias, ou seja, fornecer informações sobre o tipo e a extensão do relacionamento entre duas variáveis.

Dessa forma, para cumprir o objetivo do projeto, foi necessário investigar qualquer tipo de correlação entre as variáveis desejadas, não somente a correlação linear. Para tal, foi escolhida a correlação de distância, uma medida de dependência estatística entre duas variáveis ou vetores, que possui valor entre 0 e 1, 0 indicando que as variáveis são independentes, e 1 indicando que as variáveis são dependentes, mas não necessariamente linearmente, ao contrário de, por exemplo, a correlação de Pearson, e também não necessariamente seguindo alguma outra relação matemática mais usual, como, por exemplo, correlação quadrática ou correlação exponencial.



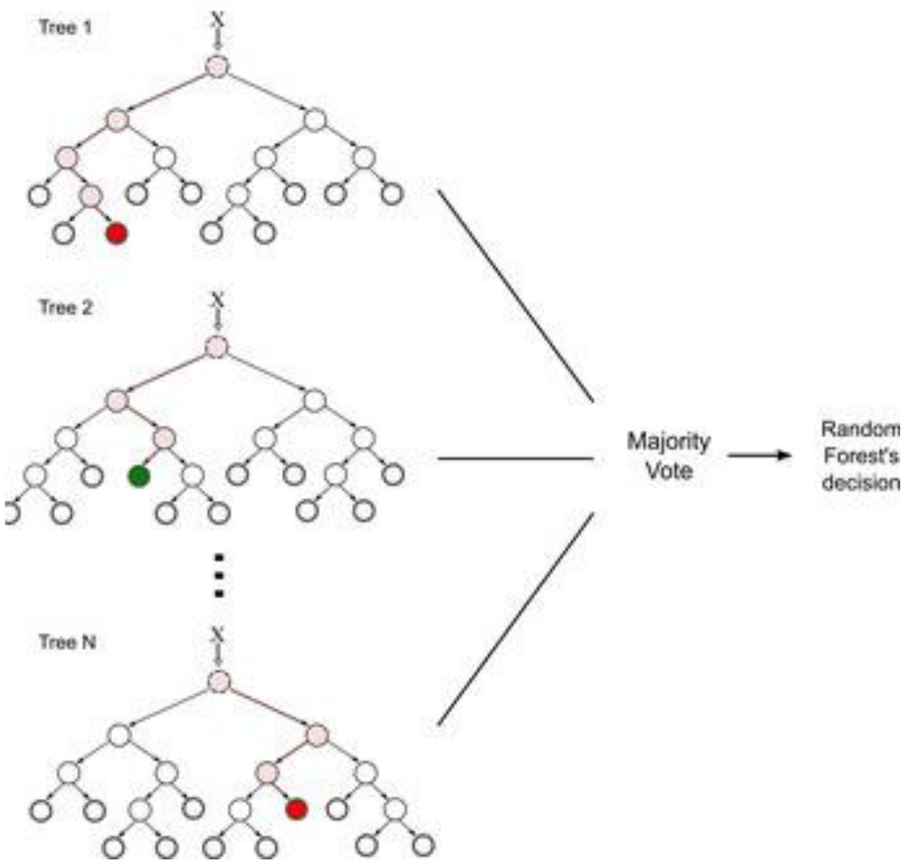


Clusterização é uma análise de agrupamentos, ou seja, uma análise na qual constroem-se clusters, coleções de objetos similares entre si, e dissimilares aos objetos de outros clusters. Neste sentido, clusterização é um método não supervisionado de machine learning, visto que não possui classes predefinidas.

Neste projeto, foi utilizada a técnica de Análise de Componentes Principais para reduzir a dimensionalidade dos dados relacionados ao perfil de comportamento dos usuários, assim como definir quais dimensões causam maior variância dos dados e, portanto, possuem maior importância durante a clusterização. Feito isso, foi usado o algoritmo de HDBSCAN (Hierarchical Density-Based Spatial Clustering of Applications with Noise), que performa clusterização de dados de acordo com a proximidade dos pontos, e de forma hierárquica, de forma a encontrar agrupamentos de usuários de comportamento semelhante.

De forma a predizer os valores das métricas de interesse, e, com isso, predizer ações de usuários escolheu-se utilizar de métodos supervisionados de machine learning, ideais para quando deseja-se descobrir uma função mapeadora de uma entrada para uma saída, de forma a aproximá-la tanto que, quando houver novas entradas, será possível prever suas saídas a partir dos resultados das medições anteriores.

Sob esta ótica, o método supervisionado de machine learning Random Forest, ou Random Decision Trees, consiste de um método de classificação e regressão, que opera por meio da construção de grande número de árvores de decisão, cada uma sob uma fração aleatória dos dados, o que resolve o problema de “overfitting”, com boa acurácia, pouca preparação prévia dos dados, e, principalmente, boa escalabilidade com o volume de dados de entrada.



Integrantes: Stefano Hellebrekers Seravalli stefanohellebrekers@gmail.com

Professor Orientador: Ricardo Luís de Azevedo Rocha rlarocha@usp.br

Co-orientador: