جامعة محمد الخامس بالرباط
Université Mohammed V de Rabat

École Nationale Supérieure d'Informatique et d'Analyse des Systèmes
Centre d'Études Doctorales en Sciences des Technologies de l'Information et de l'Ingénieur

AVIS DE SOUTENANCE DE THESE DE DOCTORAT

Madame Houda BENHAR

soutiendra publiquement sa thèse de Doctorat en Informatique

Le Samedi 05 Mars 2022 à 10H au Grand amphi à l'ENSIAS

Intitulé de la thèse

DATA PREPROCESSING IN HEART DISEASE KNOWLEDGE DISCOVERY

UNIVERSITE MOHAMMED V - RABAT
ENSIAS

Devant le Jury composé de :

Président :
Pr. Mohamed ESSAAIDI, PES, ENSIAS, Université Mohammed V de Rabat
Directeur de thèse :
Pr. Ali IDRI, PES, ENSIAS, Université Mohammed V de Rabat
Rapporteurs :
Pr. Imade BENELALLAM, PES, INSEA, Rabat
Pr. Mohammed BENKHALIFA, PES, FSR, Université Mohammed V de Rabat
Pr. Fatima Azzahra AMAZAL, PH, Faculté des Sciences, Université Ibn Zohr d'Agadir
Examinateur :
Pr. Laila CHEIKHI, PH, ENSIAS, Université Mohammed V de Rabat
Invité :
Pr. Mohamed HOSNI, PESA, ENSAM, Université Moulay Ismail de Meknès

DATA PREPROCESSING IN HEART DISEASE KNOWLEDGE DISCOVERY

Abstract: The increasing amount of data produced by various biomedical and healthcare systems has led to a need for methodologies related to knowledge discovery in databases. Data mining is the mathematical core of KDD offering a set of powerful techniques that allow the identification and extraction of meaningful information, patterns, associations or relationships embedded in large data sets. DM can greatly benefit doctors and patients, particularly in the case of diseases with high mortality and morbidity rates, such as heart disease. Nevertheless, extracting useful information from raw medical data is a challenging task. In fact, medical datasets are, in general, composed of missing values and noise, along with inconsistent, irrelevant, redundant and high-dimensional data. This might have a significant impact on the analysis and interpretation of data, thus leading to suboptimal decisions or hindering the research outcomes derived. A rigorous preparation of data is, therefore, required to deal with data imperfectness in order to improve its quality so as to satisfy the requirements and improve the performances of DM techniques.

This thesis aims to: First, perform a systematic map of studies regarding the application of data preparation for KDD in medicine to summarize the available research in this area. Second, based on the findings of the first systematic map, conduct a systematic map and literature review study to review and summarize the current evidence on the use of data preprocessing in heart disease prediction. Third, investigate the use of univariate filter methods with different threshold choices for feature selection for heart disease classification. Fourth, evaluate and compare the performances of univariate and multivariate filter feature selection, and hyperparameters' optimization on HD classification. Fifth, propose a new ensemble feature selection approach based on univariate filters for HD classification.

Most chapters of this thesis have been disseminated in international journals and highly refereed conference proceedings.

Keywords: Data mining, data preprocessing, cardiac datasets, heart disease, classification techniques, feature selection.