Population Health Management exploiting Machine Learning Algorithms to identify High-Risk Patients

INTRODUCTION OF PROBLEM

- Nowadays, the population in industrialized countries is getting older and older and the number of people aged 65+ years is expected to grow over the next decades, becoming around the 30% of the overall population by the 2060.

  Additionally, increases of more than 50% are projected in the number of older people affected by most relevant individual diseases (e.g. hypertension, diabetes, stroke, respiratory, etc.) and multi morbidities over the next 20 years [1]. The provision and funding of the health care services for this growing group of "complex" patients, with one or more chronic conditions, have become an important challenge.

ABSTRACT

- **Population aging and the increase of chronic conditions incidence and prevalence produce a higher risk of hospitalization or death. This is particularly high for patients with multimorbidity leading to a great consumption of resources.**
- **Identifying as soon as possible high-risk patients becomes an important challenge to improve health care service provision and to reduce costs.**
- **Nowadays, population health management, based on intelligent models, can be used to assess the risk and identify these "complex" patients. The aim of this study is to validate machine learning algorithms (Naïve Bayes, Cart, C5.0, Conditional Inference Tree, Random Forest, Artificial Neural Network and LASSO) to predict the risk of hospitalization or death starting from administrative and socio-economic data. The study involved the residents in the Local Health Unit of Central Tuscany.**

LITERATURE REVIEW

1. PRIVACY-PRESERVING SVM USING NONLINEAR KERNELS ON HORIZONTALLY PARTITIONED DATA.

- Traditional Data Mining and Knowledge Discovery algorithms assume free access to data, either at a centralized location or in federated form.
- Increasingly, privacy and security concerns restrict this access, thus derailing data mining projects. What we need is distributed knowledge discovery that is sensitive to this problem.
- This paper proposes a privacy-preserving solution for support vector machine (SVM) classification, PP-SVM for short. Our solution constructs the global SVM classification model from the data distributed at multiple parties, without disclosing the data of each party to others

2. ON THE DESIGN AND QUANTIFICATION OF PRIVACY PRESERVING DATA MINING ALGORITHMS

- A recently proposed technique addresses the issue of privacy preservation by perturbing the data and reconstructing distributions at an aggregate level in order to perform the mining.
- This method is able to retain privacy while accessing the information implicit in the original attributes. The distribution reconstruction process naturally leads to some loss of information which is acceptable in many practical situations.
- This paper discusses an Expectation Maximization (EM) algorithm for distribution reconstruction which is more effective than the currently available method in terms of the level of information loss.
- Specifically, we prove that the EM algorithm converges to the maximum likelihood estimate of the original distribution based on the perturbed data. We show that when a large amount of data is available, the EM algorithm provides robust estimates of the original distribution.

3. DATA PRIVACY THROUGH OPTIMAL K-ANONYMIZATION

- Data de-identification reconciles the demand for release of data for research purposes and the demand for privacy from individuals.
- This paper proposes and evaluates an optimization algorithm for the powerful de-identification procedure known as k-anonymization.
- A k-anonymized dataset has the property that each record is indistinguishable from at least k - 1 others. Even simple restrictions of optimized k-anonymity are NP-hard, leading to significant computational challenges.
- We present a new approach to exploring the space of possible anonymizations that tames the combinatorics of the problem, and develop data-management strategies to reduce reliance on expensive operations such as sorting.
- Through experiments on real census data, we show the resulting algorithm can find optimalk-anonymizations under two representative cost measures and a wide range of k.

4. PRIVACY PRESERVING NAIVE BAYES CLASSIFIER FOR HORIZONTALLY PARTITIONED DATA

- The problem of secure distributed classification is an important one.
- In many situations, data is split between multiple organizations.
- These organizations may want to utilize all of the data to create more accurate predictive models while revealing neither their training data databases nor the instances to be classified. The Naive Bayes Classifier is a simple but efficient baseline classifier.
- In this paper, we present a privacy preserving Naive Bayes Classifier for horizontally partitioned data.

LIMITATION OF EXISTING SYSTEM

- Participators need frequent information exchange with the central agent and hence the communication cost is high.
- The decision model is built on the central agent whose aim is to provide a decision support for all participators, but it ignores the data distribution differences between the participators.
- In this regard, it is clear that there is a conflicting objective of maintaining patient privacy and having sufficient data for modeling and decision making.

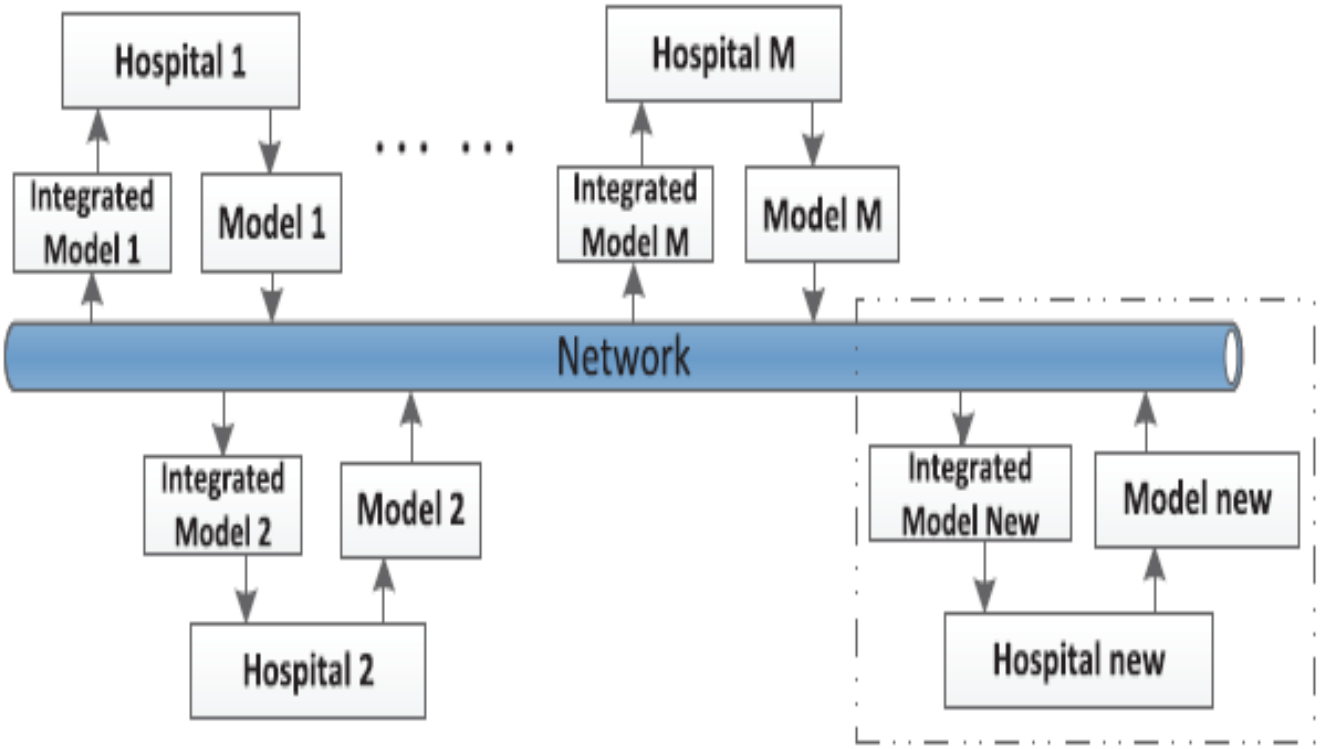
OBJECTIVES OF PROPOSED SYSTEM

- population health management, based on intelligent models, can be used to assess the risk and identify these "complex" patients.

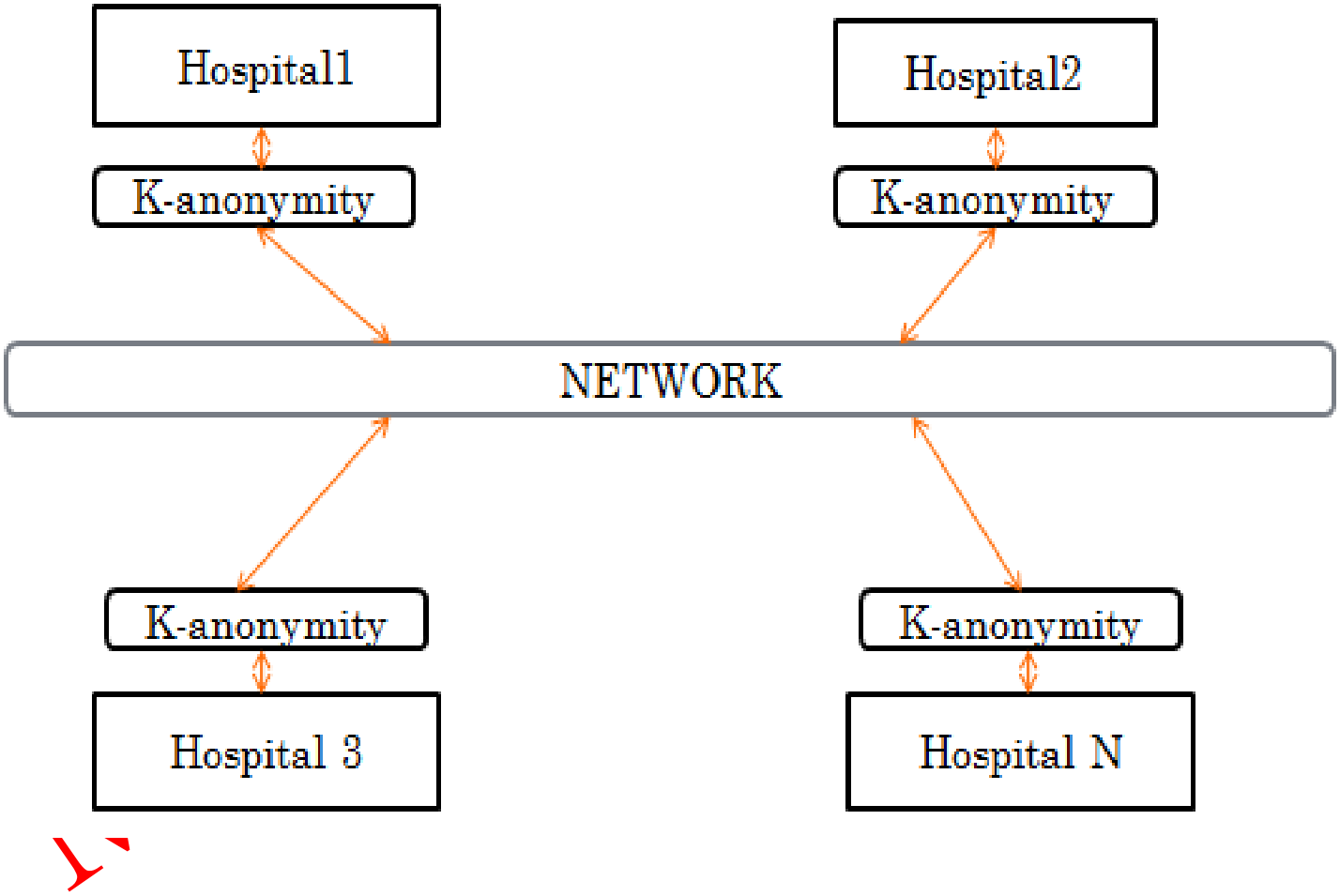
  The aim of this study is to validate machine learning algorithms (Naïve Bayes, Cart, C5.0, Conditional Inference Tree,Random Forest, Artificial Neural Network and LASSO) to predict the risk of hospitalization or death starting from administrative and socio-economic data.

- The study involved the residents in the Local Health Unit of Central Tuscany.

EXISTING ARCHITECTURE DIAGRAM



PROPOSED ARCHITECTURE DIAGRAM



Implementation Detail

Implementation Phase:

- Quasi-Identifier
- K-anonymity
- Generalization

CONCLUSION

- Population is getting older and the number of people suffering from multiple chronic conditions is increasing. For GPs and healthcare providers in general, it becomes crucial to identify as soon as possible the complex patients to treat them with specific program of care, in order to reduce or postpone hospitalizations or death.
- A possible solution to support this selection process is the development of population health management tools based on machine learning methods.
- This paper presents the performance evaluation of several machine learning algorithms to solve the binary classification problem of identifying high-risk patients in the population, by analyzing different sources of administrative an socioeconomic data.

REFERENCES

- [1] Andrew Kingston, Louise Robinson, Heather Booth, Martin Knapp, and Carol Jagger, "Projections of multi-morbidity in the older population in England 2035: estimates from the Population Ageing and Care Simulation (PACSim) model", *Age and Ageing*, https://doi.org/10.1093/ageing/afx201.
- [2] Progetto CCM 2015 Paziente Complesso.
- [3] Efrat Shadmi and Tobias Freund, "Targeting patients for multimorbid care management interventions: the case for equity in high-risk patient identification", *International Journal for Equity in Health, vol. 12,* article 70, 2013.