REPORT OF THE CHRONIC KIDNEY DISEASE DATASET

Prepared by:

Eya Belguith
Mohamed Dhibi
Jihen Souissi
Mohamed Magri
Ahmed Toukebri
Ramy Klouz

SUMMARY

01 About the DataSet

02 Understanding of the Metier

03 Data Understanding

- 1st article
- 2nd article

04 Data preparation

DATASET

DUE TO THE INCREASING NUMBER OF PEOPLE WITH CHRONIC KIDNEY DISEASE (CKD), EFFECTIVE PREDICTION MEASURES FOR THE EARLY DIAGNOSIS OF CKD ARE REQUIRED. WHICH IS WHY MANY MACHINE LEARNING-RELATED PIECES OF RESEARCH WERE MADE. IN OUR PROJECT, WE'LL DIVE INTO 2 ARTICLES THAT USED THE SAME DATASET IN ORDER TO FIND THE BEST METHODS TO PREDICT IN EARLIER STAGES OF CDK.

THE DATA WE'RE STUDYING WERE COLLECTED FROM 400 PATIENTS FROM --THE UCI MACHINE LEARNING REPOSITORY, SCHOOL OF INFORMATION AND COMPUTER SCIENCE, UNIVERSITY OF CALIFORNIA, IRVINE, CA, USA--.

THE DATASET COMPRISES 24 FEATURES:
11 NUMERIC, 13 CATEGORICAL, AND DIAGNOSTIC CLASS FEATURES [WHICH MAKE THE DATASET UNBALANCED]: "CKD"(250 CASES:62.5%) AND "NOTCKD"(150 CASES:37.5%) .

THE FEATURES ARE REPRESENTED LIKE BELLOW:
AGE(AGE), BLOOD PRESSURE (BP), SPECIFIC GRAVITY (SG), SUGAR (SG), RED BLOOD CELLS (RBC), CELL (PC) PUSSY, PUSS CELL CLUMPS (PCC), BACTERIA (BA), BLOOD GLUCOSE RANDOM (BGR), BLOOD UREA (BU), SERUM CREATININE (SC), SODIUM (SOD), POTASSIUM (POT), HEMOGLOBIN (HEMO), PACKED CELL VOLUME (PVC), WHITE BLOOD CELL COUNT (WC), RED BLOOD CELL COUNT (RC) , HYPERTENSION (HTN), DIABETES MELLITUS (DM), CORONARY ARTERY DISEASE (CAD), APPETITE (APPET), PEDAL EDEMA (PE) AND ANEMIA (ANE).

Statistical analysis of the dataset of numerical features.

Features	Mean	Standard deviation	Max	Min
Age	51.483	17.21	90	2
Blood glucose random	148.037	76.583	490	22
Serum creatinine	3.072	4.512	76	0.4
Blood pressure	76.469	13.756	180	50
Blood urea	57.426	49.987	391	1.5
Potassium	4.627	2.92	47	2.5
Packed cell volume	38.884	8.762	54	9
Sodium	137.529	9.908	163	4.5
Hemoglobin	12.526	2.815	17.8	3.1
White blood cell count	8406.12	2823.35	26400	2200
Red blood cell count	4.707	0.89	8	2.1

Statistical analysis of the dataset of nominal features.

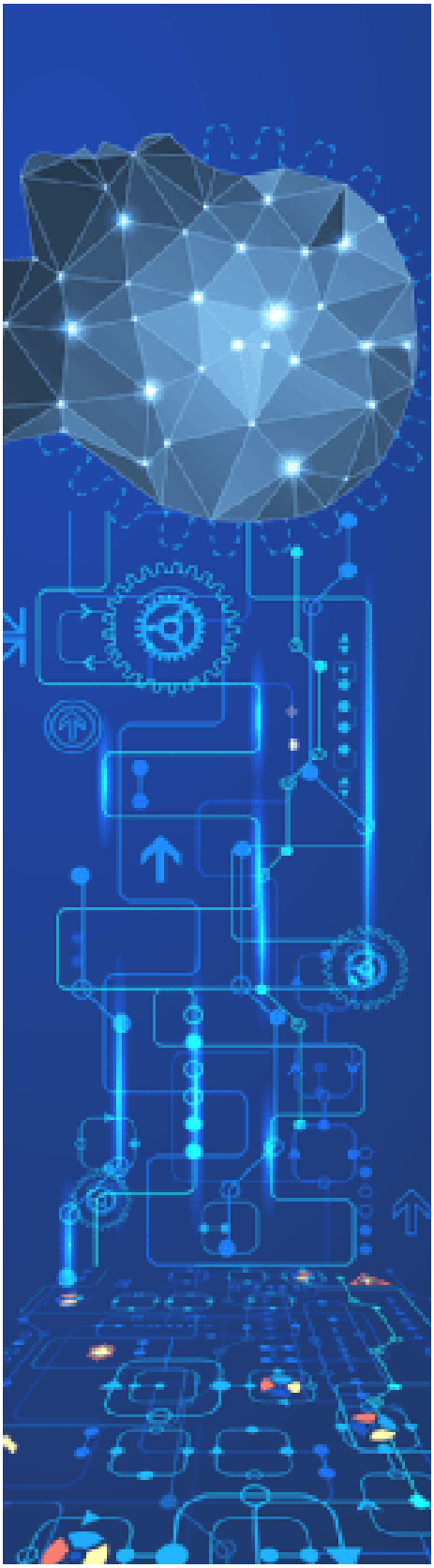
Features	Label	Count
Albumin	0	245
	1	44
	2	43
	3	43
	4	24
	5	1
Specific gravity	1.005	7
	1.01	84
	1.015	75
	1.02	153
	1.025	81
Sugar	0	339
	1	13
	2	18
	3	14
	4	13
	5	3
Pus cell	Normal	324
	Abnormal	76
Red blood cells	Normal	353
	Abnormal	47
Bacteria	Present	22
	Not present	378
Pus cell clumps	Present	42
	Not present	358
Diabetes mellitus	Yes	137
	No	263
Hypertension	Yes	147
	No	253
Edema	Yes	76
	No	324
Coronary artery disease	Yes	34
	No	366
Anemia	Yes	60
	No	340
Appetite	Good	318
	Poor	82

UNDERSTANDNG OF THE METIER

This phase consists of clearly establishing the specifications of the project:
• Clearly state the overall project objectives and business constraints.
• Translate these objectives and constraints into a machine learning problem. It is therefore a question of formulating a search for correlations.
• Prepare an initial strategy to achieve these objectives.

Context	Two studies that developed machine learning method using ensemble learning and feature selection to improve the quality of CKD diagnosis
Goals	• Achieve reproducibility of scientific experiments • Check their results if necessary • Propose improvements of the results
Machine learning based problems	what methods of classification and feature selection are more accurate?

DATA UNDERSTANDING



The understanding and preparation phases are two very important phases of a machine learning project.

The principle of understanding data is that you have to understand all the values of the DataFrame on which you are going to work, which involves analyzing the distribution of values in each column, identifying abnormal values and missing values and the analysis of the most obvious correlations.

the steps of data understanding

- What data is available?
- How much data is available?
- Do we have access to the ground truth, the values we're trying to predict?
- What format will the data be in?
- How can the data be accessed?
- Which fields are most important?
- What important metrics are reported using this data?

...

1ST ARTICAL:

DIAGNOSIS OF CHRONIC KIDNEY DISEASE USING EFFECTIVE CLASSIFICATION ALGORITHMS AND RECURSIVE FEATURE ELIMINATION TECHNIQUES

MEDICAL EXPERTS DETERMINE KIDNEY DISEASE THROUGH GLOMERULAR FILTRATION RATE (GFR), WHICH DESCRIBES KIDNEY FUNCTION

The stages of development of CKD

Stage	Description	Glomerular filtration rate (GFR) (mL/min/ 1.73 m^2)	Treatment stage
1	Kidney function is normal	≥90	Observation, blood pressure control
2	Kidney damage is mild	60–89	Observation, blood pressure control and risk factors
3	Kidney damage is moderate	30–59	Observation, blood pressure control and risk factors
4	Kidney damage is severe	15–29	Planning for end-stage renal failure
5	Established kidney failure	≤ 15	Treatment choices

USED METHODS

- PREPROCESSING

IN THIS STEP, THEY DEALT WITH OUTLIERSBY FIRST USING THE NORMALIZATION METHODE, AND TO CHECK UNBALANCED DATA: MISSING NUMERICAL FEATURES(SEPARATE OR CONTINUOUS) WERE REPLACED BY THE MEAN METHOD, AND A MODE METHOD WAS APPLIED TO REPLACE THE MISSING NOMINAL FEATURES.

- FEATURES SELECTION

THE RFE RECURSIVE FEATURE ELIMINATION METHOD WERE USED
(ALBUMIN FEATURE HAD THE HIGHEST CORRECTION (17.99%), FEATURED BY
14.34%, THEN THE PACKED CELL VOLUME FEATURE BY 12.91%, AND
THE SERUM CREATININE FEATURE BY 12.09%)
CONCERNING HOW TO FIND THE OPTIMAL NUMBER OF FEATURES, CROSS-VALIDATION WAS USED WITH RFE TO SCORE DIFFERENT FEATURE SUBSETS AND SELECT THE BEST SCORING COLLECTION OF FEATURES.

- CLASSIFICATION

SUPPORT VECTOR MACHINE(SVM), K-NEAREST NEIGHBORS (KNN), DECITION TREE, AND RANDOM FOREST ; ARE SUPERVISED ALGORITHMS THAT WERE USED TO SOLVE CLASSIFICATION AND REGRESSION PROBLEMS IN THIS ARTICAL.
EXCEPT THAT IN THIS CASE, THE RANDOM FOREST ALGORITHM OUTPERFORMED ALL OTHER APPLIED ALGORITHMS, REACHING AN ACCURACY, PRECISION, RECALL, AND F1-SCORE OF 100% FOR ALL MEASURES.

Results of diagnosing CKD using four machine learning algorithms.

Classifiers	SVM	KNN	Decision tree	Random forest
Accuracy %	96.67	98.33	99.17	100.00
Precision %	92.00	100.00	100.00	100.00
Recall %	94.74	97.37	98.68	100.00
F1-score%	97.30	98.67	99.34	100.00

2ND ARTICAL:
BOOSTED CLASSIFIER AND FEATURES SELECTION FOR ENHANCING CHRONIC KIDNEY DISEASE DIAGNOSE

AS MENTIONED IN THE 1ST ARTICAL,THE LEVEL OF GFR CAN INDICATES STAGE OF CHRONIC KIDNEY DISEASE.
-HEALTHY ADULTS HAVE 125 MLMIN PER 1.73 M2 OF GFR LEVEL.
RENAL FAILURE GFR OF LESS THAN 15 ML/MIN PER 1.73 M2.
-STAGE 1 IS INDICATED BY GFR LEVEL OF OVER 90 ML/MIN PER 1.73 M2
-STAGE 2 IS DEFINED AS A GFR OF 60-89 ML/MIN PER 1.73 M2
-STAGE 3 IS DEFINED AS A GFR BETWEEN 30-59 ML/MIN PER 1.73 M2
-STAGE 4 IS AS A GFR BETWEEN 15-29 ML/MIN PER 1.73 M2
-STAGE 5 IS DEFINED AS A GFR OF LESS THAN 15 ML/MIN PER 1.73 M2

USED METHODS

- PREPROCESSING

THE FILLING IN OF THE MISSING VALUE WAS FORMED BY STATISTICAL METHODS SUCH AS MEDIAN AND MEAN OR PROBABILITY METHOD.

- FEATURES SELECTION

FEATURES WERE SELECTED USING A CORRELATION-BASED FEATURE SELECTION (CFS) [WHICH REMOVED 7 ATTRIBUTES, LEAVING 17 ATTRIBUTES]AND ADABOOST WAS USED FOR ENSEMBLE LEARNING TO IMPROVE THE DETECTION OF CKD.

- CLASSIFICATION

KNEAREST NEIGHBOUR ALGORITHM (KNN), NAIVE BAYES AND SUPPORT VECTOR MACHINE (SVM) WAS USED AS BASE CLASSIFIER TO DETERMINE WHICH PATIENTS NEED THE MOST MEDICAL CARE (TWO CLASSES : NORMAL=WITHOUT CKD , WITH CKD).
TO ESTIMATE CLASSIFICATION PERFORMANCE, FOUR PARAMETERS ARE USED IN THIS RESEARCH: ACCURACY, PRECISION, RECALL AND F-MEASURE.
BECAUSE THE DATA ON CKD DATASET IS NOT BALANCED, PRECISION, RECALL AND F-MEASURE WERE USED TO ASSESS CLASSIFIER PERFORMANCE.

THE BEST RESULT WAS ACHIEVED BY COMBINATION OF KNN CLASSIFIER WITH CFS AND ADABOOST, WITH 0.981 ACCURACY RATE, 0.980 RECALL RATE AND 0.980 F-MEASURE RATE.
HIGHEST PRECISION RATE WAS ACHIEVED BY THE COMBINATION OF NAIVE BAYES CLASSIFIER WITH CFS AND DABOOST, WITH 0.981 PRECISION RATE.

Classification Result

Parameter	Classifiers		
	NB	kNN	SVM
	1st Method : Base		
Accuracy	0.950	0.958	0.958
Precision	0.941	0.949	0.958
Recall	0.960	0.966	0.958
F-measure	0.948	0.956	0.958
	2nd Method : CFS		
Accuracy	↑ 0.955	↑ 0.978	↑ 0.963
Precision	↑ 0.964	↑ 0.972	↑ 0.963
Recall	↑ 0.964	↑ 0.982	↑ 0.963
F-measure	↑ 0.953	↑ 0.977	↑ 0.963
	3rd Method : CFS + AdaBoost		
Accuracy	↑↑ 0.980	↑↑ 0.981	↑↑ 0.975
Precision	↑↑ 0.981	↑↑ 0.980	↑↑ 0.975
Recall	↑↑ 0.980	↑↓ 0.980	↑↑ 0.975
F-measure	↑↑ 0.980	↑↑ 0.980	↑↑ 0.975

↑ SIGN MEANS THERE IS AN INCREASE IN CLASSIFICATION RESULT
↓ SIGN MEANS THERE IS A DECREASE IN CLASSIFICATION RESULT

AS FOR DATA TRAINING AND TESTING SCHEMA IN THIS STUDY, K-CROSS VALIDATION WAS USED.

DATA PREPARATION

As our task requires, in the phase of data preparation, we included methods mentioned in both articles plus some that we found useful, such as; attaining the correlation matrice and the score of each feature, using RFE, and CFS for the feature selection, and both the Standardization and Normalization methods for the feature scaling...

(the code + detailed explanation of every step: will be found in the Notebook attached to this report)