Analysing the Effects of Different Features Selection Methods on the Classification of Kidney Diseases (CKD)

Tageldin Musa Bakhit Mohamed Noor[1], Abu Sarwar Zamani[2], Mohammed Rizwanullah[3] and Anwer Mustafa Mohamedsalih Hilal[4*]

[1]Department of Electrical Engineering, Faculty of Engineering, Ain Shams University, Cairo, Egypt
[2,3,4]Department of Computer and Self Development, Preparatory Year Deanship, Prince Sattam bin Abdulaziz University, AlKharj, Saudi Arabia.
*Corresponding Author: Anwar Hilal, Email Address:

Summary
Today, data is stored in databases in a wide range of fields. Consequently, a significant amount of data has been generated over time. The value of these massive data sets lies in the knowledge and relationships that exist between them. The data mining process involves the analysis of data to obtain knowledge and facts that assist decision makers in making the best judgments possible. The use of data mining techniques in medicine has been around for generations, which is one of the fields where data mining is most important. This study aims to establish is to determine the encroachment of feature selection algorithms on classifier accuracy (model). There are 25 features in the dataset used to diagnose Chronic Kidney Disease (CKD), was utilized in this study to see how features selection techniques affect classifier accuracy. Here, Wrapper features selection evaluators are used to select those features that improve classification accuracy. We used classical Naive Bayes and J48 classifiers in this study and when Naive Bayes is used, accuracy increases from 95% to 99.5% as a classifier with a wrapper features selection evaluator. It is not a significant degree of accuracy, when j48 classifiers are used with wrapper features selection evaluators.

Keywords: *Naïve Bayes Classifier; J48 Classifier; Data Mining; CKD*

1. Introduction

In recent years, Chronic Kidney Disease (renal failure) has been on the rise, posing a serious threat to the lives and health of many men, women, young people, and children. Renal insufficiency is a medical term that describes kidney failure in performing its tasks. Acute renal insufficiency and chronic renal insufficiency are the two kinds of kidney failure. Renal function failure can be cause by a variety of factors. Diabetes, high blood pressure, kidney inflammation (kidney glomerulonephritis), and polycystic kidney disease, a hereditary illness that causes kidneys to grow cysts and eventually fail, are the most common causes. Unknown causes about 20% of dialysis patients have no idea what the true cause of kidney failure is. These patients are frequently using the therapy for the first time after renal failure has progressed, and it is difficult to pinpoint the source of the disease at this stage. Computers have significantly improved technology in recent years, resulting in the creation of massive amounts of data. Furthermore, the development of healthcare information systems has resulted in the creation of a large number of medical databases. Data mining is a prominent topic of research that focuses on creating knowledge and managing vast amounts of heterogeneous data [1]. A data mining process is a method of obtaining useful information from enormous various types of data held in databases, data warehouses, and additional data repositories [2]. The possibilities offered by medical data mining are many, such as uncovering hidden patterns that can be used to diagnose any disease dataset. Supervisory learning is the process of classifying objects that is used to analyse medical data to discover hidden patterns [3]. In addition, Classification is use in both medical and clinical research, both of which focus on supervised learning. This study's goal is to apply classification techniques to medical science and bioinformatics. Classification techniques are design to each target class should be accurately identified in the data. The fundamental idea would be to exclude certain features from the input variables that have predictive value is limited or non-existent. There are three main categories of these techniques. Filter methods are one, Wrapper methods are another, Embedded methods comprise the third type. In general, kidney disease has to be diagnosed based on a number of tests and a thorough medical examination. In the process of diagnosing kidney failure, there are varieties of tests taken, some are not crucial to classification. Certain of these tests may take a long time to complete which may cause the patient to die. Another problem is that too many features are used in classification, resulting in poor performance and accuracy of the model. Thus, this study uses wrapper evaluation to eliminate unimportant features from classification models

Manuscript received January 5, 2022
Manuscript revised January 20, 2022
https://doi.org/10.22937/IJCSNS.2022.22.1.96

to improve performance and accuracy. This study is also significant because it used features selection evaluator to reduce data set thus improving the accuracy of classification models and performance because it used statistical approaches to calculate correlation between these features and removed irrelevant features.

2. Literature Review

This section contains reviews of a variety of technical and review publications on data mining strategies for predicting Chronic Kidney Disease. Many scholars have utilized various data mining approaches to forecast the future. S. Dhayanand et al. [4] the classification approach utilized to classify four categories of renal disorders in this study. Comparison of the Naive Bayes and Support Vector Machine (SVM) Algorithms for classification rely on Execution time and accuracy of classification are considered performance factors. The results of the study indicate that we can conclude that the SVM performs suitable in classifying the data, yields accurate results, and therefore considered as a better classifier when comparing to Naive Bayes. In this case, Naive Bayes probably qualifies the data with the least number of executions.

Lambodar Jena et al. [1] The same dataset was used here is the study to data mining algorithms for predicting chronic kidney disease although the researchers used various classifiers (j48 (99%), naive Bayes (95%), Multilayer perception (99.75%), and SVM) (62 %), Decision Table (99 %) and Conjuctive Rule (94 %) when compared to all other classifiers, the multilayer perception method has a higher classification accuracy of 99.75%. The interesting thing is that all algorithms except SVM perform poorly, with classification accuracy exceeding 90%. As a result, Multilayer Perceptron showed good performance when applied to chronic kidney disease data. There is an open problem in this study if you are able to use features selection algorithms to increase accuracy of the modes and reduce the time execution required to build them.

Naganna Chetty et al. [5] as part of this study, the researcher developed various classification algorithms, Wrappersubset attribute evaluators, and Best First Search (BFS) approaches for predicting and classifying CKD and non-CKD patients. The (BFS) predicted to have good accuracy. Sequential Minimal Optimization (97.75%) followed by IBK (95.75%) and then Naive Bayes (95%) on prediction accuracy based on the original dataset and the same classifier, IBK (Implements K-nearest neighbour) (100%) followed by Naive Bayes (99%), and then SMO (Self-organizing Map) (98.25%). In contrast, IBK classifiers perform better.

Mohammad Ayesha et al. [3] uses Naive Bayes Classifier to build a model for Chronic Kidney Disease and to reduce the attributes, it uses four attribute evaluators. The Attribute Evaluator used (WrapperSubsetEval attribute evaluator with SMO (Self organizing Map) classifier and BFS, WrapperSubsetEval attribute evaluator with IBK (Implements the K-nearest neighbor) classifier and BFS WrapperSubsetEval attribute evaluator with Naïve bayes classifier and best first search and OneR attribute evaluator with Naïve Bayes). It was also founded that using OneR algorithm, we reduced the number of attributes in the dataset by 80% and improved accuracy by 12.5% as compared to using the existing system. To avoid progression of Chronic Kidney Disease to the next stage, the system we propose extracts actions must be taken applicable to each stage so that treatments can be taken accordingly.

S. Ramya et al. [6] Here's a way was developed Predicting kidney function failure based on test results from the patient's medical report using four classification techniques. A comparison was also made between the following four techniques: Back Propagation, Neural Network, Radial Basis Function, and Random Forest. As they noted in their study, RBF (Radial Basis Function) was found to be more accurate for predicting chronic kidney disease.

3. Data Mining and Data Mining Classification

This section describes data classification, feature selection algorithm, wrapper method, Naïve Bayes and j48 classifier.

3.1 Data Classification

An important data-mining task is classification, which is the act of proposing a classification functions or classification models (also referred to as classifiers). In a classification model, data are assigned classes based on their attributes. There are several ways to construct classifications: Nave Bayes, Support Vector Machines, Multi-Layer Perceptron, Logistic Regression, Decision Trees, and Random Forests. As seen in Figure 1, the predictive and training phases are the two phases of the overall data classification process. It consists of a number of features is derived on the basis of the training data, in the training phase (e.g. “A”, “B”, “AB” or “O”, for blood type), ordinal (e.g. “large”, “medium” or “small”), integer-valued (e.g. the frequency of occurrence of part words in an email) or real-valued (e.g. a measurement of blood pressure). It is possible to work with some algorithms exclusively with discrete data such as ID3, which requires integer- or real-valued data should discretely have divided into groups (e.g., less than 5, between 5 and 10, or greater than 10).

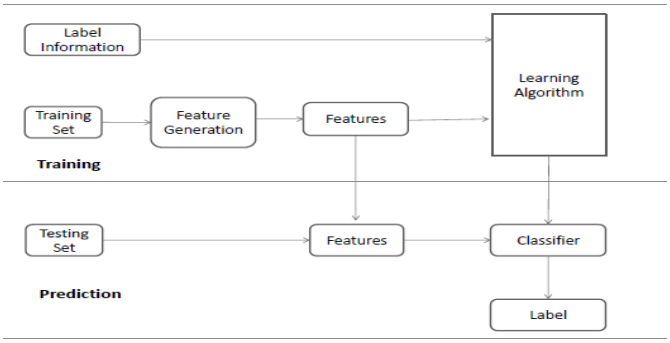

Figure 1 A General Process of Data Classification

When these extracted features used to represent data, During the learning process, the algorithm takes advantage of both learn a map function F from the label information and the data itself (or generation models like the vector space model for text data) [7]. A feature can either categorize (e.g., a classifier) or label the product as an attribute

$$f(\text{features}) \rightarrow \text{labels} \quad (1)$$

A map function (also known as a classifier) based on the feature set extracted during the training period predicts the labels of the data based on the feature set represented in the prediction phase. Training and prediction phases should use the same feature set.

3.2 Selection Features Algorithm

As it relates to data, "features," "attributes," or "variables" refer to aspects. Data collection usually begins with selecting or specifying features. There are three types of features: discrete, continuous, and nominal.

- Relevant: Those aspects influence the outcome, and their role can't be taken over by others.
- Irrelevant: Those features that have no effect on output defined as irrelevant features, and their values generated randomly for every example.
- Redundant: In regards to redundancy, any feature that can serve as a substitute for another is redundancy.

One of the most commonly used techniques feature selection is important for reducing dimensionality. According to certain criteria, out of the original details, a small subset of those relevant to the study is selected, which produces better learning performance (e.g., classification accuracy is improved through better learning), lower computational cost, and improved model interpretability [8, 9]. On the basis of if there is a training set has been labelled or not features selection algorithms are classified into three major categories: Supervisory, unsupervisory [10, 11] and semi-supervised [12, 13]. In this study, we also examine wrapper and embedded models of supervised feature selection, among which is the filter model and the wrapper model, respectively.

(i) Filter Models: Learning classifiers based on selection separated in filter models, so that there is no interaction between the Learning Algorithms that have biases and Feature Selection Algorithms have biases. Generally, distance, consistency, dependency, information, and correlation used to data about training characteristics is measured. A number of filtering methodologies have been developed over the decades, including Fisher score, the Information Gain model [15], and relief [14]. There are two steps in Filter algorithms typically used. As a starting point is to rank features according to specified criteria. The examination of features depending on whether it is univariate or multivariate. According to the univariate model, there is no correlation between features and feature spaces, but the multivariate system assesses features in batches. As a result, Multivariate schemes can handle different types of data duplicated characteristics by default. According to our results, the most popular features are picked in the second step to inspire classification models.

- Univariate Algorithm: The univariate scheme involves ranking the features are independent of the feature space, and this section describes the information gain and gain allocation algorithm to give this algorithm the advantage of being independent of the classifier.

(a) Information Gain: One of the most well-known methodologies for feature selection is information gain. This function estimates the exchange of information ith feature fi and a class label Case and measures the reliance between characteristics and labels.

$$IG(fi,C) = H(fi) - H(fi \mid C) \quad (2)$$

Where H(fi) denotes H(fi) entropy and H (fi| C) denotes fi entropy after viewing C:

$$H(f_i) = -\sum_{j} p(x_j) log_2(p(x_j)),$$
$$H(f_i|C) = -\sum_{k} p(c_k) \sum_{j} p(x_j|c_k) log_2(p(x_j|c_k)) \quad (3)$$

The information gain of a feature indicates its relevance.

(b) Gain Ratio: What can be done to reduce its bias by modifying the information gain? By choosing an attribute, Gain ratio takes branch size and number into consideration. A split accounted for in terms of its intrinsic data. The branching entropy is intrinsic to a system's information. When intrinsic information increases, attribute value tends to decrease.

$$\text{Gain Ratio(Attribute)} = \frac{\text{Gain(Attribute)}}{\text{Intrinsic} - \text{Info(Attribute)}} \quad (4)$$

- Multivariate Algorithm: It uses a batch-based approach to evaluate features. Multivariate schemes are naturally able to handle redundant features; consequently, their benefits based on their independence from classifiers and better computational complexity. Correlation-based Feature Selection (CBFS), and Fast Correlation-based Feature Selection (FCBFS) are included here rather than wrapper methods.

(a) Correlation-based Feature Selection (CBFS):

Redundancies may be classified differently based on their severity within the subset of features, CBFS searches to find the most relevant ones. An evaluator attempts subsets of features can be found by doing that have high individual intercorrelations are low between classes. Subset evaluators use numeric measures, including entropy under condition, to guide the iterative search and select features that are most closely related to class. Multivariate filters eliminate the problem of interactions between features in univariate filters, as well as the downside of single-variate filters.

For example, filters. According to CFS, the importance of each feature is determined by considering the individual predictive abilities as well as in terms of redundancy among features, the degree of redundancy. In addition to estimating correlations between attributes and classes, correlation coefficients utilized to establish inter-correlations among features. An increasing number of features have greater importance corresponding to one another and classes, as the correlation increases, the effect of the correlation decreases. Using CBFS, a search strategy known as best-first search is typically combined with forward selection, backward elimination, bidirectional search, and genetic search to identify the best subset of features. CBFS is modelled using an equation.

$$r_{zc} = \frac{k\overline{r_{zi}}}{\sqrt{k + k(k-1)\overline{r_{ii}}}} \quad (5)$$

The correlation between rzc and the summed feature subsets is defined by k, and rzi is the average correlation between rzi and the class variable, and rii is the average intercorrelation between rzi and the class variable.

(b) Fast Correlation Based FS (FCBF):

Yu and Liu, et al. [16] A symmetrical uncertainty measure is also used by FCBF. However, Search algorithms differ greatly. It is based on the idea of "predominance." A correlation is a relationship between two attributes X* and the target Y is predominant if and only

$$\text{If } \rho y,x^* \geq \delta et \ \nexists X(X \neq X^*), \rho x,x^* < \rho y,x^* \quad (6)$$

Predictive factor is considered interesting when, in the absence of another predictor that is stronger correlated to the target attribute (delta is the parameter which evaluates this one), This attribute correlates significantly with the target attribute.

(ii) Embedded Model:

Embedded models combine feature selection and classifier construction, have many advantages over conventional wrapper models: they incorporate the interaction between the classification model and the filter model, and they require far less computation than traditional wrapper models Kudo et al [17].

Embedded methods can be grouped into three categories. A pruning method that includes all features is one of these are used then delete some features after training a model by setting corresponding coefficients to 0, while still monitoring the model's performance, Support Vector Machines (SVM) are used in recursive feature elimination Guyon et al [18]. Second, there are models like ID3 that include a mechanism to select features Quinlan et al [19] and C4.5 Quinlan et al [20]. There is a third type of regularization model, which minimizes fitting errors while forcing coefficients to be small or, in some cases, exact zeroes. A feature whose coefficient is close to 0 is then eliminated Ma and J. Huang et al [21]. Increasing attention is being given to regularization models because of their high performance.

(iii) Wrapper Methods:

Wrapper approaches evaluate the variable subset Performance measurement is goals and objectives and predictor is the black box. Because evaluating 2N subsets is an NP-hard problem, suboptimal subsets are discovered using search algorithms that heuristically choose a subset. There are numerous search algorithms available that may be used to discover variables that are divided into subsets that optimizes the objective function, which is classification performance. To evaluate different subsets of features based on the given selection number, Branch and Bound used a tree structure. However, the search would become exponentially more complex for more features. The computation requirements for exhaustive search methods can get prohibitive for large datasets. As a result, simplified algorithms such as sequential search or evolutionary algorithms such as Genetic Algorithm (GA) or Particle Swarm Optimization (PSO), which produce local maximum methods employed are based on the results. These results are typically good and can be computed efficiently. Wrapper methods are broadly classified into Sequential Selection Algorithms and Heuristic Search Algorithms Irish Survey et al [22].

3.3 Naïve Bayes Classifier

Using Bayes' theorem (from Bayesian statistics) with strong (naive) independence assumptions, a Naive Bayes classifier is a simple probabilistic classifier. "Independent

feature model" is a better way to describe the underlying probability model. Although this restricted applied to real world applications, individuality assumption is rarely true, hence the characterization as naive, Typically, the algorithm produces good results and learns quickly from various supervised classification problems. The naive Bayes classifier has the advantage that it requires little training data to determine the parameters required for classification (means and variances of the variables). The entire covariance matrix need not be determined since independent variables assumed. What the Bayes theorem means and how it is applied demonstrated in the following paragraph:

Let H be some hypothesis that the tuple X is a member of a particular class C, X to be a tuple of data.
P(H/X) - is the posterior probability of H conditioned on X.
P(H) - is the prior probability of H.
P(X) - is prior probability of X.

$$(H/\mathrm{X}) = \frac{(H/\mathrm{X})\,P(H)}{P(\mathrm{X})} \quad (7)$$

The following example illustrates how a class label can be predicted using nave Bayesian classification. In table 1 below, the same training data shows how we may predict a tuple's class label using nave Bayesian classification. There are four attributes in data tuples: student, age, income, and credit rating. There are two distinct values for the class label attribute, buys computer (namely, yes and no). Let C1 correspond to the class buys computer = yes and C2 correspond to buys computer = no. The tuple we wish to classify is X = (age = youth, income = medium, student = yes, credit rating = fair).

Table 1 Tuples with Class Labels from All Electronics' Customer Database.

RID	Age	Income	Student	Credit rating	Class_buys Computer
1	Youth	High	No	Fair	No
2	Youth	High	No	Excell ent	No
3	Middle-a ged	High	No	Fair	Yes
4	Senior	Medium	No	Fair	Yes
5	Senior	Low	Yes	Fair	Yes
6	Senior	Low	Yes	Excell ent	No
7	Middle-a ged	Low	Yes	Excell ent	yes
8	Youth	Medium	No	Fair	No
9	Youth	Low	Yes	Fair	Yes
10	Senior	Medium	Yes	Fair	Yes
11	Youth	Medium	Yes	Excell ent	Yes
12	Middle-a ged	Medium	No	Excell ent	Yes
13	Middle-a ged	High	Yes	Fair	Yes
14	Senior	Medium	No	Excell ent	no

For I = 1, 2, we must maximize P (XjCi)P(Ci). The prior probability of each Class, P (Ci), can be calculated using the training tuples:
P (yep, I Purchased a computer) = 9/14 = 0.643
P = 5/14 = 0.357 (Purchase computer = no).
For I = 1, 2, we compute the following conditional probabilities to compute PXjCi):
P = 2/5 = 0.400 (credit rating = fair, j purchase computer = no).
P (student = yes j purchase computer = yes) = 6/9 = 0.667
P (student = yes j purchase computer = no) = 1/5 = 0.200
P (student = yes j purchase computer = no) = 1/5 = 0.200
P (j purchase computer = yes) = 6/9 = 0.667
P (credit rating = fair) (credit rating = fair; j does not purchase a computer) = 2/5 = 0.400
We get at P (Xj purchase computer = yes) using the probabilities stated above. =
P (age = youth / computer purchased = yes)
P (income = medium / computer purchase = yes)
P (student = yes / computer purchased = yes)
P (credit rating = fair / computer purchase = yes)
= 0.222×0.444×0.667×0.667 = 0.044.
Similarly,
P (Xj Purchased computer = no) = 0:600×0:400×0:200×0:400 = 0.019.
We compute P(XjCi)P(Ci) to discover the class Ci that maximizes P(XjCi)P(Ci).
P (Xj Purchased computer = yes), P (Purchased computer = yes) = 0.044×0.643 = 0.028
P (Xj Purchased computer = no), P (Purchased computer = no) = 0.019×0:357 = 0.007

Therefore, the naïve Bayesian classifier predicts Purchase computer = yes for tuple X.

3.4 J48 Classifier

J48 is a simple decision tree of C4.5 for classifying data. It is a supervised classification method. Using it, you can create a small binary tree. This is a univariate decision tree. This algorithm extends the ID3 algorithm. A Divide and Conquer approach is used to classify the data in this classifier. By using training sample values, it divides the data into ranges according to those values in the data.
Algorithm: Generate decision tree. Generate a decision tree from the training tuples of data partition D.

Input

- D is a data partition containing all training tuples and their class labels;

- This list of attributes consists of a number of potential candidates;
- An attribute selection method determines the partitioning criteria which will best partition a given tuple of data into individual classes based on how their attributes rank. The criteria consist of either a splitting point or a splitting subset as well as the splitting attribute.

Output

Using a Decision Tree

Method

(1) create a node N;
(2) In D, if the tuples are all of the same class, C, evaluates to
(3) The leaf node N is labelled with the class C;
(4) A list of empty attributes signifies
(5) Assign class D to N as a leaf node; // Class D represents the majority vote
(6) Determine the "best" splitting criteria by applying the Attribute _selection_ method (D, attribute_ list);
(7) Put a split criterion on node N;
(8) Multi-way splits are allowed if the splitting attribute is discrete-valued and binary trees are not restricted
(9) attribute _list ← attribute _list −splitting _attribute; // remove splitting attribute
(10) j: tuples are split by splitting criterion, and tuples are grown by producing subtrees for each part
(11) In Dj, let Dj be the number of "tuples" satisfying outcome j; // and let Dj be the partition if Dj is empty then
(12) Label node N with the majority class leaf in D;
(13) Attach node N to the node returned by Generate decision tree (Dj, attributes); otherwise return 0
(14) return N;

4. Proposed Model

The proposed methodology presented in this section. Three steps are involved in the methodology:

Step1: applying classification model by using original dataset.

Step2: Applying feature selection or (data prepossessing) to reduce the data.

Step3: applying classification model by using reduced datasets.

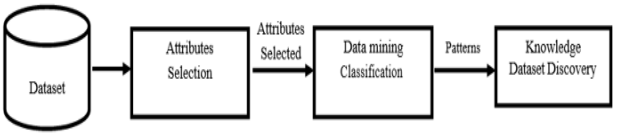

Figure 2 Proposed Framework for Mining Patterns

Basically, this framework consists of datasets with methods used to select best attributes, then using data mining classification to extract patterns leading to knowledge discovery.

The figure 3 below illustrates a methodology for selecting attributes. The process of eliminating the attributes of less importance from a dataset reduces its dimension. Here we are using best first search methods for wrapping subsets of features from Naive Bayes and j48 classifiers.

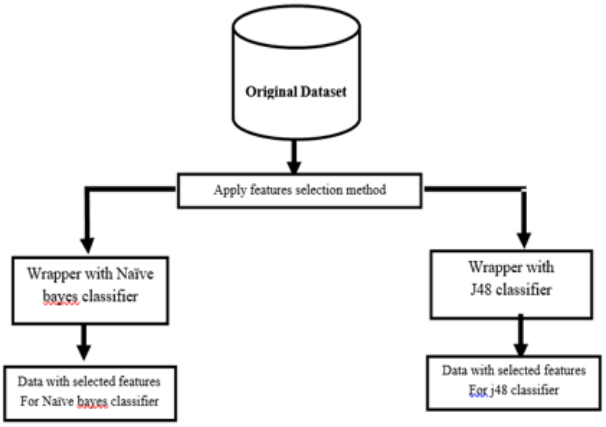

Figure 3 Framework for Attribute Selection

4.1 Material and Tools

Based on the findings of this study we used the Waikato Environment for Knowledge Analysis (Weka) program to experiment and test with this program and we provide a brief description of it in the following paragraph New Zealand's University of Waikato developed WEKA, a tool for preparing and using data for research. There are a number of machine learning algorithms that analysing data is a function of data mining. It can either be accessed from your own Java code or directly applied to a dataset. Weka is a data-reprocessing and analytics tool that includes tools for classification, regression, clustering, association rules, and visual representation. It can also be used to develop new machine learning schemes.

4.2 Dataset Used

The dataset was taken from the UCI machine-learning repository. In this dataset special for Chronic Kidney Disease (CKD) there are 25 variables, of which 1 is a class variable. There are 13 nominal variables, 11 numerical variables and 400 instances (250 CKD, 150 not-CKD). The table 2 below gives more information about this dataset.

Table 2 Description of Attribute in the Chronic Kidney Disease Dataset

S. No	Attribute	Description	Type	Permissible Values
1	Age	Age	Numerical	Age in years
2	BP	Blood Pressure	Numerical	In mm/Hg
3	SP	Specific Gravity	Nominal	(1.005, 1.010, 1.015, 1.020, 1.025
4	AL	Albumin	Nominal	(0,1,2,3,4,5)
5	SU	Sugar	Nominal	(0,1,2,3,4,5)
6	RBC	Red Blood	Nominal	Normal, Abnormal

		Cells		
7	PC	Pus Cell	Nominal	Normal, Abnormal
8	PCC	Pus Cell Clumps	Nominal	Present, Not Present
9	BA	Bacteria	Nominal	Present, Not Present
10	BGR	Blood Glucose Random	Numerical	In Mgs/dl
11	BU	Blood Urea	Numerical	In Mgs/dl
12	SC	Serum Creatinine	Numerical	In Mgs/dl
13	Sod	Sodium	Numerical	In mEq/l
14	Pot	Potassium	Numerical	In mEq/l
15	Hemo	Hemoglobin	Numerical	In gms
16	Pcv	Packed Cell Volume	Numerical	In cells/cumm
17	Wbbc	White Blood Cell Count	Numerical	In cells/cumm
18	Rbcc	Red Blood Cell Count	Numerical	Millions/cmm
19	Htn	Hypertensio n	Nominal	Yes, No
20	Dm	Diabetes Mellitus	Nominal	Yes, No
21	Cad	Coronary Artery Disease	Nominal	Yes, No
22	Appet	Appetite	Nominal	Good, Poor
23	Pe	Pedal Edema	Nominal	Yes, No
24	Ane	Anaemia	Nominal	Yes, No
25	Class	Class	Nominal	Ckd, notckd

4.3 Data Pre-Processing

Getting started with classification methods means first cleaning and modifying the data. This process is called pre-processing. Operationally, several things happen during this pre-processing step, the evaluation of missing values is also included, removing noisy data like data balancing, removing outliers, and normalizing. In the World as it is, the values that are frequently lacking. When a record is missing a value, we can remove the entire record containing that value, a process called Case Deletion. However, it has been determined that missing 5% of a data set with 30 variables (spread randomly among attributes and records) would mean omitting approximately 80% of the data. We have compared and evaluated different data imputation algorithms as an alternative to removing records that had missing values. In most cases, data missing from records can be inferred from the median or mid-range of the observed values across the data set to fill in those values. When doing Mode Imputation for nominal attributes, a value is substituted for the one that is missing with based on the average value of the attribute throughout the dataset. In this dataset, the following data pre-processing steps have been applied:

- Merge any nominal attribute that multiple values such as (sugar (Su ((0,1,2,3,4,5))) In the small range (2 or 3 outcomes).
- Discretize any numerical attribute that multiple distinct values to three categorize.
- Ignore the missing value for each attribute.
- Applied the features selection algorithms to select best features.

5. Experiments and Results

The purpose of the chapter is to explain the experiments, which focus on the effect of selecting features on the accuracy of the model. Explain the steps taken to build the model and the conclusions derived from these experiments in the next couple of paragraphs.

5.1 Model Build

A wrapper features selection algorithm was applied in this study using the Weka program, and we built the model using a J48 and naive Bayes classifier. Using classification accuracy as a guide, compare the results for these classifiers. Accuracy of classification refers to comparing the accuracy of results obtained from classifiers. We will describe in this section how precise calculations are made by model and how model accuracy is calculated.

- Classification algorithms are measured by their accuracy in terms of the result is determined by dividing the number of instances correctly classified by Count of all instances within the dataset.

$$\text{Accuracy} = \frac{TP+TN}{TP+FP+TN+FN} \tag{8}$$

Where TP – True positive, FP False positive, TN true Negative FN False Negative.

- TP Rate: A high true-positive rate is the ability to detect it. The true-positive rate is also known as sensitivity.

$$TPR = \frac{TP}{TP+FN} \tag{9}$$

The correlation between modules that were correctly classified and the classification of a given number of modules as fault-prone is used to determine precision. An error rate is the percentage of units incorrectly the prediction of a faulty outcome.

$$\text{Precision} = \frac{TP}{TP+FP} \quad (10)$$

5.2 The First Experiment

We have employed the naive Bayes (NBC) and (J48) classifiers methods to classify the dataset contains all features (dataset used for original analysis). (CKD) dataset has 13 nominal attributes, 11 numerical attributes, and 1 class of 400 cases (250 CKD, 150 notckd) of chronic kidney disease. According to the class label, the distribution of patients (CKD or notckd) is shown in Figure 4.

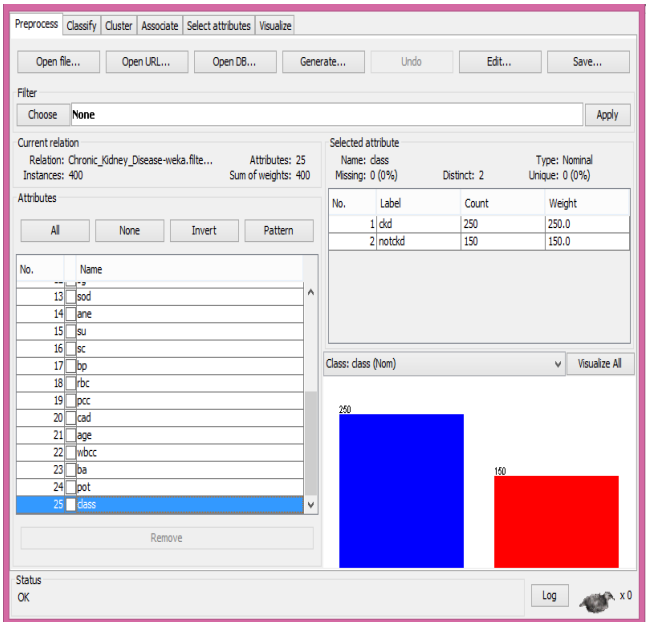

Figure 4. The distribution of the patient based on the class label (CKD or notckd).

The figure above shows the results of the first experiment before reducing the dataset by using J48 and naive Bayes cross validation test. The j48 classifier taken 99% and only (4 instances) incorrectly classified and naive Bayes has, taken 95% and only (20 instances) incorrectly classified.

The Table 4 Shows Graphical Representation of Classification Accuracies for J48 and Naïve Bayes (Nbc) Clarifier

Classifier	Total of Instances	Correctly Classified Instances	Incorrectly Classified Instances	Accuracy
J48	400	396	4	99%
Naïve Bayes	400	380	20	95%

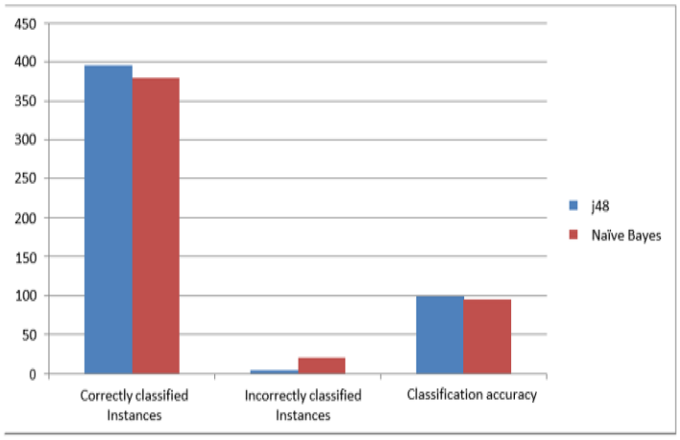

Figure 5 Results of Classification based on Original Dataset

5.3 Reducing Dataset

The dataset has been reduced using the Wrapper attribute value and best-first search method, along with Naive Bayes (NBC) and J48 classifiers. As shown in table 4, the attribute evaluator has reduced and Naive Bayes (NBC) Wrapper Subset Evaluator selects only (5) attributes (hemo, al, sc, su, and wbcc) from (25) total of attributes with 80% attributes reduction. and The Wrapper Subset Evaluator withJ48 selects only (11) (hemo, rbcc, htn, dm, bgr, appet, pe, bu, sg, sod and sc) attributes from (25) total of attributes with 54% attributes reduction. Shows the result of attributes reduction using Wrapper attribute evaluator and Best First Search method with Naive Bayes (NBC) and J48 classifiers.

Table 5 Representation the result of attributes reeducation

Attributes evaluator using BFS Method	Initial Attributes	Selected Attributes	Attributes Reduction ((%)
Wrapper Subset evaluator with J48 Classifier	25	11	54%
Wrapper Subset evaluator with Naïve Bayes Classifier	25	5	80%

Figure 6. illustrates a Graphical representation of data reduction with the wrapper features selection evaluator, which uses Nash Bayes (NBC) and J48 classifiers to select the best features for use in the evaluation.

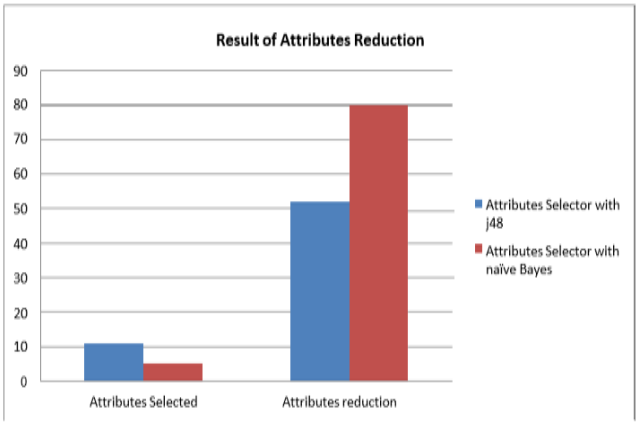

Figure 6. Results of Attributes Reduction

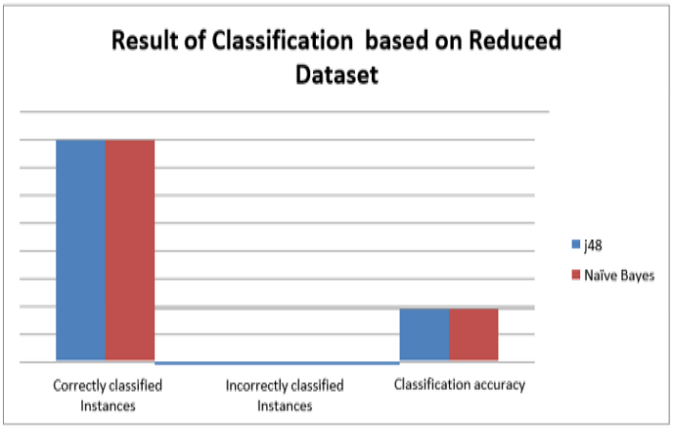

Figure 7. Results of Classification based on Original Dataset

5.4 The Second Experiment

We used (J48) and naive Bayes (NBC) classifiers after reducing the dataset with the wrapper attributes selection evaluator (using the reduced dataset) discussed in the previous section, and then followed the pre-processing below:

- Merge any nominal attribute that multiple values such as (sugar (Su ((0, 1, 2, 3, 4, 5))) in the small range (2 or 3 outcomes).
- Discretize any numerical attribute that multiple distinct value to three categorizes. Table 5 below shows the result of second experiment by using j48 and naïve Bayes (NBC) used cross validation test after dataset reduced.

Table 6. Results of Classification based on Reduced Dataset

Classifier	Total of Instances	Correctly Classified Instances	Incorrectly Classified Instances	Accuracy
J48	400	396	4	99%
Naïve Bayes	400	398	2	99.5%

Based on the results in Table 6, observe j48 classifier classified 396 cases of kidney disease correctly and only 4 cases incorrectly from (400) instances of kidney disease, achieving 99% classification accuracy.

In (400) instances of kidney disease, the Naive Bayes classifier correctly classified 398 instances, and incorrectly classified 2 instances. In addition, the accuracy of the Naive Bayes classifier increased by 5.5%, but the J48 Classifiers did not affect the result. Using the wrapper method evaluator, the Naive Bayes classifier is better than the J48 clarifier. Figure 8 presents a graphic representation of classification accuracy, as well as observing that both the Naive Bayes and the J48 classifiers decided to use certain features for the kidney classification. The (hemo hemoglobin and sc hemoglobin serum creatinine). Both these tests are vital for diagnosing kidney disease, as well as observing that Classifier J48 does not have any significant influence.

6. Conclusion

Chronic-Kidney-Disease was classified using a nave Bayes classifier and J48 classifier and a wrapper attribute evaluation model to reduce dataset. Using the UCI Dataset for CKD, we shall study the effects of this evaluator on the accuracy model. Among the 25 attributes, there is one class attribute, 13 nominal attributes and 11 numerical attributes, along with 400 instances (250 CKDs, 150 notckd). Wrapper Subset Evaluation with Naive Bayes selects only (5) attributes (hemo, al, sc, su, and wbcc) from 25 attributes with an 80% attributes reduction and a classification accuracy of 99.5%. It is better to use the reduced data set after the reduction than the original dataset before the reduction. Since the J48 algorithm uses an internal evaluator to select the best features, the precision of J48 classifier was not affected when using wrapper algorithm. Contributions will focus on the features of (hem (hemoglobin) and Sc (serum creatinine)) as they are very important features of classification (CKD) chronic kidney diseases, then using nave Bayes, with the wrapper method, which gives the best accuracy and performance.

Acknowledgement

The authors would like to thank the Deanship of Scientific Research at Prince Sattam Bin Abdulaziz University, Alkharj, Saudi Arabia for the assistance.

References

[1] Lambodar Jena , "Distributed Data Mining Classification Algorithms for Prediction of Chronic- Kidney-Disease", International Journal of Emerging Research in Management &Technology November 2015.

[2] Anwer Mustafa Mohamedsalih Hilal, Abu Sarwar Zamani, Muhammad Shahid Ghulam Farid and Mohammed Rizwanullah," A Better Prediction for Higher Education Performance using the Decision Tree" IJCSNS International Journal of Computer Science and Network Security, VOL.21 No.4, April 2021.

[3] Mohammad Ayesha ," Extraction of Action Rules for Chronic Kidney Disease using Naïve Bayes Classifier " , IEEE International Conference on Computational Intelligence and Computing Research2016.

[4] Dr. S. Vijayarani, Data Mining Classification Algorithms For Kidney Disease Prediction, International Journal on Cybernetics & Informatics (IJCI) Vol. 4, No. 4, August 2015.

[5] Naganna Chetty, Kunwar Singh Vaisla and Sithu D Sudarsan, "Role of attributes selection in classification of Chronic Kidney Disease patients", International Conference on Computing, Communication and Security (ICCCS), 4-5 Dec 2015, pp 1-6.

[6] S.Ramya, and Dr. N.Radha , "Diagnosis of Chronic Kidney Disease Using Machine Learning Algorithms", International Journal of Innovative Research in Computer and Communication Engineering, Vol. 4, Issue 1, January 2016.

[7] Jiliang Tang, Salem Alelyani and Huan Liu, " Feature Selection for Classification: A Review".

[8] Yvan Saeys, Inaki Inza and Pedro Larranaga, "A review of feature selection techniques in bioinformatics", Bioinformatics, Volume 23, Issue 19, August 2007, pp 2507-2517.

[9] Zheng, Lijuan, Hongwei Wang, and Song Gao, "Sentimental feature selection for sentiment analysis of Chinese online reviews", International Journal of Machine Learning and Cybernetics, 2015.

[10] L. Song, A. Smola, A. Gretton, K. Borgwardt, and J. Bedo. Supervised feature selection via dependence estimation. In International Conference on Machine Learning, 2007.

[11] J. Weston, A. Elisseff, B. Schoelkopf, and M. Tipping. Use of the zero norm with linear models and kernel methods. Journal of Machine Learning Research, 3:1439–1461, 2003.

[12] J.G. Dy and C.E. Brodley. Feature selection for unsupervised learning. The Journal of Machine Learning Research, 5:845–889, 2004.

[13] Z. Xu, R. Jin, J. Ye, M. Lyu, and I. King. Discriminative semi-supervised feature selection via manifold regularization. In IJCAI' 09: Proceedings of the 21th International Joint Conference on Artificial Intelligence, 2009.

[14] H. Peng, F. Long, and C. Ding. Feature selection based on mutual information: criteria of max-dependency, max-relevance, and min-redundancy. IEEE Transactions on Pattern Analysis and Machine Intelligence, pages 1226–1238, 2005.

[15] M. R. Sikonja and I. Kononenko. Theoretical and empirical analysis of Relief and ReliefF. Machine Learning, 53:23–69, 2003.

[16] Lei Yu, Huan Liu, "Feature Selection for High-Dimensional Data:A Fast Correlation-Based Filter Solution", Proceedings of the Twentieth International Conference on Machine Learning (ICML-2003), Washington DC, 2003.

[17] Mineichi Kudo, Jack Sklansky, "Comparison of algorithms that select features for pattern classifiers", Pattern Recognition Society. Published by Elsevier, Volume 33, Issue 1, January 2000, Pages 25-41.

[18] I. Guyon, J. Weston, S. Barnhill, and V. Vapnik. Gene selection for cancer classification using support vector machines. Machine learning, 46(1-3):389–422, 2002.

[19] J. R. Quinlan. Induction of decision trees. Machine learning, 1(1):81–106, 1986.

[20] J. R. Quinlan. C4.5: Programs for Machine Learning. Morgan Kaufmann, 1993.

[21] S. Ma and J. Huang. Penalized feature selection and classification in bioinformatics. Briefings in bioinformatics, 9(5):392–403, 2008.

[22] We broadly classify the Wrapper methods into Sequential Selection Algorithms and Heuristic Search Algorithms. Irish Survey of Student Engagement. (2015). The Irish Survey of Student Engagement (ISSE), 2014. [dataset]. Version 1. Irish Social Science Data Archive. SN: 0030-01. www.ucd.ie/issda/data/isse/isse2014.