USING DATA MINING FOR PHARMACEUTICAL CLASSIFICATION OF FLOWERS: AN APPLICATION OF THE CRISP-DM METHODOLOGY

António Graça
30000497@students.ual.pt

Pedro Simões
30007732@students.ual.pt

Afonso Romeiro
30007229@students.ual.pt

Abstract — Data mining is a machine learning technique that can be used to discover patterns and classify data from large and complex datasets. One of the applications of data mining is in the pharmaceutical field, where it can help to diagnose diseases, predict outcomes, and improve treatments. In this paper, we propose a method for using data mining for pharmaceutical classification of flowers, based on the CRISP-DM methodology. CRISP-DM stands for Cross-Industry Standard Process for Data Mining, and it consists of six phases: business understanding, data understanding, data preparation, modelling, evaluation, and deployment. We use the Iris dataset as an example of flower data that can be classified into three species based on four attributes: sepal length, sepal width, petal length, and petal width. We apply a neural network model to classify the Iris data, and we compare its performance with other models such as decision tree and k-nearest neighbours. We also discuss the potential benefits and challenges of using data mining for pharmaceutical classification of flowers, such as improving the identification of medicinal plants, discovering new compounds, and dealing with noisy and imbalanced data.

I. BACKGROUND

A. Pharmaceutical Classification Of Flowers

Each of the different species of Iris flowers have medical applications that can and are used in pharmaceutical companies. The Setosa species possesses a rhizome that is used as an ingredient in various medicines. However, all parts of Iris setosa are poisonous, and should therefore be processed in a safe way.
The Virginica species, on the other hand, has been used for medicinal practices for a long time. The Cherokee tribes used to pound this plant to a paste, so it could then be used as a salve for the skin. The root of this species can also be used to make an infusion to treat liver problems or a decoction to treat bladder problems.
Finally, the Versicolor species has no medical use, given its roots and leaves are highly poisonous. However, some people have used it as a "magical plant", with people carrying the root (or rhizome) to get 'financial gain', and it was even placed in cash registers to increase business.

B. Bussiness Intelligence & Data Mining

According to Turban et al. (2010), BI is an umbrella term that includes architectures, tools, databases, applications, and methodologies with the goal of using data to support decisions of business managers. DM is a BI technology that uses data-driven models to extract useful knowledge (e.g., patterns) from complex and vast data (Witten and Frank, 2005).
The Cross-Industry Standard Process for Data Mining (CRISP-DM) is a popular methodology for increasing the success of DM projects (Chapman et al., 2000). The methodology defines a non-rigid sequence of six phases, which allow the building and implementation of a DM model to be used in a real environment, helping to support business decisions. According to Turban et al. (2010), BI is an umbrella term that includes architectures, tools, databases, applications, and methodologies with the goal of using data to support decisions of business managers. DM is a BI technology that uses data-driven models to extract useful knowledge (e.g., patterns) from complex and vast data (Witten and Frank, 2005). The Cross-Industry Standard Process for Data Mining (CRISP-DM) is a popular methodology for increasing the success of DM projects (Chapman et al., 2000). The methodology defines a non-rigid sequence of six phases, which allow the building and implementation of a DM model to be used in a real environment, helping to support business decisions.

CRISP-DM defines a project as a cyclic process, where several iterations can be used to allow results to be more tuned towards the business goals. After identifying the goal to achieve (Business Understanding phase), the data needs

to be analysed (Data Understanding) and processed (Data Preparation). According to Witten and Frank (2005), data has concepts (what needs to be learned), instances (independent records related to an occurrence) and attributes (which characterize a specific aspect of a given instance). The Modelling phase builds the model that represents the learned knowledge (e.g., given an instance, the model can be used to predict the target value that represents the goal defined). Next, the model is analysed in the Evaluation phase, in terms of its performance and utility. For instance, in classification tasks (to predict a discrete target), common metrics are the confusion matrix (Kohavi and Provost 1998) and the Receiver Operating Characteristic (ROC) curve (Fawcett 2005). If the obtained model is not good enough for use to support business, then a new iteration for the CRISP-DM is defined. Else, the model is implemented in a real time environment (Deployment phase).

C. Data Mining on Pharmaceutical Classification Of Flowers

Data mining is the process of discovering patterns and insights from large and complex datasets. One of the applications of data mining is in the medical field, where it can help to diagnose diseases, predict outcomes, and improve treatments. One of the datasets that is commonly used for data mining in the medical field is the Iris dataset, which contains measurements of four features (sepal length, sepal width, petal length, and petal width) of 150 iris flowers belonging to three species (setosa, versicolor, and virginica). The Iris dataset can be used for classification tasks, which aim to assign a label to each instance based on its features. For example, a classification model can be trained to predict the species of an iris flower given its measurements. This can help to identify rare or endangered species, or to detect anomalies or diseases in the flowers. Classification models can also be used to group similar instances together based on their features, which can reveal hidden patterns or relationships in the data. For example, a clustering model can be used to find subgroups of iris flowers that have similar characteristics or behaviours. This can help to understand the diversity and evolution of the iris species, or to discover new varieties or hybrids.

II. Materials And Methods

A. Pharmaceutical Classification Of Flowers Data

The Iris dataset is a famous multivariate dataset that was introduced by the British statistician and biologist Ronald Fisher in his 1936 paper The use of multiple measurements in taxonomic problems as an example of linear discriminant analysis. It contains 150 samples from three species of Iris flowers: Iris setosa, Iris versicolor and Iris virginica. Each sample has four features measured in centimetres: the length and the width of the sepals and petals. The dataset is often used as a benchmark for various statistical classification techniques in machine learning, such as support vector machines. The dataset is available for download from various sources, such as Kaggle and UCI Machine Learning Repository (being the latter the one used).

B. Computational Environment

All experiments reported in this work were conducted using the 'caret' library (Max Kuhn 2007), which is an open-source package for the R tool that facilitates the use of data mining, and modelling techniques. In this paper, we used three data mining 'caret' models: LDA, CART, KNN, SVM, and RF, while the rattle tool was used for graphical data exploration, in Data Understanding phase.

III. Experiments

In this section, we explain the experiments performed and analyse the results obtained over a set of three CRISP-DM iterations.

A. Project viability and goal definition

One of the applications of data mining is pharmaceutical classification of flowers, which aims to identify the medicinal properties and potential uses of different flower species based on their characteristics and features.

"The Iris dataset is a well-known and widely used dataset in data mining and machine learning. It contains 150 instances of three classes of iris flowers: Iris setosa, Iris versicolor, and Iris virginica. Each instance has four features: sepal length, sepal width, petal length, and petal width. The Iris dataset can be used to demonstrate the basic concepts and techniques of data mining for pharmaceutical classification of flowers."

The project viability and goal definition of using data mining for pharmaceutical classification of flowers using the Iris dataset are as follows:

- The project is viable because it can provide valuable information and insights for the pharmaceutical industry, botany, agriculture, and trade activities. By using data mining techniques, such as feature extraction, feature selection, feature engineering, dimensionality reduction, and classification algorithms, it is possible to identify the medicinal properties and potential uses of different iris flowers based on their features and characteristics. For example, some iris flowers may have anti-inflammatory, antibacterial, antiviral, or antioxidant effects that can be useful for treating various diseases and conditions.

- The project goal is to choose a data mining model that can accurately and efficiently classify iris flowers into their respective classes based on their features and characteristics. The model should also be able to explain the rationale behind its predictions and provide confidence scores or probabilities for each class. The model should be evaluated using appropriate metrics, such as accuracy, and its confusion matrix. The model should also be compared with other existing models or methods for pharmaceutical classification of flowers using the Iris dataset or similar datasets.

B. Variable and instance selection

Variable and instance selection are important steps in data mining for pharmaceutical classification of flowers using the Iris dataset. Variable selection aims to reduce the number of features or attributes that are used to classify the flowers, while instance selection aims to reduce the number of samples or examples that are used to train the classifier. Both steps can improve the accuracy, efficiency, and interpretability of the classification model. Given that the dataset used in this case is quite small, there is no need remove any variable, since they are all necessary for the problem at hand. Nonetheless, the label, as selected, is the class, the species of the flower (setosa, versicolor, virginica).

IV. Data Understanding

Data understanding is the process of exploring and analysing data to gain insights and identify potential problems or opportunities. It is an essential step in any data science project, as it helps to define the objectives, scope, and methods of the analysis*. One of the most widely used datasets for data understanding is the Iris dataset, which contains measurements of four features (sepal length, sepal width, petal length and petal width) of 150 iris flowers belonging to three different species (setosa, versicolor and virginica). The Iris dataset is useful for demonstrating various data understanding techniques, such as descriptive statistics, data visualization, correlation analysis and clustering.*

A. Exploratory Data Analysis: Univariate analysis of Iris Data set

1) What is Exploratory Data Analysis?

Exploratory data analysis (EDA) is a process of analyzing data by using simple concepts from statistics & probability and presenting the results in easy-to-understand pictorial format.

2) What is our Objective?

The Iris flower data set consists of 50 samples from each of three species of Iris Flowers: Iris Setosa, Iris Virginica and Iris Versicolor. The data set has four features measured from each sample: sepal length, sepal width, petal length and petal width in centimeters. The goal is to classify the Iris flower into one of the three species given these four features.

For all the species, the respective values of the mean and median of its features are found to be pretty close. This indicates that data is nearly symmetrically distributed with very less presence of outliers. Box plot (explained later) is one of the best statistical tools used for outlier detection in the data.

species	sepal_length	sepal_width	petal_length	petal_width
setosa	0.352490	0.381024	0.173511	0.107210
versicolor	0.516171	0.313798	0.469911	0.197753
virginica	0.635880	0.322497	0.551895	0.274650

Image 1 — Computing the standard deviation (or variance) is an indication of how widely the data is spread about the mean.

a) Box Plots and Violin Plots

Box plot*, also known as a box and whisker plot, displays a summary of a large amount of data in five numbers — minimum, lower quartile (25th percentile), median (50th percentile), upper quartile (75th percentile) and maximum data values.*

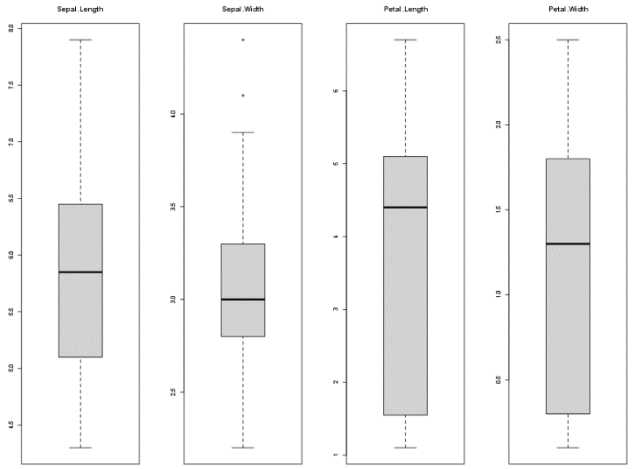

Image 2 — Boxplot of the four features.

The isolated points that can be seen in the boxplots above are the outliers in the data. Since these are very few in number, it wouldn't have any significant impact on our analysis.

b) Violin Plot

A violin plot plays a similar role as a box and whisker plot. It shows the distribution of data across several levels of one (or more) categorical variables (flower species in our case) such that those distributions can be compared. Unlike box plot, in which all of the plot components correspond to actual data points, the violin plot additionally shows the kernel density estimation of the underlying distribution.

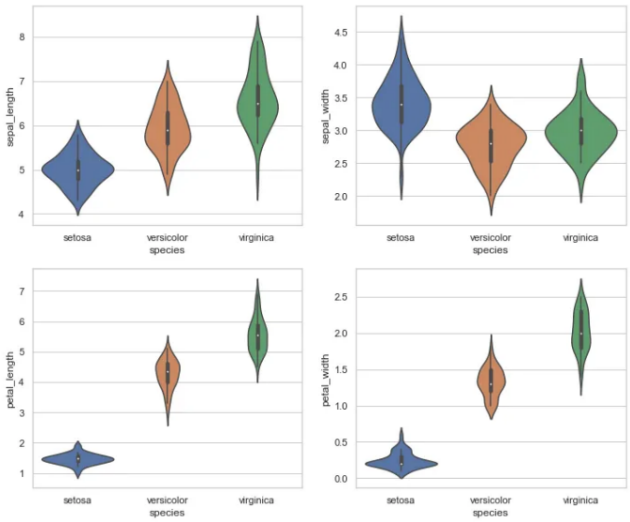

Image 3 — Violin plots comparing the class species with the four features.

Violin plots typically are more informative as compared to the box plots as violin plots also represent the underlying distribution of the data in addition to the statistical summary.

c) ***Probability Density Function (PDF) & Cumulative Distribution Function (CDF)***

Uni-variate as the name suggests is one variable analysis. Our ultimate aim is to be able to correctly identify the specie of Iris flower given it's features — sepal length, sepal width, petal length and petal width. Which among the four features is more useful than other variables in order to distinguish between the species of Iris flower? To answer this, we will plot the probability density function (PDF) with each feature as a variable on X-axis and it's histogram and corresponding kernel density plot on Y-axis.

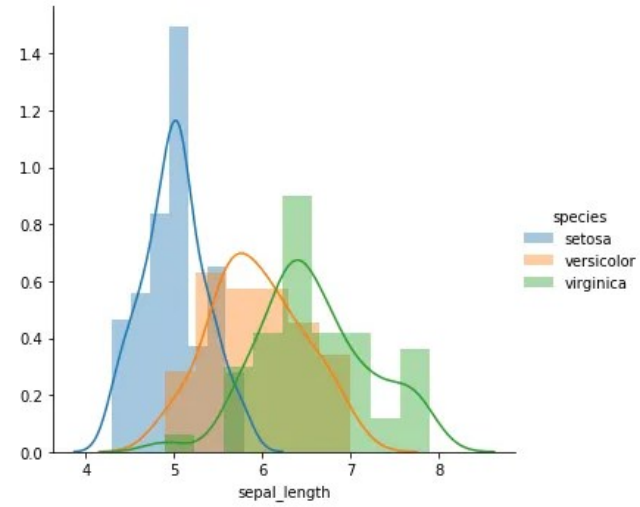

Image 4 — Classification feature: Sepal Length

The density plot alongside reveals that there is a significant amount of overlap between the species on sepal length, so it wouldn't be a good idea to consider sepal length as a distinctive feature in our univariate analysis.

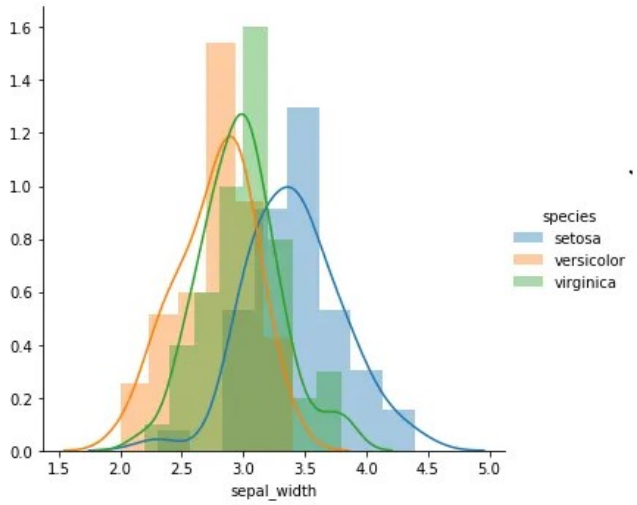

Image 5 — Classification feature: Sepal Width

With sepal width as a classification feature, the overlap is even more than sepal length as seen in Plot 1 above. The spread of the data is also high. So, again we cannot make any comment on the specie of the flower given its sepal width only.

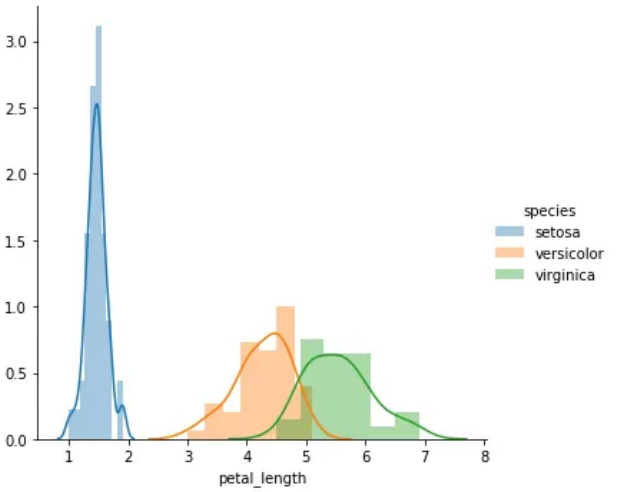

Image 6 — Classification feature: Petal Length

The density plot of petal length alongside looks promising from the point of view of univariate classification. The Setosa species are well separated from Versicolor and Virginica, although there is some overlap between the Versicolor and Virginica, but not as bad as the above two plots.

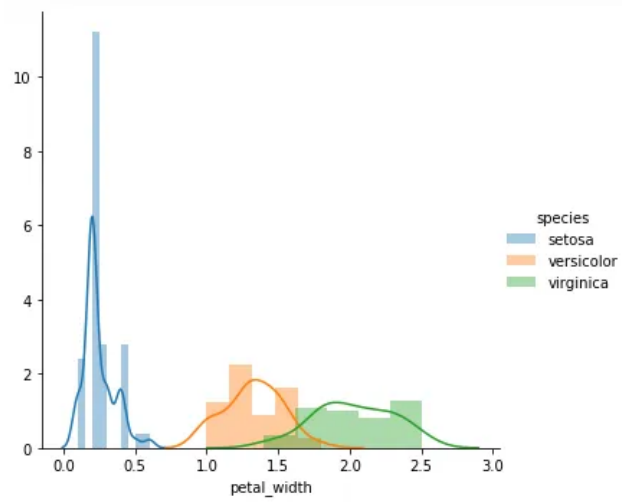

Image 7 — Classification feature: Petal Width

The density plot of petal width alongside also looks good. There is slight intersection between the Setosa and Versicolor species, while the overlap between the Versicolor and Virginica is somewhat like that of petal length.

To summarize, if we must choose one feature for classification, we will pick petal length to distinguish among the species. If we must select two features, then we will choose petal width as the second feature, but then again it would be a wiser to look at pair-plots (bi-variate and multivariate analysis) to determine which two features are most useful in classification.

We have already established above how petal length could stand out as a useful metric to differentiate between the species of Iris flower. From our preliminary investigation, below pseudo-code can be constructed. (Note that this estimation is based on the kernel density smoothed probability distribution plots obtained from histograms.)

Although the Setosa is clearly separated, there is a small overlap between the Versicolor and Virginica species. The reason why we intuitively considered 4.8 mark to distinguish between Virginica and Versicolor is because from the density plot, we can clearly see that although not all, but majority of the Versicolor flowers has petal length less than 4.8 while majority of the Virginica flowers has petal length greater than 4.8.

With this preliminary analysis, it is quite possible that some Versicolor flowers whose petal length is greater than 4.8 will get incorrectly classified as Virginica. Similarly, some Virginica flowers whose petal length happen to be less than 4.8 will get incorrectly classified as Versicolor.
Is there some way to measure what proportion or what percentage of Versicolor and Virginica flowers will be incorrectly classified with above analysis? That's where Cumulative distribution plots comes into the picture!

d) ***Using Cumulative Distribution Function (CDF) plots to quantify the proportion of misclassified flowers in above analysis***

The area under the plot of PDF over an interval represents the probability of occurrence of random variable in that interval. In our analysis, petal length is the random variable.

Mathematically, CDF is an integral of PDF over the range of values that a continuous random variable takes. CDF of a random variable evaluated at any point 'x' gives the probability that a random variable will take a value less than or equal to 'x'.

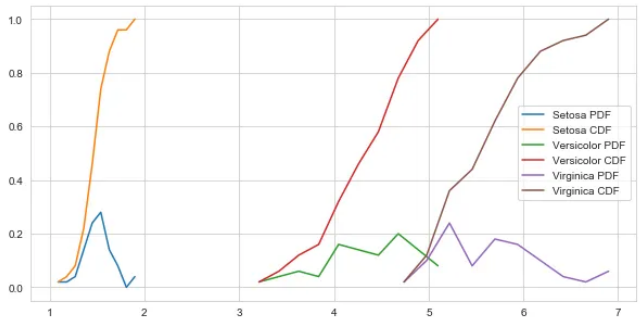

Image 8 — PDF and CDF of Iris flower species

From the above CDF plots, it can be seen that 100% of the Setosa flower species have petal length less than 1.9. Near about 95% of the Versicolor flowers have petal length less than 5, while about 10% of the Virginica flowers have petal length less than 5. So, we will incorporate our newly found insights into our previously written pseudo-code to construct a simple univariate 'classification model'.

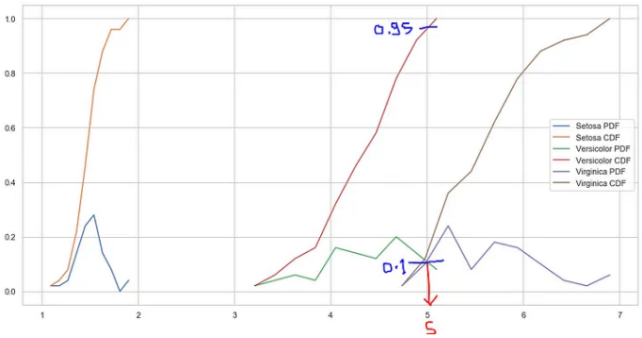

Image 9 — Cumulative distribution plot of petal length for various species

Thus, by using the cumulative distribution plot, we get a better picture and robust understanding of distribution leading to formulation of simple univariate classification model.

V. RESULTS

The iris dataset is one of the most widely used and well-known datasets in the field of machine learning and statistics. It was introduced by Ronald Fisher in 1936 as an example of linear discriminant analysis, a technique for classifying objects based on a linear combination of their features. The dataset consists of 150 samples of three different species of iris flowers: Iris setosa, Iris virginica and Iris versicolor. Each sample has four features measured in centimetres: the length and width of the sepals and petals. The goal of the analysis is to find a way to distinguish the species from each other based on these features.

In this chapter, we will present the results of applying various machine learning methods to the iris dataset, such as k-means clustering, support vector machines, decision trees and neural networks. We will compare the performance of these methods in terms of accuracy, precision, recall and F1-score. We will also visualize the data and the models using different techniques, such as scatter plots, confusion matrices and metro maps. We will discuss the advantages and disadvantages of each method and provide some insights into the characteristics of the iris dataset and its implications for machine learning.

Statistical Model	***1st Run***		***2nd Run***		***3rd Run***	
	Min.	***Max.***	***Min.***	***Max.***	***Min.***	***Max.***
LDA	*83.33%*	*97.50%*	*91.60%*	*97.50%*	*91.67%*	*97.50%*
CART	*91.66%*	*95.83%*	*83.33%*	*95%*	*91.67%*	*96.67%*
KNN	*91.66%*	*98.33%*	*91.67%*	*96.67%*	*91.67%*	*99.16%*
SVM	*83.33%*	*95.83%*	*83.33%*	*93.33%*	*83.33%*	*95.83%*
RF	*83.33%*	*95%*	*83.33%*	*95.83%*	*83.33%*	*95.83%*

The table shows the accuracy of different classification models. The accuracy of each model is calculated by dividing the number of correct predictions by the total number of predictions. The highest accuracy is achieved by KNN model in the 2nd run with 98.33% and the lowest accuracy is achieved by SVM model in the 1st and 3rd run with 83.33%.

VI. CONCLUSION

In this paper, we performed a data analysis on the iris dataset, which contains 150 observations of three species of iris flowers. We applied various statistical and machine learning techniques to explore the data, such as descriptive statistics, correlation analysis, principal component analysis, clustering, and classification. We found that the iris dataset is well-structured and easy to work with, and that the three species are clearly distinguishable by their sepal and petal measurements. We also demonstrated how to visualize the data and the results of our analysis using different types of plots. Our paper provides a comprehensive and practical guide for anyone who wants to learn more about data analysis using the iris dataset.

VII. REFERENCES

[1.] Aptéa, C. and Weiss, S. 1997. "Data mining with decision trees

[2.] and decision rules", Future Generation Computer Systems

[3.] 13, No.2-3, 197–210.

[4.] Chapman, P., Clinton, J., Kerber, R., Khabaza, T., Reinartz, T.,

[5.] Shearer, C. and Wirth, R. 2000. CRISP-DM 1.0 - Step-by-step

[6.] data mining guide, CRISP-DM Consortium.

[7.] Coppock, D. 2002. Why Lift? – Data Modeling and Mining,

[8.] Information Management Online (June).

[9.] Cortes, C. and Vapnik, V. 1995. "Support Vector Networks",

[10.] Machine Learning 20, No.3, 273–297.

[11.] Cortez, P. 2010. "Data Mining with Neural Networks and

[12.] Support Vector Machines using the R/rminer Tool". In

[13.] Proceedings of the 10th Industrial Conference on Data

[14.] Mining (Berlin, Germany, Jul.). Springer, LNAI 6171, 572

[15.] 583.

[16.] Cortez, P. and Embrechts, M. 2011. "Opening Black Box Data

[17.] Mining Models Using Sensitivity Analysis". In Proceedings

[18.] of IEEE Symposium on Computational Intelligence and Data

[19.] Mining (Paris, France), 341-348.

[20.] Fawcett, T. 2005. "An introduction to ROC analysis", Pattern

[21.] Recognition Letters 27, No.8, 861–874.

[22.] Hu, X. 2005, "A data mining approach for retailing bank

[23.] customer attrition analysis", Applied Intelligence 22(1):47

[24.] 60.

[25.] Kohavi, R. and Provost, F. 1998. "Glossary of Terms", Machine

[26.] Learning 30, No.2–3, 271–274.

[27.] Ling, X. and Li, C., 1998. "Data Mining for Direct Marketing:

[28.] Problems and Solutions". In Proceedings of the 4th KDD

[29.] conference, AAAI Press, 73–79.

[30.] Li, W., Wu, X., Sun, Y. and Zhang, Q., 2010. "Credit Card

[31.] Customer Segmentation and Target Marketing Based on Data

[32.] Mining", In Proceedings of International Conference on

[33.] Computational Intelligence and Security, 73-76.

[34.] Ou, C., Liu, C., Huang, J. and Zhong, N. 2003. "On Data Mining

[35.] for Direct Marketing". In Proceedings of the 9th RSFDGrC

[36.] conference, 2639, 491–498.

[37.] Page, C. and Luding, Y., 2003. "Bank manager's direct

[38.] marketing dilemmas – customer's attitudes and purchase

[39.] intention". International Journal of Bank Marketing 21,

[40.] No.3, 147–163.

[41.] Turban, E., Sharda, R. and Delen, D. 2010. Decision Support and

[42.] Business Intelligence Systems – 9th edition, Prentice Hall

[43.] Press, USA.

[44.] Williams, G. 2009. "Rattle: a data mining GUI for R", The R

[45.] Journal, 1(2):45-55.

[46.] Witten, I. and Frank, E. 2005. Data Mining – Pratical Machine

[47.] Learning Tools and Techniques – 2nd edition, Elsevier, USA.

[48.] Zhang, H. 2004. "The Optimality of Naïve Bayes", In

[49.] Proceedings of the 17th FLAIRS conference, AAAI Press.