Chapter 18

Improving Decision Support Systems with Data Mining Techniques

Adela Bâra and Ion Lungu

Additional information is available at the end of the chapter

http://dx.doi.org/10.5772/47788

1. Introduction

1.1. Definition and characteristics of decision support system

The Decision Support System concept goes back a long time, the definition varies depending on the evolution of information technologies and, of course, on the point of view of those who issues such a definition.

Looking through several definition we can find that Moore and Chang defined the DSS as "an extensible system, capable of ad-hoc analysis and decision modeling, focused on future planning and used at unplanned and irregular timestamps" [10]. Also Carlson and Sprague cited by [3] define decision support systems as being "interactive systems that help decedent makers use data and models in resolving unstructured and semi-structured economical problems".

In 1998 Turban defines a decision support system as "an interactive, flexible and adaptable system, exclusively designed to offer support in solving unstructured or semi-structured managerial problems, aiming to improve the decisional process. The system uses data (internal and external) and models, providing a simple and easy-to-use interface, thus, allowing the decision maker control over the decision process. The DDS offers support in all decision process's stages". [9]

In this context, studies show that the process of defining a Decision Support System has started from the idea of how the objectives of a DDS can be achieved, how a DDS's components can be identified, the features that are provided to the end user and from the perception of what such a system is capable of doing (offering support in decision making processes, in solving structured and unstructured problems).

Holsapple and Whinston, in the [5], specify five characteristics of a decision support system: contains a knowledge base that describes certain facets of the decision maker's universe (for

example, the way certain activities of the decision-making process); has the ability of purchasing and managing descriptive knowledge, as well as other types of knowledge (procedures, rules etc.); has the ability of presenting ad-hoc knowledge in a periodic report format; has the ability of selecting a subset of knowledge for viewing purposes or for deriving other knowledge, mandatory in the decision making process; is able to interact directly with the decision maker, allowing choosing flexible solution and knowledge management.

In conclusion, considering all the definitions mentioned above, some of the most important characteristics of the DDSs are: uses data and models; enhances the learning process; grows the efficiency of the decision making process; offers support in the decision making process and allows the decision maker control over the entire process; offers support in all stages of the decision making process; offers support for decision makers in solving structured or un-structured problems; offers support for a user or for a group of users etc.

1.2. Data mining techniques in decision support systems

In order to make a decision, the managers need knowledge. In case of massive data amounts, issues may occur because of data analysis and necessary knowledge extract. Data is analyzed through an automated process, known as Knowledge Discovery in data mining techniques.

Data mining can be defined as a process of exploring and analysis for large amounts of data with a specific target on discovering significantly important patterns and rules. Data mining helps finding knowledge from raw, unprocessed data. Using data mining techniques allows extracting knowledge from the data mart, data warehouse and, in particular cases, even from operational databases.

In this context, data mining gets an important role in helping organizations to understand their customers and their behavior, keeping clients, stocks anticipation, sale policies optimization as well as other benefits which bring a considerable competitive advantage to the organization.

The main purpose of these techniques is to find patterns and hidden (but relevant) relations that might lead to revenue increase. The essential difference between data mining techniques and the conventional database operation techniques is that, for the second ones, the database becomes passive and is only being used for large amounts of data population, therefore helping in future finding of that specific data. Alternatively, the database is not passive anymore, being able to serve useful information regarding the business plans put in discussion.

Regarding data mining studies, two major types of them exists. One of them is represented by the hypothesis testing, which assumes exposing a theory regarding the relation between actions and their results. The second type of study is represented by the knowledge discovery. For this type of analysis, relations between data warehouse existing data are tracked. This can be done by using data viewing tools or by using fundamental statistical analysis, such as correlation analysis.

Data mining techniques reside from classic statistical calculation, from database administration and from artificial intelligence. They are not a substitute for traditional statistical techniques, but an extension of graphical and statistical techniques.

Data mining uses a large variety of statistical algorithms, shape recognition, classification, fuzzy logic, machine learning, genetic algorithms, neural networks, data viewing etc., from which we can mention regression algorithms, decision algorithms, neural networks, clustering analysis.

Regression algorithms. Regression represents a basic statistical method. In the case of data mining, it is also an important analysis tool, used in classification applications through logical regressions as well as forecasted reports measured using the least square or other methods. Non-linear data can be transformed into useful linear data and analyzed using linear regressions. The universal test for data mining classification is the coincidence index matrix. It is primarily focused on data classification abilities of the model. For continuous regressions, class inflection points must be identified. The applications of the methods into solving business problems are multiple.

Decision trees. In data mining technology, decision trees represent rules tree-view structures, also known as joining rules. The trees' creation mechanism of the trees consists in collecting all the variables the analyst assumes might help the decision making and analyzing them considering their influence into result estimation.

The algorithm automatically determines which of the variables are the most relevant, based on the ease of data sorting. The decision tree algorithms are applied in Business data mining in areas like: loan request classification, applicants ranking for various positions.

Neural networks. This is one of the most commonly used data mining method. It consists of taking sets of observations and placing them in a relational system through arc-connected nodes. This idea derives from the way neurons act inside the human brain. Neural networks are usually structured in at least three layers, having a constant structure allowing reflection of complex non-linear relations. Each entry data has a node in the first layer, while the last layer represents the output data – the result. In order to classify the neural network model, the last layer (containing the output) has a corresponding node for each category. In most of the cases, this type of networks also have a mid node layer (hidden) which adds complexity to the model. The obtained results are compared to the targeted ones, and the difference is re-entered in the system for node's cost adjustments. The process keeps looping until the network correctly classifies the input data (at a tolerance level).

Clustering analysis. One of the most general forms of this type of analysis allows the algorithm to determine the number of subsets. Partitioning is mainly used for defining new variable categories, which divide raw data in a precise number of regions (k-means clustering). Considering a random number of centers (k), data is associated to the center which is the closest to it. The basic principle of this analysis is to identify the average characteristic for different indicators in sets of data. Thus, new observations can be measured by reporting the deviation from the average. This analysis is often the base

technique applied in a data mining study, being used in client segmentation and, implicitly, taking a segment-oriented action.

1.3. The process of developing a DSS using data mining techniques

Developing Decision Support Systems involves time, high-costs and human resources efforts and the success of the system can be affected by many risks like: system design, data quality, and technology obsolescence. The decision support systems objective is to assist the managers and executives to make decision regarding the benefit of investment, budgeting cash flows and financial planning, especially in the case of public funds.

Presently, many institutions invest in building organizational data warehouses and data marts in order to increase the performance and the efficiency of the analytical reporting activity. Also, there are several expensive tools and software that can be used to analyze the trends and to predict some future characteristics and evolution of the business. Some of these tools analyze data from the statistic perspective or by using neural networks. In our opinion, in order to build an efficient decision support system there must be combined several techniques and methods that can improve the performance and the accuracy of the analysis from two major perspectives: historical data and forecasts. This requirement can be obtain by combining data warehousing, OLAP, data mining and business intelligence tools for analyzing and reporting into a flexible architecture that must contains: A data model's level where an ETL process must be apply to clean and load data into a data warehouse or data marts; An application level with analytical models where multidimensional reporting like OLAP and data mining techniques can be combined to for historical and forecast analysis; An interface level where dashboards and reports can be build with business intelligence tools.

In the chapter it will be presented the consideration regarding to design the DSS's architecture and there will be described the methods and ways for data mining integration into a data warehouse environment. In the paper [6], the authors propose a series of development stages for business intelligence systems: *feasibility study, project planning, analysis, design, development and release into production.*

These stages can be adapted and applied in decision support systems, but during the development cycle it is mandatory that differences between general system modeling and decision support systems modeling must be treated separately, in order to obtain a successful business requirements of implementing the specifications.

Stage 1. The feasibility study consists of identifying the requirements and business opportunities and proposing solutions of improving the decision making process. Each of the proposed solutions must be justified by the implied costs and benefits.

Stage 2. Project planning consists of evaluating project sustainability possibilities, indentifying existent infrastructure components and future needs. The result of these activities concludes with *the project plan*. After its validation and approval, the effective start of the project can begin.

Stage 3. Business requirements analysis. *This stage focuses on detailing and analyzing on priority the initial requirements* of the organizational management team. Usually, the requirements are indentified based on interviews conducted by managers and the project staff. These requirements might suffer slight changes during the project, but the development team must make the managers aware of the capabilities and limitations of a DSS, therefore reducing the risk of un-feasible business requirements to occur.

Data analysis – the biggest challenge of a decision support system development project – consists of identifying necessary data, analyzing its content and the way it relates to other data. Data analysis is focused on business analysis rather than system analysis performed in traditional methodologies. It is preceded by a data cleaning activity.

Data cleaning implies transforming and filtering data sources in order to be used in building the destination module – the analysis module. This process is done by: identifying necessary data from the functional modules; analyzing the content of the selected data sources; selecting the appropriate data for the project; implementation of data filtering related specifications; selecting the tools to be used in the filtering / cleaning process. During the source selection process, a few key aspects must be taken into consideration: data integrity, precision, accuracy and data format. These facets are critical in regards to the success of the new ETL process.

Metadata analysis is an important activity in which all the identified requirements would be transformed depending on the metadata structure, and stored in a metadata dictionary. A metadata dictionary contains contextual information on the data implied in the project. The system analysis phase can end by *building a prototype* which will be presented to the managers and project staff for functional specifications' validation. The existence of quick development tools allow building new interfaces based on the analysis model.

An important step in this stage is *choosing the technologies used* in the prototype's development and, later on, in the final system. Based on a comparative analysis over advantages and disadvantages brought by each of the technologies on the project, different approaches might be taken into consideration: usage of data warehouses, including OLAP (Online Analytical Processing) functionalities, usage of knowledge extract algorithms, data source integration tools or, on a final phase and assuming a parallel approach on building the system has been taken, usage of applications integration tools.

Stage 4. System design. *Database / data warehouse design*. According to the system's requirements, the necessary data will be stored both on a detailed level as well as on aggregate level, therefore relational, object-oriented or multi-dimensional data storage approaches might be taken. During this sub-phase, the logical data model is refined and detailed and the physical model of the new system is developed in order to satisfy the reporting and analysis requirements of the managers.

While on "Data analysis", the process has been oriented to data sources (*data-in* or *data-entry*) coming from operational modules, in this phase the targets or data destinations (*data-out*) are set aiming on reports, analysis and queries. Therefore, a list of best practices must be taken into consideration:

Due to the above mentioned aspects, we recommend that the storage, management and data processing solution to consist of a centralized data warehouse on an organizational level. Following logical and physical criteria, the data warehouse can be divided into data marts on departmental level, thus being easier to maintain and developed by separate teams, following the same set of specifications.

The ETL (extract / transform / load) process design – this phase is the most complex one in the project's lifecycle and is directly dependant on the data sources' quality.

We recommend the integration of all the destination databases in a single environment and building the ETL process on it, avoiding a separation of each destination module, thus mitigating the risk of distinct data marts. The strategy of building data marts in the same environment is also viable, but only on the condition that these are already integrated. The important fact here is that the ETL process must remain the same for all levels (the *share one coordinated process* principle).

The design of the ETL process needs a series of pre-requisite stages: preliminary processing of data sources, in order to have a standardized format, data reconciliation and redundancy and inconsistency elimination of data.

The steps to be taken in creating an ETL process are the following:

1. *Creation of transformation specifications* (mapping) of the sources in regards to the specific destinations. This may be done as a matrix or as transformation diagrams.
2. *Choosing and testing the ETL tools to be used*. At the moment, a series of ETL process modeling and implementation tools exist on market, but choosing one of them would depend on the features they provide and on the support of data source integration inside the same transformation process.
3. *The ETL process design* – several extract and transform operators are used, depending on the data model (sorting, aggregation, joining, dividing operators, etc.). The process can be split into sub-processes that would run separately in order to minimize the execution time. The execution flow of the process will be modeled using flow diagrams.
4. *ETL programs design*. Depending on the program in which the data is loaded, three phases of data loading are applied:
   - initial load – the initial load of destinations with current operational data
   - historical load – the initial load of destinations with archived historical data
   - incremental load – regular loading of destinations with current data coming from operational systems
5. *Choosing the environment for running the ETL process* – represents the decision over using a dedicated server / machine or the process would be divided and run decentralized. The decision depends on the available resources and on the processing time, as well as on the timelines that the process is scheduled to run.

The results of these activities is materialized in the data mapping documentation, the flow diagram / diagrams of the ETL process, the transformation programs documentation and the process execution specifications.

Metadata repository design – if the repository is acquired and a predefined template is used, then, in this sub-stage, slight changes may occur according to the requirements identified in the metadata analysis sub-stage, but, if the option has been to build a proprietary repository, then the metadata logical model will be implemented for the new system, based on the data storage options: a relational, object-oriented or multidimensional model will be implemented.

If the option was for building a proprietary warehouse, we consider that centralization and standardization of it would represent a good strategy into a more facile administration. The activities performed in this stage are materialized in the detailed logical model and the metadata physical model.

Stage 5. Building the system. The technologies that are used for decision support systems' development are part of the business intelligence technologies category and consist of: technologies for data warehouse data organization, OLAP (On-Line Analytical Processing) analysis systems, data mining algorithms, extract, transform and load (ETL) tools, CASE (Computer-Aided Software Engineering) modeling tools and web technologies.

Stage 6. System implementation. Represents the stage when the system is being delivered, training sessions are held for implied managers / business owners, the necessary technical support is provided, data loading procedures are run, the application is installed and the performance is being tracked.

The stage ends with the release of the system into production (commercial go-live) and with the delivery of the utilities and final project documentation, the user guides and presentation manuals for the application.

1.4. The DSS architecture

Depending on the requirements indentified in the analysis phase, all of these technologies can be merged and combined, creating reliable decision support system architecture. The field literature (for example in [2], [3], [4]) proposes a typical decision support system architecture that contains distinct levels which use the above mentioned technologies in order to be created.

In [3], the definition provided by Bonczek and Holsapple, the main components of a decisional system are emphasized: a DSS is described as being "a system composed of three interacting modules: *the user interface* (Dialog Management), *the data management component* (Data Management), *the model management component* (Model Management)". In [4] it is identified four core components that form a decision support system: *the interface*, often considered to be the most important component, *the database system* which includes all the databases and the database management systems (DBMS) of the organization, *the model system* containing the analytical, mathematical and statistical models and *the communication component*, composed of the core network and the mobile devices.

The DSS architecture can also be seen from a development level point of view, from bottom to top, pyramidal, having three layers: bottom-tier middle-tier and top-tier, the connection

of all these layers being made on the telecom layer. Thus, DSS architecture might be composed of the following levels:

Level I (bottom-tier) – data management. It is composed of data, metadata, DBMS (database management systems), data warehouses, data dictionaries and metadata dictionaries. At this level, data coming from several different systems must be integrated and the main techniques used for this process are replication, federalization or data migration, together with data warehouse loading. Data coming from operational databases and external sources are extracted using interface-type applications, known as gateways, running on DBMS and allowing client-applications to generate server-side executable SQL code. During the extract and data processing, different tools can be used – filtering, predefined procedures etc. Data cleaning and transformations is strongly dependent on the data sources and on the quality of it. Several sources may be put in discussion: files, databases, e-mails, internet and unconventional sources.

This sub-stage focuses on implementing the ETL designed requirements, running and testing them.

At this architectural level, in order to load data in a data warehouse, a series of tasks is mandatory:

- Collecting and extracting data from the data sources that have been identified during the analysis phase, according to the management's business requirements. A source data warehouse (staging area) can be created in order to load all the necessary data which then must be processed and loaded into another destination warehouse. This process, most often, transforms raw data for compliance with the internal format of the warehouse;
- Data cleaning and transformation to assure data accuracy and to confirm data can be used for analysis;
- Loading data in the destination warehouse.

This process is extremely important for the success of the future decision support system, thus a faulty design could lead to the failure of the DSS. Subsequently, a method for data refresh must be taken into consideration as time passes. Therefore, the ETL process must be automatically run on accurate timestamps defined during the design and analysis stages. The destination warehouse data will be used at higher levels of the system. If the chosen approach has been to build a new data warehouse (custom development) instead of purchasing a predefined solution, the metadata dictionaries used by the components and utilities of the system must be build and adapted to the solution, integrated and the connection interfaces between the user and the metadata centralized dictionary and the dictionaries used by each component separately must also be developed. Thus, the structure of the dictionary is created and the data will be loaded according to the logical and physical models already designed.

Level II (middle-tier) – model management or analysis level. This is the level where data is processed and the necessary information for decision making is extracted. This level

contains data analysis, simulation and forecast models, in order to respond to the high level business requirements. On this level, the core components are: *the model base, the model database management system, the meta-models, the model management and execution server*.

For creating this level, *OLAP technology* can be used. It is based on multidimensional data representation and allows quick and interactive data analysis by using roll-up, drill-down, slice or dice operations.

If the business requirements requires, at this level, *knowledge extract algorithms using data mining algorithms* can be designed and implemented. These algorithms assure data transformation into knowledge using statistical analysis or artificial intelligence techniques and allowing the identification of correlations, rules and knowledge in order to support the decision making process. In order to integrate the analysis and models resulted from different sub-system types, several application integration technologies could be used: application servers that implement middleware models, service-oriented architecture (SOA), Java platforms.

Extracting knowledge from data (Data Mining) – very often, the success of a DSS is determined by the discovery of new facts and data correlations and not by building reports that just presents data. In order to fulfill these requirements, data mining techniques must be applied, together with knowledge extract from the organizational data, such as: clustering, forecasting, predictive modeling and classification.

In this context, an analysis regarding the data mining applicability domains, the specific algorithms to be used and the teams that would develop these initiatives must be done as a mandatory task.

The tasks to be performed are:

1. *Domain and applicability objective setting for data mining techniques* – specific requirements that cannot be resolved through other methods are analyzed, as well as the opportunity of applying data mining techniques and the way they would solve the problem
2. *Data collecting* – a validation is preformed in order to check the existence of the loaded data; if not, new data sources are set and previous steps regarding data design need re-validation
3. *Data consolidation and cleaning* – when data is not loaded in the correct format, cleaning processes are applied. The ETL processes implemented in Stage II may change due to these additional requirements.
4. *Data setup* – data mining algorithms use setup steps by formatting and loading data, in order to be compliant to the agreed techniques.
5. *Building the analytical model* – the step focuses on implementing the data mining algorithms and the specifications for the learning and testing stages
6. *Result interpretation* – according to the agreed requirements and objectives, a validation of the results must be performed. The measured values are interpreted and it is decided whether they can be used by managers.
7. *Result validation* – the measured values are compared to the expected ones, accepted deviations and errors are set based on statistics or comparative analysis and why these

deviations have occurred. If the result can be used, they are presented to the managers (business owners) for the final validation.

8. *Analytical model monitoring* – the performance of the model is tracked in a timeline already defined in the requirements phase.

After applying data mining techniques, a complete database is obtained and will be used by specific programs, as well as the analysis model specifications.

Level III (top-tier) – The interface or the presentation layer. Represents the level where the interaction with the users takes place, where the managers and the persons involved in the decision making process can communicate with the system and can analyze the presented results. The user interface must be specially designed so that this type of users will easily interact with the system. This level is composed of *queries and reports generating tools, dynamic analysis tools* (data viewing using different perspectives, post-implementation evolution analysis, forecasting, data correlations), *data publishing and data presenting tools* in a simple, intuitive and flexible way for the end-users. On this level, the *human resource* can be found, represented by *decision makers* which interact with the system through its interfaces. In the last few years, a growing share in the development of decision support system interfaces is taken by the *portal-based web technologies*. Business Intelligence portals hold the most important position in creating specialized, flexible, user-friendly and accessible interfaces, allowing users a good end-to-end experience, nice graphical appearance, report integration options and graphic tools, obtained in the previous stages.

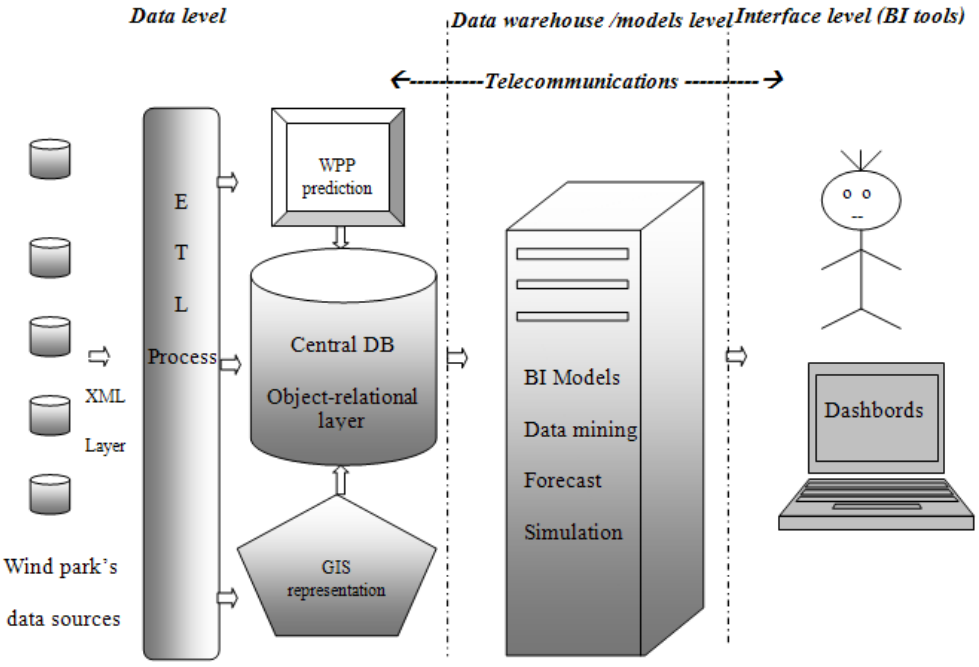

Figure 1. The DSS architecture used for WPP systems

Level IV – Telecommunications. Represents the level that allows interconnecting all the previous levels and may contain web servers, computer networks, communication devices, distributed platforms, GRID technologies and mobile communication platforms.

Based on these steps and DSS' architecture, in the following sections we propose a conceptual model that can be implemented in the case study in which the National Power System's (NPS) activity was analyzed, especially the wind power pants' production and wind energy integration into the NPS.

The architecture is explained in [1] where all the components are presented in detail. The entire case study consider the characteristics of wind power plants, the ways to integrate the energy produced and the impact on decision-making system from the technical point of view (the characteristics and impact on reserve power), financial, commercial, environmental, legal (by analyzing the issues raised by access to the electric wind power). The next section present only the methods applied to predict and determine the wind energy based on data mining techniques and the results obtained for some of the algorithms.

2. Case Study: applying data mining for forecasting the wind power plants' energy production

2.1. The problem formulation: the characteristics of wind energy production

Setting a wind farm is particularly important because its energy production depends to a several meteorological factors of that site. The main criteria for determining the location of wind generating units are wind speed and direction, orography and terrain conditions, environmental conditions, distance from the electricity grid, substations and connection conditions, access to equipment and personal interference human activities (tourism, proximity to settlements, roads, railways, airports), electromagnetic interference, conditions on land use and relationships with local authorities the power of its consumers.

But the main natural factor, wind speed, records significant fluctuations even within hours. Optimal range of wind speeds that produce wind turbines' energy ranges from 3 to 25 m / s If the wind speed falls below this limit or exceeding 27 m / s, the turbines stop. Even for an area such as Dobrogea (where the area is windy) the wind speed in certain areas of land varies significantly. Therefore, to determine if the location is suitable it is necessary to measure the meteorological factors, such as wind speed and direction, temperature and pressure, etc.

From the above mentioned in the investment phase is particularly important to determine the energy produced by wind sources in order to choose the type of wind generator and location of each production station and also in the operational phase to achieve good production forecasts. But forecasts in wind power plants (WPP) still records significant deviations from the real values of energy products due to the inability of present systems to correctly estimate the wind speed. The problem becomes more complex because the forecasts obtained are used to establish the energy resources necessary to cover any gaps in the energy system. It is well known that rapid availability of power is questionable or

expensive. If the forecast of wind energy from wind sources is more accurate the more it reduces the power system reserves. The role of a good prognosis is particularly important because it reduces the costs of ensuring safe operation National Power System (NPS) and therefore has no significant increases in energy prices as a result of these reserves.

But the most important problem, which depends on the amount of energy reserves, and return on investment in wind power, is the component on which these wind power plants are based, namely the wind. From Figure 2 one can see large fluctuations of the wind recorded by an anemometer within 24 h at a height of 50 m

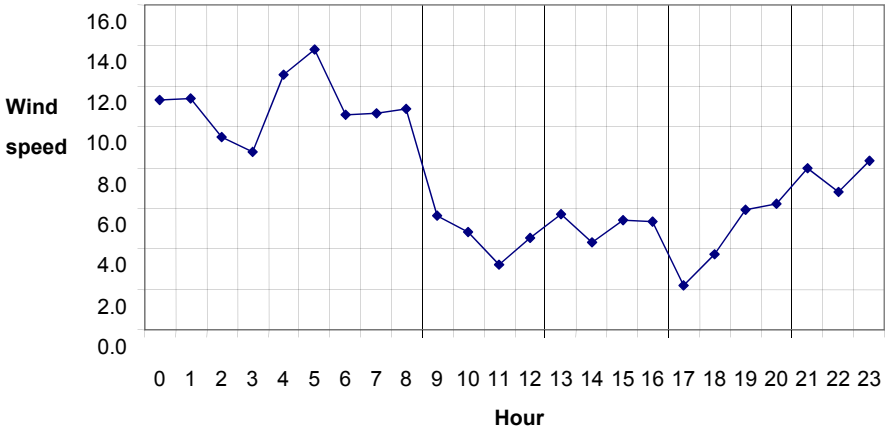

Figure 2. Wind speeds measured within 24 h in a location in Dobrogea area

Wind energy production is conditioned by other factors, some of which are characterized by low predictability, such as the effect of shading, soil orography, the power, losses to the point of connection, etc. Currently, there are several informational systems used for the prediction of energy but the accuracy of these systems is still quite low. This can be seen analyzing the notifications forwarded by the wind power energy production within 7 days (Figure 3).

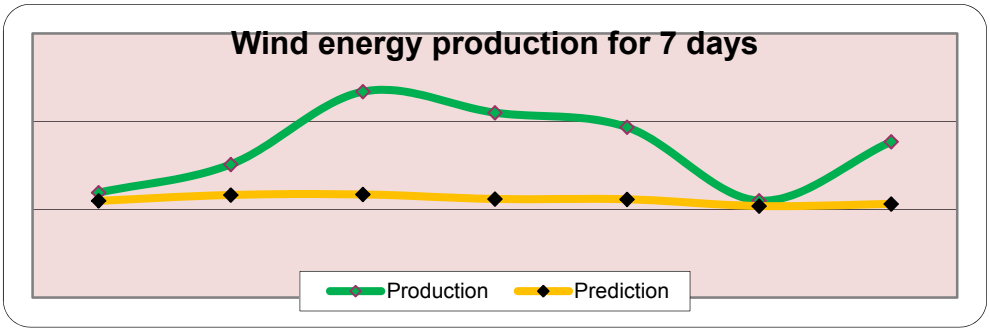

Figure 3. Comparison of energy prediction and actual wind production (source [8])

These problems regarding the low prediction accuracy and of data integration from various equipment and local systems, energy efficiency analysis, lead to the need to develop solutions for a better predictive power as products, but also to support decision making in this area.

A better prediction cannot be achieved by classical statistical methods, and this is the reason for requiring the use of modern techniques like data mining. In these considerations we have been analyzed in the following section, in detail, the main algorithms that can be applied to predict the wind.

2.2. Proposing an effective model for predicting wind energy

For building and testing the data mining algorithms we'll use Oracle Data Miner (ODM) software tool developed by Oracle Corporation that provides a friendly interface for data analysis and validation results. Oracle Data Miner provides several tools (wizards) for the data processing and for the stages of preparation, training, testing and evaluation required in data mining technology.

Oracle Data Miner implements the following types of algorithms [7]:

- Predictive models or supervised training: classification algorithms, regression algorithms and selection of important attributes;
- Descriptive models or unsupervised learning: clustering, association rules, extraction algorithms for new attributes based on existing ones;
- Models for multimedia (TEXT) and bio-informatics (BLAST).

For all the algorithms the data preparation is required. In our case study the data were measured and recorded from 10 to 10 minutes between 09.11.2009 - 28.02.2010. The values recorded at height of 50 m count 16,037 records. The minimum value recorded in this period is 0 m / s, maximum value of 24.8 m / s and average 6.9 m / s.

From the set consisting of the 16,037 records of wind speed at 50 m height, about 2500 were lower or equal to 3.5 m / s - start speed of a wind generator (GGE). In approx. 1100 cases wind recorded a speed exceeding 12 m / s Approximately 8,800 measured values were lower than average speed, and about 7200 values above average wind speed (6.9 m / s).

For the algorithms, the source table entries were imported and divided into three sets for each stage that will be completed. Thus these sets of records will be inserted in three tables, namely: *wind_build*, *wind_test*, *wind_apply* (Figure 4).

Each table contains information on different time intervals in which measurements are made as follows: for table for learning are considered records of the period 09.11 - 15.01 (about 10,000 records), the table for the testing process are considered records of the period 16.01 - 15.02 (about 4600 records), and for evaluation the table contains records in period 16.02 - 23.02 (about 1100 records). Tables can be viewed directly in Oracle Data Miner interface by accessing data sources.

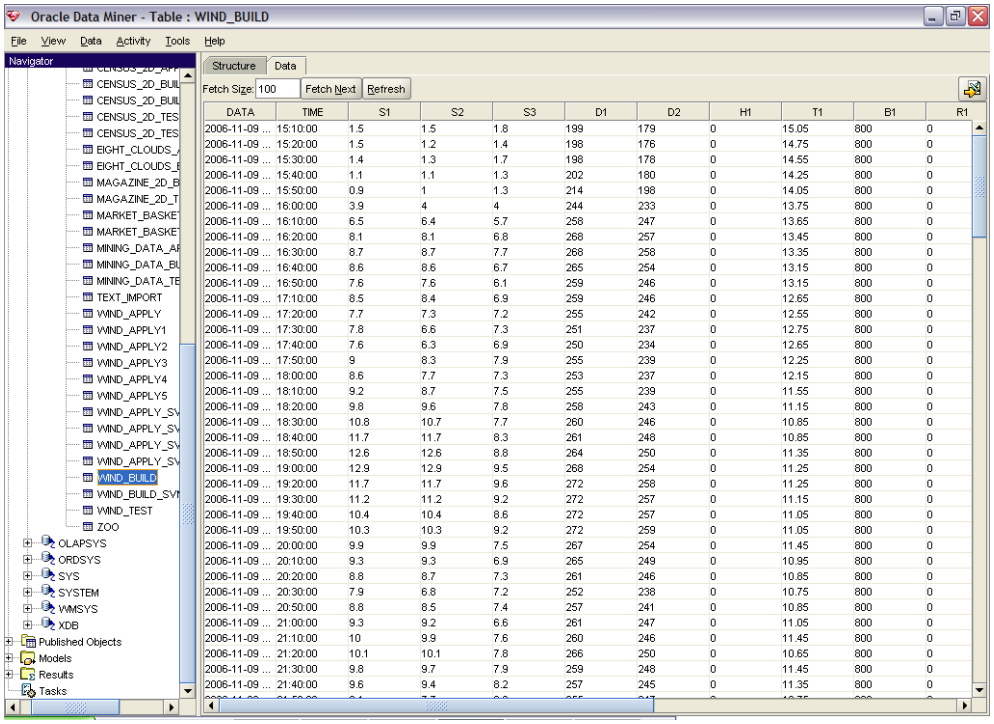

Figure 4. The records prepared for data mining

After the data preparation step, we applied the following algorithms: Naïve Bayes with an error rate of about 8%, decision tree with a 1% error rate, regression with error rate of 43% and after analyzing the results we modified the regression model and obtained a significant increase in prediction accuracy from 57.68% to 93.72%. The steps and the results are presented in the next sections.

2.2.1. Naïve Bayes results

We'll further present how to apply the Naïve Bayes (NB) algorithm on the measured data to analyze the target attribute E-01. Attribute E-01 has two values: 1 – the turbine produces energy when wind speed is within the range 3.5 to 25 m / s and 0 otherwise.

By applying the NB algorithm will forecast whether or not the turbine will produce energy depending on weather conditions.

It will go through three distinct phases:

- The learning stage consists in applying the algorithms NB on the *wind_build* table data set and build the analysis model;
- The test phase, the model built in the previous step is tested on the table *wind_test*;

- The validation phase, the model built on data set is applied on *wind_apply* table to check the results obtained from the algorithm.

The learning stage. In this stage we applied the templates of analysis to build the model on the table *wind_build*.

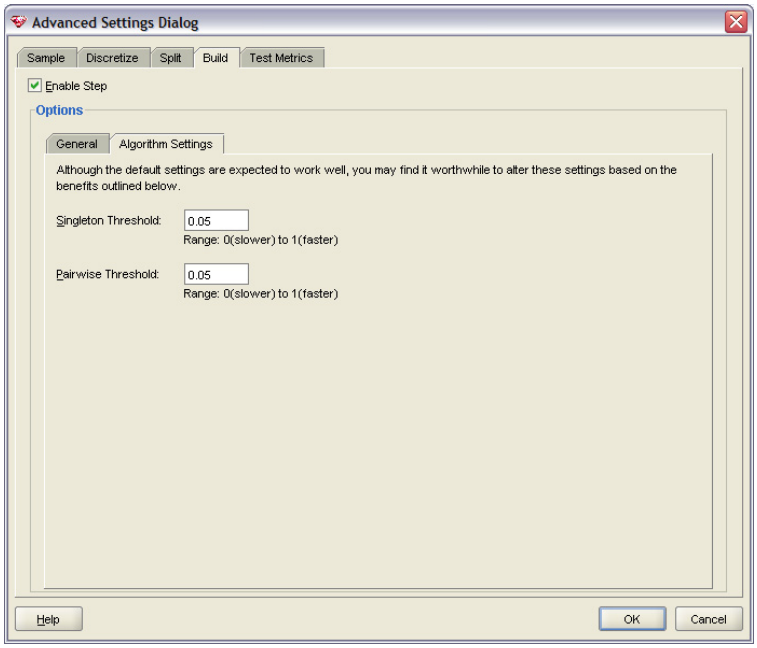

Figure 5. Setting the parameters for the implementation of Naive Bayes algorithm

For this scope a number of steps are required to configure the Oracle Data Miner algorithm parameters, such as setting the model name (*NB_wind_build*), source table, and minimum thresholds for the interpretation of outliers (Figure 5). In our case, is considered the minimum threshold of 5% - only data above this threshold will be considered on the basis of learning (for example, large fluctuations of wind speed are observed at high variations of temperatures whose incidence is rare).

We obtained a 88.6% accuracy of predictions with the NB algorithm.

The test phase. After building the NB model for our data set, we applied the testing algorithm. The results are shown in Figure 6.

The validation phase. To validate the results we considered three sets of validation data (figure 7): table wind_apply with data during 16-23.02.2010 period, wind_apply1 table with data from 24 h (from 02/24/2010) and wind_apply2 table with data from 24h (02/25/2010).

The accuracy of predictions obtained for the three sets is: 91% for table wind_apply (104 erroneous predictions of 1152 records), 99% for table wind_apply1 (1 prediction error of 144 records), 91% for table wind_apply2 (13 of erroneous predictions of 144 records).

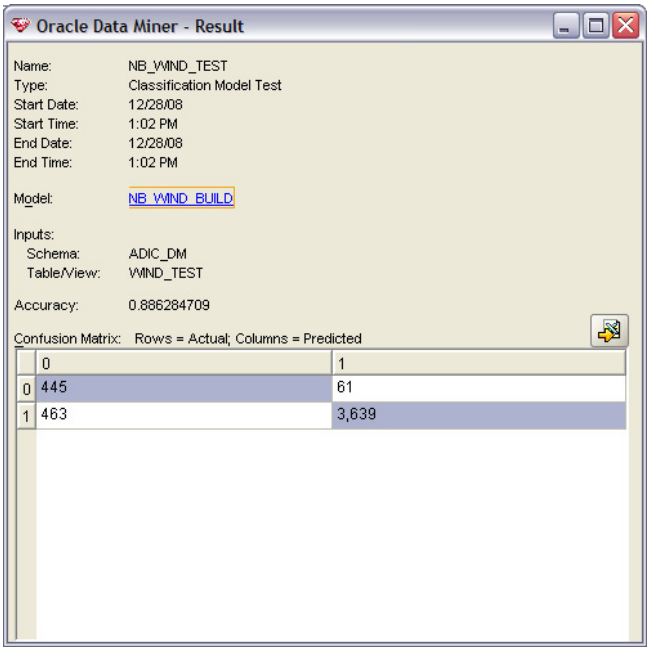

Figure 6. The results of NB algorithm on the test set

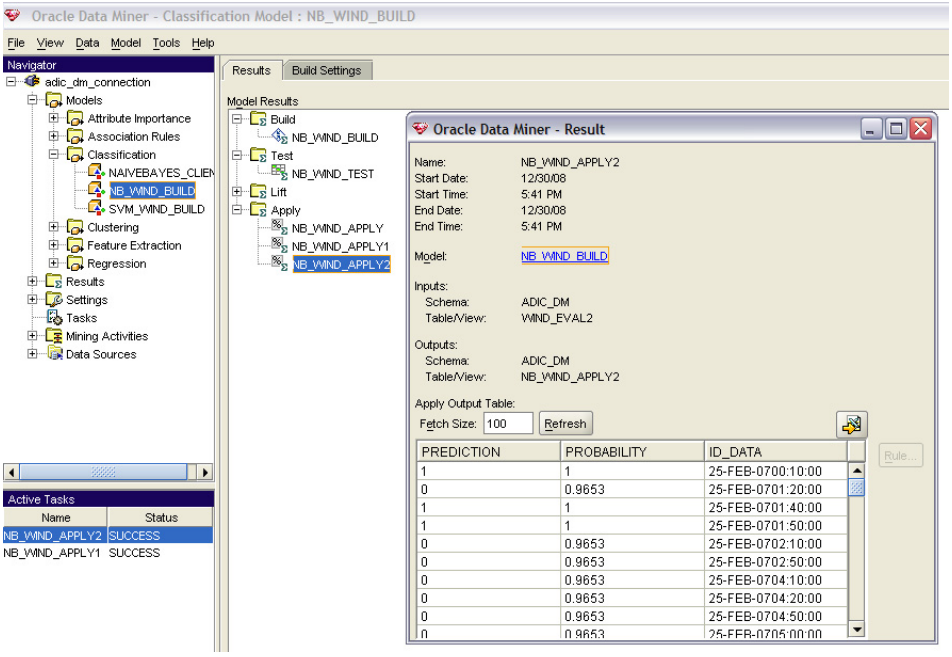

Figure 7. The results of NB algorithm

In conclusion, the error rate resulting from the application of NB algorithm is less than 8% which is considered to be satisfactory.

2.2.2. Decision Tree results

Another prediction algorithm applied to E_01 variable is the Decision Tree. After building and testing the model on the data sets, following the same steps presented in the previous section, we obtain an accuracy of 99.48% (Figure 8), higher than that obtained by applying the NB algorithm.

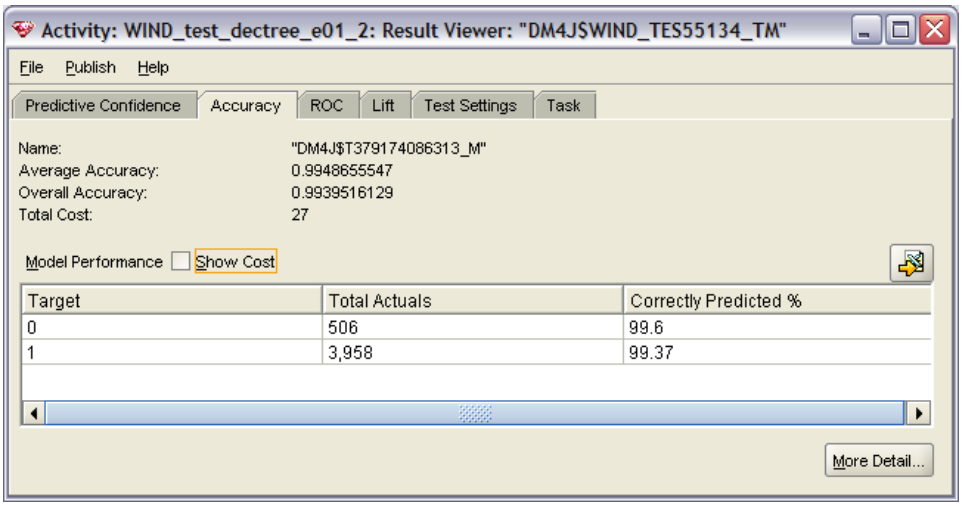

Figure 8. The results obtained by applying the decision tree algorithm

The results detailed on the three tables: the set from table wind_apply recorded an error of only 0.6%, for table wind_apply1 no errors were detected and the table wind_apply2 recorded an error of 0.7%.

In conclusion the results obtained by applying the Decision Tree algorithm are better than those obtained with the NB algorithm. But to get a real energy prediction of turbine's actual values is necessary to apply other algorithms where the target attribute has discrete values, not only values 0 or 1.

2.2.3. The regression results

On the initial data set we introduced column E which is the amount of power produced by wind speed (S2) measured at 50 m cubed. The values in this column will be the target attribute for the regression algorithm. We applied the regression on the data sets, following the same steps (preparation, learning, testing and applying) and the results obtained from the algorithm have an accuracy of only 57.68% (Figure 9) which gives no confidence to achieve rigorous forecasts.

We observed significant errors between actual values and the predicted values ranges from ± 250 kW (figure 9), which would mean that if the current value of the power produced is 50 kW then the algorithm will predict a value of 300 kW, the difference between them being unacceptable.

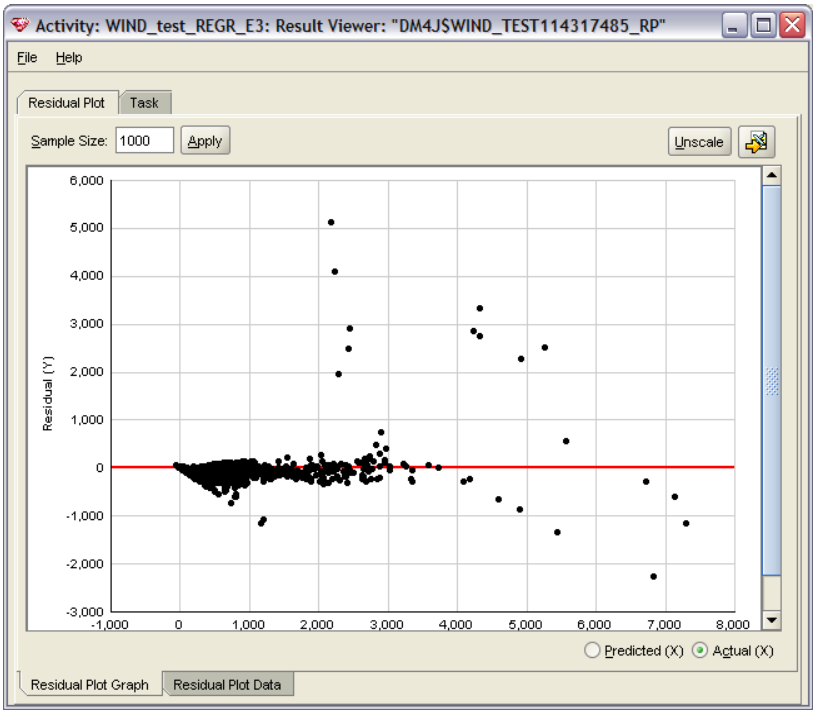

Figure 9. Deviations registered for the Regression model

Consequently, the regression model should be applied to an attribute with a low degree of scattering depending on meteorological factors. Thus, we introduced E_PRAG attribute for grouping values into intervals depending on power produced by wind velocity of 0.5 m / s. For example, we found that at wind speeds between 0 and 3.5 m / s there is 0 kW power output, at speeds ranging from 3.5 to 4 m / s power output is 43 kW, etc.. These thresholds are defined in accordance with the power characteristics of the turbines.

After building the regression model on these thresholds it shows a significant increase in prediction accuracy from 57.68% to 93.72%. Applying the model on the test set (the table wind_test) and observing the diagram in Figure 10 we found that the variation of residual value y, the deviation between the actual value and predicted value lies within ± 50 kW, which is an acceptable deviation. By placing the cursor on any point on the diagram (capture in figure 10) one can view the following information: the current value (in our case 1500 kW), the prediction (1492 kW) and the deviation Y (7.5 kW).

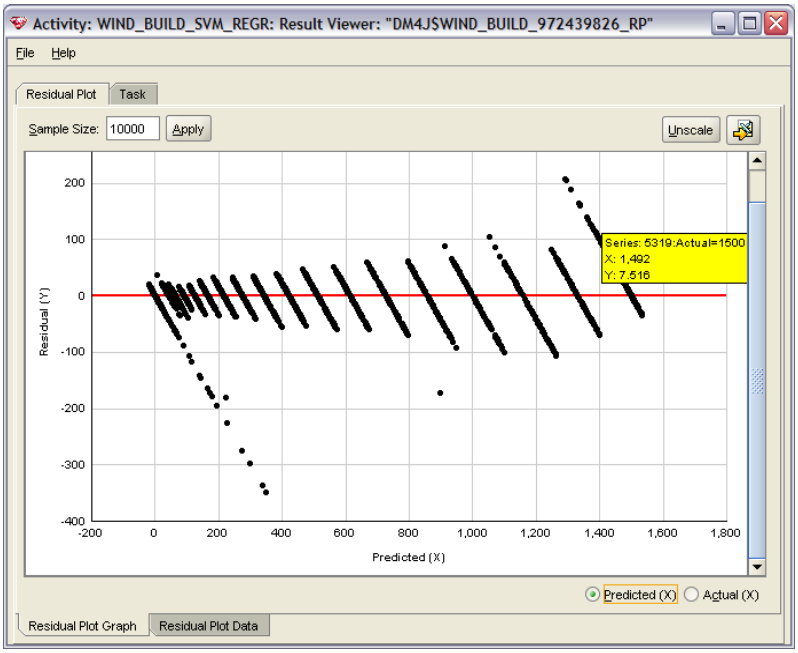

Figure 10. The regression results on E_Prag attribute with thresholds values

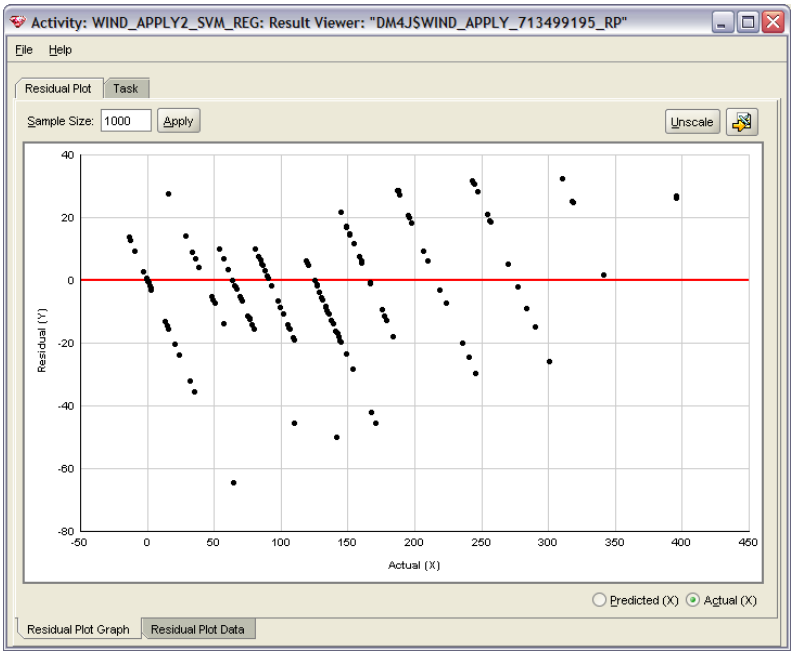

Figure 11. The deviations recorded for the E_PRAG attribute forecast for one day

At the evaluation process for the regression model for E_PRAG attribute, the results are presented in Figure 11. It is noted that deviations are located within ± 20 kW which means that at the current value of 1000 kW the results in a forecast is between 980 kW and 1020 kW, which is an acceptable error.

Summarizing the results of the evaluation phase in Table 1 (only for the schedule 0:00 to 5:00), it shows the actual and predicted values for E_PRAG attribute every 10 minutes.

Date/Time	E actual	E prog	E_prag actual	E-prag prog
25-02-201000:00:00	357.91	779.1	343	341.1
25-02-201000:10:00	250.05	694.7	216	245.7
25-02-201000:20:00	140.61	617.5	125	170.4
25-02-201000:30:00	140.61	620	125	170.7
25-02-201000:40:00	110.59	612.7	91	141.2
25-02-201000:50:00	54.87	651.1	43	56.9
25-02-201001:00:00	0	506.4	0	15.5
25-02-201001:10:00	0	695.4	0	64.5
25-02-201001:20:00	0	771.3	0	20.4
25-02-201001:30:00	79.51	618.4	64	109.5
25-02-201001:40:00	140.61	627.9	125	167.2
25-02-201001:50:00	125	622.3	125	153.3
25-02-201002:00:00	0	695.7	0	23.6
25-02-201002:10:00	0	782.9	0	14.6
25-02-201002:20:00	0	766.1	0	13.1
25-02-201002:30:00	0	798.9	0	32
25-02-201002:40:00	0	755.6	0	2.5
25-02-201002:50:00	0	739.6	0	0
25-02-201003:00:00	0	797.1	0	35.4
25-02-201003:10:00	0	682.3	0	0.8
25-02-201003:20:00	0	657.3	0	0
25-02-201003:30:00	0	733.9	0	1.6
25-02-201003:40:00	0	720.1	0	0
25-02-201003:50:00	68.92	602.6	64	66.1
25-02-201004:00:00	0	643.9	0	0.3
25-02-201004:10:00	0	506	0	0
25-02-201004:20:00	0	473.2	0	0
25-02-201004:30:00	42.88	465.8	43	15.5
25-02-201004:40:00	0	454	0	3
25-02-201004:50:00	54.87	599.1	43	38.7
25-02-201005:00:00	91.13	554.5	91	89.6

Table 1. Comparison of actual and estimated values from the regression algorithm

Applying data mining algorithms for the prediction power of WPP, notable results were achieved in particular by setting the thresholds for E_PRAG. Thus the data mining algorithm was able to learn and to establish better dependence between variables and the prognosis is much closer to actually measured values.

Finally was done the forecasting model for wind power plants produced energy, which can be applied successfully in a DSS prototype according to the architecture presented in Part I.

Author details

Adela Bâra and Ion Lungu
Academy of Economic Studies, Bucharest, Romania

Acknowledgement

This paper is a result of the research project PN II, TE Program, Code 332: "Informatics Solutions for decision making support in the uncertain and unpredictable environments in order to integrate them within a grid network", financed within the framework of People research program.

3. References

[1] Bâra A., Velicanu A., Botha I., Oprea S. V., Solutions for the Data Level's Representation in a Decision Support System in Wind PowerPlants, MAMECTICS'11 International Conference, 1-3 July 2011, Iasi, Romania, Publisher WSEAS Press, ISBN 978-1-61804-011-4
[2] Lungu I, Bara A – Executive Information Systems, ASE Publishing House, Bucharest, 2007
[3] Muntean M. – Initiation in OLAP technology: theory and practice, ASE Publishing House, Bucharest, 2004
[4] Power D.J. - Decision Support Systems: Concepts and Resources, Cedar, Falls, IA: DSSresources.com, http://dssresources.com/dssbook/
[5] Holsapple C.W, Whinston A.B - Decision Support Systems: A knowledge – Based Approach, West Publishing Company, 1996
[6] Moss L., Atre S. – Business Intelligence Roadmap – The complete project lifecycle for decision-support applications, Addison-Wesley, 2004
[7] ORACLE Corporation – Oracle Data Mining Concepts 10g, www.oracle.com, 2010
[8] www.transelectrica.ro, wind production reports, 2011
[9] Turban, E. - Decision Support Systems and Intelligent Systems, 5th ed.,Englewood Cliffs, New Jersey, Prentice Hall, 1998
[10] Watson H., R. Kelly Rainer, Chang E. Koh - Executive Information Systems: A framework for Development and a Survey of Current Practices. MIS Quarterly, Vol. 15, No. I, March, 1991.
[11] Ackerman T., Wind Power in Power Systems, Wiley, 2005

[12] Bâra A., Lungu I., Oprea S. V., Public Institutions' Investments with Data Mining Techniques, Journal WSEAS Transactions on Computers, Volume 8, 2009, ISSN: 1109-2750, http://www.worldses.org/journals/computers/computers-2009.htm

[13] Burton T., Sharpe D. - Wind Enegy Handbook, Wiley, 2001

[14] Oprea S – Renewable resources integration into the GRID. PhD Thesis, Politehnica University, Bucharest, 2009

[15] Simona Vasilica Oprea, Adela Bâra, Victor Vlăducu, Anda Velicanu - *Data Level's Integrated Model for the National Grid Company's Decision Support System*, Recent Advances in Computers, Communications, Applied Social Science and Mathematics, International Conference on Computers, Digital Communications and Computing (ICDCCC'11), Barcelona, Spain, 15-17 septembrie 2011, pp 167-172, ISBN: 978-1-61804-030-5, published by WSEAS Press, http://www.wseas.us/books/2011/Barcelona/ICICIC.pdf