# Proposal for VanHack Artificially Intelligent Recruiter

This document is aimed to provide a proposal for implementing a machine lerning technique to implement an Articial Intelligent Recruiter.

## Motivation

The motivation for this proposal is to give answers to the following questions:

* How do we know who is the best person for a job?
* What can we do to automatically determine the best VanHacker for a position, even if they didn't apply?

## Objective

We are proposing the development of an artificially intelligent recruiter that can understand what a job description is looking for and predict / recommend the best VanHacker for the job from the database. The goal is to be able to bring companies great talent fast and to get VanHackers hired even faster.

## State of Art

The VanHack database is a plain database, there are only plain tables. There are no relatioship between tables, no views, no stored procedures or functions. Besides that, the data are bad structured. There are many useless and inconsistent informations that makes the task of applying data mining techniques impossible for the moment. In Figure 1, we can notice the Entity-Relationship diagram.

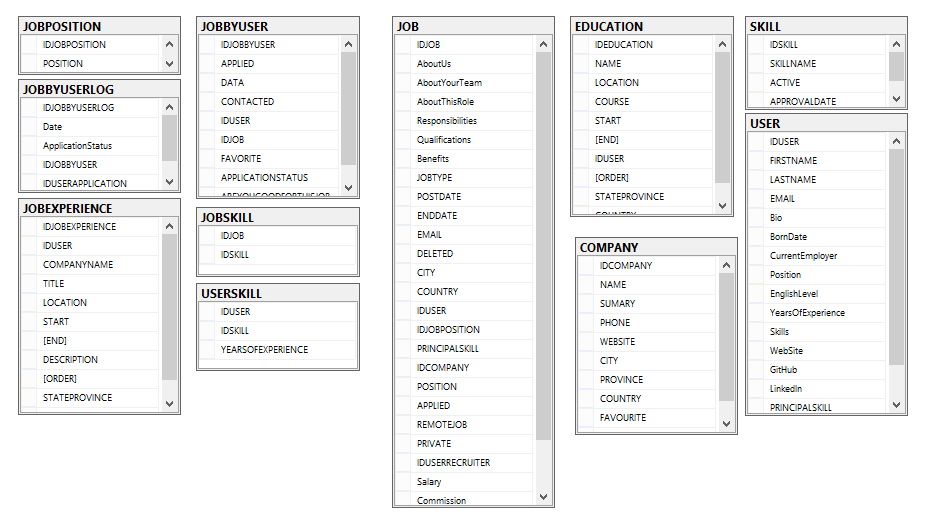


Figure 1 - ER Diagram for Hackathon Database

In Figure 1, we can see that the Job table is the most importante table in the database, because it stores informations of Jobs and there are many informations about the Jobs opportunities. The tables SKILL, JOBSKILL, USERSKILL, JOBPOSITION, JOBBYUSERLOG, JOBBYUSER, JOBEXPERIENCE, EDUCATION, COMPANY AND USER are master data informations.

Although, in this database, it is not clear to check the positions applied by the users because there are some repeated fields in correlated tables like JOBBYUSERLOG.ApplicationStatus and JOBBYUSER.APPLIED. In addition, there are many null datas in columns.

Regarding the history of hired people, there is a Spreadsheet that must be traslated into a database table. In Figure 2, we can see some informations about the history of sucess applications, there are no registers of failed applications.

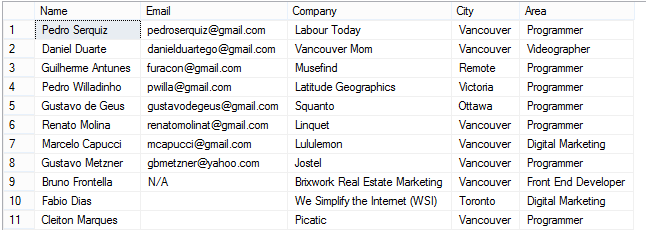


Figure 2 - ApplicationHistory spreadsheet

In Figure 2, we can also identify that there is no information about the date of user has been hired by the company, neither the job position applied by the user. So, lacks many importante informations to make the data stored in hackathon database useful for the business and for the recruiting process.

Therefore, due to the problems identified above we cannot proceed to implement a machine learning technique for the VanHack Artificially Intelligent Recruiter. Instead, we go to propose a methodology to implement the machine learning technique.

## Background Information

### Data Mining

Data mining is the process of discovering actionable information from large sets of data. Data mining uses mathematical analysis to derive patterns and trends that exist in data. Typically, these patterns cannot be discovered by traditional data exploration because the relationships are too complex or because there is too much data.

These patterns and trends can be collected and defined as a data mining model. Mining models can be applied to specific scenarios, such as:

* Forecasting: Estimating sales, predicting server loads or server downtime
* Risk and probability: Choosing the best customers for targeted mailings, determining the probable break-even point for risk scenarios, assigning probabilities to diagnoses or other outcomes
* Recommendations: Determining which products are likely to be sold together, generating recommendations
* Finding sequences: Analyzing customer selections in a shopping cart, predicting next likely events
* Grouping: Separating customers or events into cluster of related items, analyzing and predicting affinities

Building a mining model is part of a larger process that includes everything from asking questions about the data and creating a model to answer those questions, to deploying the model into a working environment. This process can be defined by using the following six basic steps, provided in Figure 3.

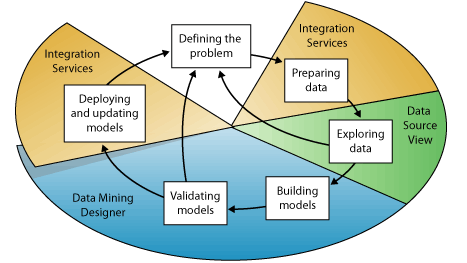


Figure 3 - Data Mining Process

The process illustrated in Figure 3, is cyclical, meaning that creating a data mining model is a dynamic and iterative process. After you explore the data, you may find that the data is insufficient to create the appropriate mining models, and that you therefore have to look for more data. Alternatively, you may build several models and then realize that the models do not adequately answer the problem you defined, and that you therefore must redefine the problem. You may have to update the models after they have been deployed because more data has become available. Each step in the process might need to be repeated many times in order to create a good model.

#### Defining the Problem

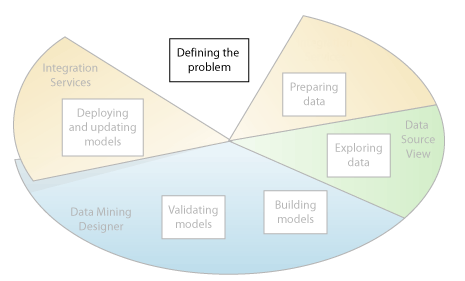


Figure 4 - Defining the problem

The first step, represented in the Figure 4, defining the problem, consists of analyzing business requirements, defining the scope of the problem, defining the metrics by which the model will be evaluated, and defining specific objectives for the data mining project. These tasks translate into questions such as the following:

* What are you looking for? What types of relationships are you trying to find?
* Does the problem you are trying to solve reflect the policies or processes of the business?
* Do you want to make predictions from the data mining model, or just look for interesting patterns and associations?
* Which outcome or attribute do you want to try to predict?
* What kind of data do you have and what kind of information is in each column? If there are multiple tables, how are the tables related? Do you need to perform any cleansing, aggregation, or processing to make the data usable?
* How is the data distributed? Is the data seasonal? Does the data accurately represent the processes of the business?

To answer these questions, you might have to conduct a data availability study, to investigate the needs of the business users with regard to the available data. If the data does not support the needs of the users, you might have to redefine the project.

You also need to consider the ways in which the results of the model can be incorporated in key performance indicators (KPI) that are used to measure business progress.

#### Preparing Data

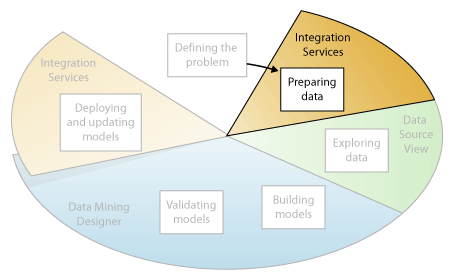


Figure 5 - Preparing the data

The second step in the data mining process, as highlighted in the Figure 5, is to consolidate and clean the data that was identified in the Defining the Problem step.

Data can be scattered across a company and stored in different formats, or may contain inconsistencies such as incorrect or missing entries. For example, the data might show that a customer bought a product before the product was offered on the market, or that the customer shops regularly at a store located 2,000 miles from her home.

Data cleaning is not just about removing bad data or interpolating missing values, but about finding hidden correlations in the data, identifying sources of data that are the most accurate, and determining which columns are the most appropriate for use in analysis. For example, should you use the shipping date or the order date? Is the best sales influencer the quantity, total price, or a discounted price? Incomplete data, wrong data, and inputs that appear separate but in fact are strongly correlated all can influence the results of the model in ways you do not expect.

#### Exploring Data

The third step in the data mining process, as highlighted in the Figure 6, is to explore the prepared data.

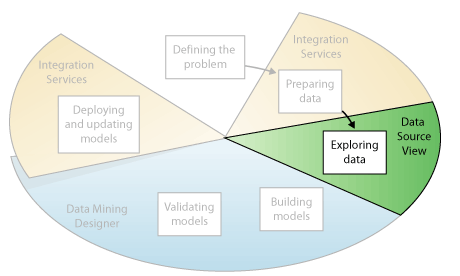


Figure 6 - Exploring the data

You must understand the data in order to make appropriate decisions when you create the mining models. Exploration techniques include calculating the minimum and maximum values, calculating mean and standard deviations, and looking at the distribution of the data. For example, you might determine by reviewing the maximum, minimum, and mean values that the data is not representative of your customers or business processes, and that you therefore must obtain more balanced data or review the assumptions that are the basis for your expectations. Standard deviations and other distribution values can provide useful information about the stability and accuracy of the results. A large standard deviation can indicate that adding more data might help you improve the model. Data that strongly deviates from a standard distribution might be skewed, or might represent an accurate picture of a real-life problem, but make it difficult to fit a model to the data.

By exploring the data in light of your own understanding of the business problem, you can decide if the dataset contains flawed data, and then you can devise a strategy for fixing the problems or gain a deeper understanding of the behaviors that are typical of your business.

#### Building Models

The fourth step in the data mining process, as highlighted in the Figure 7, is to build the mining model or models. You will use the knowledge that you gained in the Exploring Data step to help define and create the models.

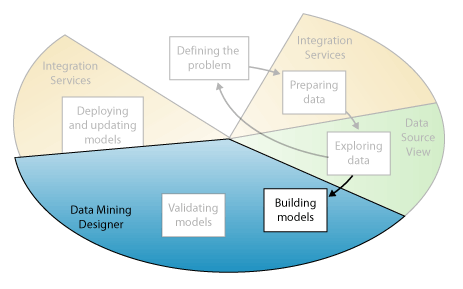


Figure 7 - Building Models

You define the columns of data that you want to use by creating a mining structure. The mining structure is linked to the source of data, but does not actually contain any data until you process it. When you process the mining structure, aggregates and other statistical information are generated and can be used for further analysis. This information can be used by any mining model that is based on the structure.

Before the structure and model is processed, a data mining model too is just a container that specifies the columns used for input, the attribute that you are predicting, and parameters that tell the algorithm how to process the data. Processing a model is often called training.

Training refers to the process of applying a specific mathematical algorithm to the data in the structure in order to extract patterns. The patterns that you find in the training process depend on the selection of training data, the algorithm you chose, and how you have configured the algorithm.

Now that you generated the data mining model, the problem is translated into adjusting the parameters for each algorithm. In addition, you can apply filters to the training data to use just a subset of the data, creating different results. After you pass data through the model, the mining model object contains summaries and patterns that can be queried or used for prediction.

It is important to remember that whenever the data changes, you must update both the mining structure and the mining model. When you update a mining structure by reprocessing it, Analysis Services retrieves data from the source, including any new data if the source is dynamically updated, and repopulates the mining structure. If you have models that are based on the structure, you can choose to update the models that are based on the structure, which means they are retrained on the new data, or you can leave the models as is.

#### Exploring and Validating Models

The fifth step in the data mining process, as highlighted in the following diagram, is to explore the mining models that you have built and test their effectiveness.

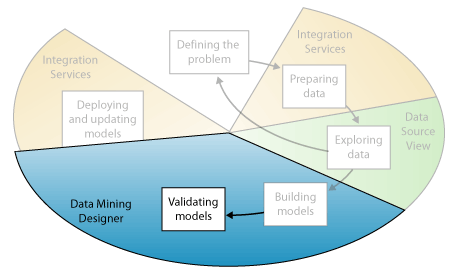


Figure 8 - Exploring and Validating models

Before you deploy a model into a production environment, you will want to test how well the model performs. Also, when you build a model, you typically create multiple models with different configurations and test all models to see which yields the best results for your problem and your data.

Then, you separate your data into training and testing datasets so that you can accurately assess the performance of all models on the same data. You use the training dataset to build the model, and the testing dataset to test the accuracy of the model.

To verify whether the model is specific to your data, or may be used to make inferences on the general population, you can use the statistical technique called cross-validation to automatically create subsets of the data and test the model against each subset.

#### Deploying and Updating Models

The last step in the data mining process, as highlighted in the following diagram, is to deploy the models that performed the best to a production environment.

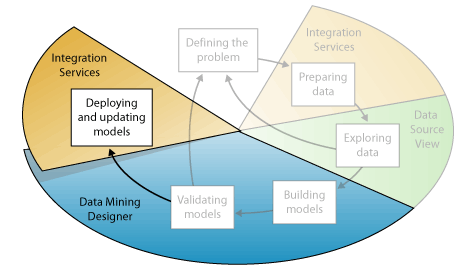


Figure 9 - Deploying and Updating Models

After the mining models exist in a production environment, you can perform many tasks, depending on your needs: create predictions, which you can then use to make business decisions; retrieve statistics, rules, or formulas from the model; embed data mining functionality directly into an application; create a report that lets users directly query against an existing mining model; update the models after review and analysis. Any update requires that you reprocess the models; update the models dynamically, as more data comes into the organization, and making constant changes to improve the effectiveness of the solution should be part of the deployment strategy.

## Methodology

The proposed methodology consists of applying the data mining technique to implement the VanHack Artificially Intelligent Recruiter.

## System Architecture

The proposed architecture is based on the .NET framework technology. It consists of six components: a database instance with two databases, a raw data web application, a data mining core service, an optimization algorithm, a logging service and a WCF integration service.

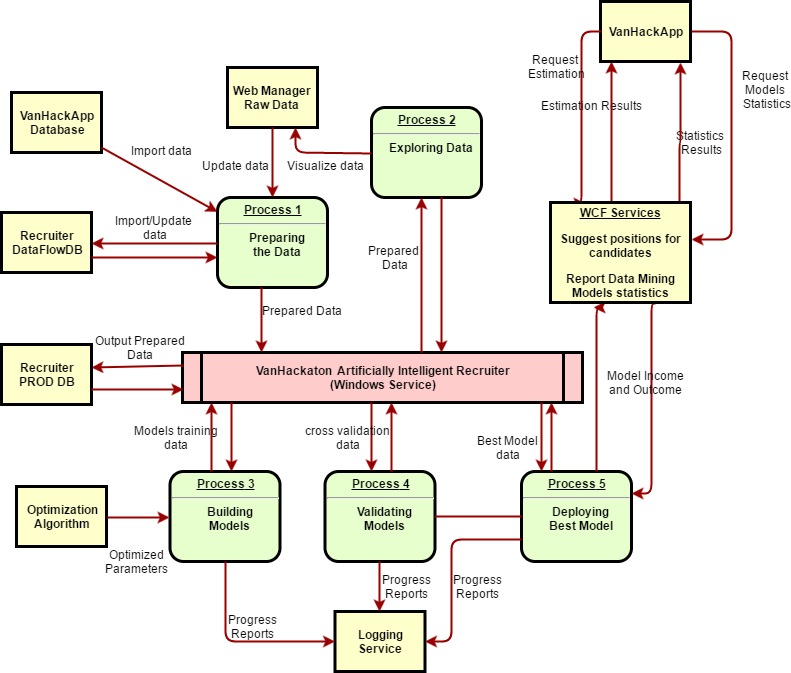


Figure 10 - Recruiter System Architecture

In Figure 10, we propose a scalable system architecture to the Artificial Intelligent Recruiter System. The diagram represents a data flow diagram. In the green boxes, we see the 5 main process of the Data Mining technique. In the yellow boxes, we see applications and databases of the system, except the VanHackApp box. The red box is the core application to be developed.

The components are described in details below.

### Database

The database instance to be used is a Microsoft SQL Server version database. Several jobs and stored procedures will be implemented to fetch, maintain and update the master data from the VanHackApp database instance to the VanHackIntelligentRecruiter database instance.

After the master data is stored in the VanHackIntelligentRecruiter database instance, it is necessary to implement the preparation stored procedures to treat the data. For this occasion, it is important to create two databases. One for Dataflow and another for the Production enviroment.

The Dataflow database are used for preparing and exploring data. You must create a web application to visualize tables and update its values before the preparing procedures run.

The Production enviroment database stores the post-preparation data to be used by the data mining algorithm. This database also stores models informations and support the building models, validating models and deploying best model steps.

### Web Manager Raw Data Application

The web manager raw data is a simple ASP.NET application that exposes the data tables and ables the data mining specialist to manage data before the automatic preparing stored procedures run. This tool is essential to the second step, exploring data.

### Optimization Algorithm

The optimization algorithm aims to iteratively optimize the estimation models parameters. It is a component of the VanHackathon Artificially Intelligent recruiter wndows service. It should never stop running until a error limiar is reached and no more data arrives at the production enviroment database.

There are common algorithms that can be implemented like Particle Sworm Optimization (PSO), Genetic Algorithms (GA) and Fish School Search (FSS).

### Logging Service

The Logging service aims to log error, information, trace levels of the entirely system. It is a component of the component of the VanHackathon Artificially Intelligent recruiter and can use the database of files to store and maintain the logs informations.

### WCF Services

The WCF Services are the integration mechanism to integrate the VanHackathon Artificially Intelligent recruiter with the VanHackApp. Through the methods, interface and objects exposed by the WCF service, it is easy to use the intelligent recruiter.

### VanHackathon Artificially Intelligent Recruiter

The core module of the machine learning system is the VanHackathon Artificially Intelligent recruiter that is a Windows Service application that automates the complete process of data mining described in previous sections. It also contains the Logging service and Optimization Algorithm.

This modules contains many implementation of classification/estimation algorithms such as: Multi-Layer Perceptron (MLP), Support Vector Machines(SVM) and Radial Basis Function(RBF) Artificial Neural Networks and other conventional algorithms like Linear Regression Models(LRM) and Tree Regression Models(TRM).

For the preparing step, there are several algorithms to be implemented some of them are Principal Component Analysis(PCA), Apriori and Random Forest that are focused in determining the best variables that represent the data source tables columns.