1.6. Features of the Prototype.

1.6.1 Techniques

*1.6.1.1 Prediction Systems*

Predicting a system is usually done by learning from the past for which historical data is obtained and analyzed to study the resulting pattern in the specific sector. Predicting any event requires knowledge about past performance. Data from the past is used mainly to learn the patterns that existed. Historical data provides information on the specific pattern of learning the data. Estimation of some variable of interest at some specified future date. Usually, it is based on statistical and time series forecasting methods. Learning from the past provides knowledge about future to some extent.

*1.6.1.2 Prediction Techniques*

Prediction of an event requires vague, imperfect and uncertain knowledge. Complexity in a prediction system is its intrinsic characteristic. Various Artificial Intelligence (AI) techniques have been utilized in realizing a prediction system. The AI based prediction models can be classified into four groups: models based on neural networks, fuzzy logic, genetic algorithm and expert systems.

Under these categories, Time Series models, Classification model, Naïve Bayes Classifier are widely used techniques.

Fuzzy Modeling

In [fuzzy mathematics,](https://en.wikipedia.org/wiki/Fuzzy_mathematics) fuzzy logic is a form of [many-valued logic](https://en.wikipedia.org/wiki/Many-valued_logic) in which the [truth values](https://en.wikipedia.org/wiki/Truth_value) of variables may be any [real number](https://en.wikipedia.org/wiki/Real_number) between 0 and 1 both inclusive. It is employed to handle the concept of partial truth, where the truth value may range between completely true and completely false. By contrast, in [Boolean logic,](https://en.wikipedia.org/wiki/Boolean_algebra) the truth values of variables may only be the [integer](https://en.wikipedia.org/wiki/Integer) values 0 or 1.

The term fuzzy logic was introduced with the 1965 proposal of [fuzzy set theory](https://en.wikipedia.org/wiki/Fuzzy_set_theory) by [Lotfi](https://en.wikipedia.org/wiki/Lotfi_A._Zadeh) [Zadeh.](https://en.wikipedia.org/wiki/Lotfi_A._Zadeh) Fuzzy logic had, however, been studied since the 1920s, as [infinite-valued logic](https://en.wikipedia.org/wiki/%C5%81ukasiewicz_logic)—notably by [Łukasiewicz](https://en.wikipedia.org/wiki/Jan_%C5%81ukasiewicz) and [Tarski.](https://en.wikipedia.org/wiki/Alfred_Tarski)

Fuzzy logic is based on the observation that people make decisions based on imprecise and non-numerical information. Fuzzy models or sets are mathematical means of representing [vagueness](https://en.wikipedia.org/wiki/Vagueness) and imprecise information (hence the term fuzzy). These models have the capability of recognizing, representing, manipulating, interpreting, and utilizing data and information that are vague and lack certainty.



|  |  |
| --- | --- |
| TEAM-ZEROBUG | 12 |

**Online Diabetes Prediction with Data Mining**

The problem of fuzzy system identification is the problem of eliciting IF-THEN rules from raw input and output data. This proceeds through two steps:

1. Clustering
2. Specification of input-output relations (IF THEN rules)

Clustering algorithms are used extensively not only to organize and categorize data, but are also useful for data compression and model construction. By finding similarities in data, one can represent similar data with fewer symbols.

Classification Model

The classification model is the simplest of the several types of predictive analytics models. It puts data in categories based on what it learns from historical data. Classification models are best to answer yes or no questions, providing broad analysis that’s helpful for guiding decisive action.

Also, classification is a typical supervised learning task. We use it in those cases where we have to predict a categorical type, that is if a particular example belongs to a category or not (unlike regression, which is used to predict continuous values).

Naive Bayes Classifier

Naive Bayes classifiers are a family of simple ["probabilistic classifiers"](https://en.wikipedia.org/wiki/Probabilistic_classification) based on applying [Bayes'](https://en.wikipedia.org/wiki/Bayes%27_theorem) [theorem](https://en.wikipedia.org/wiki/Bayes%27_theorem) with strong (naïve) [independence](https://en.wikipedia.org/wiki/Statistical_independence) assumptions between the features. They are among the simplest [Bayesian network](https://en.wikipedia.org/wiki/Bayesian_network) models. But they could be coupled with [Kernel density estimation](https://en.wikipedia.org/wiki/Kernel_density_estimation) and achieve higher accuracy levels.

Naive Bayes classifiers are highly scalable, requiring a number of parameters linear in the number of variables (features/predictors) in a learning problem. [Maximum-likelihood](https://en.wikipedia.org/wiki/Maximum-likelihood_estimation) training can be done by evaluating a [closed-form expression,](https://en.wikipedia.org/wiki/Closed-form_expression) which takes [linear time,](https://en.wikipedia.org/wiki/Linear_time) rather than by expensive [iterative](https://en.wikipedia.org/wiki/Iterative_method) [approximation](https://en.wikipedia.org/wiki/Iterative_method) as used for many other types of classifiers.

Naive Bayes is a simple technique for constructing classifiers: models that assign class labels to problem instances, represented as vectors of [feature](https://en.wikipedia.org/wiki/Feature_vector) values, where the class labels are drawn from some finite set. There is not a single [algorithm](https://en.wikipedia.org/wiki/Algorithm) for training such classifiers, but a family of algorithms based on a common principle: all naive Bayes classifiers assume that the value of a particular feature is [independent](https://en.wikipedia.org/wiki/Independence_(probability_theory)) of the value of any other feature, given the class variable.



|  |  |
| --- | --- |
| TEAM-ZEROBUG | 13 |

**Online Diabetes Prediction with Data Mining**

Survival or Duration Analysis

Survival Analysis is used to estimate the lifespan of a particular population under study. It is also called 'Time to Event' Analysis as the goal is to estimate the time for an individual or a group of individuals to experience an event of interest. Survival analysis is a branch of [statistics](https://en.wikipedia.org/wiki/Statistics) for analyzing the expected duration of time until one or more events happen, such as death in biological organisms and failure in mechanical systems.

More generally, survival analysis involves the modelling of time to event data; in this context, death or failure is considered an "event" in the survival analysis literature – traditionally only a single event occurs for each subject, after which the organism or mechanism is dead or broken. Recurring event or repeated event models relax that assumption. The study of recurring events is relevant in [systems reliability,](https://en.wikipedia.org/wiki/Systems_reliability) and in many areas of social sciences and medical research.

1.6.2. Comparison of most used Prediction Techniques

Table 1.1 – Prediction Techniques Comparison

|  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
| **Techniques** |  |  |  | **Advantages** | | |  |  |  | **Disadvantages** | | | |  |
|  |  | | | | | | | |  |  |  |  |  |  |
|  | ✓ The structure of Fuzzy Logic Systems | | | | | | | |  | Fuzzy | logic | is | not | always |
|  |  | is easy and understandable. | | | | |  |  |  | accurate, so the results are | | | | |
|  | ✓ Fuzzy | | | logic is | | widely | used | for |  | perceived | | based | | on |
|  |  | commercial and practical purposes. | | | | | | |  | assumption, so it may not be | | | | |
|  | ✓ It helps you to control machines and | | | | | | | |  | widely accepted. | | | |  |
|  |  | consumer products. | | | | |  |  | Fuzzy systems don't have the | | | | | |
|  | ✓ It may not offer accurate reasoning, | | | | | | | |  | capability | | of | machine | |
|  |  | but the only acceptable reasoning. | | | | | |  |  | learning as-well-as neural | | | | |
|  | ✓ It | | helps | | you | to deal | with | the |  | network | | type | | pattern |
|  |  | uncertainty in engineering. | | | | |  |  |  | recognition. | |  |  |  |
|  | ✓ | Mostly robust as no precise inputs | | | | | | |  | Validation and | | | Verification | |
| Fuzzy Logic |  | required. | | |  |  |  |  |  | of a fuzzy knowledge-based | | | | |
|  | ✓ It | | can | be | programmed | | to in | the |  | system | needs | | extensive | |
|  |  | situation when feedback sensor stops | | | | | | |  | testing with hardware. | | | | |
|  |  | working. | | |  |  |  |  |  | Setting | exact, | | fuzzy rules | |
|  | ✓ It can easily be modified to improve or | | | | | | | |  | and, membership functions is | | | | |
|  |  | alter system performance. | | | | |  |  |  | a difficult task. | | |  |  |
|  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |



|  |  |
| --- | --- |
| TEAM-ZEROBUG | 14 |

**Online Diabetes Prediction with Data Mining**

|  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
|  | ✓ inexpensive | | sensors can be used | | | | |  | Some | fuzzy | | time logic is | | |
|  |  | which helps you to keep the overall | | | | | |  | confused | | with | | probability | |
|  |  | system cost and complexity low. | | | | | |  | theory and the terms. | | | | |  |
|  | ✓ It provides a most effective solution to | | | | | | |  |  |  |  |  |  |  |
|  |  | complex issues. | |  |  |  |  |  |  |  |  |  |  |  |
|  |  |  |  |  | | |  |  |  | | | | | |
|  | ✓ | Probabilistic |  | Approach, | | | gives |  | The assumptions of logistic | | | | | |
|  |  | information |  | about | | statistical | |  | regression. | | |  |  |  |
|  |  | significance of features. | | | |  |  | Need to manually choose the | | | | | | |
|  | ✓ Simple to | | understand, fast and | | | | |  | number of neighbors’ ‘k’. | | | | | |
|  |  | efficient. |  |  |  |  |  |  | Not | appropriate | | | for | non- |
|  | ✓ Performant, not biased by outliers, not | | | | | | |  | linear problems, not the best | | | | | |
|  |  | sensitive to overfitting. | | | |  |  |  | choice for large number of | | | | | |
| Classification | ✓ High performance | | | | on non – | | linear |  | features. | |  |  |  |  |
|  | problems, not biased by outliers, not | | | | | | Not the best choice for large | | | | | | |
| Model |  | sensitive to overfitting. | | | |  |  |  | number | | of | features, more | | |
|  |  |  |  |  |
|  | ✓ Interpretability, | | | no | need | for | feature |  | complex. | |  |  |  |  |
|  |  | scaling, works on both linear / non – | | | | | | Poor | | results on very small | | | | |
|  |  | linear problems. | |  |  |  |  |  | datasets, | | overfitting | | | can |
|  | ✓ Powerful | | and |  | accurate, | | good |  | easily occur. | | |  |  |  |
|  |  | performance | on | | many | problems, | |  | No |  |  | interpretability, | | |
|  |  | including non – linear. | | | |  |  |  | overfitting can easily occur, | | | | | |
|  |  |  |  |  |  |  |  |  | need to choose the number of | | | | | |
|  |  |  |  |  |  |  |  |  | trees manually. | | | |  |  |
|  |  | | | | | | |  | |  | | | | |
|  | ✓ When assumption of independent | | | | | | | Main | | imitation of Naive | | | | |
|  |  | predictors holds true, a Naive Bayes | | | | | |  | Bayes is the assumption of | | | | | |
|  |  | classifier performs better as compared | | | | | |  | independent | | |  | predictors. | |
|  |  | to other models. | |  |  |  |  |  | Naive | | Bayes | | implicitly | |
|  |  |  |  |  |  |  |  |  | assumes that all the attributes | | | | | |
|  | ✓ Naive Bayes requires a small amount | | | | | | |  | are mutually independent. In | | | | | |
|  |  | of training data to estimate the test | | | | | |  | real life, it is almost | | | | | |
|  |  | data. So, the training period is less. | | | | | |  | impossible that we get a set | | | | | |
|  |  |  |  |  |  |  |  |  | of predictors which | | | | | are |
|  |  |  |  |  |  |  |  |  | completely independent. | | | | | |
|  | ✓ Naive Bayes is also easy to | | | | | | |  |  |  |  |  |  |  |
|  |  | implement. |  |  |  |  |  | If categorical variable has a | | | | | | |
|  |  |  |  |  |  |  |  |  | category in test data set, | | | | | |
| Naïve Bayes |  |  |  |  |  |  |  |  | which was not observed in | | | | | |
|  |  |  |  |  |  |  |  |  | training data set, then model | | | | | |
|  |  |  |  |  |  |  |  |  | will assign a 0 (zero) | | | | | |
|  |  |  |  |  |  |  |  |  | probability | | | and | will | be |
|  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |
|  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |
|  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |
| TEAM-ZEROBUG | |  |  |  |  |  |  |  |  |  |  |  |  | 15 |

**Online Diabetes Prediction with Data Mining**

|  |  |  |  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
|  |  |  |  |  |  |  | unable to make a prediction. | | | | | |
|  |  |  |  |  |  |  | This is often known as Zero | | | | | |
|  |  |  |  |  |  |  | Frequency. To solve this, we | | | | | |
|  |  |  |  |  |  |  | can | use | the | smoothing | | |
|  |  |  |  |  |  |  | technique. | | One | of | | the |
|  |  |  |  |  |  |  | simplest | |  | smoothing | | |
|  |  |  |  |  |  |  | techniques is called Laplace | | | | | |
|  |  |  |  |  |  |  | estimation. | |  |  |  |  |
|  |  |  | | | |  |  |  |  | |  | |
| Survival | ✓ | Good method to understand basics | | | |  | Can | only | compare | | limited | |
| or |  | and produce descriptive results. | | | |  | number of groups. | | | |  |  |
| Duration |  |  |  |  |
| ✓ | Life table: good for | | large | data and | Does not allow inclusion of | | | | | | |
| Analysis |
|  | crude measurement of event times. | | | |  | multiple | | covariates | | | and |
|  |  |  |
|  | ✓ Flexible | | model, | often | initial |  | multivariate controls. | | | |  |  |
|  |  | exploratory choice in analyses. | | | |  | Less | appropriate | | for | testing | |
|  | ✓ Allows | | inclusion | of | multiple |  | hypotheses | | about | |  | time |
|  |  | covariates, multivariate analysis. | | | |  | dependence. (i.e. how hazard | | | | | |
|  |  |  |  |  |  |  | varies over time) | | |  |  |  |
|  |  |  |  |  |  |  |  |  |  |  |  |  |