Abstract - Diabetes is always a major concern and challenge across the world. The main factor of diabetes is high level complications that varies with every patient. So the analysis of the data on various advances in health science leads to significant production of data. These data lead to a high throughput genetic data and clinical information. The main focus of this research is to generate outcomes with various analytic patterns based on the existing datasets. These outcome will assist the diabetic practitioner to prescribe the patients based on their genetics and clinical information. To have an effective analytic model this paper focuses on building an Incremental approach that identifies the factors for type 1 juvenile patient’s data and apply analytics through Machine learning algorithm to monitor glucose level in the patients. The model will be trained to adapt new learnings based on the data to analyse the glucose requirement level after following the diabetes prescription regime. The first step is to identify all the potential factors that should be a part of analysis in juvenile data sets. ML model are used to classify the data and group them based on potential achievable factors. The model will use incremental approach to learn new patterns and to generate analytical outcomes. Further this will assist for better medication and prescription for preventing and curing juvenile diabetes.

Keywords - Juvenile diabetes, Machine learning, genetic data, Incremental approach, training set, clinical information, Decision tree algorithm.

I. INTRODUCTION

As per World Health Organization, out of world’s 347 million people with diabetes, from 5 to 10% have type 1 diabetes (T1D), the critical kind. In T1D, the main concern is the production of insulin, an essential hormone needed to convert food into energy and to regulate glucose level. T1D cannot be prevented or cured, but it can be effectively treated with external supplies of insulin and managed glucose control. T1D diabetes occurs when body’s immune system attacks and destroys certain cells in pancreas. T1D is usually diagnosed in children and young adults. Scientists do not yet know exactly what causes T1D but they believe that autoimmune, genetic and environmental factors are involved. The risk of T1D is higher that virtually all other severe chronic disease of childhood. According to Juvenile Diabetes Research foundation as many as several millions may have T1D. Each year over 13,000 children are diagnosed with diabetes. Glucose control helps to delay complications that are related to diabetes. In this paper we describe a methodology that assists the practitioners to prescribe the potential T1D patients based on factors that can be the cause. An automatic prediction model that can warn the people of imminent changes on their glucose level will enable them to take preventive actions.

II. PROPOSED METHODOLOGY

Machine learning is useful when there is a large amount of example data and when the rules for prediction are unclear. In the case of predicting T1D, though physicians were able to intuitively estimate the risk of the disease, they weren’t able to analyze the factors that can be relied upon. Machine learning algorithm will make the model to learn with new findings that can assist the physicians to analyze the factors which will assist them for diagnosing. While constructing the ML model, there are many approaches that can build the system. Classification approach works well with analyzing the factors for T1D as majority of the data will have a pattern of categorial labels. By classifying the data based on the factors the algorithm can define various rules base on the situation. Given a training set this methodology can classify the data with the given a set of values , the model can predict if the patient will be prone to T1D symptoms and with what severity. Hence the prediction becomes a binary (yes/no) classification problem. From a computational standpoint, classification problems are easier and more efficient to solve than any other data model. So the research focuses on creating a classification model to transform the data to analysis of factors that will influence the medication of juvenile diabetic.

Data cleansing involves transforming raw data into a form that is easily readable without ambiguity for a machine learning algorithm. The algorithm should identify any problems such as ambiguity, improper data and missing values in a given data set and correcting them. Data cleansing has advantages and disadvantages. The data cleansing preprocess the data that makes it ready for the algorithm to apply its techniques to understand the patterns that gets generated with modified structure pattern.

III. PROPOSED OVERALL MODEL

The data which is fetched with multiple iterations can be processed and stored as processed data.
Factors that are broadly considered with multiple iterations are:
- Family History
- Genetical Information
- Age
- Economic factors
- Location

While collecting the training sets if some data are not available from the patient information IL model will generate according to its learning process to support the data. The result set will be derived from the knowledge that has been derived with various scenarios. The categorical values that can be considered in these factors is shown in the following table:

<table>
<thead>
<tr>
<th>Factors</th>
<th>Categorical values</th>
</tr>
</thead>
<tbody>
<tr>
<td>Age</td>
<td>Immediate, First family, Siblings</td>
</tr>
<tr>
<td>Family</td>
<td>Father, Mother</td>
</tr>
<tr>
<td>Genes</td>
<td>East, West, South, North, Summer, Winter</td>
</tr>
<tr>
<td>Location</td>
<td>High, Moderate, Average, Poor</td>
</tr>
</tbody>
</table>

The Machine learning model includes techniques that can derive a predictive pattern based on the factors. The techniques that can be used include, Decision trees, Naïve bayes and Neural networks.

The algorithm should identify the Causes and the factors related to the cause that influences Juvenile diabetes. In this framework expert systems and model based schemes provide mechanisms for the generation of hypothesis from patients data. In missing information the expert system can generate using generalization or mapping techniques.

Building intelligent systems that are able to deal with incremental learning is one of key research avenues in machine learning, data mining and knowledge discovery. Adaptivity can take different forms, but the most important one is certainly incrementality. Such systems are continuously updated as more data becomes available over time. It will help to integrate intelligence into knowledge learning systems.

### IV. PROCESS OF THE MODEL

Factors identified from table 1 are applied with the algorithm.

1. The rules generated makes the system to learn to new patterns and behaviour with the existing factors.
2. Factors should lead to diagnoses of juvenile data. So, Rule 1: if Age>10-15, and family history>immediate and genes>mother what is the possibility for the child to be diabetic.
3. Rule 2: if Age>10-15, and family history>Sibling and genes>father what is the possibility for the child to be diabetic.

Multiple rules can be generated for a boolean response as diabetic ‘yes’ or ‘no’.

### V. DECISION TREE ALGORITHM

#### C4.5 Decision Trees

A decision tree is a learning mechanism, which uses a “divide and conquer” strategy to classify instances. It consists of leaf nodes labeled with a class and test nodes that connect two or more subtrees. Every test node computes a result based on some particular attribute, and this outcome decides along which subtree the example should go. Every instance starts from the root of the tree and traverses down to the leaf, depending on the attribute tests at each node until it reaches a leaf, whereby the class label of the leaf, determines the class prediction of the example. In order to build a decision tree, at every node an attribute needs to be chosen that would offer the best split among the training data. The decision is based on the value of information gain or gain ratio (i.e. the...
attribute that most effectively splits the training examples according to their class labels). The attribute that has the best information gain at any node, is used to split that node. The decision tree, could also be geometrically represented as partitioning the x-dimensional data space (defined by x attributes) by decision planes corresponding to the tests at each node. Real data in most cases contain noisy data, and the divide and conquer approach to building trees tends to build such discrepancies into the classifier, which evidently leads to lower prediction accuracy on unseen test data. To overcome such overfitting of the training 19 data, decision trees are typically pruned. Smaller, pruned trees are not only easy to understand but also are generally more accurate on test data.

The splitting criteria used by C4.5 at each node of the decision tree is an information-based measure, the gain ratio, which takes into consideration different probabilities of test outcomes. Let D be a dataset and T be a particular test that has outcomes T1, T2, T3, ..., Tk which is used to partition D into subsets D1, D2, D3, ..., Dk, such that Di contains the set of examples having the outcome Ti for the test T. Now, if C denotes the number of classes and p(D,j), the proportion of cases in D that belong to the jth class, then the residual uncertainty about the class to which a case in D belongs is expressed as

\[
\text{info}(D) = - \sum_{j=1}^{C} p(D,j) \times \log_2(p(D,j))
\]

and the corresponding information gain from a test T with k outcomes is mathematically represented as:

\[
\text{gain}(D, T) = \text{info}(D) - \frac{1}{K} \sum_{i=1}^{K} \frac{|D_i|}{|D|} \times \text{info}(D_i)
\]

Again, the potential information obtained by partitioning a set of examples is based on which Di subset the example lies and is known as the split information, defined mathematically as:

\[
\text{split}(D, T) = - \sum_{i=1}^{K} \frac{|D_i|}{|D|} \times \log_2(|D_i||D|)
\]

CONCLUSIONS

In case the instances belong to the same class the tree represents a leaf so the leaf is returned by labeling with the same class. The potential information is calculated for every attribute, given by a test on the attribute. Then the gain in information is calculated that would result from a test on the attribute. The proposed Decision model with Tree classifier has been successfully used for markable improved classification of diabetic dataset.

Further results showed that the performance of the model can be further improved by selecting appropriate factors using feature selection algorithms. The Performance analysis has to be done for 70-30 ratio (training-test) partitioning method using Classification accuracy and information gain measures. The classification accuracies obtained by the proposed methods using Tree algorithm is one of the best results to derive rules for juvenile diabetes based on the factors selection.

REFERENCES