FINAL PROJECT REPORT
INTELLIGENT CHILD CARE MATCHER AND WAITLIST OPTIMIZER

TEAM MEMBERS

ANG PAU HUANG, EDWIN
CHEOK MEI LI (MAVE)
PADMAPRIYA MATHIVANAN
PRIYANSH MISHRA

TABLE OF CONTENTS

1 INTRODUCTION

1.1 EXECUTIVE SUMMARY

Choosing a child care centre for your little one is an important decision. If you are putting your child into care for the first time, it may be a stressful experience since you want what's best for your child, but you don't know where to start. It's your child's early experiences that will have a lasting impact on their future habits and resilience. This is why choosing a quality child care centre is essential.

Various aspects in choosing the best option, Determining the type of childcare service we are looking for, to Find out their philosophy and values, the teaching methods and learning activities that the child care centres encompass, The location of the centres, What meals are given to the children, most importantly the child's safety also How to tell if a childcare centre isn't right for your child? Possibly by the reviews that we read on various forums to prevent ending up in a wrong childcare that is not suitable for your little one.

Our team, comprising of 4 members, 2 among us are new parents, have faced these challenges and found the process hectic while choosing the best childcare for our little one however to simplify this decision process have chosen to create a recommendation system for the parents to key in their preferences and get the recommendations in one click without need to surf across various forum discussion to get the reviews and other factors like choosing food and distance etc.

Our system aids for both the Parents in choosing the child care centres also the Child care centre to optimize their high waitlist for admission. We have performed the knowledge acquisition by short interviews through emails to the child care centres on our product. We have used the techniques and knowledge gained in the lecturers and hand-on to build this system we incorporate the KIE DROOLS for building the recommendation system and PYTHON FLASK for a UI integrating both the systems, Google APIs for distance and sentiment analysis of scrapped reviews.

We had regular online meetings to discuss our progress, our team had challenging time sharing knowledge and helping out without physically meeting to accomplish tasks together due to the Covid19 situation. Overall, we have good learning from each of the team member. We are overwhelmed with joy and gratitude to our lecturers and would like to share the insights of this project.

1.2 BUSINESS PROBLEM BACKGROUND

During the search for a suitable childcare centre for their child, parents in Singapore will often identify a list of potential childcares based on factors like distance and fees. Parents do also often conduct a lot of online research on past parents' reviews on potential childcare centres. It can be challenging for a parent to be able to identify a childcare centre that meet his/her Preferences.

Figure 1: Parents asking reviews on Facebook

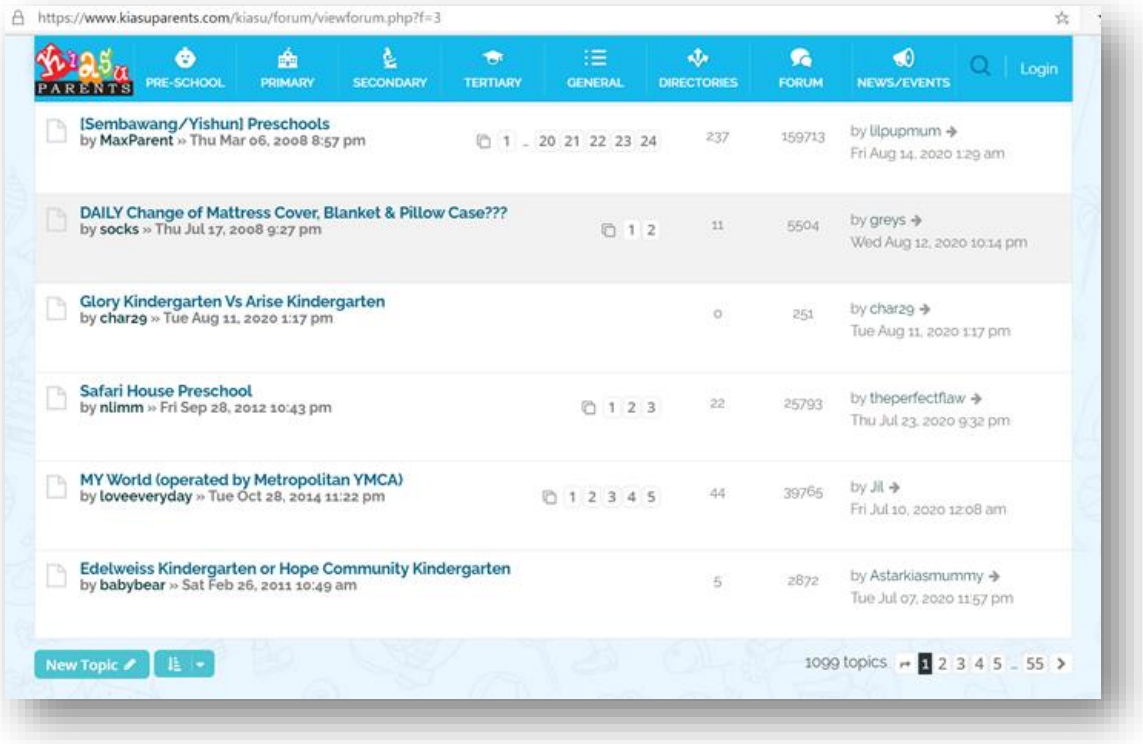


Figure 1: Parents asking for reviews on KiasuParents

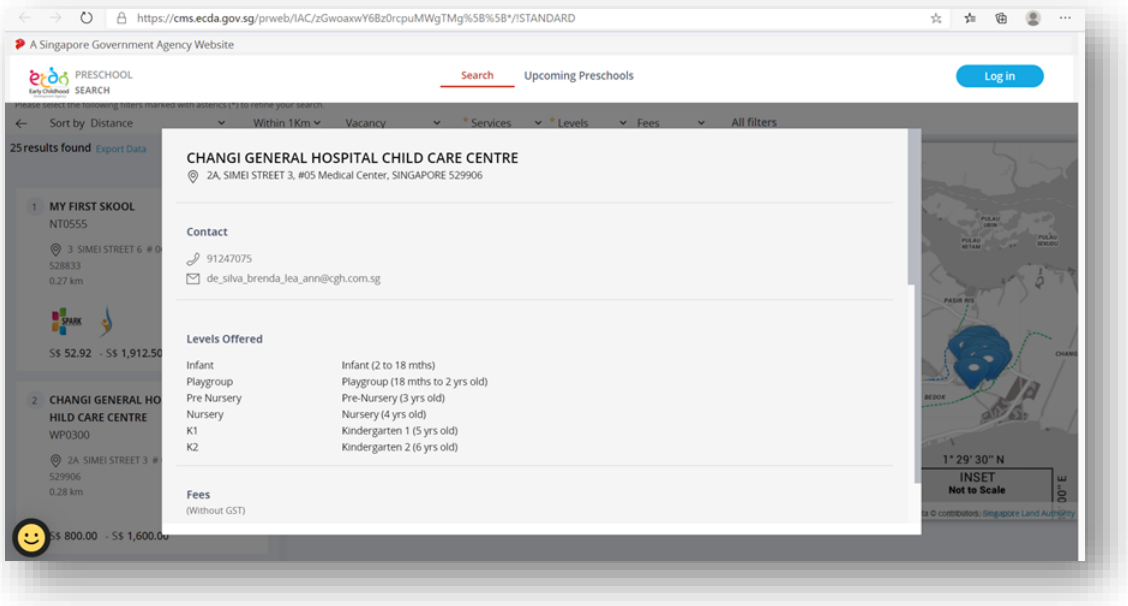
Separately, a recurring problem among childcare centre operators today is the need to manage an inflated waitlist of prospective parents which does not reflect true demand[1]. One reason is parents who had secured childcare places in a centre would normally not remove their names on waitlist of other centres. Even if Early Childhood Development Agency (ECDA), the regulatory and developmental agency for the early childhood sector, is able to coordinate the waitlists to reduce inflation issues, the problem remains that some parents may already have found alternative childcare methods, like domestic helpers and did not bother to withdrawing their applications. The resources spent in managing waitlists are high, which includes contacting and planning for parent's visits to childcare centres.

This is evident from a feedback from a pre-school principal. She mentioned that parents take time getting back to the preschools on the parents' interest, even when they are interested. Furthermore, challenges of managing a waitlist has worsened since the COVID pandemic, as pre-schools need to either conduct virtual tours or can only allow parents to enter the pre-school when the children are not around, such as in the evenings or Saturdays.

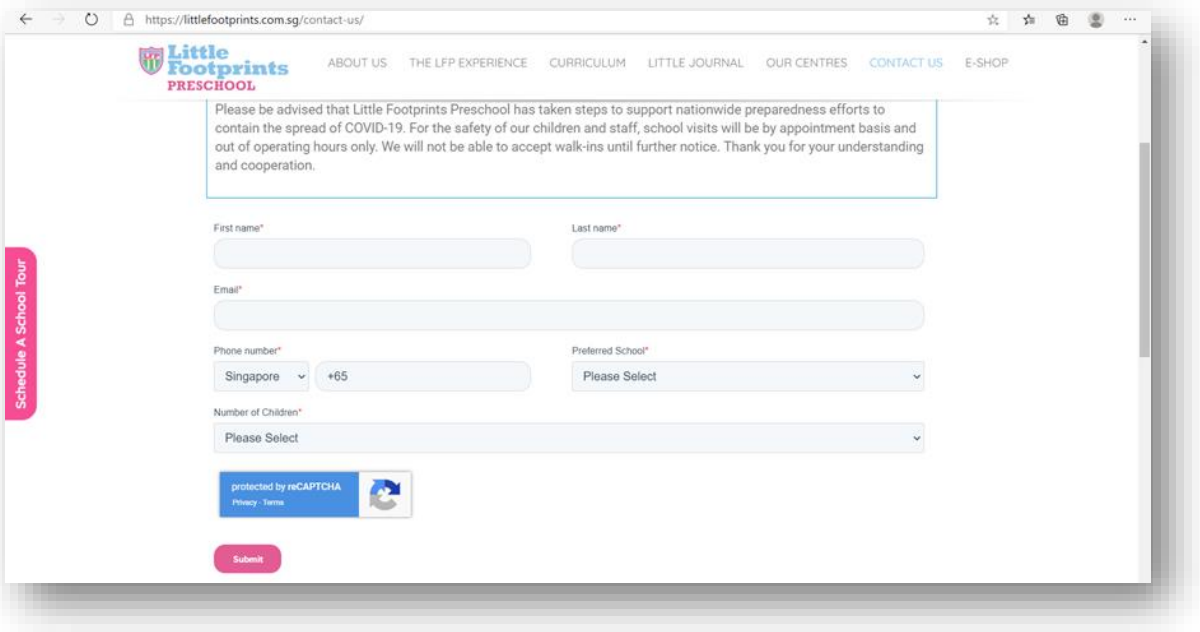
[1] https://www.asiaone.com/singapore/why-childcare-supply-and-demand-dont-add

It is also important for us to understand what the current solutions are for parents to register interest. Both solutions do not help parents quickly glean reviews on the childcares nor help childcares differentiate parents' interests (besides by interest registration date).

a. Registering through ECDA website. This website allows parents to register up to 10 childcares, without allowing parents to rank the childcares. It provides only some information on the childcare centres such as levels offered by the childcare, the fees etc.



b. Alternatively, parents can go straight to the childcare centres to register interest.



As the number of childcare centres and enrolments in Singapore increase, our system offers solutions toward these problems.

Preschool Enrolment Statistics Singapore 2019

Year	2012	2013	2014	2015	2016	2017	2018	April 2019
Total no. of child care centres	1,016	1,083	1,143	1,256	1,342	1,419	1,495	1,517
Total no. of child care centre places	92,779	101,597	109,694	123,327	137,278	149,803	167,421	171,660
Total Enrolment	75,530	73,852	83,928	95,414	103,221	110,826	119,195	119,333
Enrolment in full-day programme	65,826	65,650	75,518	86,898	95,357	103,404	111,838	111,725

Source: ECDA[3]

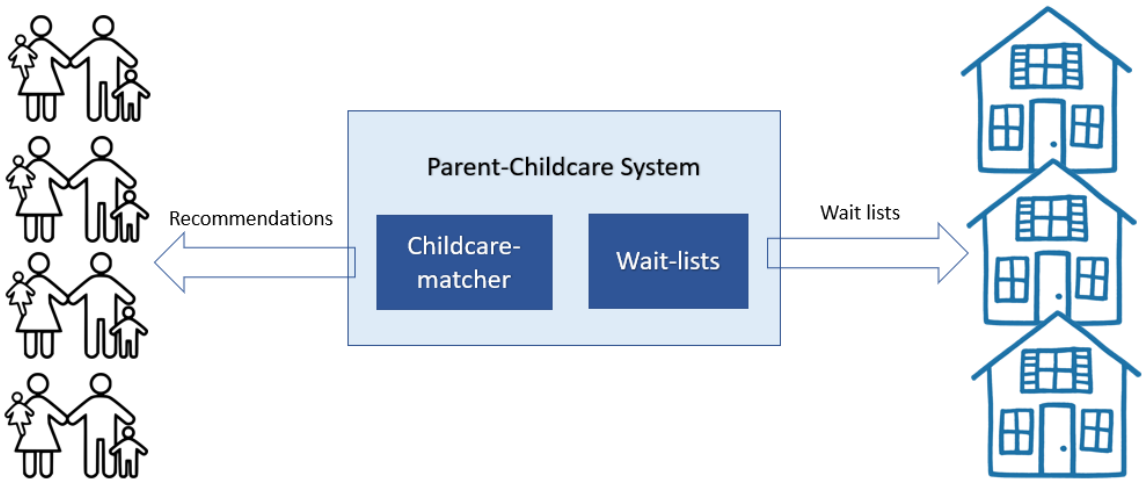
[3] M890651 - Number, Capacity And Enrolment In Child Care Centers

1.3 PROJECT OBJECTIVES

This project aims to deliver a system that provides:

1. Parents with childcare recommendations, along with sentiment analysis based on actual parents' reviews from our system. Parents can then rank and submit their interest on the childcares through our system.

2. Childcare centres with an optimised waitlist of parent interests. which include information like likelihood of prospective parent taking up placements. Childcares can then better optimise their resources on managing waitlists. In view of PDPA, only essential information on parents will be included in waitlists.



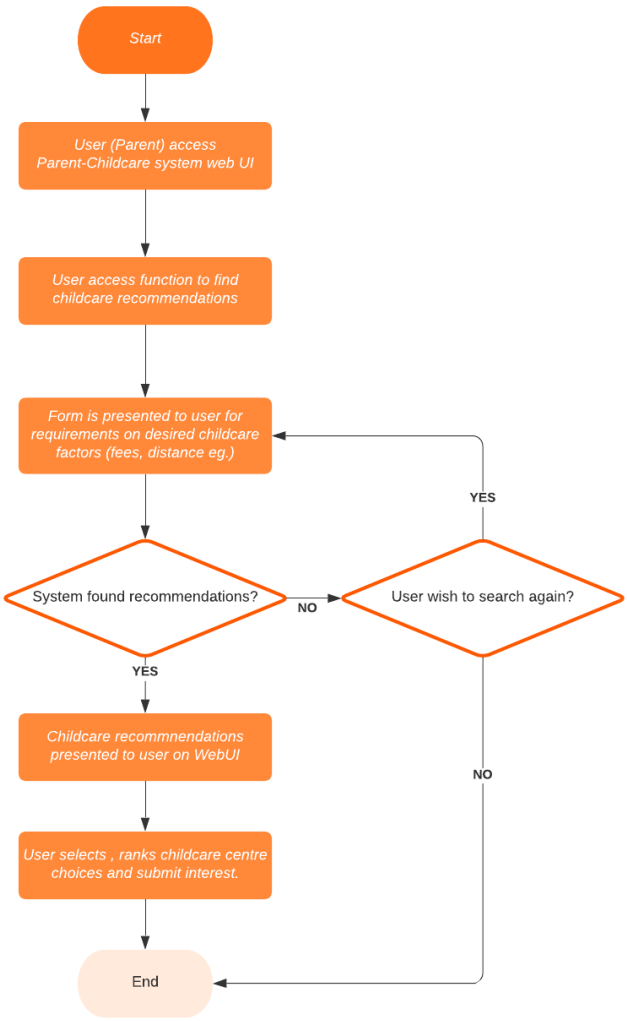
1.4 KNOWLEDGE ACQUISITION:

Source of information	Information explored	Acquisition methods
Project Model I – Childcare Matcher		
ECDA	Child care centre data	Extracted data based on certain features
Kaisu parents portal / Forum	Child care reviews and comments/discussions	Web scrapping to extract the comments grouped by school names
Distance database	Postal sector information of Singapore from Wikipedia. Distance of postal section using google API with constraints 1, 2 and 3 km and postal code of all Child care centres by web scrapping addresses.	Web scrapping and extraction of information
Project Model II – Waitlist Optimizer		
Discussion with Childcare centres	Approached the child care centre to understand about their challenges with waitlist and whether our approach would solve their challenge	Through email

2 SYSTEM DESIGN

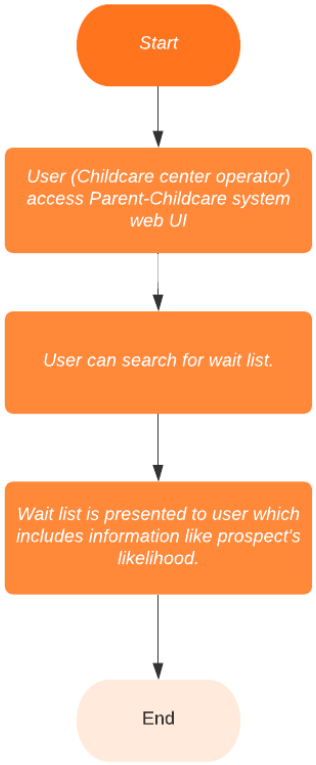
2.1 USER WORKFLOW: PARENTS

Parents are able to access a website on the internet. User friendly forms will be presented to parents to gather requirements. Our system then uses the inputs to produce recommendations for the parents. If parents found a suitable childcare centre among the recommendation, they may rank the childcare to indicate their preference level and submit their interest to our system on the website.



2.2 USER WORKFLOW: CHILDCARE CENTRES

Childcare centre operators are able to access a website on the internet. The operators may search (based on dates for instance) and retrieve waitlists of prospective parents. When waitlist of parent interests are presented to the system users, it will also include information like likelihood of prospective parent taking up placements.



2.3 SYSTEM ARCHITECTURE

Below figure shows the overall architecture of the system. For System I – Child care recommendation, the process is broken into I and II for better description of the subsystem features. This overall illustration shows both the subsystems (1. Child care recommender 2. Child care waitlist optimizer) and interfaced front end and back end systems.

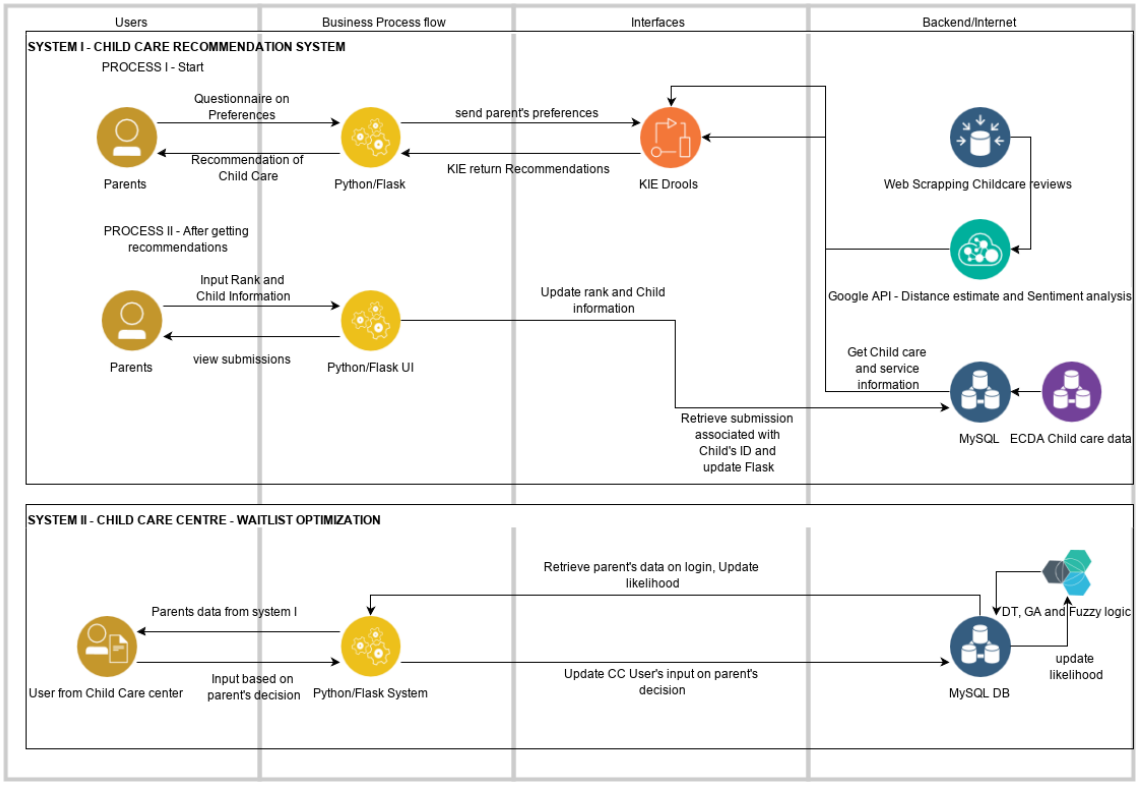
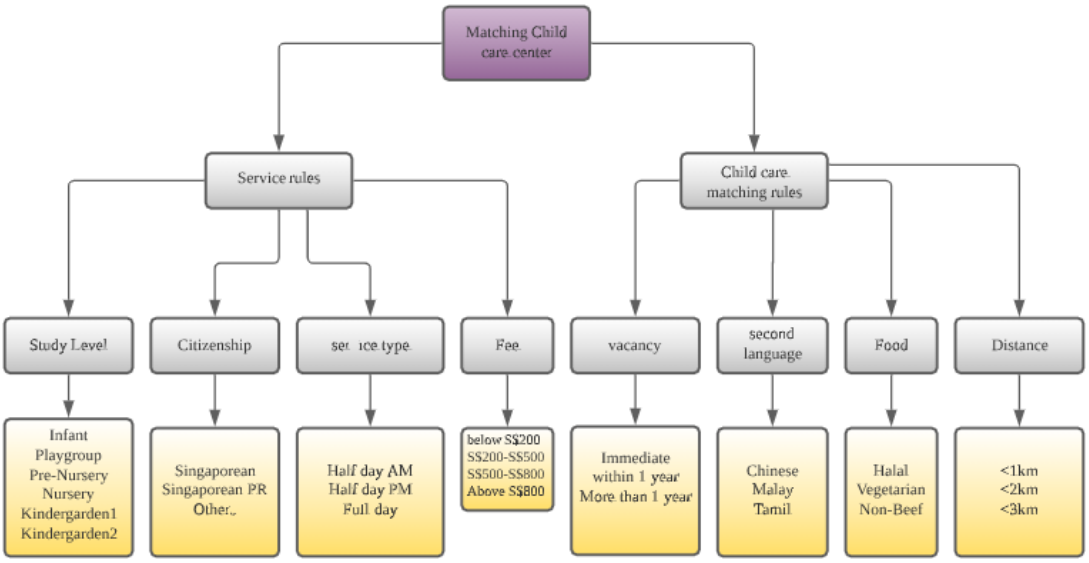


Figure – System Architecture subsystems – CC recommender and CC Waitlist optimizer

2.4 INFERENCE DIAGRAM

The inference diagram gives a decision situation on various factors considered while recommendations of the child care centres. The decision is divided into various sub goal levels and given in a tree structure where the root node is the final recommendations and the leaf / end nodes are the data we gather from the user (parents) as their preferences through questionnaire.



2.5 DESIGN: CHILDCARE MATCHER FOR PARENTS

The Business Process is shown in the diagram below. Child care data is taken from ECDA (Early Childhood Development Agency) loaded into MySQL. Kaisu parents' forums crawled for reviews and comments and loaded to KIE. Distance database is created using google API and the system has been integrated with rules engine in KIE. The BPMN business process together with drools rules are used in KIE Workbench tool.

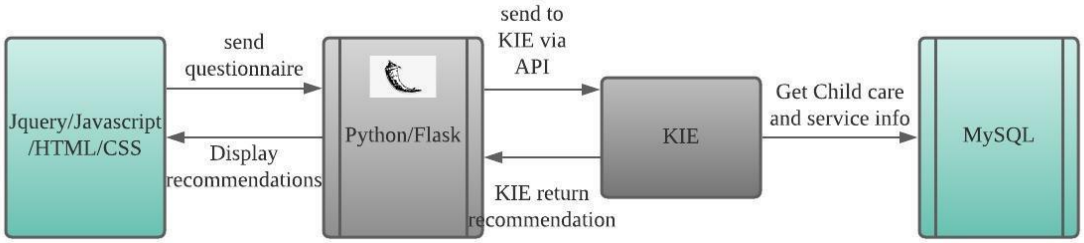
This system I – child care matcher can be divided into two sub processes, (I) parents start to input their preferences in the form of questionnaire through user interface using Python/Flask. This data is then sent to KIE Drools engine which gets the child care and service information from the database and returns the recommendation to parents via Python/Flask. (II) after receiving the appropriate recommendation by our smart system, Parents provide a submission through ranking and provide their child's information as well. They can also view previous submission made in the system.

Limitations:

- The ideal location field in the questionnaire only allows postal codes
- As a part of validation, every field in the questionnaire is mandatory
- Child care centre receiving these details from System I may have duplicates in the ranking input of the user (Parents).

Future Improvements:

- Mapping of scrapped reviews/comments to the centre code instead of the child care centre names
- Sentiment analysis score to be specific to unique comment and ignore replies to threads in comments/reviews
- Allowing any one to use a child's IC to register to childcare is a risk. Improve security and authenticity.

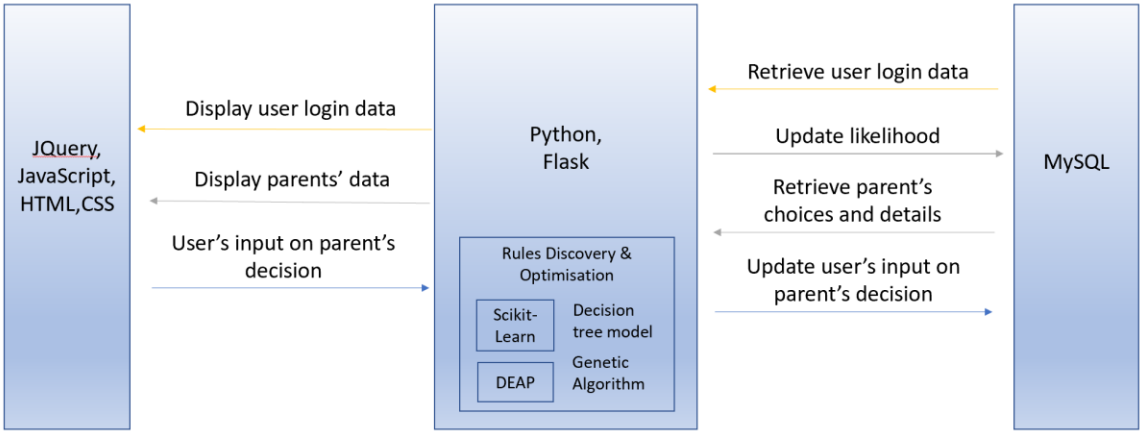


2.6 DESIGN: WAITLISTS FOR CHILDCARE CENTRE

Python and Flask is used as the main framework for development and deployment. The user (childcare centre operator) interface with the system via a HTML frontend throughout the entire process. The frontend displays informationn on waitlists and collects inputs from the user.

Childcare centre user, parent's and wait-list information are stored in MySQL database. Dataset from database are used for rules discovery and optimisation functions. Scikit-Learn and Distributed Evolutionary Algorithms (DEAP) libraries are used for rules discovery and optimisation.

The following diagram illustrates the interaction between the components of the system.



3 SYSTEM SOLUTION & IMPLEMENTATION

3.1 CHILDCARE MATCHER FOR PARENTS

3.1.1 RULE BASED SYSTEM

Filtering based on rules on source of data/input provided by the user(parents). There are service level rules where filtering based on Filter by study level, by citizenship, by type of service (full day, half day AM or half day PM), by fee range and child care level rules Filter by childcare vacancy, by second language, by food and by distance and less stringent rules to filter by childcare vacancy.

3.1.2 SENTIMENT ANALYSIS

Reviews/comments in forums and various other sources are of high importance when it comes to choosing the right pre-school for our kids. Going about surfing the websites of schools and various other forums to read the pros and cons of a school is hectic and time consuming for the parents.

Research with comments/reviews play a very significant role for reasons:

- They help parents decide on the service
- They prove the institution's reputation
- They give an insight of their trustworthiness
- Getting to know about other facilities, infrastructure and maintenance
- Also, about student's welfare etc.

Hence, they have to be taken into consideration no matter how time consuming and tiring they are. In order make this process easy for the parents hunting for child care, we have scrapped comments/reviews from various KiasuParent and have their sentiments analysed and sentiment magnitude score given out using Google API calls. Finally, these sentiments are stored in database. We have also identified a threshold value from overall reviews that we scrapped and divided our comments/reviews based on how positive or negative they are. In addition to the list of recommended childcares for the user, our system also gives the reviews for each school recommended, if available.

In this way the user finds it very useful to have the comments/reviews in one place for their recommendations that helps them choose better childcares based on their preferences.

Future Work/Improvements:

- To scrape better and lot more of reviews/comments
- Restructure them and organize them in more efficient way
- Have lot more of reviews/comments with each school

3.2 WAIT-LISTS FOR CHILDCARE CENTRE

3.2.1 KNOWLEDGE BASED SYSTEM

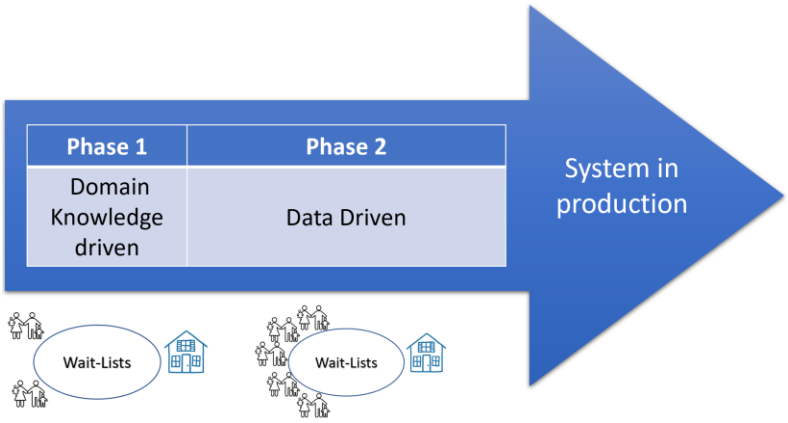
The system's objective is to provide childcare centres with waitlist of parent interests. A key information that childcare centres operators are looking at is the likelihood of prospective parent taking up placements. Hence the core system's function here is to predict this likelihood.

There are 3 classes of likelihood to be predicted by this system: High, Medium and Low.

For each childcare centre account, our system behaves differently at two phases:

1. **Phase 1**: During this time, labelled data is scare, especially when the system is early in production stage. This is because there is small number of parents using the system to submit their interests (for each childcare centre) at this stage. Domain expert knowledge are represented as fuzzy rules and used to classify likelihood.

2. **Phase 2**: When there are more labelled data (i.e., due to higher number of parents using our system, data driven approach is adopted. Decision tree models are set up to discover rules leading to likelihood classification. Genetic algorithm is also used to find more optimal rules leading to highest likelihood (i.e. category "High likelihood").



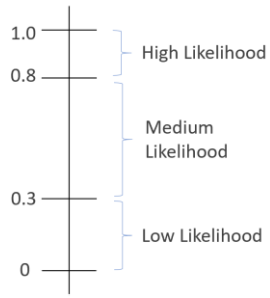
3.2.2 RULE DEFINITION, DISCOVERY, AND OPTIMISATION

Labelled data are scare in early phase of system. When parents use the system to search for childcare centres, their entries in the system forms the data analysed by our system. Parent entries completed with their decision to accept placement in a childcare are referred to as labelled data here.

A threshold of 50 past entries with parents' decisions on whether to take up an offer is used by our system, whereby when the number of labelled data falls below the threshold, our system behaviour is in Phase 1.

3.2.2.1 PHASE 1: FUZZY RULES

The 3 classes of likelihood are categorised into High, Medium and Low here as follows:



Functions are designed to represent domain expert knowledge in fuzzy rule manner, and are used to define likelihood at this stage:

1. Difference (months) between parents' registration of interest and child's enrolment date. When the gap in time between registration and enrolment date is bigger, there is higher the likelihood of parent accepting placement in childcare centre. This is because parents who are sure of their childcare choice tend to register earlier, rather than at the very last minute.

> Fuzzy logic illustration:
>
> IF registration date is 11 months away from enrolment date, THEN parent is most likely to take up offer (100%).
>
> For instance, parent register in Jan 2020 for Dec 2020 vacancy. 11 months is used because a current solution, the ECDA's website, only allows for parents to register up to a year before enrolment date.

2. Difference (months) between the date childcare centre operator is looking at waitlist and parents' registration date. The smaller the gap in time here means that there is higher likelihood of parent prospect accepting placement in childcare centre. If the parent only very recently registered their interest, there is a higher likelihood that the parent will take up the offer

Fuzzy logic illustration:

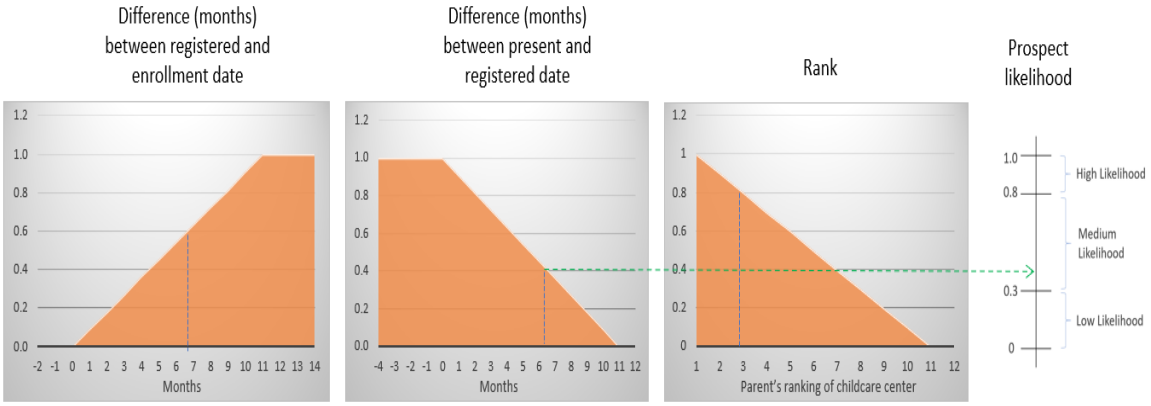
IF parent registered child in Jan 2020 (for later enrolment date) and childcare contacted the parent within Jan 2020, THEN parent is most likely to take up offer (100%).

3. Rank of childcare centre by parent. When parents submit their entries into our system to register their interest in multiple childcare centres, they rank their preference for each childcare centre. A lower number represents a higher rank. A higher rank leads to a higher likelihood of parent prospect accepting placement for that childcare centre.

Fuzzy logic illustration:

IF parent indicate a childcare is rank 1, THEN parent is most likely to take up offer (100%).

The following diagram illustrates the design of functions which is used in fuzzy inference. The minimum output among the three functions is used to define the likelihood of a parent accepting placement in a childcare centre.



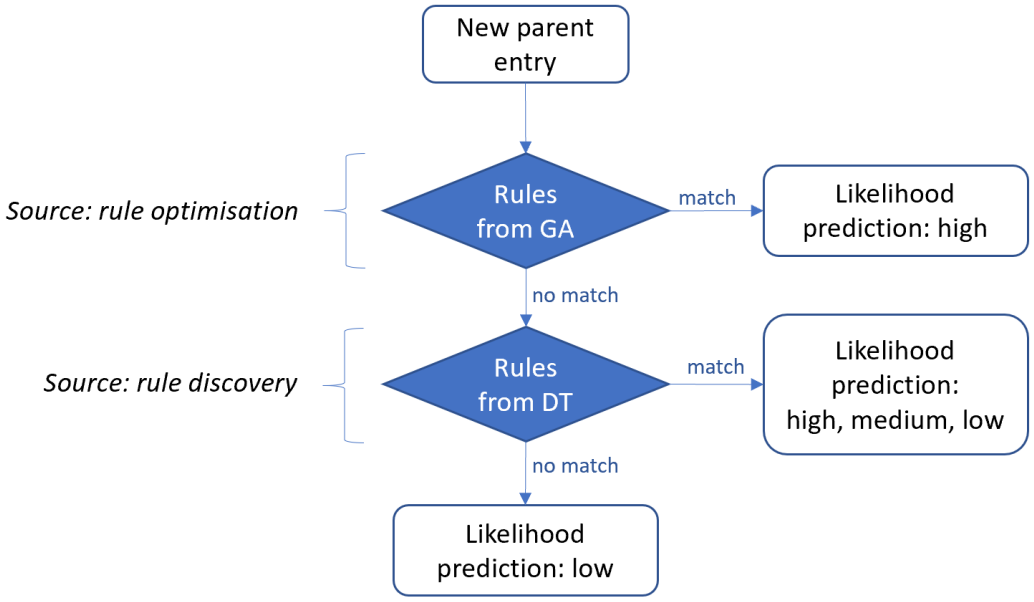
Fuzzy logic illustration:

IF parent indicate a childcare is rank 10 (10%) AND childcare contacted parent the same month he/she registered (100%) and parent is

registering child to attend only 11 months later (100%), THEN parent is only 10% likely to take up offer from this childcare.

3.2.2.2 PHASE 2: RULES DISCOVERY

When the number of labelled data[4] is above threshold (50), our system is data driven and in Phase 2. Prediction of likelihood is done using decision tree (DT) and genetic algorithm (GA) as follows.



The rules set from the decision tree regressor model is the core rule set used to determine prospect likelihood for all likelihood classes. The GA is an add-on, to find more optimal rules leading to "high" likelihood.

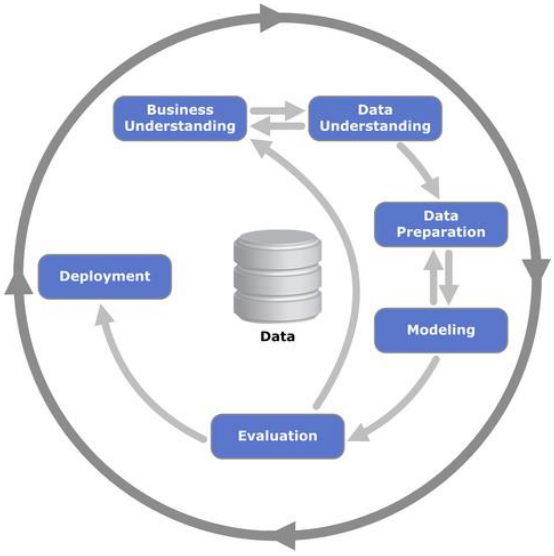
[4] Labelled data refers to parent entries completed with their decision to accept placement in a childcare.

Phase 2 is data-driven in the sense that when the real world environment changes, or when there are differences in parent's consideration for different childcare centres, these differences are captured from past labelled data, and used in rule discovery.

Decision tree models are used to discover rules leading to likelihood categories of high, medium and low. GA is also used to find optimal rules leading to high likelihood. The 3 classes of likelihood are categorised into High, Medium and Low here using ranges:

Likelihood	Range
High	2 and above
Medium	1 to 2
Low	0 to 1

CRISP DM, framework is used to perform the rules discovery from labelled data.



1. Business Understanding: The present business problem is that childcare centre operators need to manage an inflated waitlist of prospective parents which does not reflect true demand. The business objective for the system here is to estimate the likelihood of a prospect accepting placement. The operator may use this information in planning and managing of prospect waitlist.

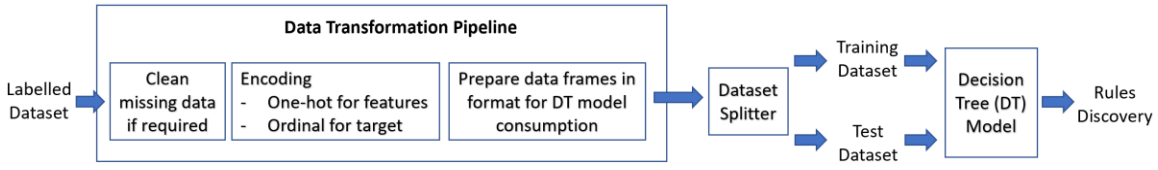
2. Data Understanding: ECDA website was studied to understand features that parents look for when searching for childcare centre. Knowledge from ddomain expert (i.e. parents) was gathered. This collective understanding is used to create a sample dataset. The format of dataset match the format actual data to be collected when system is in production.

   Real data about prospects was considered to be collected via surveys. A sizable dataset for decision tree model development is required. This means a sizable survey is required. Financial cost was a limitation during this project, hence data is simulated and used for model development. When the system is operational, there will be real data on prospects when they provide inputs into the systems.

3. Data Preparation: There is a need for one-hot-encoding of data. Decision tree models will be used set up using Scikit-Learn library in Python. Scikit-Learn has a limitation that only numeric data inputs are accepted for a decision tree set up.

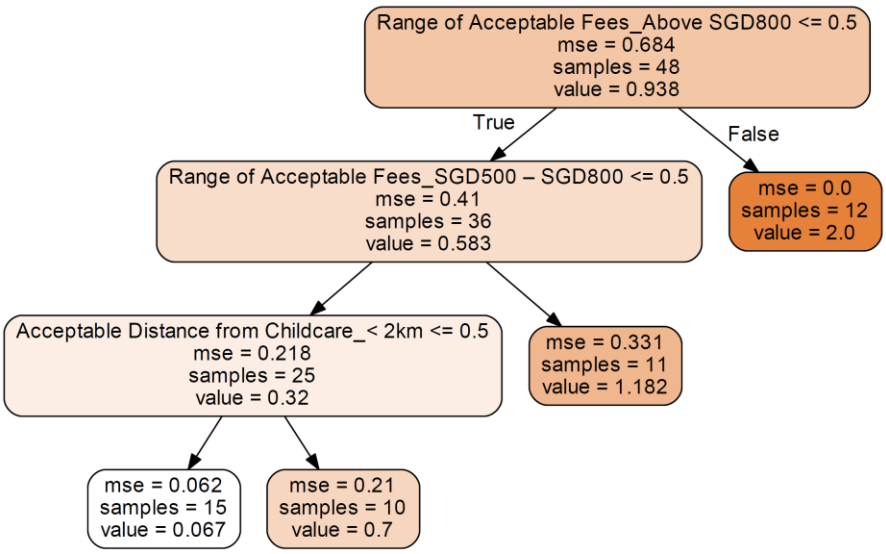
   The labelled values for the dataset is made up prospective parent final decisions which are categorised as “Reject offer”, “No contact” and “Accept offer”. These responses represents likelihood of accepting placement, and are encoded ordinally.

   A data transformation pipeline had been prepared for data preparation activities. After data transformation, the dataset is partitioned into 2 sets for training and testing.



4. Modelling: Decision tree Regressor model was set up for the discovery of three set of rules leading to 3 classes of likelihood (i.e. Low, Medium, High) of prospective parent accepting childcare centre placement.

   The model produces decision tree diagrams and rules are also displayed in text form. The diagrams are used for interpretation of the rules output. The rules discovered here are used to determine likelihood of parents accepting placement. The following shows examples of model outputs.




```
|--- Range of Acceptable Fees_Above SGD800 <= 0.50
|   |--- Range of Acceptable Fees_SGD500 – SGD800 <= 0.50
|   |   |--- Acceptable Distance from Childcare_< 2km <= 0.50
|   |   |   |--- value: [0.07]
|   |   |--- Acceptable Distance from Childcare_< 2km >  0.50
|   |   |   |--- value: [0.70]
|   |--- Range of Acceptable Fees_SGD500 – SGD800 >  0.50
|   |   |--- value: [1.18]
|--- Range of Acceptable Fees_Above SGD800 >  0.50
|   |--- value: [2.00]
```

5. Evaluation: Scikit-Learn's GridSearchCV library was used to find the best set of hyperparameters (tree depth, number of samples per leaf) to form the model. Mean-square error was used as metric in the selection.

   Using test dataset, the decision tree model was also evaluated based on mean-square error between it's prediction and true label values.

6. Deployment: After the evaluation was successful, the code was integrated with the overall system.

3.2.2.3 PHASE 2: RULES OPTIMISATION

When childcare operator manages their waitlist, one task is to allocate resource (manpower eg.) to look into prospective parents with "high" likelihood of accepting placement.

Rules are discovered by decision tree model for 3 classes of likelihood (ie. Low, Medium, High). The model adds a label of numeric value to indicate the likelihood. Genetic algorithm (GA) is used in our system to find optimal rule sets that leads to "High" likelihood value of 2 or higher.

Likelihood	Range
High	2 and above
Medium	1 to 2
Low	0 to 1

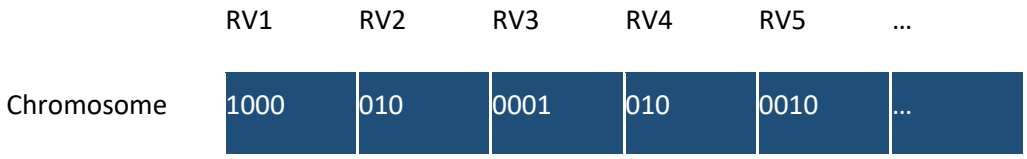
The following describes how GA approach is designed used to perform this.

1. Chromosome design:

- When the decision tree is being trained using training dataset, the features values of the dataset are one-hot encoded. These feature values represent rule variables. The following shows a partial example:

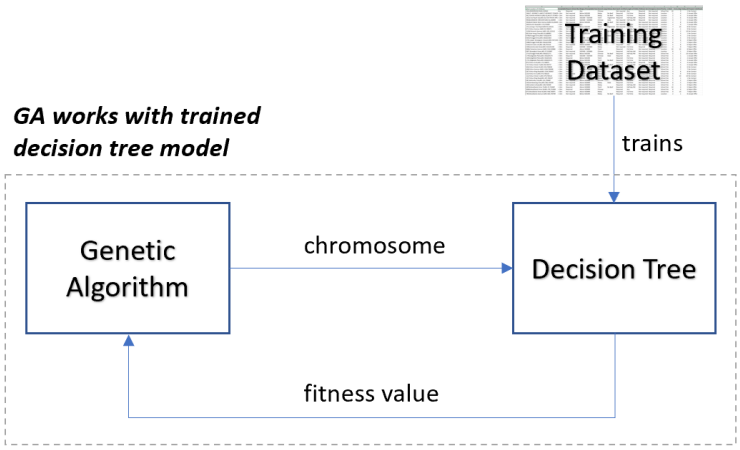
Description	Values	Encoded values
Rule variable (RV) 1: *Acceptable distance from childcare*	{ <1km, <2km, <3km, Any }	{ 1000, 0100, 0010, 0001 }
Rule variable (RV) 2: *Second Language Taught*	{ Chinese, Malay, Tamil }	{ 100, 010, 001 }

- The chromosome design for GA use matches the format of that decision tree model use. Hence, rule variables are represented by sets of binary numbers concatenated together to form the chromosome being fed into GA,



1. Fitness Value:

- The trained decision tree regressor model described earlier (section 3.2.2.2) is used here. When a GA generated solution is passed into the decision tree model, one of the leaf nodes is activated and the model outputs a prediction value.



- The fitness value used here is a weighted sum of: **leaf node sample coverage + decision tree regressor prediction value**.

- A specific leaf node in decision tree is activated when the decision tree model is used to provide a likelihood prediction. Sample coverage refers to the proportion of training samples captured by the leaf node relative to total number of training samples used during decision tree training.

- GA is looking for chromosomes (set of rule variables) with higher fitness values in terms of higher sample coverage and prediction value.

1. Constraint:

   - Each rule variable in the chromosome is one-hot encoded. The constraint here that GA qualifies as a valid chromosome/solution is that there can only be one '1' value per rule variable.

2. Selection:

   - Tournament selection method is used to select parents from the population. The winner of each tournament is used for crossover and mutation.

3. Crossover and mutation:

   - Random one-point crossover and random flip-bit mutation functions from Distributed Evolutionary Algorithms (DEAP) library is used.

4. How GA is used by system:

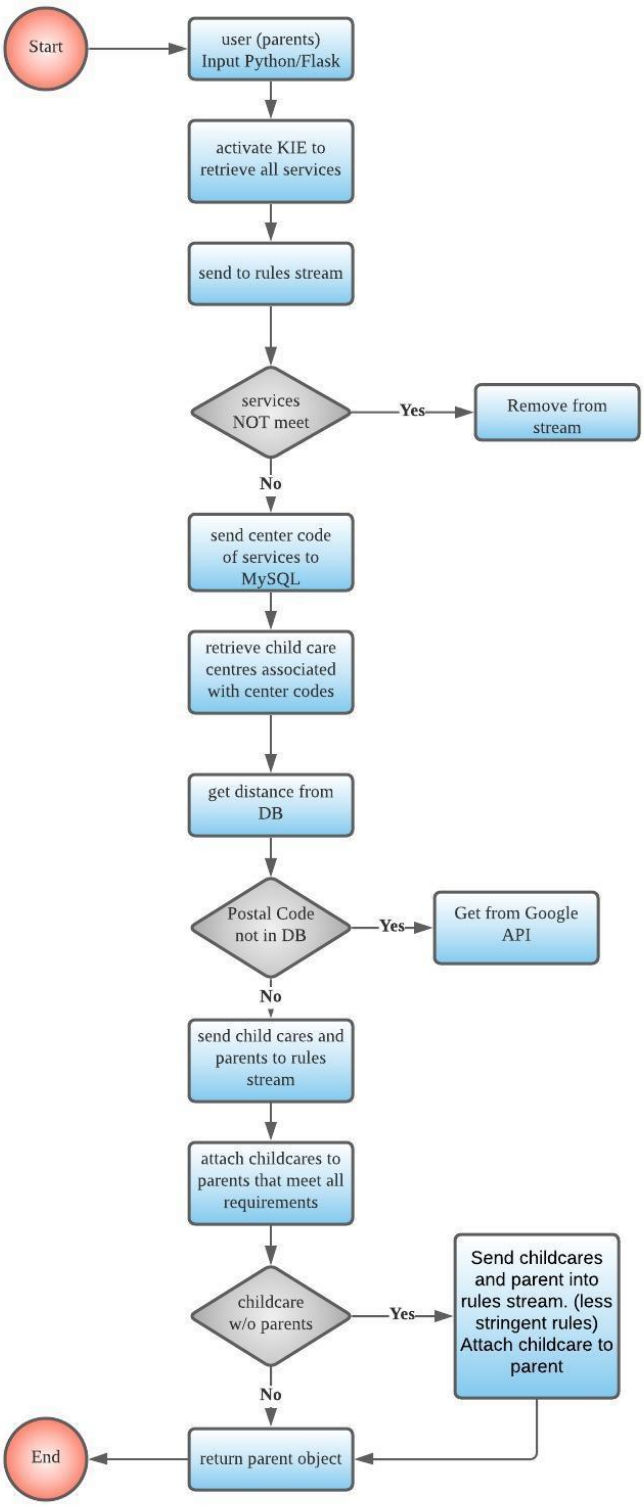
   - GA generates a population of solutions. A set of solutions with fitness values equal or above the fitness value of 2.0 is extracted.

   - When the system is in live production mode and receiving data inputs from parents, the data input is encoded into chromosome format and checked for matches with this set.

   - If there is a match, the data input is classified as "High" likelihood. If there are no matches, rules from decision tree will be used to predict likelihood.

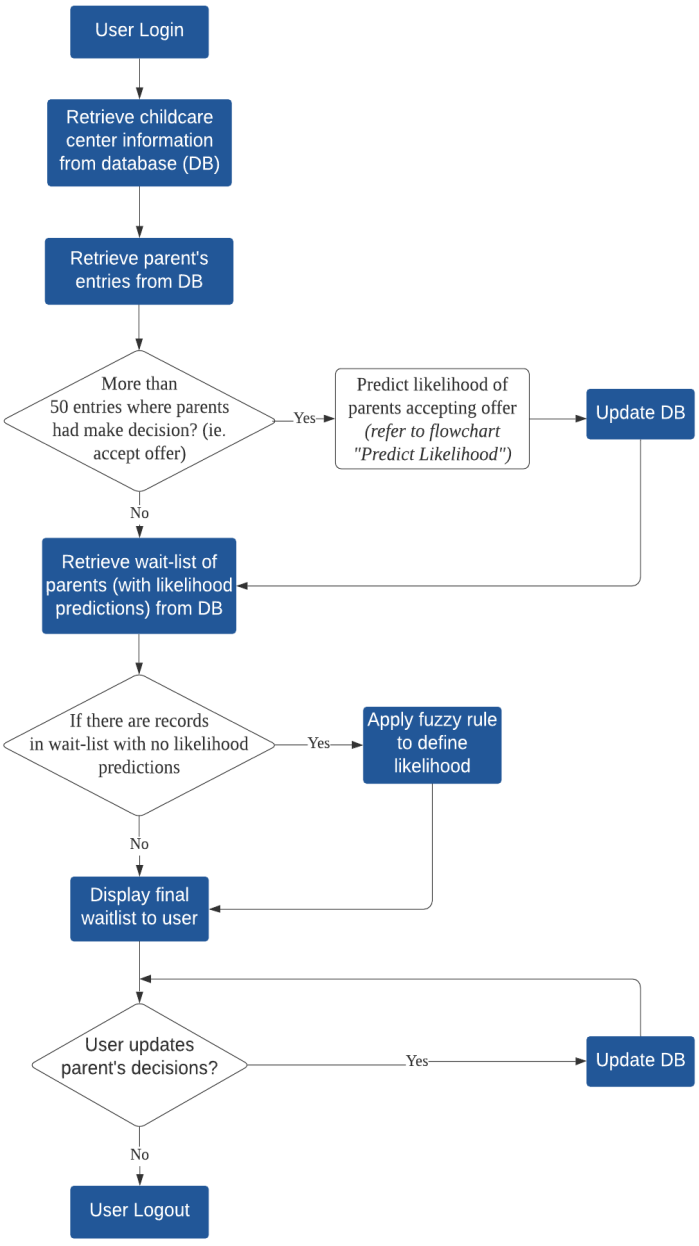
4 SYSTEM FUNCTIONAL DETAILS

4.1 SYSTEM DATA FLOW: CHILDCARE MATCHER FOR PARENTS

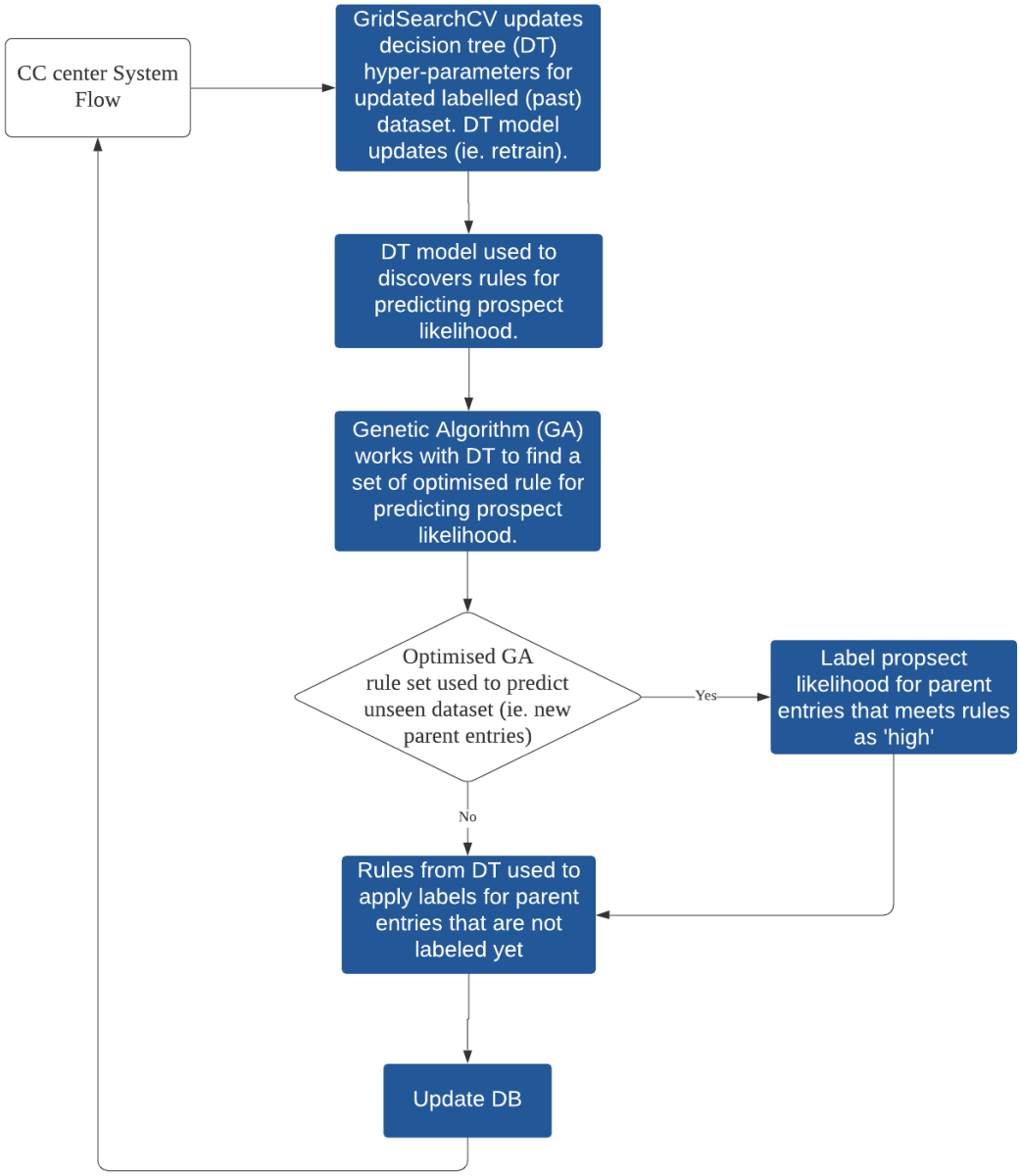

Start
user (parents) Input Python/Flask
activate KIE to retrieve all services
send to rules stream
services NOT meet
Yes
Remove from stream
No
send center code of services to MySQL
retrieve child care centres associated with center codes
get distance from DB
Postal Code not in DB
Yes
Get from Google API
No
send child cares and parents to rules stream
attach childcares to parents that meet all requirements
childcare w/o parents
Yes
Send childcares and parent into rules stream. (less stringent rules) Attach childcare to parent
No
End
return parent object

4.2 SYSTEM DATA FLOW: WAITLIST FOR CHILDCARE CENTRES

4.2.1 WAIT-LIST GENERATION AND DISPLAY



4.2.2 PREDICT LIKELIHOOD



5 CONCLUSION

Our team had a very good time working together. Knowledge gathering payed way to lot of learning and understanding of the real time challenges as new parents. The structured data collection introduced us to new skills and interviews of child care centres helped build our network and interactions and mainly getting to know the pros and cons of our proposal.

Building the system introduced many challenges, KIE drools was new and a learning for our team. Working on this project allowed us to learn from each other who are experts in certain areas of development.

Overall this system is of great value add for new parents and one stop place to get recommendations and also we look at the other end of the system helping child care centres to optimize the waitlist to promote better quality in terms of the benefits out of our product.

6 APPENDICES

APPENDIX A: INTERVIEW OF THE CHILD CARE CENTER

Survey - National University Of Singapore - Project exploration Inbox ×

Priya Parthiban <padmaparthi04@gmail.com> Fri, Aug 7, 11:37 PM
to rosieta.littlegemspreschool

Greetings Ms. Rose
I'm doing a project in Intelligent systems with NUS.
As a part of our Project - I approach you to understand the challenges that puts off when it comes to managing your wait list in order to get parents enrolled in your institution.
We are in process of developing a system for Child Care center wherein they can better manage their wait list. To brief out on our system,

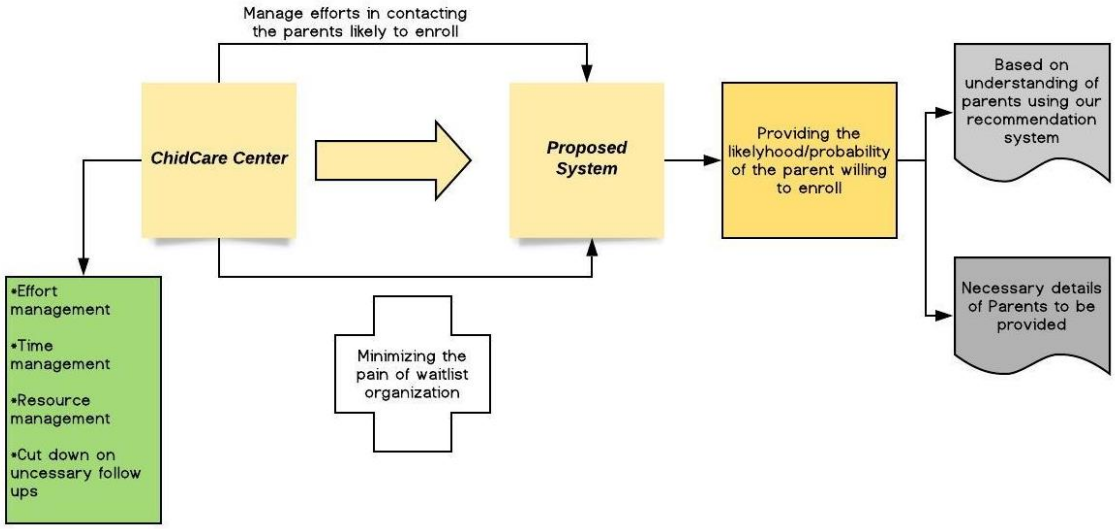
1. We have a initial Child care recommendation system where in the parents register and give their preferences/needs based on their inputs we recommend a list child care suitable for their needs.
2. Based on the inputs provided by parents we come up with a probability/likelihood factor that the parent may/may not be interested to enroll in the particular Child care center.
3. This probability factor and necessary details of the parent to be provided to you which enhances your chances of optimizing time to contact the parents who are more likely to enroll in your center.

Please find below basic flow of the process for your understanding.
Please feel free to let us know

1. If such a process can help reduce your challenges in contacting/follow up of parents ?
2. whether you would be interested to give our model a test try?
3. Challenges that you can share with us in managing wait list based on the current system ?

Thanks for the Time and Regards
Priya
Student - NUS -ISS

***Information in this email would be used for review and audit purposes for the project*



Rosieta Abdul Mutalip <rosieta.littlegemspreschool@gmail.com> Aug 13, 2020, 11:52 AM
to me

Dear Priya,

I would like to try out the new system if it is possible. It can probably help us in the future waiting list for parents. As of now, the challenges that we faced are time consuming as we need to have a virtual tour for parents to view the centre or they can only come down to have a tour when the children are not around in the evening or on Saturdays. They take time to get back to us whether they are interested or not. So it would be good if your system is better and let's see from there.

Looking forward to hearing from you.

Warm Regards,
Rosieta
Principal

LITTLE GEMS PRESCHOOL PTE LTD
3 Ang Mo Kio St 62 #01-09/10 Link@AMK S(569139)
T: 6253 7009 (AMK) HP(92217141)
W: www.littlegemspreschool.sg