Answers :

Queston1:-

Part -a

URL of the webpage: https://archive.ics.uci.edu/ml/machine-learning-databases/00529/

URL for the site:
https://archive.ics.uci.edu/ml/datasets/Early+stage+diabetes+risk+prediction+dataset.

Part -b

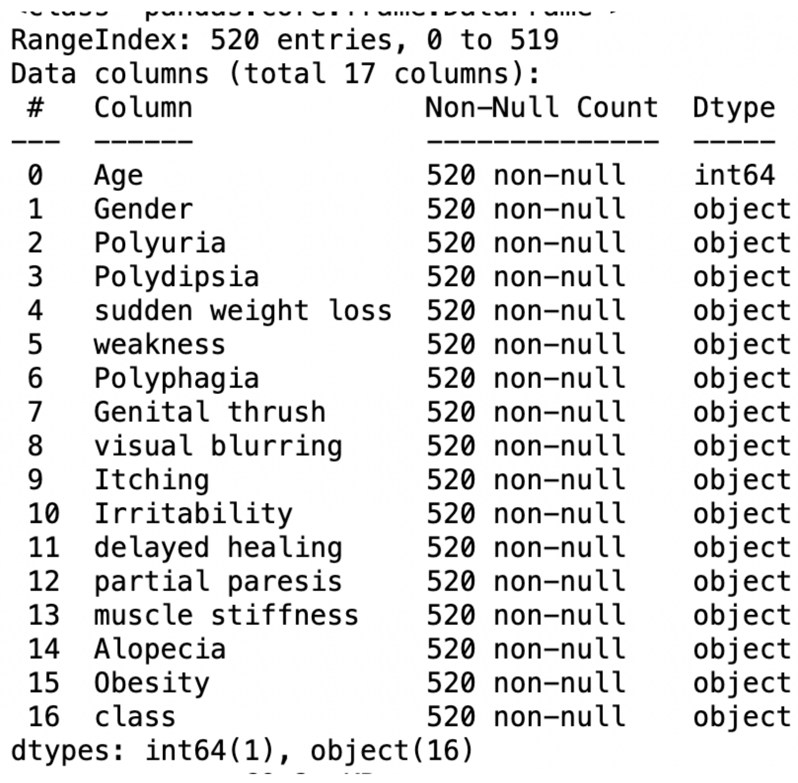
Brief description of data set:

--The data set is used for the early-stage diabetes risk prediction, it is one CSV file containing all the information

-- Objects: we have a data set of size and length (520, 17)

--Attributes: We have 17 attributes for the above data set. which is a combination of different types namely Gender(objects), age(Integer), others are binary either yes or no, and in the last column name class it is positive/negative (objects)

Please find the below information related to the given dataset, which defines how many objects and attributes are present in the dataset and their types,



```
RangeIndex: 520 entries, 0 to 519
Data columns (total 17 columns):
 #   Column              Non-Null Count  Dtype 
---  ------              --------------  ----- 
 0   Age                 520 non-null    int64 
 1   Gender              520 non-null    object
 2   Polyuria            520 non-null    object
 3   Polydipsia          520 non-null    object
 4   sudden weight loss  520 non-null    object
 5   weakness            520 non-null    object
 6   Polyphagia          520 non-null    object
 7   Genital thrush      520 non-null    object
 8   visual blurring     520 non-null    object
 9   Itching             520 non-null    object
 10  Irritability        520 non-null    object
 11  delayed healing     520 non-null    object
 12  partial paresis     520 non-null    object
 13  muscle stiffness    520 non-null    object
 14  Alopecia            520 non-null    object
 15  Obesity             520 non-null    object
 16  class               520 non-null    object
dtypes: int64(1), object(16)
```

Part -c:

- Early diabetes identification is usually preferred for a clinically significant result due to the existence of a relatively extended asymptomatic phase.
- From the data set we have based on the symptoms like sudden weight loss, Irritability, obesity, etc a person can be classified whether he has diabetes or not
- Only by careful evaluation of both common and uncommon symptoms, which can be identified at various stages from disease onset to diagnosis, is an early diagnosis of diabetes possible.

Part -D:-

- This knowledge is quite useful, if in the future if a doctor gets a patient and he has some of the symptoms lets say 6 symptoms mentioned in the table, then he has a high risk of getting diabetes or he is already suffering from that
- The applications of data-mining techniques in the selected articles were useful for extracting valuable knowledge and generating new hypotheses for further scientific research/experimentation and improving health care for diabetes patients.
- The results could be used for both scientific research and real-life practice to improve the quality of health care for diabetes patients.

Part a :

URL of the webpage:
https://www.kaggle.com/datasets/divyansh22/flight-delay-prediction?resource=download

URL for the site:
https://www.kaggle.com/datasets/divyansh22/flight-delay-prediction?resource=download

Part -b

Brief description of data set:

--This dataset contains all the flights in the month of January 2020.

– There are more than 400,000 flights in the month of January throughout the United States. The features were manually chosen to do a primary time series analysis.

– There are several other features available on their website.

-- Objects: we have a data set of size and length (1191331, 23)

--Attributes: We have 23 attributes for the above data set. which is a combination of different types like DAY_OF_MONTH, DAY_OF_WEEK, **OP_UNIQUE_CARRIER, OP_CARRIER_AIRLINE_ID, OP_CARRIER, TAIL_NUM, OP_CARRIER_FL_NUM, ORIGIN_AIRPORT_ID**

	tipo
DAY_OF_MONTH	int64
DAY_OF_WEEK	int64
ORIGIN	object
DEST	object
DEP_TIME	float64
DEP_DEL15	float64
DEP_TIME_BLK	object
ARR_TIME	float64
ARR_DEL15	float64
CANCELLED	float64
DIVERTED	float64
DISTANCE	float64

year	int64

Part -c:

- This data could well be used to predict the flight delay at the destination airport specifically for the month of January in upcoming years as the data is for January only.
- This file contains all the flights starting from 1st January 2020 till 31st January 2020. There are around 1191331 rows in this file and 23 feature columns indicating the features of the flight including information about the origin airport, destination airport, airplane information, departure time, and arrival time.
- By using the current month data set related to flight delays and their arrival times, we can predict the upcoming flight delays or arrival times.

Part -D:-

- It will be pretty useful for the people who travel a lot from one place to another and plan their journey according to the predictions.
- From this data set, we can even tell which career is maintaining the timeline and arriving on time at the airport.
- We can even say when a flight is canceled due to what by checking the cancelled column in the data set

Question 2:-

Part -a:- Paper session, title, authors, and affiliations.

- Paper session: Research Track Full Papers
- Title: Learning Models of Individual Behavior in Chess
- Authors: Reid McIlroy-Young, Russell Wang, Siddhartha Sen, Jon Kleinberg, Ashton Anderson

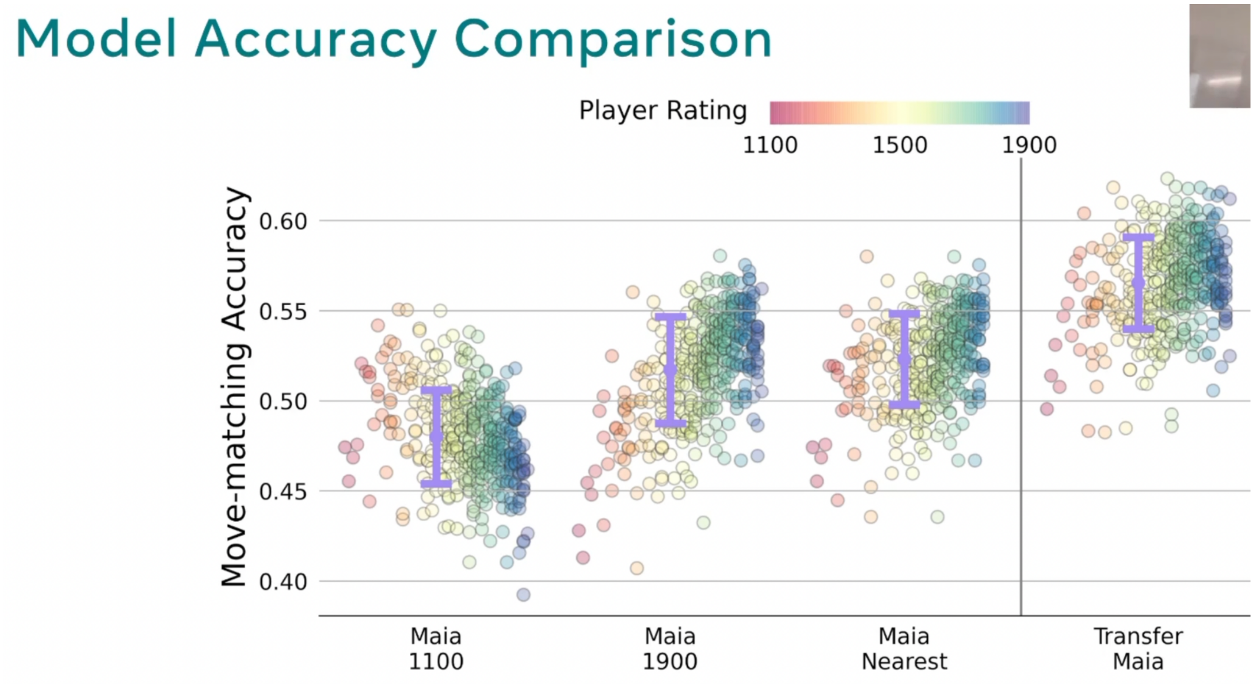
Part -b :- What problem is addressed? Why is the problem important and challenging?

- We all would have played online chess or even playing now with a AI bot. Which is only capable of doing some ideal moves which follows the basic rules that are restricted moves or actions.

- But the problem here is it only follows the set of instructions which are pretty common moves done which are aggreated moves. In this research, the aim is to predict the individualize chess moves for a player.
- The challenge of anticipating human actions, as opposed to anticipating ideal actions, has gotten a lot of attention in the quest to create AI systems that are focused on humans.
- This solution is the updated version of the previous one named maia which is a neural chess engine model which will follow some set of rules to predict the next moves knowing only their skill levels
- The challenge here is to predict the moves of a individual player who has their own style of play rather than the following the steps. I.e tricks, new moves something like that.

Part -c:- A high-level, brief description of the proposed solution (no need to include the details)?

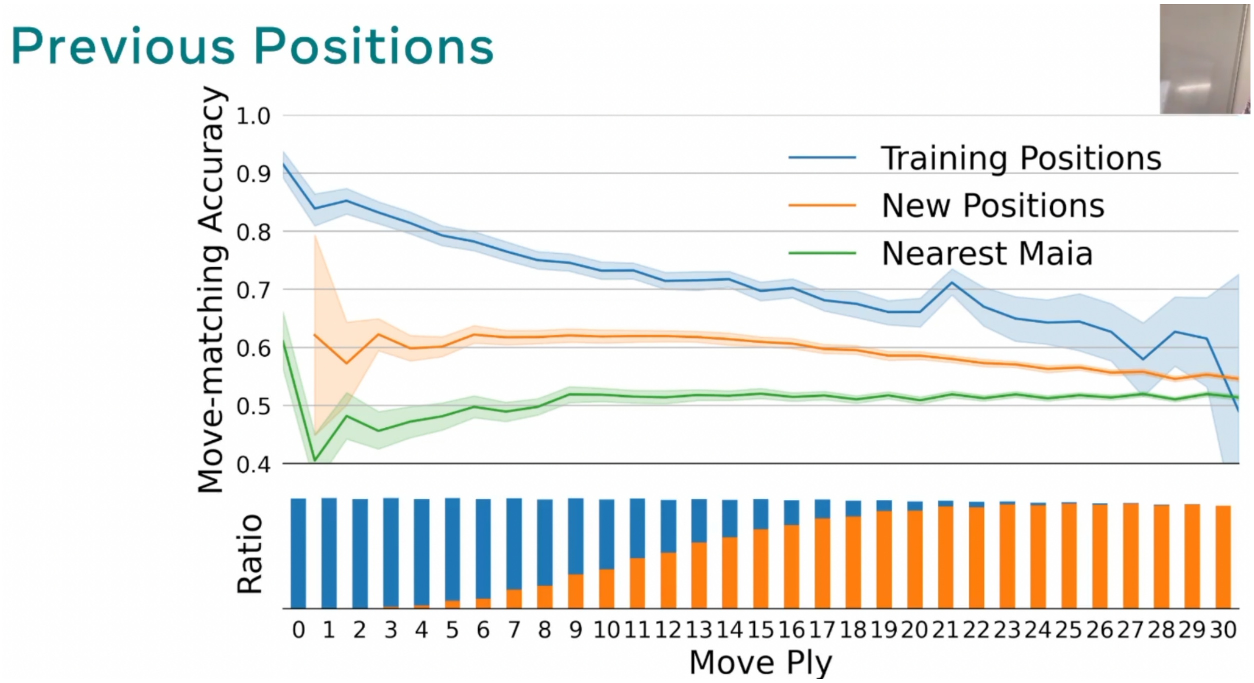
- The value that each specific person may receive from interacting with these systems is potentially constrained by the existing work's emphasis on capturing human activity in an aggregate sense.
- Selected players from the blitz(one of the popular chess games) who having more than 20000+ games
- Final methodology is to use the training set, use transfer learning to convert a maia model into a individualized model
- IN accuracy comparison, individualized model outperformed all maia models achieving 4-5 % higher accuracy per player



The final results and conclusions will be found in the part d.

Part -d:- How is the proposed solution evaluated?(data sets, metrics, etc.)

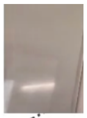
Data set information :

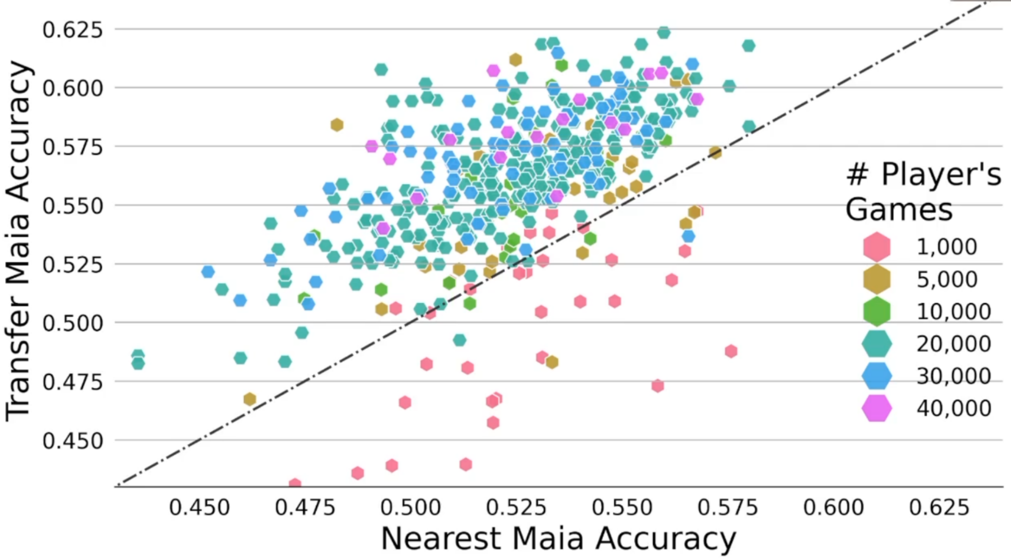
1. Over 20000 rated games in blitz(3-7minutes) time frame.
2. Low variance of rating over time.
3. Rating over 1000 and 2000(medium skilled).
4. For each player divide their game into
    a. 80% training
    b. 10% validate
    c. 10% test
5. From this, they have come up with IQR to find the high skill and low skill for the current approach. What is the difference of accuracy a individual player can achieve.
6. Final solution sample of 400 players
7. Final Results from above results and procedure
    a. Examining 400 players using 100 sets each has the following results
        i. 98.4% top 1 accuracy
        ii. 98.5 % 1st 5 moves removed
        iii. 55% 1st 10 moves removed
8. Transfer models perform very well on positions they have seen before, but still outperform maia model in novel positions - for move matching accuracy
9. 

Previous Positions

10. Sample size effects play a major role in the data mining, which can alter the final predictions in the accuracy or move-in probabilities. Please find the below for the sample size effects and its accuracy effects

Sample Size Effects





11.