Capstone Project Proposal

Jimmey Jose

Business Goals

Project Overview and Goal What is the industry problem you are trying to solve? Why use ML/AI in solving this task? Be as specific as you can when describing how ML/AI can provide value. For example, if you're labeling images, how will this help the business?	**Problem** Answer script valuation by teachers/valuators is a very time consuming process in which the steps of summing up the marks/scores and inputting into report cards or online/digital systems is a monotonous and repetitive task prone to many errors; Eliminating this process of mark summation and system input is the problem this product is trying to solve. **Goal** Build a product which can consume scanned evaluated answer scripts and give an output report of the summed up annotated marks grouped by the student identification and by question answer reference number/alphabet. **Why use ML/AI in this product?** Identifying annotated marks/scores grouped by answer numbers from answer scripts is the core part of this product without which we would not be able to extract the data to solve this problem; this is where Machine Learning adds value in this product.
Business Case Why is this an important problem to solve? Make a case for building this product in terms of its impact on recurring revenue, market share, customer happiness and/or other drivers of business success.	An average teacher would take somewhere between 30 seconds to 3 minutes to sum up and input into the final scores into their logs. This process involves the following steps 1. Sum up all annotated marks 2. Write the marks on ths sheet 3. Input final score into their log by writing or using a computer; there is also a probability of this process being repeated multiple times Automating the above 3 steps will save the answer

	script valuer a lot of time depending on how many scripts he/she has to value. Below is an illustration of how much effort a system like this can save using an example A school will conduct least 4 main exams a year for a minimum of 4 subjects. If the valuator spends 2 mins on average for this activity, there is a saving of 32 mins (4*4*2 mins) per student in a year. A school with 1000 students will have a saving of 533 hours (32 min * 1000) in a year.
Application of ML/AI What precise task will you use ML/AI to accomplish? What business outcome or objective will you achieve?	ML will used to accomplish the following tasks 1. Identify student identification info from the first sheet of the answer script 2. Identify valuator annotated Marks/Scores on the answer sheet. 3. Identify question answer reference number/alphabet and the start and end of each answer associated with it 4. Map the student id to the list of marks/scores which is also mapped to the question answer reference number The above inputs extracted from the ML/AI model is the basis for this product which will eliminate the process of summing up and inputting marks/scores for answer scripts.

Success Metrics

Success Metrics What business metrics will you apply to determine the success of your product? Good metrics are clearly defined and easily measurable. Specify how you will establish a baseline value to provide a point of comparison.	The following metrics will showcase the accuracy of the product and the level of intervention required by students or teachers / valuators. 1. Overall percentage of answer scripts in which no manual intervention was required from students or valuators in a single examination set (interventions based on missed mark annotations or reconsideration of marks is not considered here) 2. Overall reduction in time spent by the valuator; this will be a summation of the amount of time spent before the product was used for these

	activities minus any time spent on interventions after the product was introduced **Baseline setting for each of the above metric** 1. The higher the first metric, the better; the baseline for this should be at least 80 % when we start of else there is risk of losing users in the beginning itself. In the long run, the aim should be to get this to 99 to 100% 2. The baseline for the second metric is again subjective to the valuator. Ideally this should be lower than the time the valuator spends combined in summing, inputting and other interventions

Data

Data Acquisition Where will you source your data from? What is the cost to acquire these data? Are there any personally identifying information (PII) or data sensitivity issues you will need to overcome? Will data become available on an ongoing basis, or will you acquire a large batch of data that will need to be refreshed?	**Data Source** Partnerships will have to be established with educational institutions who will be the customers for this product to get source data and pilot programmes. The main data inputs will be evaluated answer scripts of various types. These answer scripts can be scanned and returned. **Personally Identifying Information** One of the key data points they system needs to capture is the student identifier data which can be the name or roll number based on the data privacy requirements specified by the institution **Cost and availability of data** Partnerships with institutions should ideally enable us to get the continually get training data at no cost. The models should ideally be validated with data from every institute we engage to identify and correct by training with institute/subject specific data.

	Estimated size and type of training data Training the models will require at least 500 labeled answer scripts for the criteria noted below. Ideally we should be able to mix in answer scripts satisfying the different criteria to lower the data size. The ideal minimum size for each criteria is also mentioned. 1. Answer scripts mixed between different institutes for every subject or similar subject for which the training model will be used. Ideal minimum data set size - 1000 answer scripts split between 2 to 3 institutes for each subject or similar subject 2. Answer scripts mapping different kinds of question answer reference types. Ideal minimum data set size - 600 answer scripts split between 3 different question answer reference types. 3. Answer scripts with a combination of different types of stationary used - ink color, types of paper etc. Ideal minimum data set size - 1000 scripts with upto 10 different combinations. 4. Answer scripts with all possible types of annotated marks. Ideal minimum data set size - 500 scripts 5. Unevaluated answer scripts - 500 scripts The data set will need at least 2500 answer scripts if we intend to train from only 1 subject (or similar subject as mentioned in point 1). Typically each answer script will be at least 2 to 3 pages or more meaning the training data set size will be a minimum of around 7500 scanned pages. The numbers used are based on best guesses considering the data type which will need to be trimmed and refined more as we start training the data and learning more about it.
Data Source Consider the size and source of your data; what biases are built into the data and how might the data be improved?	**Biases** The following varied characteristics in the data should be considered as they could potentially introduce bias 1. Varied handwriting styles which might result in poorer identification of certain styles due to biases in available OCR from the following perspectives

	a. Annotated marks made by teachers b. Student identifier and question answer references made by students 2. Different language answer scripts can introduce bias based if the model is not trained with similar data 3. Varying stationary like ink color or answer script paper type may result in very different looking answer scripts can result in noise being trained and lead to bias
Choice of Data Labels What labels did you decide to add to your data? And why did you decide on these labels versus any other option?	1. **Student Identifier -** This will identify the student by name, roll number, class or any other variable decided by the institute management 2. **Question Answer Reference** This will identify the questions which were attempted by the student 3. **Answer Block -** This will identify the starting and ending point of the answer by the question answer identifier marked earlier 4. **Annotated Marks -** This will identify the marks given by the valuator for each answer identified by the labels 2 and 3 The above four labels represent the data to be extracted with its grouping which makes it meaningful. They capture a students marks for a single answer script grouped by each question answer reference.

Model

Model Building How will you resource building the model that you need? Will you outsource model training and/or hosting to an external platform, or will you build the model using an in-house team, and why?	This model should be ideally custom build leveraging existing open source libraries because of the following considerations. 1. Need to deploy at least 2 separate ML models for data extraction which is difficult to outsource to an external platform a. Model to identify and classify locations of Mark annotations, Student ID info and Question answer references along with its start and end point b. NLP models for doing OCR of mark annotation, Student IDinfo and question answer references

	2. Requirement to go through multiple images for a single data set making it very expensive if its outsourced as the billings are generally by page/size of data which will be huge here
Evaluating Results Which model performance metrics are appropriate to measure the success of your model? What level of performance is required?	As discussed above this product will need at least the two models mentioned above. 1. The first model which is being used for identifying and classifying the data locations should use recall as the main metric to assess performance as we need to identify all possible data points here to move forward. We need to aim for a high recall even at the cost of a low threshold here. We should be aiming for a recall > .9 here 2. The second model which is an NLP model should use precision as the main metric. Here we need to aim for reasonably high precision at a reasonable threshold.. Our aim here is to identify the object content using OCR, discard any wrong data picked up by the first model and extract the required data because of which a low precision on the data extracted will generate wrong reports and increase manual intervention; it is more important to get it right than capture all here. We should be aiming for a precision level >.8 here.

Minimum Viable Product (MVP)

Design What does your minimum viable product look like? Include sketches of your product.	The features in the MVP will be limited to the following: 1. ML model to extract marks grouped by student IDs and question answer reference numbers/alphabet 2. A webapp for with teacher and student logins to enable the following a. Teachers/valuators can view the consolidated marks report by student and class and the scanned answer scripts to cross verify and make corrections b. Students can view their scanned answer scripts and consolidated marks

	Wireframes for teacher and student webapp are in the appendix at the end of this document; the wireframes are also attached separately without the descriptions
Use Cases What persona are you designing for? Can you describe the major epic-level use cases your product addresses? How will users access this product?	**Personas** 1. Teachers / Valuators 2. Students **Epic Level Use Cases** 1. Marks extraction grouped by students IDs 2. Marks grouping by the question answer reference number/alphabet 3. Data collation and report creation at the student, class and school level 4. Basic Webapp for teachers to view consolidated marks scanned answer scripts by student and class 5. Basic Webapp for students to view their scanned answer scripts and consolidated marks
Roll-out How will this be adopted? What does the go-to-market plan look like?	**Pre-launch plan** 1. Testing Phase (Pre-launch) As mentioned in the data acquisition plan, we will need to establish partnerships with a few educational institutes in which the system can be tested with every script being checked both manually and using the product to showcase the effectiveness and time saving ability. 2. Pilot Phase (Pre-launch) Once the testing phase is over, the product will have to deployed free at the partner institutes for collecting data on savings and for demos for other institutes during launch. 3. Launch This product should be launched and publicised through school boards and universities which have huge sway on large number of institutes. The effectiveness of the product can be showcased using the time and cost saving data collected during pilot and also through live demos at pilot institutes. Since there will be a lot of scepticism around the accuracy of the product in the initial phase, it should be deployed for free or at deep discounts during the growth phase. 4. Post Launch

	The effort and time saving should ideally give the product a huge push through word of mouth which should be supported by a strong web presence to quickly deploy and close sales .

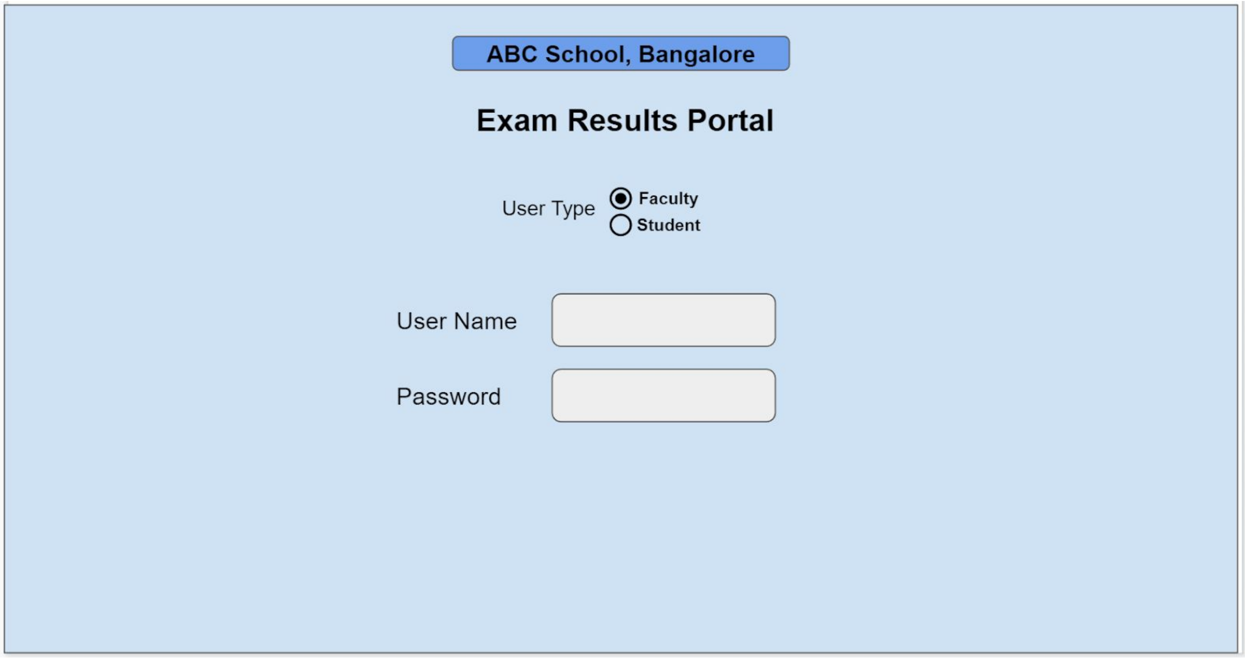
Post-MVP-Deployment

Designing for Longevity How might you improve your product in the long-term? How might real-world data be different from the training data? How will your product learn from new data? How might you employ A/B testing to improve your product?	This MVP is the first stage on top of which an end to end platform can be built for conducting and evaluating all kinds of exams. A few of the features that can make this a game changer in the education space are: 1. An easy digital annotation platform which can be used for valuing answer scripts on any mobile device 2. Enabling auto valuation or valuation helper for objective type questions which are single words and descriptive 3. Question answer bank for easy creation and deployment of standardized exam papers across boards and universities **Testing** (Training Data Vs Real Data) Answer script patterns are more or less the same across the world with students answering against the question answer reference number/ alphabet. This ensures that the training data which is collected from the partner institutes will be in line with answer scripts from other institutes. However, differences in the type of stationary used (paper, pen color etc.), Language of answer scripts , Varying subjects and Handwriting styles can have significant impacts on the data extraction ML model accuracies. This can be countered by the active learning approach build into the product described in the below section. **Active Learning** The best approach would be to mandate additional training with institute, subject or class specific data before deployment in new institutes; the product should have features which enable the institute to easily upload

	this data which can be used to train. 10 % of this new data should be used for manual validation to ensure higher confidence levels. **Versioning** The foot in the door to educational institutes using the MVP should open up opportunities to deploy the first version in multiple institutes from which more data can be obtained for refining the model for future versions. The first few iterative versions will focus only on improving the existing model and not adding new features. We should move on to adding new features only after our existing models accuracy ratios have significantly improved
Monitor Bias How do you plan to monitor or mitigate unwanted bias in your model?	Biases introduced in the ML models can be monitored by looking at the failure data in which manual intervention was required. **Bias Mitigation Plan** 1. Consider developing separate ML models for any answer script sets which have high failure rate 2. Continually train existing models with data based on analysis of answer scripts sets which have failures 3. Consider deploying uniform exam stationary across institutes to reduce noise related biases 4. Continually update the model with the latest open source OCR libraries to ensure even uncompatible handwritings are understood reducing bias introduced by handwriting

Appendix - Wireframes

1. **Login Screen**



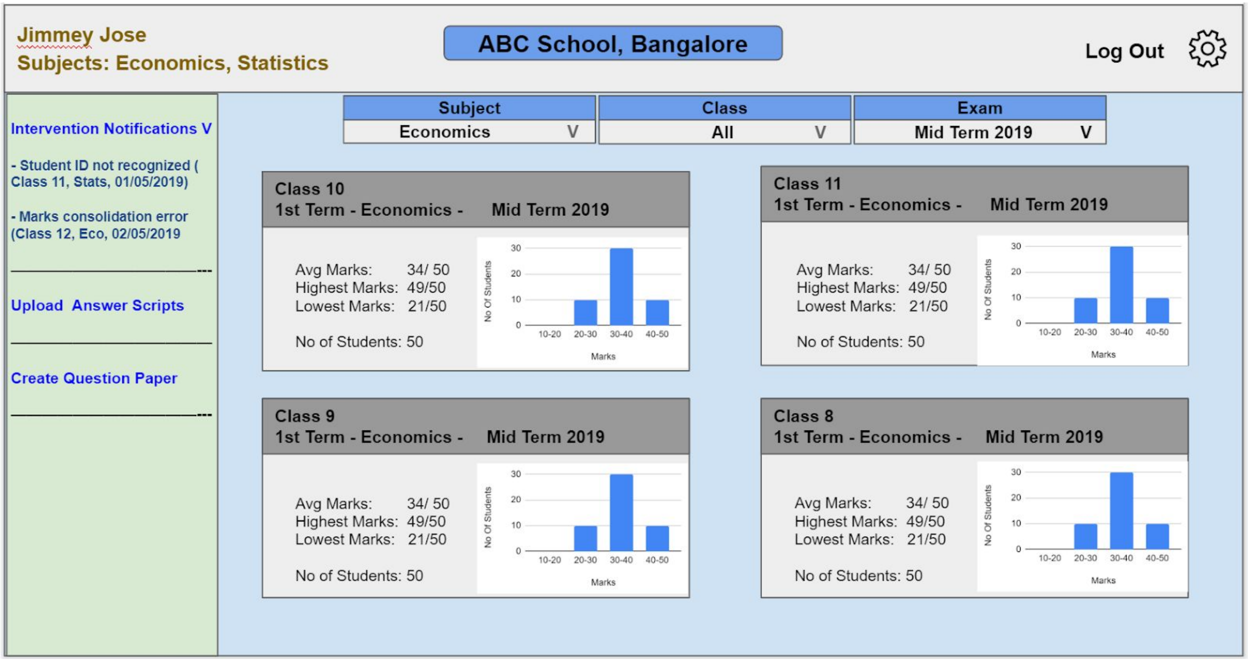
ABC School, Bangalore

Exam Results Portal

User Type ● Faculty ○ Student

A simple combined login screen for both teachers / valuators and students into the exams results portal webapp in which they can view the consolidated exam results info . The radio button option is not necessarily required; it was added to easily depict that both teachers / valuators and students can login.

2. Teacher/ Valuator Dashboard



Jimmey Jose
Subjects: Economics, Statistics

ABC School, Bangalore

Log Out

Intervention Notifications V

- Student ID not recognized (Class 11, Stats, 01/05/2019)

- Marks consolidation error (Class 12, Eco, 02/05/2019

Upload Answer Scripts

Create Question Paper

Subject: Economics V

Class: All V

Exam: Mid Term 2019 V

Class 10
1st Term - Economics - Mid Term 2019

Avg Marks: 34/ 50
Highest Marks: 49/50
Lowest Marks: 21/50

No of Students: 50

Class 11
1st Term - Economics - Mid Term 2019

Avg Marks: 34/ 50
Highest Marks: 49/50
Lowest Marks: 21/50

No of Students: 50

Class 9
1st Term - Economics - Mid Term 2019

Avg Marks: 34/ 50
Highest Marks: 49/50
Lowest Marks: 21/50

No of Students: 50

Class 8
1st Term - Economics - Mid Term 2019

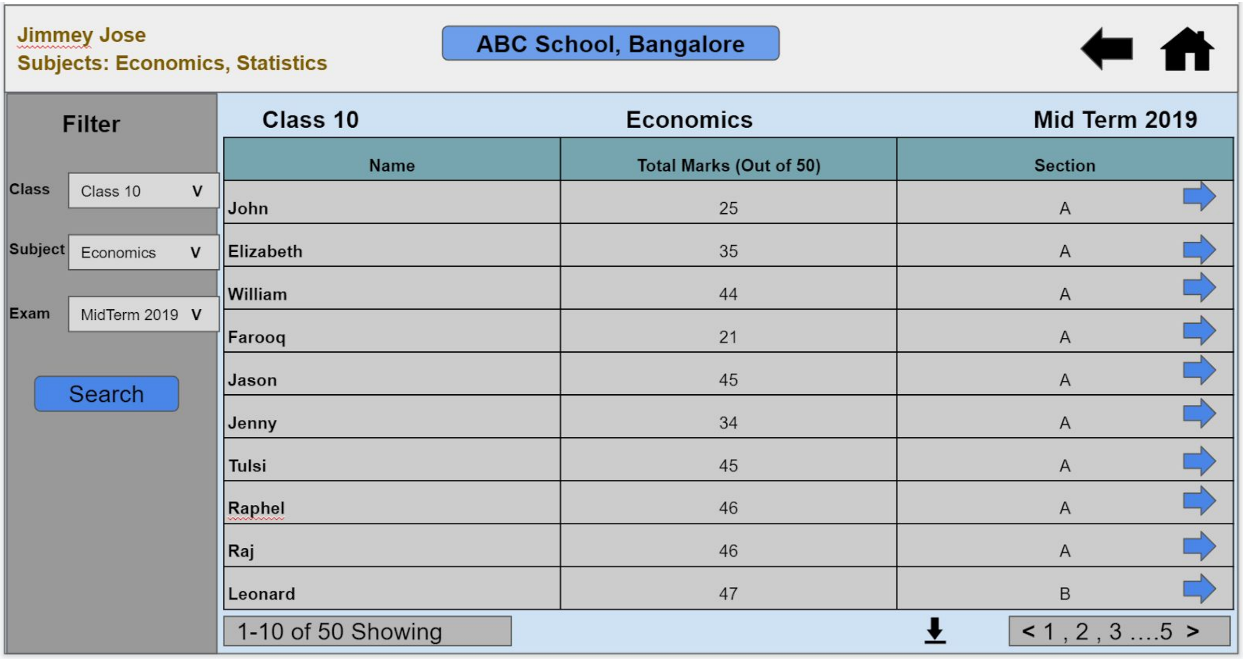
Avg Marks: 34/ 50
Highest Marks: 49/50
Lowest Marks: 21/50

No of Students: 50

The teachers / valuators login will login the user to the above the dashboard in which they will be able to get a birds eye view of all the answer scripts they had evaluated. On login the user will be able to view and do the following:

1. View snapshots about the answer scripts that were uploaded; by default they should show the latest ones. Each snapshot will show the average, highest and lowest marks along with a bar graph showing the marks vs no of students distribution
2. Filter and view snapshots by subject, class and exam time
3. View any intervention notifications which can be clicked on and navigated to make the required changes
4. Upload any new valuated answer script sets
5. The user name and the info will be visible on the left top corner

3. Snapshot Detail View



Jimmey Jose
Subjects: Economics, Statistics

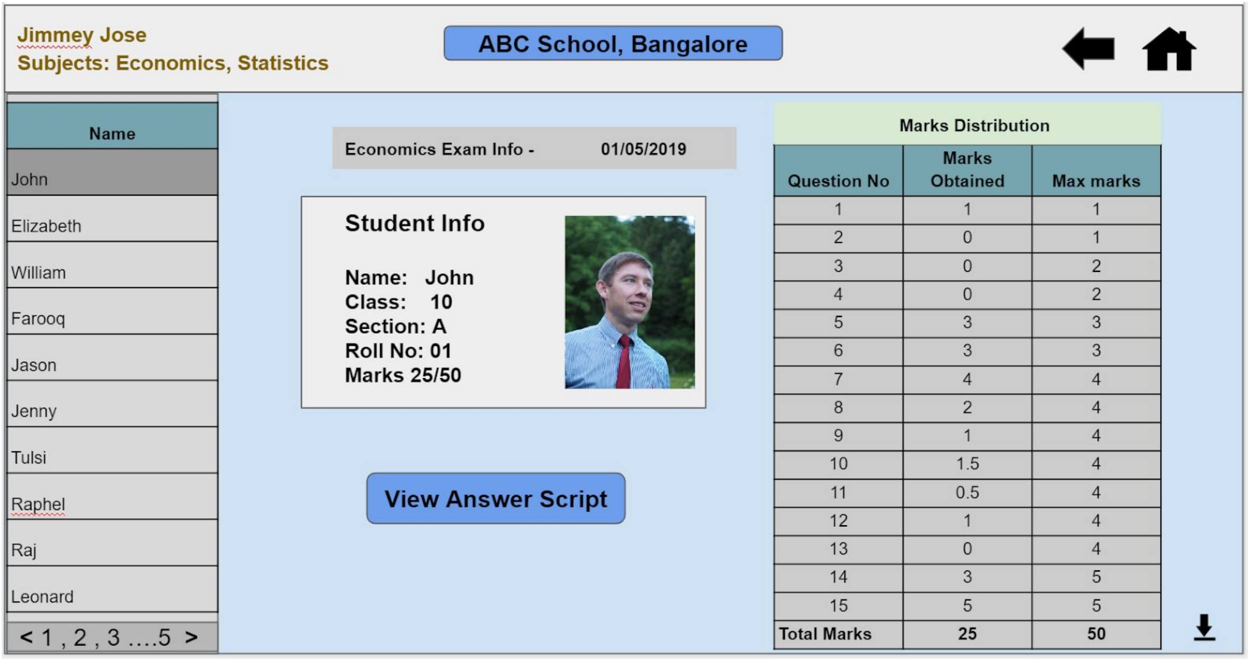
Class: Class 10
Subject: Economics
Exam: MidTerm 2019

Class 10 Economics Mid Term 2019

Name	Total Marks (Out of 50)	Section	
John	25	A	➡
Elizabeth	35	A	➡
William	44	A	➡
Farooq	21	A	➡
Jason	45	A	➡
Jenny	34	A	➡
Tulsi	45	A	➡
Raphel	46	A	➡
Raj	46	A	➡
Leonard	47	B	➡

The teachers / valuators can navigate to the detailed data set related to the snapshot by clicking on the snapshot. This detail view will show the summary of the answer script data set which includes the student identification information (name here) and total marks obtained.
The teachers / valuators can download this data set as an excel file and easily navigate to other snapshot detail views using the filter on the side.

4. Student Detail View For Teachers



Jimmey Jose
Subjects: Economics, Statistics

Name
John
Elizabeth
William
Farooq
Jason
Jenny
Tulsi
Raphel
Raj
Leonard

Economics Exam Info - 01/05/2019

Student Info

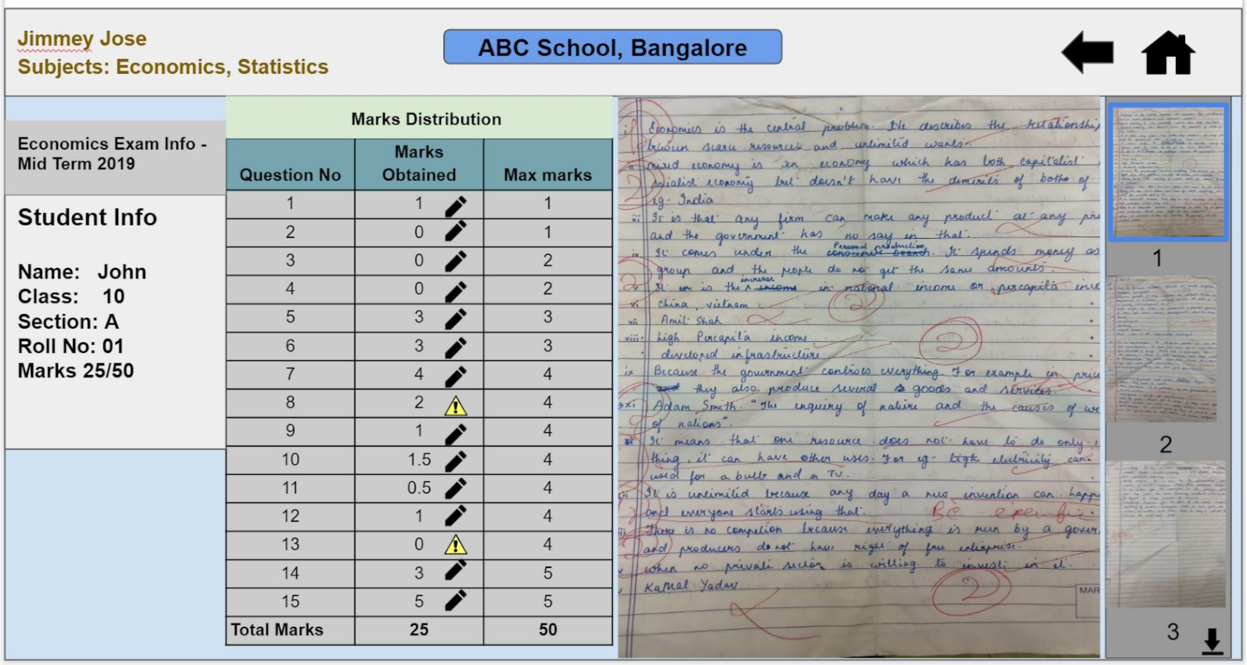
Name: John
Class: 10
Section: A
Roll No: 01
Marks 25/50

Marks Distribution

Question No	Marks Obtained	Max marks
1	1	1
2	0	1
3	0	2
4	0	2
5	3	3
6	3	3
7	4	4
8	2	4
9	1	4
10	1.5	4
11	0.5	4
12	1	4
13	0	4
14	3	5
15	5	5
Total Marks	25	50

The teachers / valuators will be able to view the specific students detailed marks card in this screen with the breakup of marks scored for each question. The teachers / valuators can navigate to any of the other students data in the same view and also navigate to view the detailed answer script

5. Student Answer Script and Edit View For Teachers



Jimmey Jose
Subjects: Economics, Statistics

ABC School, Bangalore

Economics Exam Info - Mid Term 2019

Student Info

Name: John
Class: 10
Section: A
Roll No: 01
Marks 25/50

Marks Distribution

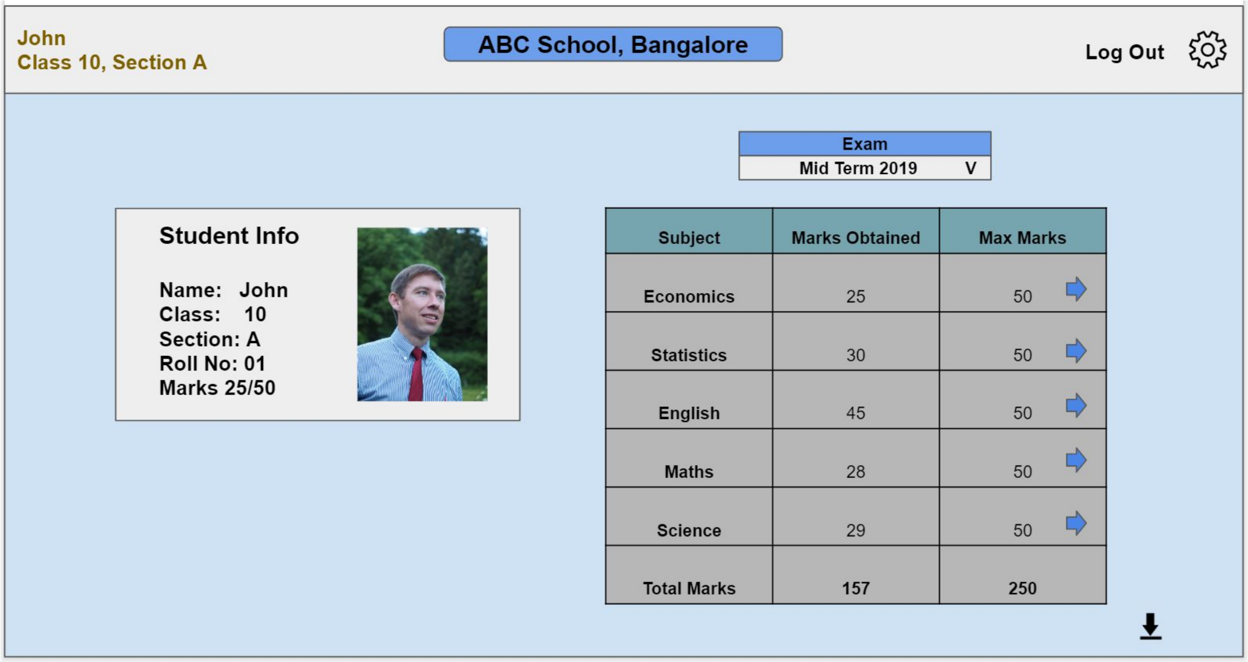
Question No	Marks Obtained	Max marks
1	1 ✎	1
2	0 ✎	1
3	0 ✎	2
4	0 ✎	2
5	3 ✎	3
6	3 ✎	3
7	4 ✎	4
8	2 ⚠	4
9	1 ✎	4
10	1.5 ✎	4
11	0.5 ✎	4
12	1 ✎	4
13	0 ⚠	4
14	3 ✎	5
15	5 ✎	5
Total Marks	25	50

i) Economics is the central problem. It describes the relationshi
between scarce resources and unlimited wants.
ii mixed economy is an economy which has both capitalist
socialist economy but doesn't have the demerits of both of
eg. India
iii It is that any firm can make any product at any pri
and the government has no say in that.
iv It comes under the ~~consumer branch~~ Personal production. It spends money as
group and the people do not get the same amounts.
v It ~~on~~ is the increase ~~income~~ in national income or percapita inc
vi China, vietnam
vii Amit Shah
viii High Percapita income
developed infrastructure
ix Because the government controls everything. For example in pric
~~and~~ they also produce several goods and services.
xi Adam Smith "The inquiry of nature and the causes of we
of nations".
It means that one resource does not have to do only
thing, it can have other uses. For eg. light electricity can
used for a bulb and a TV.
It is unlimited because any day a new invention can happ
and everyone starts using that. Be specific
There is no compition because everything is run by a gover
and producers do not have right of free enterprise.
when no private sector is willing to invest in it.
Kamal Yadav

1
2
3

The teachers / valuators will be able to view the scanned answer script here for any revalidation. The pencil icon can be clicked to make any necessary corrections to the marks. The yellow action sign signifies any action requirement either by the system or from students to verify or revise the marks; the icon can be clicked on to make the necessary changes. The answer script can also be downloaded if required.

6. Student Dashboard



John
Class 10, Section A

ABC School, Bangalore

Log Out

Exam
Mid Term 2019 V

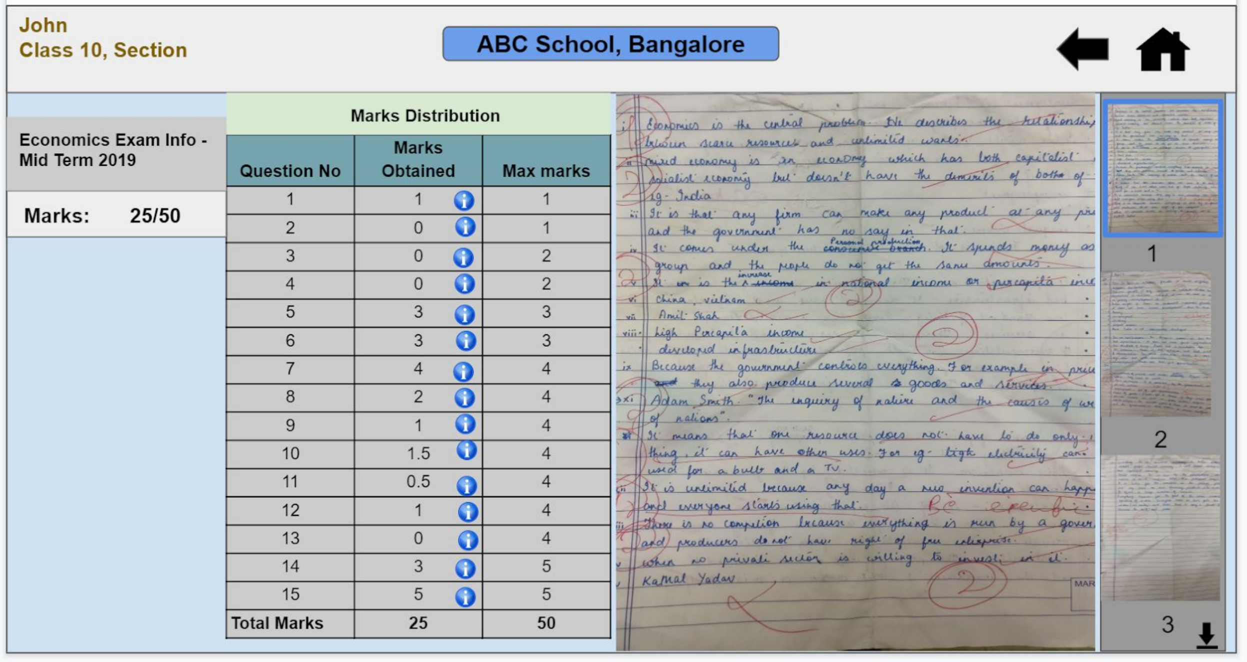
Student Info

Name: John
Class: 10
Section: A
Roll No: 01
Marks 25/50

Subject	Marks Obtained	Max Marks
Economics	25	50 ➡
Statistics	30	50 ➡
English	45	50 ➡
Maths	28	50 ➡
Science	29	50 ➡
Total Marks	157	250

This is the student login dashboard in which they will be able to view a consolidated report card for all marks uploaded using the product. They can use the blue arrows to navigate the subject detail view with the answer scripts.

7. Student Subject Detail View



John
Class 10, Section

ABC School, Bangalore

Economics Exam Info - Mid Term 2019

Marks: 25/50

Marks Distribution

Question No	Marks Obtained	Max marks
1	1 ⓘ	1
2	0 ⓘ	1
3	0 ⓘ	2
4	0 ⓘ	2
5	3 ⓘ	3
6	3 ⓘ	3
7	4 ⓘ	4
8	2 ⓘ	4
9	1 ⓘ	4
10	1.5 ⓘ	4
11	0.5 ⓘ	4
12	1 ⓘ	4
13	0 ⓘ	4
14	3 ⓘ	5
15	5 ⓘ	5
Total Marks	25	50

i) Economics is the central problem. It describes the relationship
between scarce resources and unlimited wants.
ii mixed economy is an economy which has both capitalist
socialist economy but doesn't have the demerits of both of
eg- India
iii It is that any firm can make any product at any pri
and the government has no say in that.
iv It comes under the ~~consumer branch~~ Personal production. It spends money as
group and the people do not get the same amounts.
v It ~~in~~ is the increase ~~income~~ in national income or percapita inc
vi China, vietnam
vii Amit Shah
viii High Percapita income
developed infrastructure
ix Because the government controls everything. For example in pric
~~and~~ they also produce several goods and services.
xi Adam Smith "The enquiry of nature and the causes of we
of nations".
It means that one resource does not have to do only 1
thing, it can have other uses. For eg- light electricity can
used for a bulb and a TV.
It is unlimited because any day a new invention can happ
and everyone starts using that. Be specific
iii There is no competion because everything is run by a gover
and producers do not have right of free enterprise.
v when no private sector is willing to invest in it.
Kamal Yadav

The student will be able to view the detailed marks distribution along with the evaluated answer scripts in this screen. They can raise any concerns with respective valuator using the blue action button specific to any of the questions. The student can also download the answer script if required.