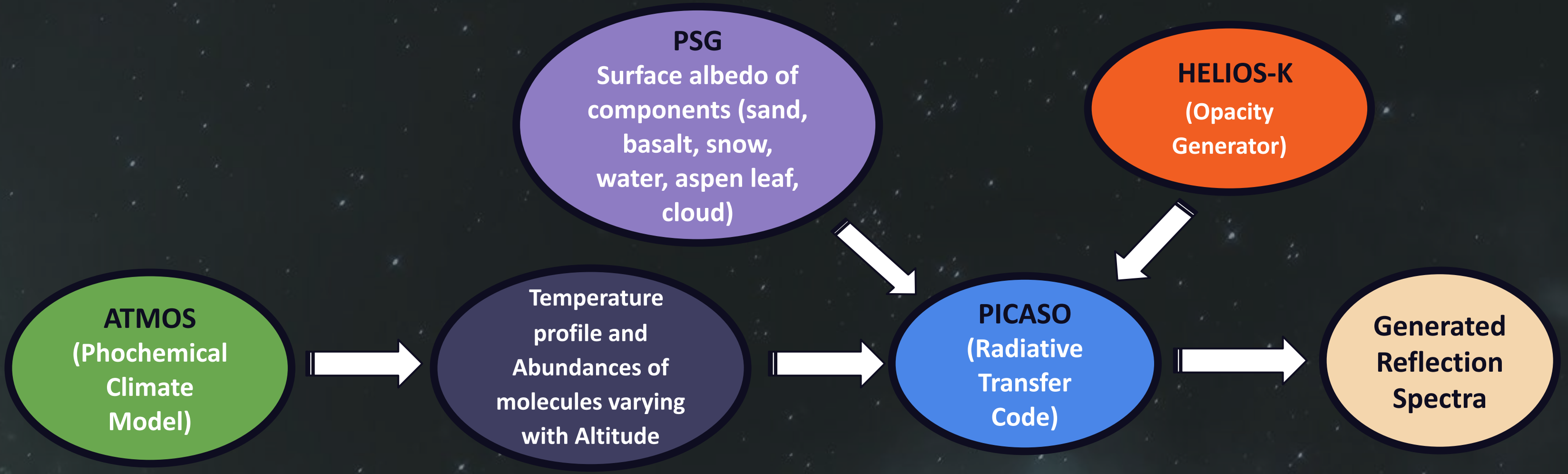
Exoplanetary Surface Composition Prediction using ML

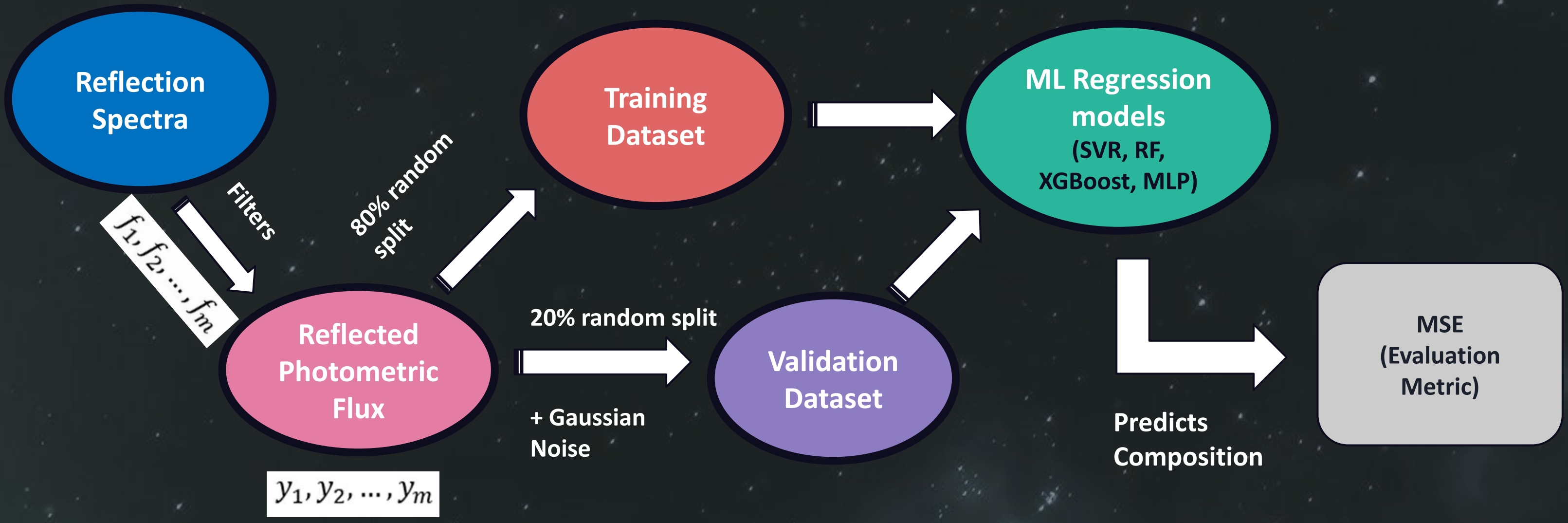
Goals Achieved

- Applying Neural Network (MLP)
- Calculating the Opacities from HELIOS-K
- Applying XGBoost Regression
- Augmentation of Noise in 20% dataset for Validation set
- Performance of the ML models with Noisy data
- Response of ML models on the dataset with varying hyperparameters

Data Generation



ML Implementation



Atmos: A coupled climate-photochemical model

Photochemical Model

- Generates an initial atmospheric state
- **User-specified boundary conditions:**

  gas mixing ratios or fluxes and deposition velocities, the stellar spectrum, the total atmospheric pressure, the initial temperature-pressure profile

- 233 chemical reactions and includes 50 chemical species, 9 of which are short-lived
- Output:

  altitude-dependent abundances of H_2O photochemically produced in, or transported to, the stratosphere, CO_2, O_3, CH_4, O_2, N_2, and C_2H_6.

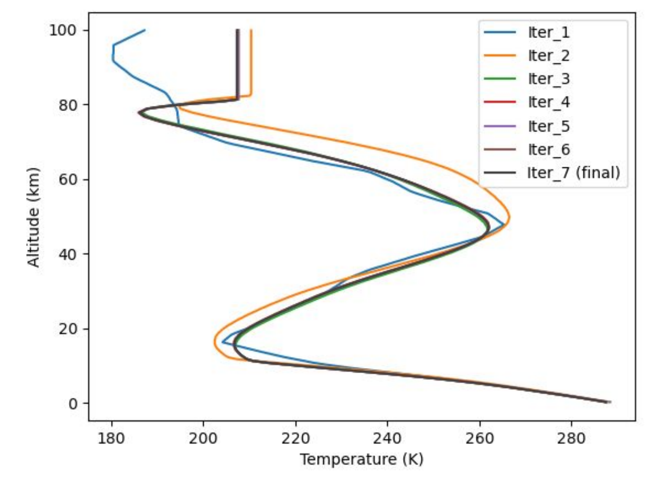
Climate Model

- The tropospheric temp. calculated by following a wet adiabatic lapse rate to the altitude at which the stratospheric temperature is reached
- **Input:**

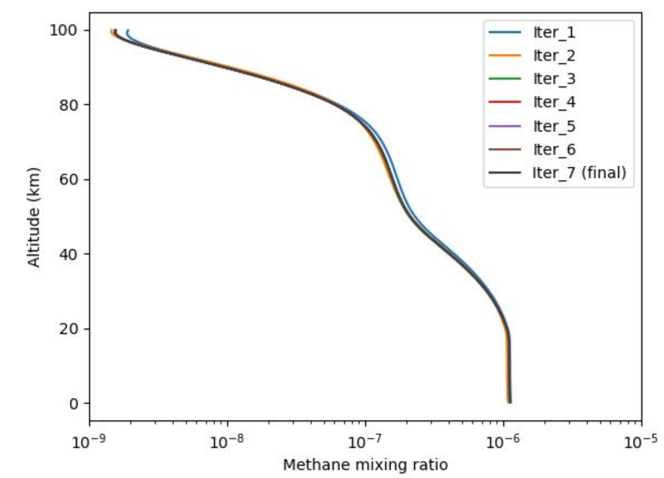
  the number of steps to run the model, pressure at the surface, pressure at the top of the atmosphere, surface temperature, surface albedo, solar constant and surface gravity.

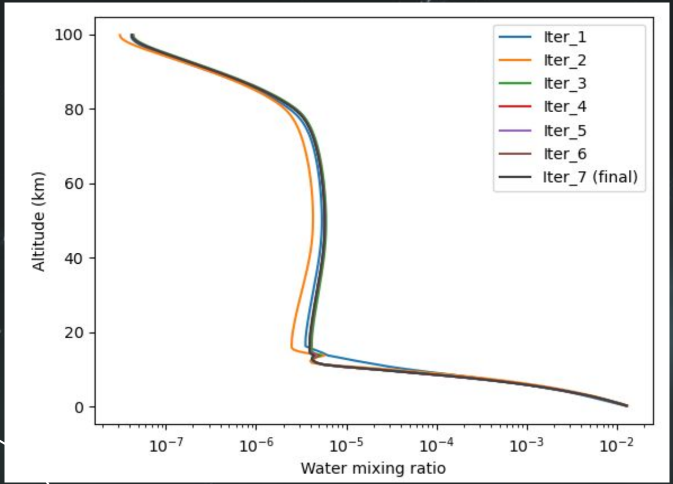
- **Output:**

  altitude, temperature, water mixing ratio



Plots showing Atmos finding Convergence for Modern Earth conditions





PICASO : A Radiative Transfer Model

- PICASO is an atmospheric radiative transfer model to produce the reflection spectra. The original documentation was used as reference for using PICASO in our codes.
- PICASO can be used for obtaining transmission, emission and reflection spectra.
- Using PICASO for obtaining the reflection spectra of exoplanets with a certain wavelength dependent albedo function for its surface components.
- The main equation used in PICASO is the radiative transfer equation given below,

PICASO takes the following inputs:

- Basic planetary properties (planet mass, radius, stellar spectra)
- PT profile and abundances (Obtained from ATMOS)
- Cloud profile (angle-scattering albedo, asymmetry and total extinction) (not used in our case).
  Instead albedo of the cloud is obtained from a model and is used along with other albedo functions for finding effective albedo.
- Surface albedo: can be average surface albedo or wavelength dependent surface albedo (in our case).

$$\boxed{I\left(\tau_i, \mu\right) = I\left(\tau_{i+1}, \mu\right) e^{\delta\tau_i/\mu} - \int_0^{\delta\tau_i} S\left(\tau', \mu\right) e^{-\tau'/\mu} d\tau'/\mu}$$

Data- Set Description

- The data set provided in the paper which we will be trying to reproduce with PICASO and ATMOS consists of reflection spectra for various surface combinations.
- Filters are applied on this spectra to obtain the photometric flux. In our case we took nine photometric flux values which was found to provide us with a good accuracy rate.

Surface combinations

	cloud	snow	sand	seawater	basalt	veg
0	0.00	0.00	0.00	0.00	0.00	1.00
1	0.00	0.00	0.00	0.00	0.05	0.95
2	0.00	0.00	0.00	0.00	0.10	0.90
3	0.00	0.00	0.00	0.00	0.15	0.85
4	0.00	0.00	0.00	0.00	0.20	0.80
...	...	...	...	...	...	...
53125	0.95	0.00	0.00	0.00	0.05	0.00
53126	0.95	0.00	0.00	0.05	0.00	0.00
53127	0.95	0.00	0.05	0.00	0.00	0.00
53128	0.95	0.05	0.00	0.00	0.00	0.00
53129	1.00	0.00	0.00	0.00	0.00	0.00

- Six surfaces were considered which are Snow, sand, basalt, cloud, vegetation and sea water
- Permutations with 5 percent steps for these six surfaces lead to a total of 53,130 different surfaces.
- The data is then divided into training and validation data with an 80-20 ratio.

Flux values for the surface combinations

	f1	f2	f3	f4	f5	f6	f7	f8	f9
0	66.450295	70.559679	60.551194	34.801215	13.230524	8.823439	6.337946	0.350842	1.599316
1	67.564830	69.209650	58.643643	33.787692	12.898762	8.762820	6.250520	0.381013	1.635478
2	68.679366	67.859620	56.736091	32.774169	12.567000	8.702201	6.163094	0.411184	1.671641
3	69.793902	66.509591	54.828539	31.760645	12.235237	8.641582	6.075669	0.441354	1.707803
4	70.908438	65.159561	52.920988	30.747122	11.903475	8.580963	5.988243	0.471525	1.743966
...	...	...	...	...	...	...	...	...	...
53125	234.117912	142.502061	90.005522	58.435347	33.283890	21.958330	16.991558	4.328158	5.985801
53126	232.629441	141.027892	89.131071	57.823231	33.000671	21.622128	16.787464	4.287007	5.882511
53127	238.863237	146.362171	92.777791	60.235069	34.058652	22.909336	17.573305	4.480611	6.353865
53128	245.008517	149.283036	93.169635	59.558788	33.361182	21.608670	16.806243	4.282940	5.877620
53129	241.769328	147.709586	93.563699	60.746116	34.688553	22.713449	17.644302	4.505732	6.178602

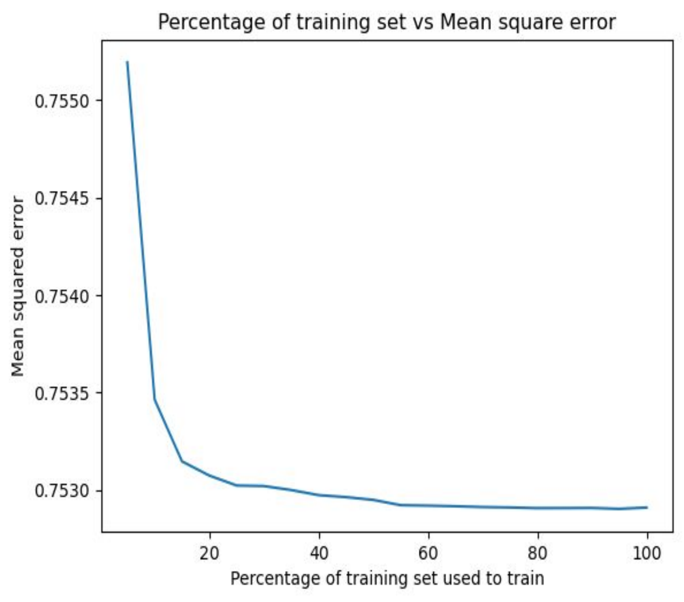
Implementing ML Algorithms

Four ml algorithms with different hyperparameters were implemented and their accuracies were compared for various S/N ratios.
The algorithms are

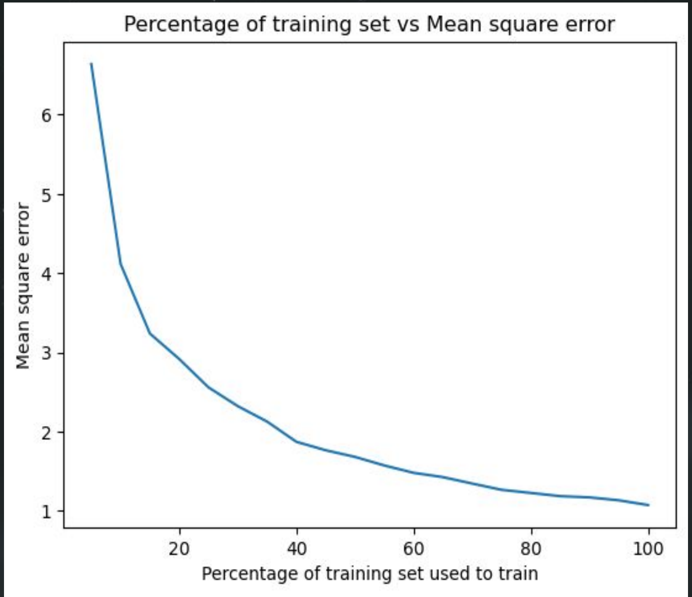
- XGBoost: performed better than SVR for signal to noise ratio<35
- MLP: performed poor compared to other algorithm probably due to the small training dataset
- SVR: performed very well when size of the training dataset was very small
- RFR: performed very well and was consistent for even very low S/N ratio
- The hyperparameters for these algorithms were also optimized aiming for a good balance between accuracy and run-time

Results

MSE for different training dataset sizes at S/N = 70

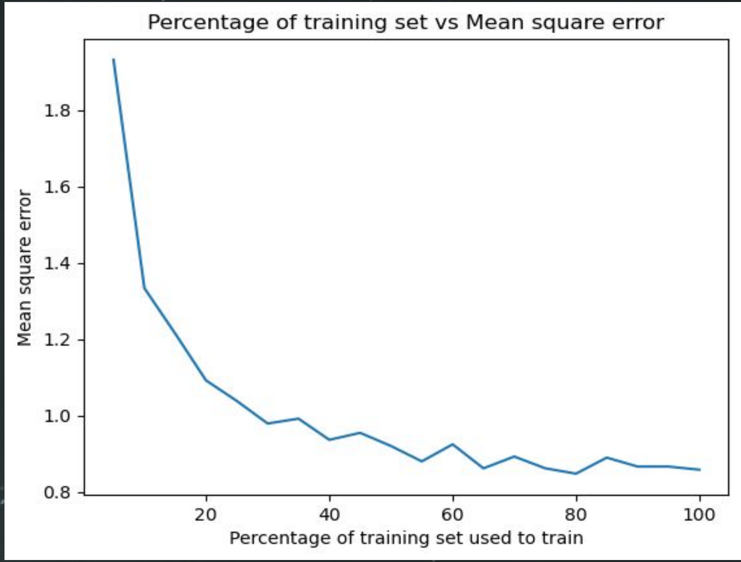


SVR

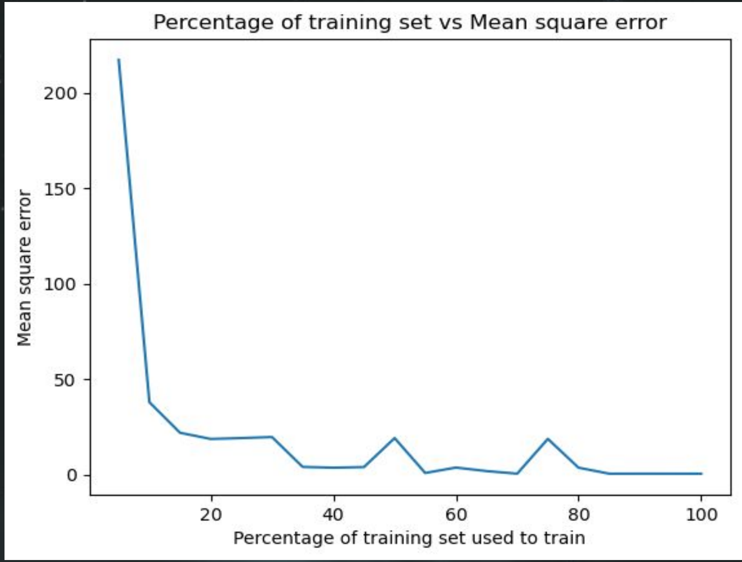


RFR

MSE for different training dataset sizes at S/N = 70

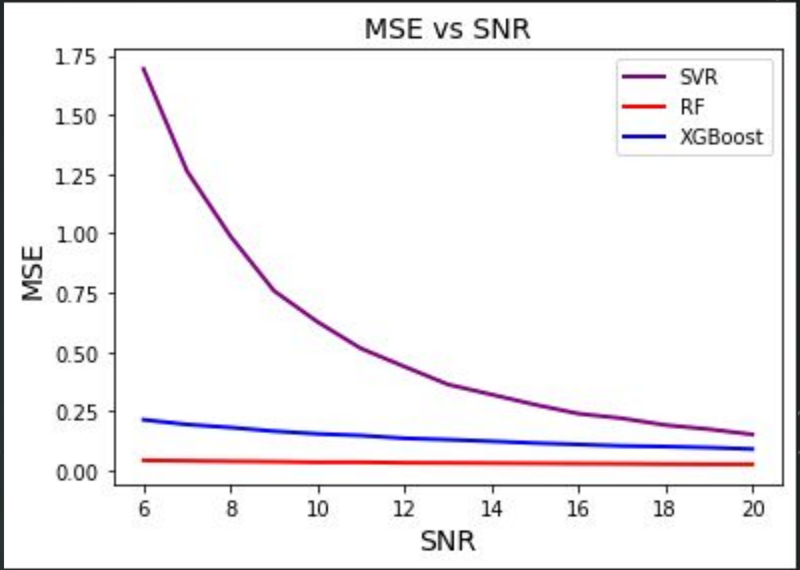


XGBoost

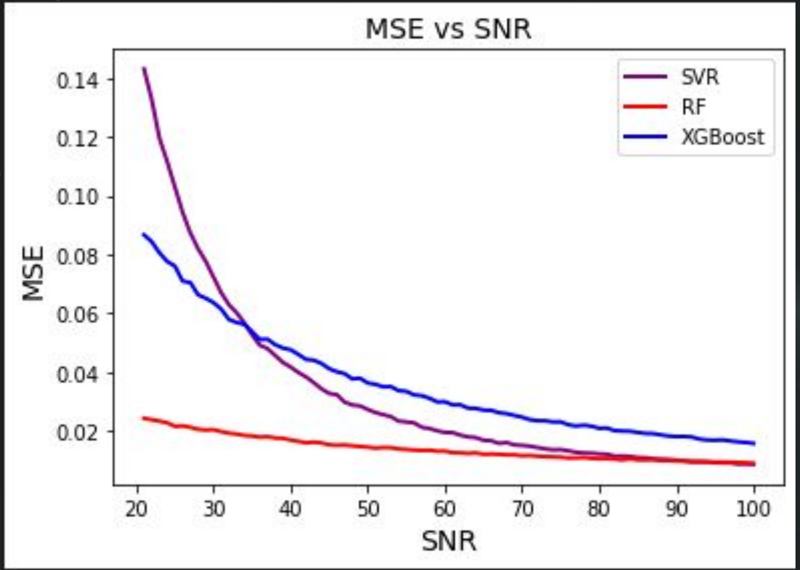


MLP

MSE of SVR, RF and XGBoost for varying S/N

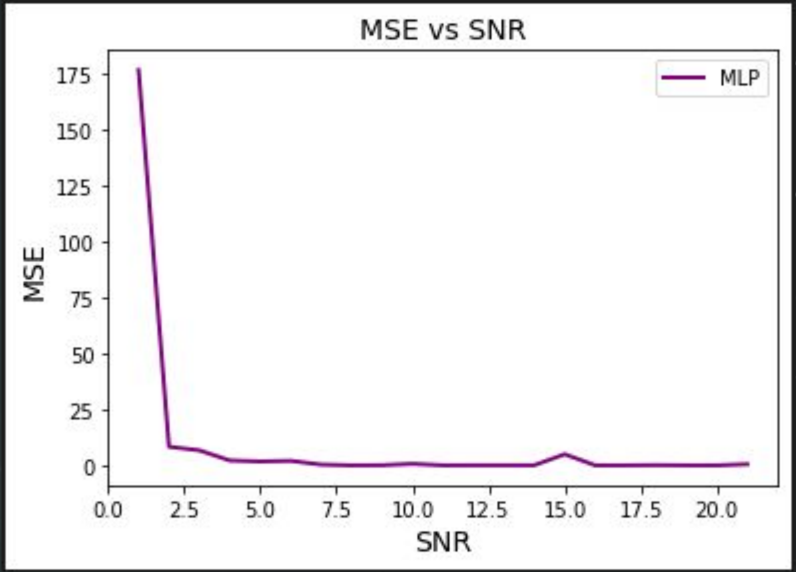


For S/N between 0 and 20

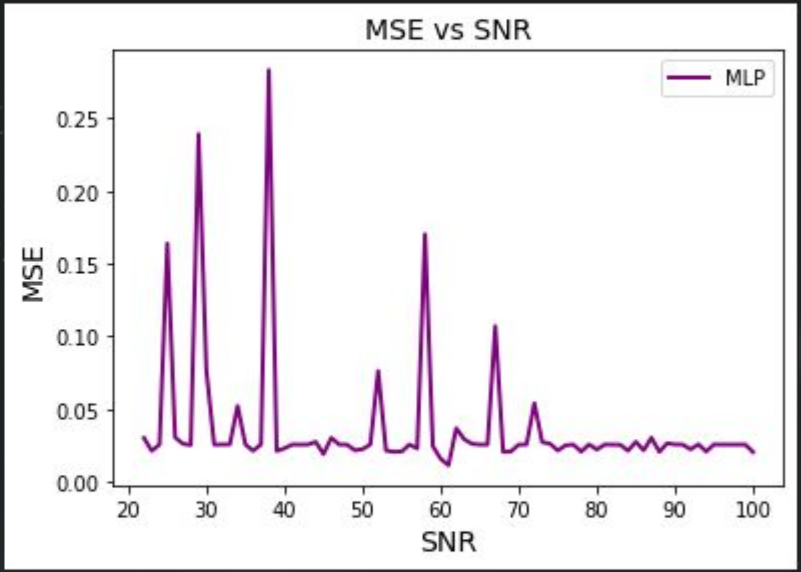


For S/N between 20 and 100

MSE of MLP for varying S/N

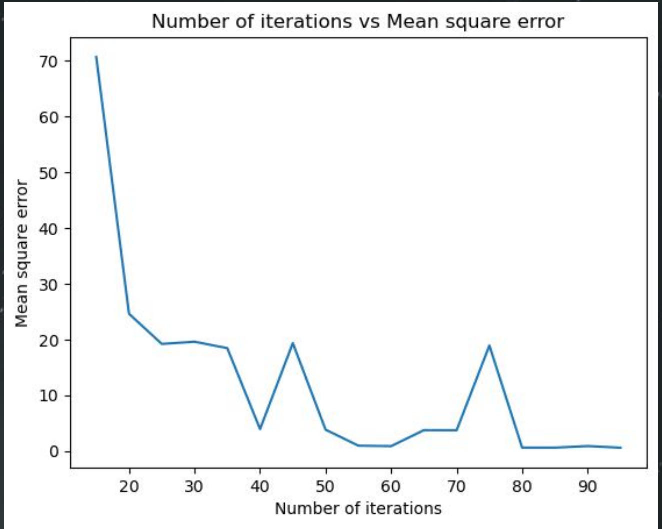


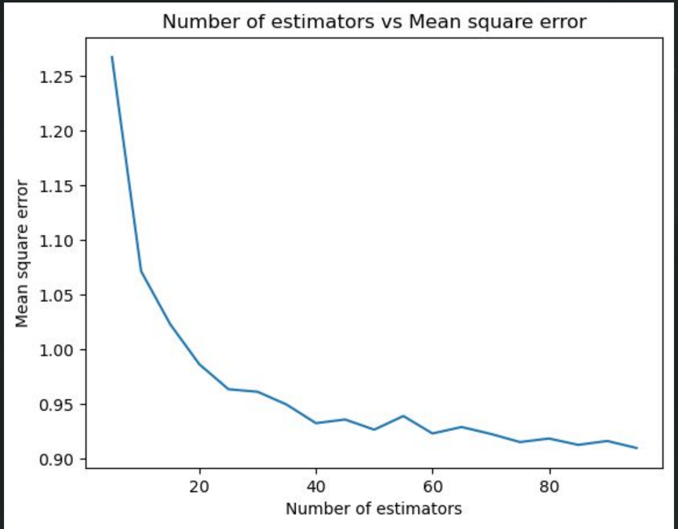
For S/N between 0 and 20

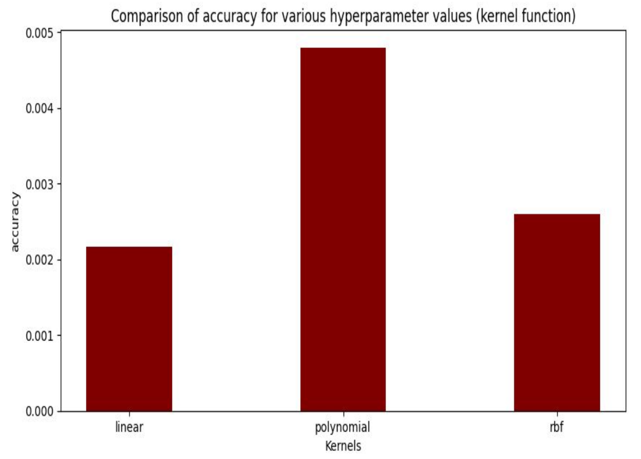


For S/N between 20 and 100

MSE for varying hyperparameters







Conclusions

- It was shown that ML algorithms can be used to predict with reasonable accuracy the surface composition of exoplanets
- The accuracies of the ML algorithms were compared and the hyperparameters were optimized for each algorithm
- MLP did not perform as good as the other algorithms due to the small dataset that was considered.
- RFR performs best consistently for reasonably large datasets with respectable accuracies even for S/N<20.
- The accuracy of the SVR using the linear kernel is found to perform very well compared to all the other algorithms which was considered when training dataset is small with reasonable amount of noise(S/N=70).

Future Plans

- Completing the opacity generation for modern earth atmospheric molecules using HELIOS-K
- Generation of Spectra using PICASO after opacities are fed into it
- Normalization of the generated data set
- Addition of the data set into the available set to further train models
- Repetition of the techniques and algorithms applied on the available data set for this new data set
- Improve the accuracy of our method through further refinement and slight tuning of the hyper-parameters as concluded earlier