UNIVERSITY OF TARTU
Institute of Computer Science
Computer Science Curriculum

Alina Vorontseva

Using Machine Learning to Find New Members of the Pleiades

Master's Thesis (30 ECTS)

Supervisor: Raul Vicente Zafra, PhD
Supervisor: João Alves, PhD

Tartu 2019

Using Machine Learning to Find New Members of the Pleiades

Abstract:
An open star cluster is a group of gravitationally bound stars that move together through space and have the same origin. One of the most famous star clusters, which can be seen with the unaided eye, is Pleiades. In order to accurately model clusters' formation and evolution, scientists need to know exact cluster members to include them into analysis. Unlike the stars that are close to the cluster core, it is hard to relate the stars that are farther from the cluster center and can therefore be confused for field stars.

In this thesis, we find Pleiades member candidates among stars with unknown membership using Machine Learning. Here we show that spectral data alone is not enough for clear membership determination, although combined with stars' positions and velocities, it produces valid results.

Our 22 suggested Pleiades member candidates have positions, velocities, abundances and atmospheric parameters similar to the Pleiades stars. Features with more predictive power are positions and abundances Fe/H, M/H and C/Fe.

The model relying on spectral features has been able to find a lot more stars with chemical composition similar to the Pleiades. The fact that some of these predicted stars are too far away (more than 20 pc from the cluster center) proves that spectral information alone is not discriminative enough to isolate the members of one particular cluster. Still, it is very useful to separate Pleiades member candidates from field stars, since the precision of the model on the test dataset is 0.957. Features that are more important for the prediction are N/Fe, C/Fe and T_{eff}.

The results obtained in this thesis will be very useful for large future sky surveys. Having many stars as possible cluster members, our model will help to carefully reduce their number for a detailed membership study.

Keywords:
Pleiades, star cluster, Machine Learning, Gaia, APOGEE

CERCS:
P520 (Astronomy, space research, cosmic chemistry), P170 (Computer science, numerical analysis, systems, control)

Masinaõppe kasutamine Plejaadide uute liikmete leidmiseks

Lühikokkuvõte:
Avatud täheparv on gravitatsiooniliselt seotud tähtede rühm, mis liiguvad läbi ruumi ja millel on sama päritolu. Üks kuulsamaid täheparve, mida võib palja silmaga näha, on Plejaadid. Klastrite moodustumise ja arengu täpseks modelleerimiseks peavad teadlased täpselt teadma klastri liikmeid, et neid analüüsi kaasata. Erinevalt klastri südamikule lähedastest tähtedest, klastri keskpunktist kaugemal asuvaid tähti on raske kategoriseerida ja neid tihti valesti loetakse iseseisvateks tähedeks.

Selles lõputöös masinõppe abil otsitakse Plejaadide potentsiaalseid liikmeid tundmatu tähtede hulgast. Näitame et spektriandmetest ei piisa liikmesuse selgeks määramiseks, kuigi koos tähtede asukoha ja kiirusega on võimalik saada õigeid tulemusi.

Meie poolt 22 soovitatud Plejaadi liikmekandidaadi positsioonid, kiirused ja atmosfääri parameetrid on sarnased Plejaadide tähtedega. Suurimate ennustatavama võimsusega omadused on positsioonid ja Fe/H, M/H ja C/Fe suhed.

Spektraalomadustel baseeruv mudel on suutnud leida palju rohkem tähti, mille keemiline koostis on sarnane plejaadidega. See fakt et mõned ennustatud tähed on kauged klastri keskpunktist (rohkem kui 20 pc), tõestab, et alinut spektraalomadused ei ole piisavalt diskrimineerivad, et eraldada ühe konkreetse klastri liikmed. Siiski sellel mudelil oli piisav täpsus katseandmetes: 0.957. Ennustamise jaoks olulisemad omadused on N/Fe, C/Fe ja T_{eff}.

Lõputöö raames saadud tulemused on kasulikud suurte tulevaste taevavaatluste jaoks. See peaks märkimisväärselt vähendama klastrikandidaatide arvu ja muutma üksikasjalikud liikmelisuse uuringud paremini hallatavaks.

Võtmesõnad:
Plejaadid, täheparv, masinõpe, Gaia, APOGEE

CERCS:
P520 (Astronoomia, kosmoseuuringud, kosmosekeemia), P170 (Arvutiteadus, arvutusmeetodid, süsteemid, juhtimine (automaatjuhtimisteooria))

Contents

1 Introduction

Nowadays, it is hard to imagine any scientific discovery that is made without collecting massive amounts of data to support its conclusions. Astronomy is no exception; moreover, it is one of the scientific disciplines with the **biggest datasets**. To exemplify, a recent image of a black hole was produced by processing more than 5 petabytes of data (Goddi et al. 2019). Many other surveys daily generate massive amounts of data, e.g. *Large Synoptic Survey Telescope* (LSST), that produces 20TB of data per night (Ivezić et al. 2019).

All this data needs to be analyzed to reveal the secrets of the universe creation and evolution, characteristics of underlying physical processes and current events in the universe. The times when astronomers reviewed each photographic plate manually are long gone. Instead, scientists are using modern methods and processing techniques to extract information from massive amounts of data in a fast, efficient and accurate way. However, sometimes the usual methods do not allow to dig deeper into the hidden dependencies and correlations in the complex astronomical data. One set of methods that can help to shed light into more complex relationships in the data is **machine learning**. It has now penetrated almost every sphere of human existence: from medical applications to user-targeted advertisements.

Small wonder astronomers also have started to use machine learning in their research. A few famous examples include a citizen science project of the galaxy morphology classification *Galaxy Zoo* (Banerji et al. 2010) and a photometric redshift estimation project (Carliles et al. 2010). One of the tasks that is explicitly well-suited for machine learning is the study of open star clusters.

Open **star clusters** are groups of stars that were born together from the same molecular cloud and are gravitationally bound to each other. Stars in a cluster have roughly the same age and chemical composition and they move together in space. Understandably, scientists are interested in obtaining the most precise cluster parameters (e.g. age, distance, chemical composition), because the latter are necessary to model the formation and evolution of the cluster. In order to obtain the most accurate results, one needs to know **exact cluster members** to include them into analysis. Cluster members must be selected with high confidence, but also with an attempt to maintain the biggest possible sample. This guarantees high-quality results and fewer introduced biases. This said, there is always a problem of defining exact members of any star cluster. Unlike the stars that are close to the cluster core, it is hard to relate the stars that are farther from the cluster center and can therefore be confused for field stars.

Incidentally, one of the most famous star clusters is **Pleiades**, which can be seen with the unaided eye on a dark sky. This star cluster is a relatively close, young open cluster (Bouy et al. 2015). It has been studied extensively by many sky surveys, including **Gaia** (Gaia Collaboration et al. 2018b) and **APOGEE** (S. R. Majewski et al. 2017). The surveys in question provide important high-quality data, that is openly available.

Numerical simulations in Converse et al. 2010 and the study of Lodieu et al. 2019 show that nowadays Pleiades are starting to form the tidal tails, i.e. elongated stellar structures of stars, formed as a result of gravitational forces. Stars from these tails are more difficult to determine with the usual methods that take into account densities of stars in space, because stars in tidal tails can appear farther away from the cluster center and have a different space velocity. Still, as they were formed together with all other stars in the cluster, they should have the same age and chemical composition. These properties can be subsequently utilized to relate stars from the cluster to their origin and distinguish them from field stars (stars that do not belong to the cluster).

So far, many authors (Sarro et al. 2014; Bouy et al. 2015; Kos et al. 2018; Gaia Collaboration et al. 2018a; Lodieu et al. 2019) have already suggested their lists of candidate members of the Pleiades cluster. In previous studies, probabilistic models were often applied to the space velocities and positions data, rarely accounting for a limited subset of abundances and photometric features or taking the data from multiple sky surveys. With our project, in contrast, **new members of the Pleiades star cluster will be found by means of machine learning**. Thus, our thesis will make use of high-quality astrometry data from Gaia DR2 and comprehensive abundance data from APOGEE DR15 in order to train a supervised machine learning model to predict whether a star belongs to the Pleiades star cluster. This model can then be used on the new data, in the pool of which the model predictions will reveal new Pleiades stars.

2 Background

In our thesis, most of the background information related to astronomical definitions, principles and theory, is based on Karttunen et al. 2017.

2.1 Astronomical measurements

There are three techniques in modern astronomy used to characterize an object: photometry, astrometry and spectroscopy. **Astrometry** is a fundamental branch in astronomy that focuses on the accurate measurement of objects' positions in the sky and in space, as well as their velocities. The object's observable position in the sky plane is defined by two coordinates in equatorial coordinate system: *right ascension* α and *declination* δ. For studies of objects in our galaxy, more natural reference plane is the plane of the Milky Way with the Sun as its origin. The galactic longitude l is measured counterclockwise from the direction of the center of the Milky Way, and galactic latitude b is measured from the galactic plane, positive northwards and negative southwards. Galactic coordinates are spherical and could be transformed into Cartesian (with X, Y, Z axes) with a simple transformation.

Stars that appear close in the sky are not necessarily close to each other in space, as they may have different distances. A measurement unit of distance in astronomy is parsec (pc), which is defined as a distance from which the separation between the Earth and the Sun would be observed at one arc second. $1\,pc$ is about 3.26 light years or 3×10^{16} meters. The distance to an object can be measured using *parallaxes* p. A parallax is an angle subtended by the radius of the Earth's orbit as seen from the star; it can be measured as the object's movement in relation to more distant non-moving stars. The relation between parallax p in arc seconds and distance r in parsecs is simply

$$r = \frac{1}{p}.$$

Apart from the parallax motion, sky objects also have *proper motions* μ, i.e. a slow movement relative to distant background stars with a constant direction, caused by the relative motion of the Sun and the stars through space. These proper motions, in turn, have two components: change in right ascension $\mu_\alpha \cos\delta$ and in declination μ_δ; total proper motion is then

$$\mu = \sqrt{\mu_\alpha^2 \cos^2\delta + \mu_\delta^2}.$$

This is a tangential component of a star velocity with respect to the Sun. The other component that is directed along the line of sight is called *radial velocity*. It shows the speed with which the object is moving from or towards us. Proper motions are measured by comparing observations from long time intervals (years or decades) that allow us to

notice the change of objects' position. Radial velocities are measured using Doppler shift of stellar spectrum.

Photometry is a study of objects' light intensity. It is often performed by measuring objects' brightness in multiple *passbands* (filters) which are sensitive only to specific wavelengths of light. The combination of these measurements can reveal some physical properties of an object, like color, luminosity, and even chemical composition and temperature. Color can be obtained as a difference between an object's brightness in different passbands. A measurement unit of an object's brightness in astronomy is *apparent magnitude* m; $1\,m$ equals $\sqrt[5]{100}$ difference in brightness ratios. Obviously the lower the value is, the brighter the star is. To illustrate, the Sun has an apparent magnitude of $-26.8\,m$, whereas the full Moon has $-12.5\,m$. With the naked eye on a dark sky people are able to see stars up to $6\,m$, while modern telescopes are capable of capturing stars over $30\,m$. Apparent magnitude depends on the distance to the object, its intrinsic brightness and the amount of dust on the line of sight. *Absolute magnitude* M is defined as the apparent magnitude at a distance of 10 parsecs from the star. Ignoring the dust, the equation which relates apparent magnitude m, the absolute magnitude M and the distance to object r is

$$m - M = 5 \lg \frac{r}{10\,pc}. \tag{1}$$

Spectroscopy focuses on measuring the spectrum of electromagnetic radiation. The strength of various absorption lines can reveal information about objects' temperature, mass and chemical composition. It is also used to determine distance and velocities. Spectra are obtained by means of a spectrograph, which disperses the light of an object, and a CCD-camera, which measures the light intensity for each wavelength to subsequently produce the intensity curve. The absorption lines appear as troughs of various sizes in the curve. The shape of the spectrum reflects the properties of the stellar atmosphere and its chemical composition.

However, computing the structure of the atmosphere is not a simple task because of the rotation, magnetic fields and inherent inability to measure some physical values directly (e.g. the radius of a star). Therefore, the global parameters that define the structure of the stellar atmosphere - chemical abundances, effective temperature T_{eff} and the gravitational acceleration at the surface g - are to be found through complex numerical simulations. *Effective temperature* is the temperature of a blackbody which radiates with the same total flux density as a star. For a range of parameter values, synthetic spectra are computed, upon which chemical abundances and other parameters can be found by comparing the observed line strengths and other spectral features with the theoretical ones.

Generally, stars are classified into groups according to their spectra. The current Harvard classification is mostly based on temperature, but for a more precise classification, one also has to take into account the luminosity of the star. This system is known as Yerkes classification and it distinguishes six luminosity classes. A luminosity class is

determined from spectral lines strongly dependent on the stellar surface gravity, which is closely related to luminosity. The diagram showing the relation of these two variables - luminosity or absolute magnitude and spectral type or effective temperature - is known as *Hertzsprung–Russell diagram* (HRD) and is one of the most important diagrams in stellar evolution studies. The schematic HRD is shown in Figure 1 below. In practice, the color or spectral classes can be used instead of effective temperature on the horizontal axis, as they are easier to measure. For the same reason, absolute magnitude, calculated with (1), is used for the vertical axis.

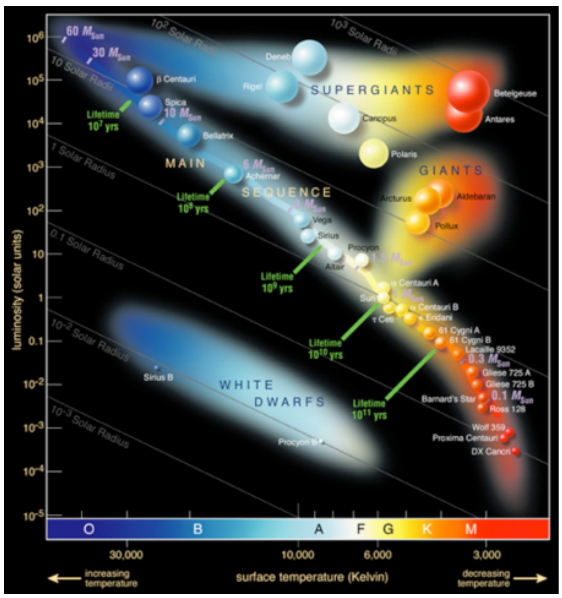

Figure 1. Schematic HRD diagram

HRD diagram shows luminosity and effective temperature relation. The densest regions on the diagram are main sequence, where stars remain for most of their evolutionary track, and white dwarfs, giants and supergiants groups, which are the end-stages of star evolution, dependent on their mass. Credits: ESO.

The distribution of stars in HRD is not uniform. Rather, there are several dense areas that represent various stellar evolution stages and where stars stay for a long time. Most stars are located at a roughly diagonal curve that is called the *main sequence* and spans from the bottom right to upper left corners of HRD. Main sequence stars are burning hydrogen in their cores and that takes the longest part of stars' lifetime. When a star is formed, its position on the main sequence depends on its mass. More massive hot stars are evolving faster and thus leave the main sequence a lot quicker than cold low-mass stars. Stars are slowly moving up and to the right along the main sequence as the amount of hydrogen in their cores decreases, until no hydrogen is left in the core. This is precisely the moment when the star leaves the main sequence and its further position on HRD depends on the mass. Typically, a white dwarf cluster is positioned about 10 magnitudes

below the main sequence and is populated with former low-mass stars (less than 0.6 Solar mass). The giant branch located above the main sequence can be divided into several branches that correspond to different phases of the stellar evolution: subgiant, horizontal, red giant, supergiant or asymptotic branch.

2.2 Star clusters

Open star clusters are groups of stars that were born together from the **same molecular cloud** and are still gravitationally bound to each other. Typically, an open cluster contains a few hundred stars, at times up to a few thousand. Stars in an open cluster **move together** through space with approximately the same velocity with respect to the Sun. As opposed to dense globular clusters, open clusters are usually dispersed as a result of A) external gravitational forces, B) the influence of the kinetic energy of the cluster members or C) the rotation of the Milky Way. However, some star clusters, e.g. Pleiades, remain quite dense.

Open star clusters are formed inside large dense interstellar clouds, when the latter begin to collapse under their own gravity. As the density of the clouds is not uniform, denser parts will eventually fragment into clumps that will become protostars. Noteworthy, it is now believed that stars are not formed individually, but rather in associations. Only the stars that are gravitationally bound to each other will form a cluster, whereas other stars from the association will be dispersed with the passage of time.

HRD for open clusters has a specific shape. It features a **narrow main sequence** which includes most of the cluster stars, amongst which only a **few giants** are present and unresolved binary stars appear as a group of stars around 1 magnitude above the main sequence. An HRD of a star cluster is a unique snapshot of what stars of the same composition and age but different masses look like. As the most massive stars evolve faster, most of the stars in any given young cluster are positioned in the upper part of the main sequence. For older clusters, on the contrary, the main sequence will be well-developed and the most massive stars will be starting to evolve off the main sequence.

One of the methods to determine the age of the cluster is by an end-point of the main sequence on HRD, where it turns to the giant branch. Younger clusters have a larger number of brighter and hotter stars of earlier spectral types. Thus, the *main sequence fitting* method can be used to determine the distance to a cluster based on the difference between apparent and absolute magnitude on the color-magnitude diagram.

2.2.1 Pleiades

The Pleiades, also known as the Seven Sisters, Melotte 22 or Messier 45, is a young open star cluster located in the constellation Taurus. It can be seen with the unaided eye, as its apparent magnitude is about 1.6 m and size is 110 arcmin. Amazingly to the

Table 1. Summary of Pleiades' distance studies

Author	Members	Distance	Method	Data
Giannuzzi 1995	1	132.4±11.3	isochrone fitting	The Eighth Catalogue of the Orbital Elements of Spectroscopic Binary Systems
Gatewood et al. 2000	7	130.9±7.4	parallaxes	Thaw/MAP
Soderblom et al. 2005	3	134.6±3.1	parallaxes	HUBBLE
An et al. 2007	39	135.5±10.2	main sequence fitting	WEBDA & Open Cluster Database
Leeuwen 2009	53	120.2±1.9	parallaxes	HIPPARCOS
Melis et al. 2014	4	136.2±1.2	parallaxes	Very Long Baseline Radio Interferometry (VLBI)
David et al. 2016	4	132.0±5.0	orbital modeling of eclipsing binaries	Kepler K2 mission
Galli et al. 2017	1210	134.4±2.9	moving cluster method	DANCe survey & Tycho-2
Gaia Collaboration et al. 2018a	1059	136.2±5.0	parallaxes	Gaia DR2

unaided eye though, it has **over a thousand members**. The Pleiades center coordinates are $(\alpha, \delta) \sim (56.75, 24.1167)$, its members have a significant proper motion ($\mu_\alpha \cos\delta$, μ_δ) $\sim$ (19.997, -45.548) mas/yr (Gaia Collaboration et al. 2018a). Lodieu et al. 2019 estimate the core size 2.00$\pm$0.25 pc, the half-mass radius 4.5 pc and the tidal radius as 11.6 pc. It is a relatively **close** cluster, although the distance to the Pleiades has been a subject of controversial discussions. Multiple theoretical methods have indicated this distance as 130-140 pc. Interestingly, parallax measurements from the *Hipparchos* satellite suggested the distance in question of only 118 pc, but new measurements serve to prove that this distance is too small. For example, in Gaia DR2 (Gaia Collaboration et al. 2018a) the distance is estimated as 136.2$\pm$5.0 pc. Some studies of Pleiades' age are presented in Table 1.

The Pleiades' **age** has also been debated, as results from different methods show conflicting results. While both Zero-Age-Main-Sequence turn-off and isochrone fitting methods estimate the age as 70–80 million years (Myr), two more methods - the model fitting and lithium depletion boundary - suggest 120-130 Myr, whereas more recent results imply a younger age of 112.5 Myr (Lodieu et al. 2019). In any case, this cluster's age is truly unique in a sense that cluster contains stars at various evolution stages - some lowest mass members are still stepping into the main sequence, while the majority of the stars are sitting on the main sequence burning hydrogen, and, still, some of the massive

stars are already turning off from the main sequence.

2.3 Sky surveys

2.3.1 Gaia

ESO Gaia mission (Gaia Collaboration et al. 2016) was launched on 19 December 2013. It aims to create a very extensive, accurate and complete **archive of stars**. The space environment and unique design of the Gaia experiment allows for an unprecedented sky coverage and **data quality**, both of which are impossible to get with ground-based facilities. Quite thrillingly, this data is expected to support solving questions about the evolutionary history of our Galaxy, its structure and origin. What is more, this data is made publicly available without limitations and can be obtained from the Gaia Archive[1].

The Second Gaia Data Release (**DR2**; Gaia Collaboration et al. 2018b) occurred on 25 April 2018. The data for it had been collected during 668 days from July 2014 to May 2016, to be more precise. From more than a whopping 1.3 billion sources with a limiting magnitude of G = 21 and a bright limit of G $\approx$ 3 Gaia measured **5 astrometric parameters**: positions on the sky (α, δ), parallaxes (p) and proper motions (μ_α, μ_δ), using the astrometric instrument. For a full astrometric processing pipeline, the author of this thesis refers the reader to Lindegren et al. 2018.

More than 7.2 million stars have also the **radial velocity** (rv) measurement, which is established by a spectroscopic instrument known as the radial-velocity spectrometer (RVS). It obtains spectra of the bright end of the Gaia sample and provides radial velocities through Doppler-shift measurements using cross-correlation for stars brighter than $G_{RVS} \approx 16$ mag and atmospheric parameters for stars brighter than $G_{RVS} \approx 12.5$ mag, where G_{RVS} denotes the integrated, instrumental magnitude in the spectroscopic bandpass. It is worth mentioning that to date no radial velocities have been determined for objects identified as emission-line stars. The effective temperatures for the sources with radial velocities are in the range of about 3,550 to 6,900 K, so there are no radial velocities for "cool" and "hot" stars. One needs to keep this all-important information in mind to properly consider the completeness of the study.

In addition to precise astrometry, there are also **photometry magnitudes** G (330 - 1050 nm), G_{BP} (330 nm - 680 nm) and G_{RP} (630 nm - 1,050 nm), and astrophysical characterizations (for instance, interstellar reddenings, surface gravities, metallicities, and effective temperatures (T_{eff}) for stars, photometric redshifts for quasars, etc.) measured by the photometric instrument.

For more details about the mission and data collection, please check Gaia Collaboration et al. 2016.

[1]The Gaia Archive is reachable from the Gaia home page at http://www.cosmos.esa.int/gaia and directly at http://archives.esac.esa.int/gaia

2.3.2 APOGEE

The Apache Point Observatory Galactic Evolution Experiment (**APOGEE**) (Steven R. Majewski et al. 2017) aims to collect high-resolution near-infrared (NIR; 1.6 μm H-band) spectra with a primary goal of mapping the distribution of the elements throughout different parts of our Galaxy. The fourteenth data release (**DR14**) (Abolfathi et al. 2018) was made public on 31st July 2017. It includes data from APOGEE-1, which is a component of the Sloan Digital Sky Survey III (SDSS-III, Eisenstein et al. 2011), and APOGEE-2, which is a component of SDSS-IV (Blanton 2017) and an extension of the project to the Southern hemisphere. This comprehensive, systematic and high-precision chemical and kinematical study focuses mostly on red giant branch stars, red clump stars, and asymptotic giant branch stars. Several more specific tasks that APOGEE aims to tackle are: 1) the first systematic determination of the 3D chemical abundance distribution; 2) determining the distribution of chemical abundances for a variety of elements in different parts of the Milky way; 3) establishing the nature of the Galactic bar(s) and spiral arms and their influence on the disk.

The first installment of the APOGEE Survey (APOGEE-1) was carried out from September 2011 to July 2014, and APOGEE2 began observing at APO (Apache Point Observatory, USA) in September 2014. Both APOGEE-1 and APOGEE-2 North utilize the wide-field (3° diameter FOV) SDSS 2.5 m telescope to obtain spectra for hundreds of stars per exposure. APOGEE-2 South is making observations on the 2.5-m du Pont Telescope at Las Campanas Observatory (Chile).

As every observed APOGEE field contains more stars than it can observe, APOGEE developed a **complex targeting strategy**. It targets stars of the "main sample", "special targets" (calibration stars, star cluster members, etc.), and a sample of early-type stars observed as telluric absorption monitors for each exposure. The target selection procedure is based on simple color and brightness cuts, in order to improve the spacial sampling of the Galaxy. Various schemes for selecting stars across the magnitude distribution have been adopted to ensure large spreads in distance representation along each line of sight. Additionally, to minimize the foreground dwarf star contamination and to favor the targeting of halo giants, additional photometric criteria have been adopted. Target selection is described in detail in Zasowski et al. 2013; Zasowski et al. 2017.

APOGEE data products include **stellar atmospheric parameters** (T_{eff}, $\log g$), **M/H** (overall metal abundance), α/M (relative α-element abundance, defined as O, Mg, Si, S, Ca, and Ti) and **individual element abundances** (namely, C, Cl, N, O, Na, Mg, Al, Si, P, S, K, Ca, Ti, TiII, V, Cr, Mn, Fe, Co, Ni, Cu, Ge, Rb, Y and Nd) as determined from the ASPCAP (APOGEE Stellar Parameters and Chemical Abundances Pipeline) analysis, with their errors, as well as the information relevant to target selection, including, for instance, coordinates, photometry, proper motions, radial velocity and assumed interstellar extinction.

3 Methods

3.1 Dataset

It should be mentioned that the creation of the dataset for this study consists of several steps. The dataset for further analysis is intended to contain stars in the 10° radius around the Pleiades, providing spectral data along with astrometry information. This data is present in different datasets, so the **first step** would be to **match these multiple datasets** together. **Second**, measurements with **high SNR** should be **removed**. Essentially, the **third step** requires **determining** which stars from the dataset belong to the **Pleiades**, which belong to field stars, and what are the stars whose membership is currently unknown. All this will be achieved by, first of all, utilizing the lists of Pleiades stars provided in the previous studies and, second of all, by analyzing the positions and proper motions of the stars. The **fourth step** is dealing with **missing values**, where our ultimate goal is to select spectral features that have the lowest number of missing values. **In the fifth step**, many dataset variations are created by different ways of data resampling. As the dataset is imbalanced and its size is rather small, in order to train the best machine learning model, we will explore multiple approaches for balancing the dataset. **Finally**, other dataset variations are created by adding more features or transforming them with the PCA. All these steps are described below in detail.

3.1.1 Gaia and APOGEE data crossmatching

The dataset is constructed from the open data from Gaia and APOGEE sky surveys. Gaia data provides accurate astrometry and parallax measurements, while APOGEE has spectral data which can be used for the machine learning model. That said, both datasets still need to be joined (**crossmatched**), so that each object in the dataset would have parameters from both datasets. This is done by sky positions, because these features are common for the selected datasets. As a result, a symmetric best match is found for each object in the datasets, and only matched pairs are kept.

As the first step, all stars within a 10° radius around the Pleiades position $(\alpha, \delta) \sim (56.75, 24.1167)$ were selected from the Gaia DR2 data, resulting in 2,286,544 stars. The filtering criteria introduced at this step require stars to have both proper motions measurements. In the next step, APOGEE DR15, which has 277,371 stars, is matched with the full Gaia DR2 data. The match was done using *gaia_tools*[2] Python library and 275,020 pairs of stars were found. Out of them, 3,050 are in the selected 10° radius around the Pleiades. The APOGEE dataset and the result of crossmatching with the Gaia data in the region of interest can be seen in Figure 2, which was generated using *Topcat*[3] software.

[2] https://github.com/jobovy/gaia_tools

[3] http://www.starlink.ac.uk/topcat/

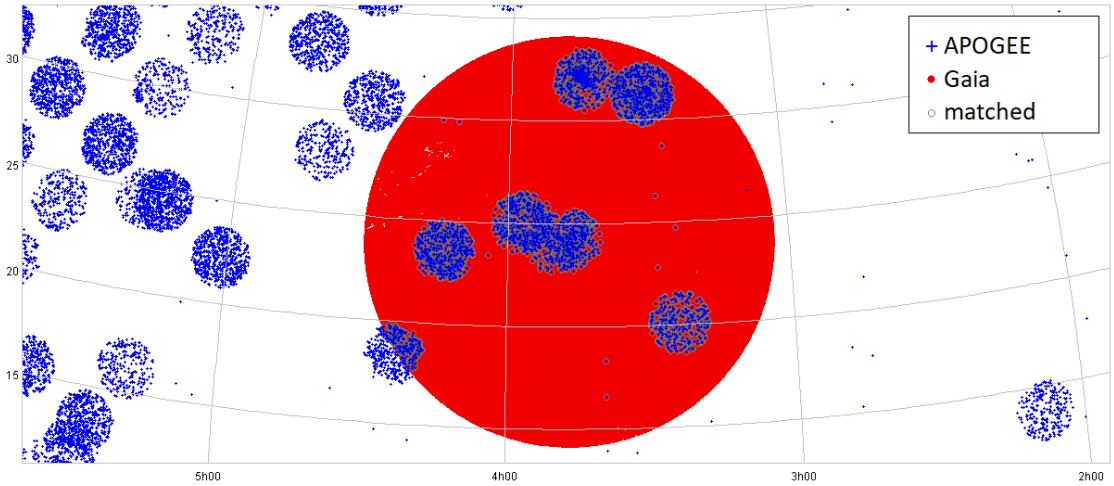

Figure 2. Crossmatching Gaia and APOGEE data

Figure shows the area of the sky around Pleiades center. Red dots are stars from Gaia in 10° area around Pleiades' center, blue dots are stars from APOGEE data. Grey dots are the stars that are present in both datasets - the result of crossmatch. After removing stars with too low SNR measurements, remaining 2,799 stars will be referred to as the base dataset.

3.1.2 Removing stars with low SNR

Objects that have inaccurate measurements may, consequently, bias the analysis and compromise the result, which makes it crystal clear that data cleaning needs to be done. Generally, it is always better to keep objects with as high as possible **signal to noise ratio** (SNR). However, the datasets created in this project are relatively small, thus a compromise between number of objects kept and SNR needs to be found. The distribution of SNR for some values can be seen in Figure 3.

It should be added that the filtering SNR thresholds used for all the datasets in this work are 1 for spectral abundances and radial velocity measurements and 2 for proper motions and parallax measurements. Values that do not satisfy the requirement of $SNR > 2$ for proper motions or parallax measurements are removed, as, in general, these values are relatively easy to measure and high error means low quality of data. Values that do not satisfy the requirement of $SNR > 1$ for spectral features and radial velocity measurements are marked as missing not to remove objects from the dataset completely, because they still can bring some value with other features. With this selection, 251 stars were removed from the dataset because of insufficiently accurate proper motions or parallaxes, and 88 radial velocity measurements and 15,423 spectral measurements are marked as missing values. **This leaves 2,799 stars from the crossmatched dataset**; further in this work, this dataset will be referred to as **the base dataset**.

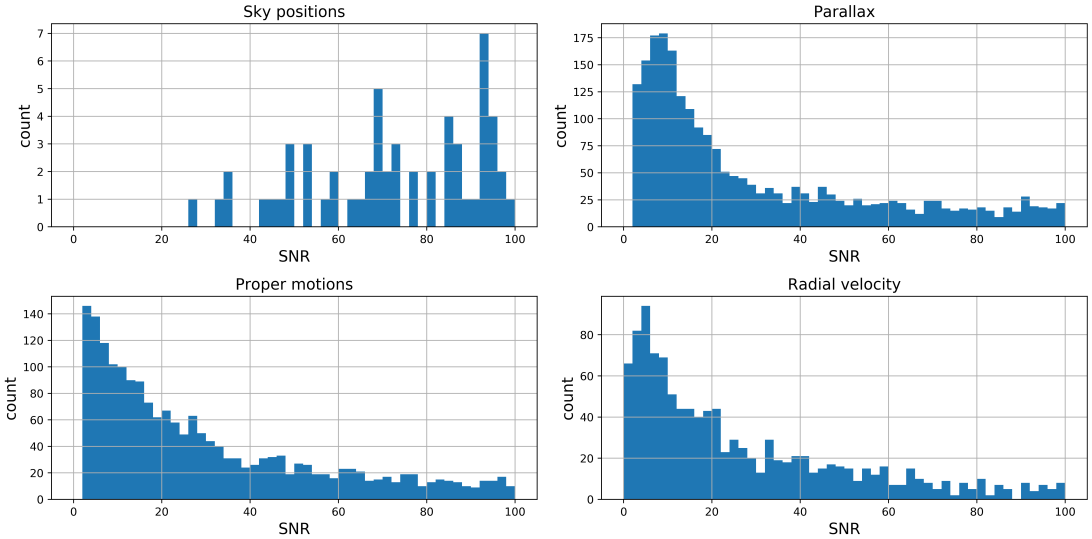

Figure 3. SNR distribution

Figure shows the histogram of SNR in range from 0 to 100 for sky positions (α, δ), parallax (p), proper motions ($\mu_\alpha \cos\delta$, μ_δ) and radial velocity (rv). For sky positions and proper motions, the minimum of two values was used for histogram. Except for sky positions, all other values seem to have a significant number of low-SNR measurements.

3.1.3 Selection of cluster members

To train a supervised machine learning model, the ground truth (**training label**) needs to be defined. In other words, for each training sample there should be an answer if this particular star **is a Pleiades star or a field star**. One can get these labels either from previous studies of the same object, where the authors supply the list of cluster members, or, alternatively, by determining the labels by a set of manually designed rules. The former approach, in general, is more robust than the latter, as it considers various factors and focuses on the quality of the selection, therefore ensuring a low false positive rate. However, such lists often are claimed to be not complete, which can introduce biases to the analysis.

For training a machine learning model, it is also vital to have a **list of field stars**. In effect, some scientists provide datasets of stars with membership probability; in this case, stars with low probability can be marked as field stars, while high probability stars can be marked as Pleiades stars. However, often these lists contain Pleiades stars only; in which case, field stars must be selected from other objects in a dataset by choosing stars that have an extremely low chance of being Pleiades stars. The approach used in this work is choosing stars that are situated more than 100 pc away from the Pleiades cluster center. As the Pleiades are assumed to have a tidal radius of 11.6 pc, according to Lodieu et al. 2019, this distance threshold is very reasonable.

In this work of ours, **four sets of labels** have been used, resulting in different dataset sizes, star selections and prediction quality. The summary information on these label sets can be found in Table 2.

First, the list of 1,326 Pleiades members from Gaia Collaboration et al. 2018a was used. The list is very accurate, although not complete. Therefore, there is a chance of finding new Pleiades members among the stars that are not in the list yet. This list can be directly matched with the base dataset by matching *source_id* column, which is present in both datasets.

Second, the dataset of Pleiades membership probability from Bouy et al. 2015 was used. It has 1,972,245 stars in $80^{\circ 2}$ region centered on the Pleiades cluster, 2,010 of which have a membership probability of more than 0.75 (this conservative selection threshold value is suggested by the authors of the list). This dataset was further crossmatched with the base dataset using *Topcat* by sky coordinates with the maximum error of 1 arc second. In the resulting dataset 16 stars with a high membership probability were situated too far from the cluster center, so they were by necessity removed from the dataset due to possible matching errors.

Third, both previous lists can be used together to mark the Pleiades and field stars, which results in more stars labeled as Pleiades.

Fourth, labels were created by applying proper motion and parallax constraints. All stars in the cluster should have similar proper motions, so it is only reasonable to use these measurements to find cluster members. Distance to the star is influencing proper motions (closer stars have higher velocities in the sky, and vice versa), so in order to obtain a better selection, labeling will be based on a new measure, which is proper motions divided by parallax. According to the Gaia Collaboration et al. 2018a, the Pleiades cluster has the following proper motions: ($\mu_\alpha \cos\delta$, μ_δ) $\sim$ (19.997, -45.548) mas/yr and a parallax 7.364 mas. The Pleiades proper motion difference (pmd) for stars with parallax (p) and proper motions ($\mu_\alpha \cos\delta$, μ_δ) is defined as following:

$$pmd = \sqrt{\left(\frac{\mu_\alpha \cos\delta}{p} - \frac{19.997}{7.364}\right)^2 + \left(\frac{\mu_\delta}{p} - \frac{-45.548}{7.364}\right)^2}$$

In this work, stars having $pmd < 0.325$ and within distance 125-145 pc are considered Pleiades.

The distribution of distances in range [100, 170] pc and radial velocities in range [-30, 30] km/s is duly shown in Figure 4.

3.1.4 Dealing with missing values

Some of the abundances and stellar parameters in the APOGEE data have missing values. The number of non-missing records per each column can be viewed in Figure 5.

In order to construct a dataset without missing spectral values, a threshold of 1,100 was set on count of rows with non-missing data. This threshold was selected heuristically

Table 2. Summary of labels sets

Labels source	**Labeled stars**	**Matched with base dataset**	**Pleiades stars**	**Field stars**	**Unknown membership**	rv **range of Pleiades**	**Distance range of the Pleiades**
Gaia Collaboration et al. 2018a	1326	375	375	1981	443	[-12.7, 18.2]	[119.9, 149.6]
Bouy et al. 2015	1972245	909	379	530	1890	[-53.0 18.2]	[111.2, 157.3]
Bouy et al. 2015 and Gaia Collaboration et al. 2018a	1972313	2426	408	530	1861	[-60.3, 42.1]	[78.9, 375.9]
Proper motions and parallax	2799	2799	360	1981	458	[-60.3, 42.1]	[111.2, 157.3]

"Labeled stars" column shows how many stars have an identification whether they belong to Pleiades or field stars for each label set. How many of them have APOGEE spectral features is visible in "Matched with base dataset". Next three columns "Pleiades stars", "Field stars" and "Unknown stars" shows how stars from the base dataset are distributed between star classes. Last two columns mark radial velocity and distance ranges of Pleiades stars in the dataset.

based on the number of non-missing values, in an attempt to keep as many rows and as many columns without missing data as possible. As a result, out of 24 spectral measurement columns, that include abundances and atmospheric parameters, only 12 have at least 1,100 non-missing values for all of the selected spectral measurements. The selected valid **columns with non-missing data** for all the stars are T_{eff}, Fe/H, M/H, Mg/Fe, α/M, N/Fe, Mn/Fe, Al/Fe, C/Fe, Si/Fe, O/Fe, Ni/Fe. This filtering step reduces the dataset to the values, listed in Table 3.

Table 3. Datasets contents after filtering out APOGEE missing values

Labels source	**Pleiades stars**	**Field stars**	**Unknown label**
Gaia Collaboration et al. 2018a	221	1437	290
Bouy et al. 2015	229	461	1258
Gaia Collaboration et al. 2018a and Bouy et al. 2015	246	461	1241
Proper motions and parallax	219	1437	292

Values in the table show the number of stars left in each group after incomplete rows (rows having missing values in selected feature columns) were removed.

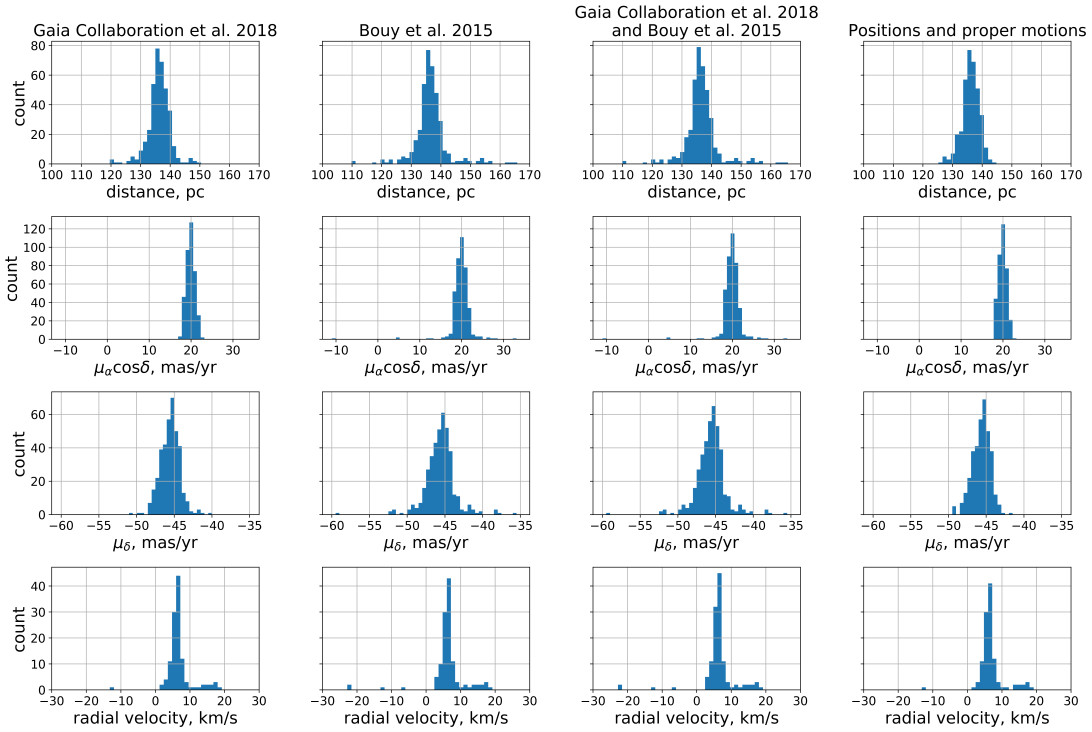

Figure 4. Distance and velocities distributions for each label set

Distance distribution of the Pleiades stars for each label set is shown in the first row. The range is from 100 to 170 pc, while most stars are concentrated in the range 130-142 pc. A small number of stars are outside the plot range for the second and third label sets. Proper motion distributions are shown in the second and third rows. Radial velocity distribution of the Pleiades stars for each label set is shown in the last row. The range is from -30 to 30 km/s, while most stars are concentrated in the range [3, -9] km/s. A small number of stars are outside the plot range for the second and third label sets.

3.1.5 Regulating imbalance

All datasets in Table 3 are **imbalanced**, which means that some class in the dataset is under-represented and has fewer samples than others. The class imbalance may lead to a worse algorithm performance on the underrepresented class. However, as the aim of this project is to accurately predict the Pleiades members and they happen to be the underrepresented class, the methods designed to tackle the problem of imbalance should be applied. There are two approaches for this: A) cost-sensitive learning and B) sampling (Prati 2009). The former one is easy to implement in some machine learning algorithms and will be discussed further in the thesis.

As for the latter, there are several approaches to sampling: oversampling the under-represented class and/or undersampling the over-represented class. Two simple and

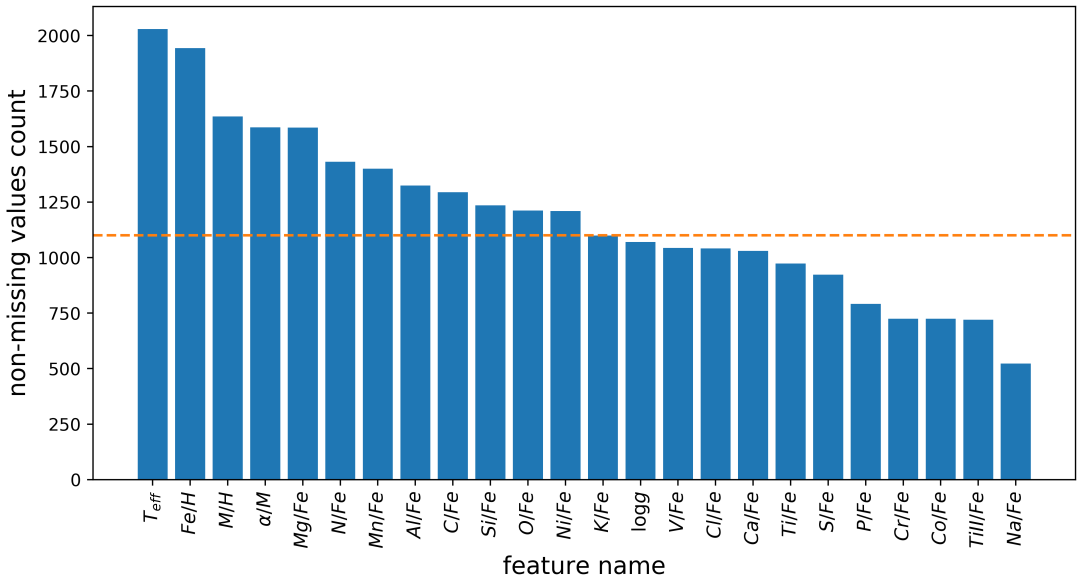

Figure 5. Number of non-missing values for APOGEE data in base dataset
The height of each bar shows the number of non-missing values for corresponding abundance. The dashed horizontal line shows the selection threshold - only the 12 columns that have more non-missing values than the threshold are kept in the dataset.

popular methods are random **oversampling** (exact copies of the random samples of the minority class are added to the dataset) and random **undersampling** (random samples of the majority class are removed from the dataset). The former has a drawback of having a biased model which leads to overfitting, and the latter would result in a significant reduction in the dataset size and, therefore, the reduction in the possibly valuable data.

There are several approaches described in Prati 2009 that can reduce the negative impact of the aforementioned weaknesses. Oversampling can be done with interpolation between close samples to avoid exact replication (e.g. **SMOTE** - Synthetic Minority Over-sampling Technique), and undersampling can benefit from removing noisy or less important samples (e.g. **ENN** - Edited Nearest Neighbor Rule that removes a sample if its label contradicts the labels of its nearest neighbors). Another way to reduce the bias in oversampling is to use the information about **uncertainities** (reported as standard deviations in APOGEE data), and this information is readily available for the APOGEE spectral features. For each star and each feature, for a reported measurement X and its uncertainty ERR_X, new data $\hat{X}$ can be sampled from the normal distribution $\hat{X} \sim \mathcal{N}(X, ERR_X)$.

Resampling techniques should be applied only to the training data, as these techniques are important only for the proper model training. Testing and evaluation data do not need resampling because they are used to evaluate the models' performance.

In our work several approaches have been tried out, creating multiple versions of the same data. Next, each of these versions is used to train a machine learning model, and finally their performance is analyzed. The composition of all obtained training datasets is listed in Table 4.

Table 4. Composition of all training datasets

Labels source	**Gaia Collaboration et al. 2018a**		**Bouy et al. 2015**		**Gaia Collaboration et al. 2018a and Bouy et al. 2015**		**Proper motions and parallaxes**	
Method	*Pleiades stars*	*Field stars*	*Pleiades stars*	*Field stars*	*Pleiades stars*	*Field stars*	*Pleiades stars*	*Field stars*
Undersampling	176	176	183	183	196	196	175	175
Undersampling with missing values	176	176	303	303	196	196	175	175
Unbalanced	176	1149	183	368	196	368	175	1149
Unbalanced with missing values	300	1584	303	424	326	424	288	1584
Exact oversampling	1149	1149	368	368	368	368	1149	1149
Exact oversampling with missing values	1584	1584	424	424	424	424	1584	1584
Oversampling in distribution	9192	9192	2944	2944	2944	2944	9192	9192
Oversampling in distribution with missing values	12672	12672	3392	3392	3392	3392	12672	12672

3.1.6 Other dataset variations

Although the main goal of ours is to have a model that uses only spectral features, it would be beneficial to compare its performance with the models that also use positions and velocities. As spectral data may be noisy, the usage of additional data might make the analysis more accurate. However, there is a risk of missing a lot of Pleiades stars that have drifted away from the cluster over time and have changed their velocity vector.

Table 4 mentions only the number of samples (rows) in the training dataset; it does not change with the number of features (columns). One version of these dataset variations has 12 features that are left after the missing values have been removed as described in Section 3.1.3 3.1.4. In the second version, for each dataset without missing values five additional features have been added: $\frac{\mu_\alpha \cos\delta}{p}$ and $\frac{\mu_\delta}{p}$ (proper motions divided by parallax) and positions in Cartesian Galactic coordinates: *X, Y, Z*. The third version is PCA calculated for the datasets without missing values from the previous two versions, which will be duly described in detail in Section 4.1. This makes **35 dataset variations** in total: 14 datasets with spectral features, 7 datasets with additional velocities and

position features, and 14 datasets with principal components as features. These will be further referred to as **dataset variations** in the thesis.

3.2 Models

To repeat, the ultimate aim of the project is to train a machine learning model that will predict a binary label (a Pleiades or a field star) using only spectral data as an input. Once operational, this model can be successfully used to find new cluster members in the Pleiades neighborhood.

In the following subsections the training process and the metrics used will be explained in detail.

3.2.1 Training

In order to find the best model that fits the data, one must run all the models on all dataset variations and then compare their performance. Due to the small size of the datasets and the simplicity of the models we are working with it will be relatively quick to train all models with all datasets variations.

As the datasets in this project are relatively small, instead of doing a simple random train/test split, **k-fold splits** will be applied in order to use all the data to the maximum extent. That means that the whole dataset with labels is randomly divided into k parts (in case of this thesis, $k = 5$). The model is then trained on four parts, and the fifth is used for testing. This process is repeated five times, each time the testing set being a different group. This results in having predictions for the whole dataset, and for each sample the prediction was made when the sample was in the test set and was not seen by the model. Further in this thesis, referrals to training and test datasets will correspond to one of the splits (where 80% of the labeled data is in the training dataset and 20% of it is in the testing dataset). This covers the steps 1-3 from the diagram of the training process (Figure 6) and is basically a cross-validation scheme often used in machine learning.

Some of the models can predict probabilities instead of the binary label, and regression models predict only a numeric value. To transform these into a binary label, a threshold will be used. Accordingly, all values below the threshold belong to one class, while the others fall into another class. The **optimal threshold** can be found by comparing the true positive rate (TPR) and the false positive rate (FPR), and both of these values will be soon described in Section 3.2.3. As TPR needs to be maximized while FPR has to be minimized, the threshold value that gives the biggest $TPR - FPR$ value will be selected. A **two-fold split** is used to find the best decision threshold in the testing phase. Namely, the hold-out group that was not previously seen by the model is split into two parts. The first part is then used to find the optimal threshold used to produce predictions for the second part. Next, the purposes of the parts swap and the threshold found on the second part is used for the predictions for the first part. For the

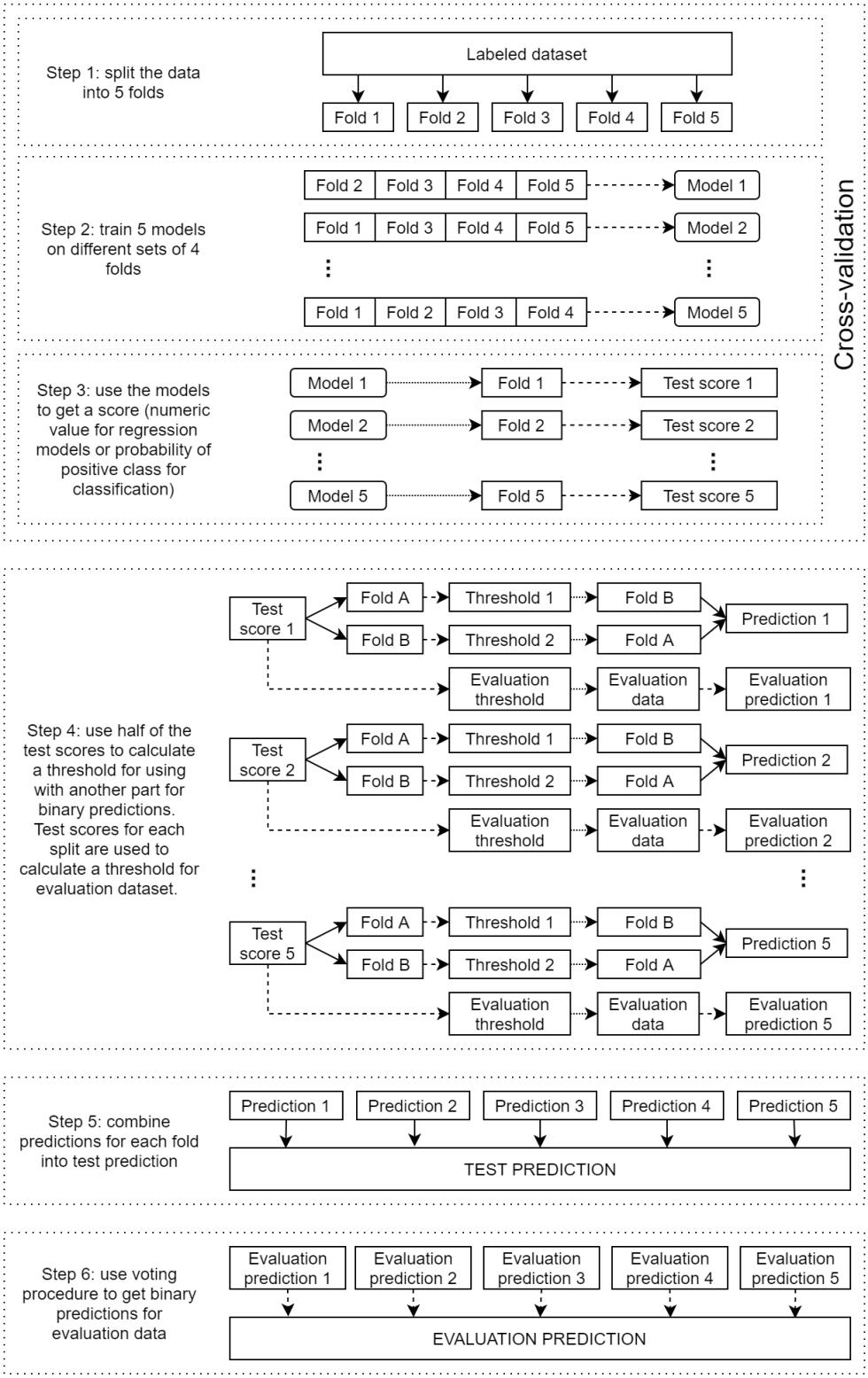

Figure 6. Diagram explaining the training and prediction process

final threshold, the one that will be used by the model to generate predictions on the evaluation set, both splits are used. This process is illustrated by step 4 in Figure 6. Finally, the predictions on each fold are **combined** into the one prediction vector (step 5 in Figure 6) and the evaluation prediction vector is formed from five the evaluation predictions by voting (step 6 in Figure 6).

3.2.2 Machine learning models

For this project, 4 regression model types and 5 classification model types have been used. Regression models are 1) Linear Regression and its variations - Lasso, Ridge and Elastic Net, 2) Support Vector Regression with linear, radial and polynomial kernels, 3) Random Forest Regression and 4) Symbolic Regression. Classification models are: 1) Logistic Regression, 2) Naive Bayes, 3) Support Vector Classification with linear, radial and polynomial kernels, 4) Random Forest Classification and 5) Symbolic Classification. These models are briefly described below.

Let $\boldsymbol{y}$ be the response variable (label), X represent the feature matrix (rows are samples, columns are feature values) and $\boldsymbol{\alpha}$ stand for the estimated coefficients. The objective of the **Elastic Net** is following:

$$\begin{gathered} \min \|\boldsymbol{y} - X\boldsymbol{\alpha}\|_2^2 + \lambda_1 \|\boldsymbol{\alpha}\|_1 + \lambda_2 \|\boldsymbol{\alpha}\|_2^2 \\ \boldsymbol{v} \in \mathbb{R}^n : \quad \|\boldsymbol{v}\|_1 = \sum_{i=1}^{n} |v_i|, \quad \|\boldsymbol{v}\|_2 = \sqrt{\sum_{i=1}^{n} v_i^2} \end{gathered} \tag{2}$$

When $\lambda_2 = 0$, equation 2 becomes the objective of **Lasso** regression; when $\lambda_1 = 0$ equation 2 becomes the objective of **Ridge** regression. If both $\lambda_1 = 0$ and $\lambda_2 = 0$, equation 2 becomes a simple **Linear** regression. Essentially, λ_1 and λ_2 are regularization terms, introduced to reduce the magnitude of the coefficients in order to avoid overfitting and keep the model more generalized.

For binary classification, the probability of instance $\boldsymbol{x} \in \mathbb{R}^n$ being in the positive class can be estimated by the **Logistic regression** model:

$$p = \frac{1}{1 + e^{-\alpha_0 - \sum_{i=1}^{n} \alpha_i x_i}}$$

Thresholds on probability are used to derive the predicted class.

Naive Bayes models assume that the features are independent, and therefore the Bayes theorem can be used to determine the probability of a sample belonging to each class. In consequence, the classification decision is then based on class y with the biggest probability of having the given set of inputs $\boldsymbol{x} \in \mathbb{R}^n$:

$$\hat{y} = \underset{\boldsymbol{y}}{\operatorname{argmax}} P(y) \prod_{i=1}^{n} P(x_i|y)$$

In case of Gaussian Naive Bayes, the continuous values associated with each feature are assumed to be normally distributed:

$$P(x_i|y) = \frac{1}{\sqrt{2\pi\sigma_y}} exp\left(-\frac{(x_i - \mu_y)^2}{2\sigma_y^2}\right),$$

where μ_y and σ_y are average and standard deviation of feature values associated with class y.

Methods that are based on **SVM** (support vector machine) are trying to find a hyperplane that best separates the samples of two classes. Given training vectors $x_i \in \mathbb{R}^n$, $i = 1, ..., l$, in two classes, and an indicator vector $\boldsymbol{y} \in \mathbb{R}^l$ such that $y_i \in \{1, -1\}$, soft-margin SVM has the following objective for fitting the model (Chang et al. 2011):

$$\begin{aligned} \min_{\boldsymbol{\alpha}} \quad & \frac{1}{2}\boldsymbol{\alpha}^T Q \boldsymbol{\alpha} - \boldsymbol{e}^T \boldsymbol{\alpha}, \\ \text{subject to} \quad & \boldsymbol{y}^T \boldsymbol{\alpha} = 0, \\ & 0 \leq \alpha_i \leq C, \qquad i = 1, ..., l, \end{aligned} \tag{3}$$

where $\boldsymbol{e} = [1, ..., 1]^T$ is the vector of all ones, Q is an l by l positive semidefinite matrix, $Q_{ij} = y_i y_j K(\boldsymbol{x}_i, \boldsymbol{x}_j)$, and $K(\boldsymbol{x}_i, \boldsymbol{x}_j)$ is the kernel function. Then, prediction $\hat{y}$ for $\boldsymbol{x}$ by is produced by:

$$\hat{y} = sgn\left(\sum_{i=1}^{l} y_i \alpha_i K(\boldsymbol{x}_i, \boldsymbol{x}) + b\right),$$
$$b = \sum_{i=1}^{n} \alpha_i y_i K(x_i, x^*) - y^*,$$

where $\boldsymbol{x}^*$ is a support vector of class y^*.

SVR (support vector regression) is defined in a similar way (Chang et al. 2011):

$$\begin{aligned} \min_{\boldsymbol{\alpha}, \boldsymbol{\alpha}^*} \quad & \frac{1}{2}(\boldsymbol{\alpha} - \boldsymbol{\alpha}^*)^T Q (\boldsymbol{\alpha} - \boldsymbol{\alpha}^*) + \epsilon \sum_{i=1}^{l} (\alpha_i + \alpha_i^*) + \sum_{i=1}^{l} z_i (\alpha_i - \alpha_i^*) \\ \text{subject to} \quad & \boldsymbol{e}^T (\boldsymbol{\alpha} - \boldsymbol{\alpha}^*) = 0, \\ & 0 \leq \alpha_i, \alpha_i^* \leq C, \qquad i = 1, ..., l, \end{aligned} \tag{4}$$

where $Q_{ij} = K(\boldsymbol{x}_i, \boldsymbol{x}_j)$. Prediction $\hat{y}$ for $\boldsymbol{x}$ by model 4 is then:

$$\hat{y} = \sum_{i=1}^{l} (-\alpha_i + \alpha_i^*) K(\boldsymbol{x}_i, \boldsymbol{x}) + b$$

The possible kernel functions for models 3 and 4 are:

$$\text{Linear kernel: } K(x, x') = x \cdot x'$$

$$\text{Gaussian (RBF) kernel: } K(x, x') = exp\left(-\frac{\|x - x'\|^2}{2\sigma}\right), \quad \sigma > 0$$

$$\text{Polynomial kernel: } K(x, x') = (x \cdot x' + r)^d, \quad r \in \mathbb{R}, d \in \mathbb{N},$$

where RBF stands for the radial basis function.

Random Forest models are, in essence, ensembles of simple decision tree models. Each decision tree is built on a new dataset: samples were selected from training data with replacement, and a random subset of features was selected on each split for this data. Prediction is then based on the majority class (for classification) or average value (for regression) in the individual trees prediction.

As for **symbolic models**, they are fit by searching the space of mathematical expressions to find the model that best matches the observed data in terms of accuracy and simplicity. Starting with a set of mathematical operators, functions and constants, they are randomly combined to form equations. Then, new equations are produced from the previous ones using the principles of genetic programming. This method has a benefit of not relying on a specific model type. Instead, it learns the structure of the model along with its parameters.

Finally, the training parameters of the models are presented in Table 5. For the used classification models except for Naive Bayes and Symbolic classifier, it is possible to set the *class_weight* parameter to *"balanced"* so that each class weight is inversely proportional to the class frequencies in the input data. This is done to introduce cost-sensitive learning to tackle the imbalance problem. This parameter is optional and will be used only when the training dataset is imbalanced.

3.2.3 Metrics

Careful selection of the metrics is absolutely crucial for comparison of all the models and selecting only the best one. The metrics that have been used in this work are accuracy, precision, recall, F1-score, Kullback–Leibler divergence and Wasserstein distance. These metrics are evaluated on test datasets to select the best models and estimated on evaluation sets to assess models' performance.

The most popular metrics to use in machine learning are accuracy, precision, recall and F1-score. Accuracy measures the proportion of correctly classified samples among all samples, precision evaluates the proportion of the correctly classified positive class samples in all predicted positive class samples. Recall gauges the proportion of the correctly classified positive class samples in all positive class samples and, finally, F1-score is a harmonic mean of recall and precision. The metrics described above are defined

Table 5. Machine learning models

Model name	**Model type**	**Parameters**
Random Forest Classifier	classification	n_estimators=20, max_leaf_nodes = 9, [class_weight='balanced']
Logistic Regression	classification	solver = 'lbfgs' , [class_weight='balanced']
SVM with radial kernel	classification	kernel='rbf', gamma=1, C=1, probability=True, max_iter=5000, [class_weight='balanced']
SVM with linear kernel	classification	kernel='linear', gamma=1, C=1, probability=True, max_iter=5000, [class_weight='balanced']
SVM with polynomial kernel	classification	kernel='poly', gamma='scale', C=1, probability=True, degree=3, max_iter=5000, [class_weight='balanced']
Naive Bayes	classification	
Symbolic Classifier	classification	function_set=['add', 'sub', 'mul', 'div', 'min', 'max', 'abs', 'neg', 'inv']
Random Forest Regressor	regression	n_estimators=20, max_leaf_nodes = 9
Linear Regression	regression	
Ridge	regression	alpha=1, max_iter=5000
Lasso	regression	alpha=0.05, max_iter=5000
ElasticNet	regression	l1_ratio=0.25, alpha=0.1,max_iter=5000
SVR with radial kernel	regression	kernel='rbf', gamma=1, C=1, probability=True, max_iter=5000
SVR with linear kernel	regression	kernel='linear', gamma=1, C=1, probability=True, max_iter=5000
SVR with polynomial kernel	regression	kernel='poly',gamma='scale', C=1, probability=True, degree=3, max_iter=5000
Symbolic Classifier	regression	function_set=['add', 'sub', 'mul', 'div', 'min', 'max', 'abs', 'neg', 'inv']

as follows (Powers 2007):

$$accuracy = \frac{TP + TN}{TP + TN + FP + FN}$$

$$precision = \frac{TP}{TP + FP}$$

$$recall = \frac{TP}{TP + FN}$$

$$F1 = \frac{2 \cdot precision \cdot recall}{precision + recall} = \frac{2 \cdot TP}{2 \cdot TP + FP + FN}$$

where TP stands for the number of true positives (positive class samples predicted correctly), TN denotes true negatives (negative class samples predicted correctly), FP displays false positives (negative class samples predicted as positives) and FN represents false negatives (positive class samples predicted as negatives).

In this thesis, stars from the Pleiades are considered as positive class, and field stars are negative class. Some of the datasets used in this work are imbalanced, so accuracy is not a good metric for them, because accuracy will be biased towards the majority (negative) class. In turn, precision and recall are well-suited metrics for the task, as both of them are independent of theoretically infinite number of true negatives and are able to measure how well the model has learned to identify the positive class (Pleiades).

The aforementioned metrics can be measured only on the test dataset, where the label for samples is known. But in order to evaluate the validity of predictions on evaluation datasets, approaches that are based on background knowledge should be used instead. It is expected that the distance, proper motion and radial velocity distributions of the predicted Pleiades stars should be similar to the distributions of the already known Pleiades stars. Two good metrics to measure the **difference in distributions** are Kullback–Leibler divergence and Wasserstein distance.

Kullback–Leibler divergence (also known as relative entropy) D_{KL} for discrete probability distributions P, Q on the same probability space X is defined as follows (Bazán et al. 2019):

$$D_{KL}(P||Q) = \sum_{x \in X} P(x) log \left(\frac{P(x)}{Q(x)} \right)$$

and shows how Q is different from P. This measure is both non-negative and asymmetric: $D_{KL}(P||Q) \geqslant 0$, $D_{KL}(P||Q) \neq D_{KL}(Q||P)$. It shows the amount of information lost when Q is used to approximate P.

In this project, for some physical values (distance, proper motion or radial velocity) and for P being the distribution of this value for the already known Pleiades stars and Q standing for the predicted Pleiades stars, $D_{KL}(Q||P)$ is a bit more important than $D_{KL}(P||Q)$. The former will penalize predictions if they are not supported by true values; in other words, minimizing $D_{KL}(Q||P)$ is an equivalent to minimizing the number of false positives. The latter will penalize predictions if they do not correspond to the true values, or in other words, minimizing $D_{KL}(P||Q)$ is an equivalent to maximizing recall. Both metrics are useful, but one needs to keep in mind the properties of distributions for the known and predicted Pleiades stars. The latter may have much fewer values and they may seem more random (e.g. while the known Pleiades stars have a well-defined distance distribution depicted in Figure 4, it might happen that none of the predicted Pleiades will have the distance close to the mode, resulting in a different, though still valid distribution).

Another metric that estimates the difference between distributions is **Wasserstein distance** (Bazán et al. 2019, also known as "the earth mover's distance"). It can be viewed as an amount of work needed to transform one distribution into another, where "work" is measured as the amount of distribution weight that must be moved, multiplied by the distance it has to be moved for. If P is the empirical distribution of dataset $X_1, ..., X_n$ and Q is the empirical distribution of another dataset $Y_1, ..., Y_n$ of the same

size, then the definition of Wasserstein distance is

$$W_1(P, Q) = \sum_{i=1}^{n} ||X_i - Y_i||,$$

where $||\cdot||$ denotes the "ground distance", which, in effect, can be any distance (Euclidean, Manhattan, etc.).

4 Results

4.1 Dimensionality reduction

Most of the datasets in this project are relatively small, so having many features may result in losses in the models' performance. However, simply removing features will result in the loss of valuable data. Instead, dimensionality reduction methods, such as **principal component analysis** (PCA), could be applied to achieve the best results. Spectral data is known to be correlated, so it is very beneficial that the principal components are orthogonal to each other. PCA is also good for visualization, as it allows to represent the data using fewer dimensions.

PCA converts a set of observations into an uncorrelated orthogonal basis set, referred to as principal components. Out of these, each successive component accounts for as much variability in the data as possible. Therefore, by selecting only a few first components, it is possible to get a sufficiently close representation of the information in the dataset. In other words, the orthogonal directions with the maximal variation of data are found. In practice, principal components are defined by eigenvectors of the covariance matrix. The vector that corresponds to the largest eigenvalue is the first component, and further in the descending order of eigenvalues.

In our thesis, PCA has been applied for each of the dataset variations without missing values from Table 4 and PCA-datasets have also been included in the whole analysis pipeline. To give just one example, PCA is presented for the training data of the undersampled dataset. The cumulative variance explained by principal components is depicted in Figure 7. Eventually, **8 principal components** have been selected, as they explain 94% of the variance in the data.

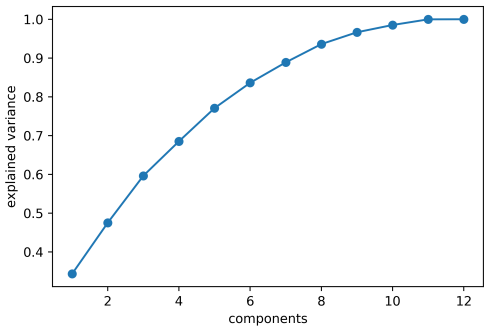

Figure 7. Cumulative variance explained by principal components
The first component explains 34% of the variance in the dataset, the second explains 13%. The first 8 components explain 94% of the variance and they form a new dataset that will be added to the analysis pipeline.

Figure 8 shows the data with the selected principal components pairs that best show the difference in the distributions of the two classes (Pleiades and field stars).

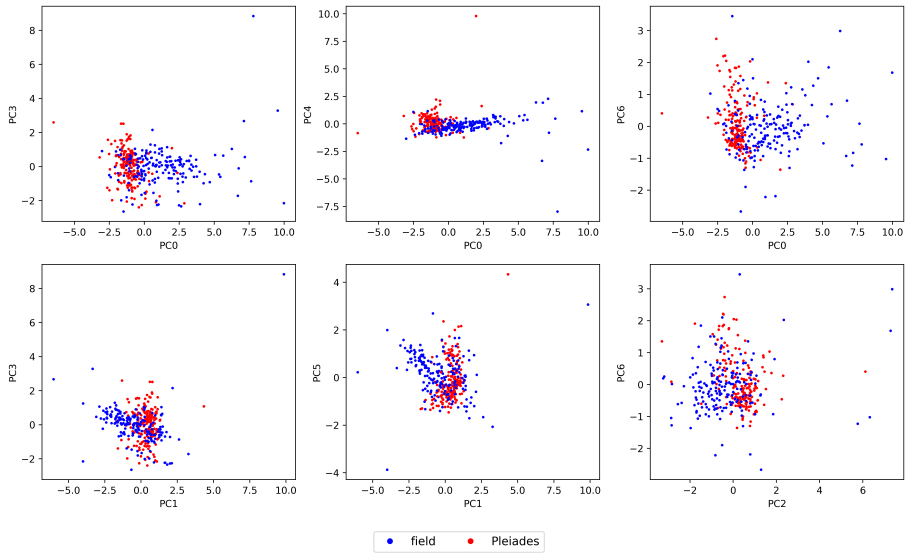

Figure 8. Dataset in principal components space
There is a clear difference in the distribution of blue (field stars) and red (Pleiades stars) in principal component space.

4.2 Selecting the best models

In this section the performance of all models on all dataset variations is going to be evaluated. The underlying aim is to select **two best models**: one that uses spectral data only and the other that, in addition to spectral data, also uses positions and velocities. The comparison will be performed for predictions obtained for the whole labeled dataset, where the prediction for each sample is made when the sample is in the test dataset (the process is described in detail in Section 3.2.1).

As there are 585 models and multiple metrics that all need to be taken into account, the selection will be done in several steps. In the first step, for each set of labels and model types (with or without positions and proper motions) 10 best models are selected. This is done by successively removing the results with low metric values on the test dataset, starting from accuracy and F1 measure, followed by precision and recall. It is very unlikely to get the Pleiades stars quite far away from the cluster, that is why the final selection of top-performing models is based on the similarity of distance distributions (Wasserstein metric). There is also a preference for the models with a smaller Wasserstein

distance between radial velocity distributions for labeled and predicted Pleiades. As the final step, for each label set two best models are selected based on the combination of all metrics, including D_{KL}. However, if some top-performing model is not directly or easily **interpretable**, an additional model will be selected for the same group.

The selected models for each group are listed in Table 6. Among the models that rely on spectral features only, Logistic regression model on the undersampled data will be selected for further analysis, as it is easily interpretable and has good performance metrics. Further on, this model will be referred to as the **spectral model**. The best model that also uses positions and velocities is Naive Bayes on unbalanced data, will, in turn, be referred to as the **enhanced spectral model**. In the next subsection the performance of the chosen models will be evaluated on evaluation set.

4.3 Predicting new Pleiades members

The dataset for predictions (evaluation dataset) consists of all other stars from the dataset that have not been seen by the model before (not in the positive or negative class). This is the dataset of objects where we might expect to find the Pleiades stars. The spectral model and the enhanced spectral model will be applied to the **evaluation data** of the respectful datasets and in this way the predictions for each star will be obtained. Positive class predictions on the evaluation dataset will mean new Pleiades stars.

Still, to properly calculate the metrics related to the distribution similarity for the evaluation dataset, it is necessary to have the same proportion of Pleiades as in the labeled data that was used to train and test the model. Otherwise the metric values will be strongly biased. And one way to avoid such biases is to evaluate the metrics on the **subset** of the evaluation dataset that has the **same distribution** of distances as the labeled data. The histogram of distance distribution for the labeled and evaluation data for the spectral model is shown in Figure 9, and same histogram for the enhanced spectral model may be found in Figure 10. Both figures show the distribution of distances of stars used for training or prediction on the distance range from 0 to 600 pc. There is a small number of stars in the labeled data outside this range, but stars with distances of more than 600 pc are immediately discarded from evaluation data.

The process described in Section 3.2.1 yields five models for each dataset variation. For generating predictions on the evaluation dataset, a **voting** scheme will be used. Each of the five models generates a binary prediction for evaluation data, and the final prediction is of the majority class. In other words, the positive class is predicted if at least three models predict that a particular star is in the positive class.

Metrics for evaluation data for the spectral and enhanced spectral models are listed in Table 7. The spectral model predicts **256 stars to be from Pleiades**, but big values of the Wasserstein distance (more than two times what it was on the test dataset) between distance distributions of the predicted and true Pleiades, as well as D_{KL} and Wasserstein distance for radial velocity distributions (at least five times the test values), suggest the

Table 6. Performance of selected models on test data

As mentioned earlier, 585 models have been trained to predict if a particular star belongs to the Pleiades. Their performance metrics have been evaluated on test predictions, and for each group (based on the label set and if the model uses positions and velocities in addition to spectral data), the best model is selected. Two models have been chosen for further analysis and prediction. They are highlighted with
the background color and will be referred to as the spectral model and the enhanced spectral model.

Labels set	**Dataset variation method**	**Model**	**Accuracy**	**Precision**	**Recall**	**Distance** $D_{KL}(T\|\|P)$	**Distance** $D_{KL}(P\|\|T)$	**Distance** W_1	**Radial velocity** $D_{KL}(T\|\|P)$	**Radial velocity** $D_{KL}(P\|\|T)$	**Radial velocity** W_1
Gaia Collaboration et al. 2018a	Undersampling	SVR with rbf kernel	0.952	0.972	0.932	3.685	29.919	5.286	1.360	0.072	0.547
Gaia Collaboration et al. 2018a	Undersampling	Logistic regression	0.928	0.957	0.896	3.685	29.919	14.940	1.344	0.066	0.475
Bouy et al. 2015	Undersampling	Random Forest classification	0.856	0.825	0.904	3.392	27.268	12.485	1.584	2.880	1.465
Gaia Collaboration et al. 2018a and Bouy et al. 2015	Undersampling	SVR with rbf kernel	0.882	0.851	0.927	3.137	20.567	7.883	1.354	1.421	1.150
Gaia Collaboration et al. 2018a and Bouy et al. 2015	Undersampling	Random Forest regression	0.854	0.827	0.894	31.034	41.888	10.364	1.391	1.990	1.399
Positions and proper motions	Undersampling	SVR with rbf kernel	0.936	0.961	0.909	3.924	32.499	7.014	0.899	0.074	0.684
Positions and proper motions	Undersampling. PCA	Random Forest classification	0.943	0.953	0.932	3.924	32.499	8.644	0.873	0.382	0.465
Gaia Collaboration et al. 2018a	Unbalanced*	Naive Bayes	0.995	1.000	0.959	3.685	29.919	0.541	1.369	0.077	0.649
Bouy et al. 2015	Undersampling*	Random Forest classification	0.948	0.936	0.961	3.392	27.268	0.924	2.001	0.688	1.033
Gaia Collaboration et al. 2018a and Bouy et al. 2015	Undersampling*	Symbolic regressor	0.862	0.868	0.854	3.137	20.567	1.962	1.149	1.511	1.482
Positions and proper motions	Unbalanced*	Naive Bayes	0.996	1.000	0.973	3.924	32.499	0.304	1.442	0.079	0.610

T and P distributions in D_{KL} metric define distributions of a given value for true Pleiades stars and predicted stars, respectfully.
* datasets that use positions and velocities in addition to spectral data

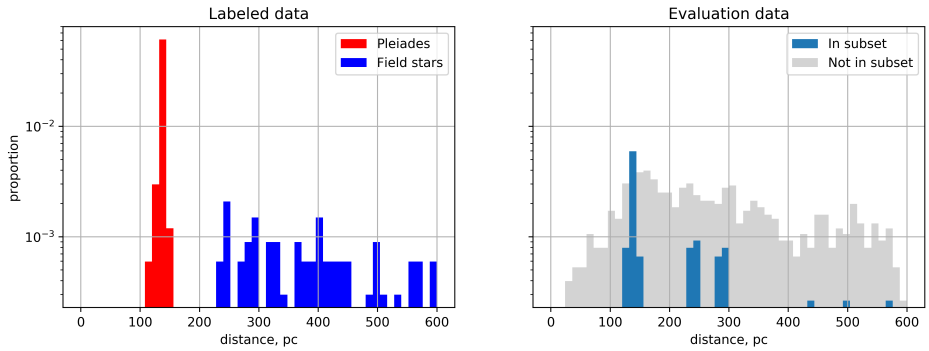

Figure 9. Distance distributions for labeled data and a subset of evaluation data for spectral model

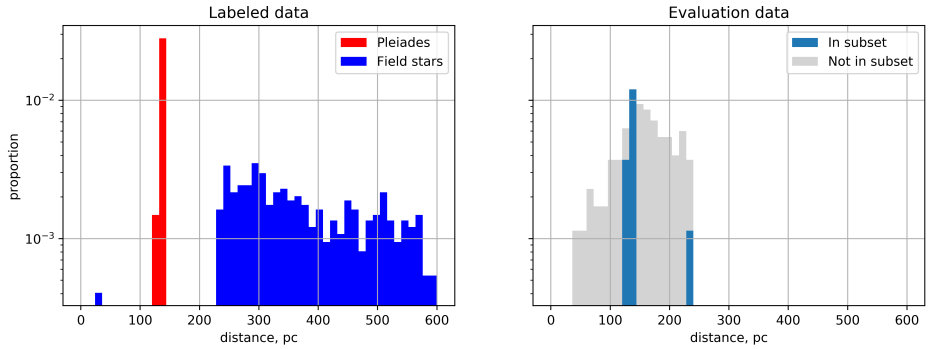

Figure 10. Distance distributions for labeled data and a subset of evaluation data for enhanced spectral model

fact of contamination (false positive prediction). In fact, as the model uses only spectral features for prediction, it is only able to find stars with similar spectral features. So, as it turns out, *spectral features alone are not able to clearly separate the Pleiades stars* as there are many more stars that have similar features.

That said, D_{KL} and W_1 metrics for distance and radial velocity of the Pleiades predicted by the enhanced spectral model have small values that are very similar to the values obtained on the test dataset. Moreover, the ranges of these values are very narrow and agree with the corresponding ranges of the Pleiades.

There are only **55 Pleiades stars** predicted by the spectral model in the distance range **from 115 to 155 pc**. All other stars predicted by the model are too far away from the cluster, so they **should not be considered as Pleiades** member candidates. The validity of these predictions will be discussed in Section 4.4. As for the enhanced spectral model, all 22 predicted Pleiades are in the aforementioned range. Other properties of the predicted Pleiades are listed in Table 8.

Table 7. Performance of selected models on a subset of evaluation data

Some stars predicted as Pleiades by the spectral model have distances and radial velocities that are not inherent to
Pleiades stars. The same values for the enhanced spectral model are very similar to Pleiades stars.

Model name	**Distance** $D_{KL}(T\|\|P)$	**Distance** $D_{KL}(P\|\|T)$	**Distance** W_1	**Radial velocity** $D_{KL}(T\|\|P)$	**Radial velocity** $D_{KL}(P\|\|T)$	**Radial velocity** W_1
Spectral model	3.68	29.92	32.73	7.29	16.33	12.14
Enhanced spectral model	3.92	32.50	1.39	4.98	0.32	1.61

T and P distributions in D_{KL} metric define distributions of a given value for true Pleiades stars and predicted stars, respectfully.

Table 8. Properties of predicted Pleiades member candidates

The spectral model has too wide ranges of proper motions and radial velocity, still suggesting some contamination. However, the properties of candidates predicted by the enhanced spectral model are very similar to true Pleiades stars.

Model name	**Predicted Pleiades**	**Distance range**	$\mu_\alpha \cos\delta$ **range**	μ_δ **range**	**Radial velocity range**
Spectral model	55	[116.31, 154.00]	[-28.7, 59.3]	[-60.4, 16.2]	[-25.8, 25.7]
Enhanced spectral model	22	[122.97, 152.39]	[16.4, 24.7]	[-49.9, -41.1]	[3.0, 7.0]

The full list of the predicted Pleiades stars can be found in Appendix (Table 11). It consists of **63 stars**: 41 are predicted with the spectral model, 8 more are predicted with the enhanced spectral model and the remaining 14 are predicted by both models. Labeled Pleiades, along with predicted Pleiades, are plotted in cartesian galactic coordinates in Figure 11. This figure also shows the **velocity vectors** for the labeled and some individual predicted Pleiades. It can be seen that some stars from the spectral model predictions have velocity vectors that are very untypical of the Pleiades. This is not the case for the enhanced spectral model predictions, where velocity vectors of the Pleiades candidates members are the same as for the labeled Pleiades.

It can be therefore seen that for enhanced spectral model predictions the predicted Pleiades are scattered around the Pleiades core, mostly stretching out along X-coordinate, and a bit along Y and Z axis. This perfectly corresponds to the results from Lodieu et al. 2019 (see Figure 5 in the corresponding article) where they observe the presence of a tail-like structure along the (X, Z) directions. However, in our study the amount of the predicted stars is too small to make any assumptions about the tidal tails.

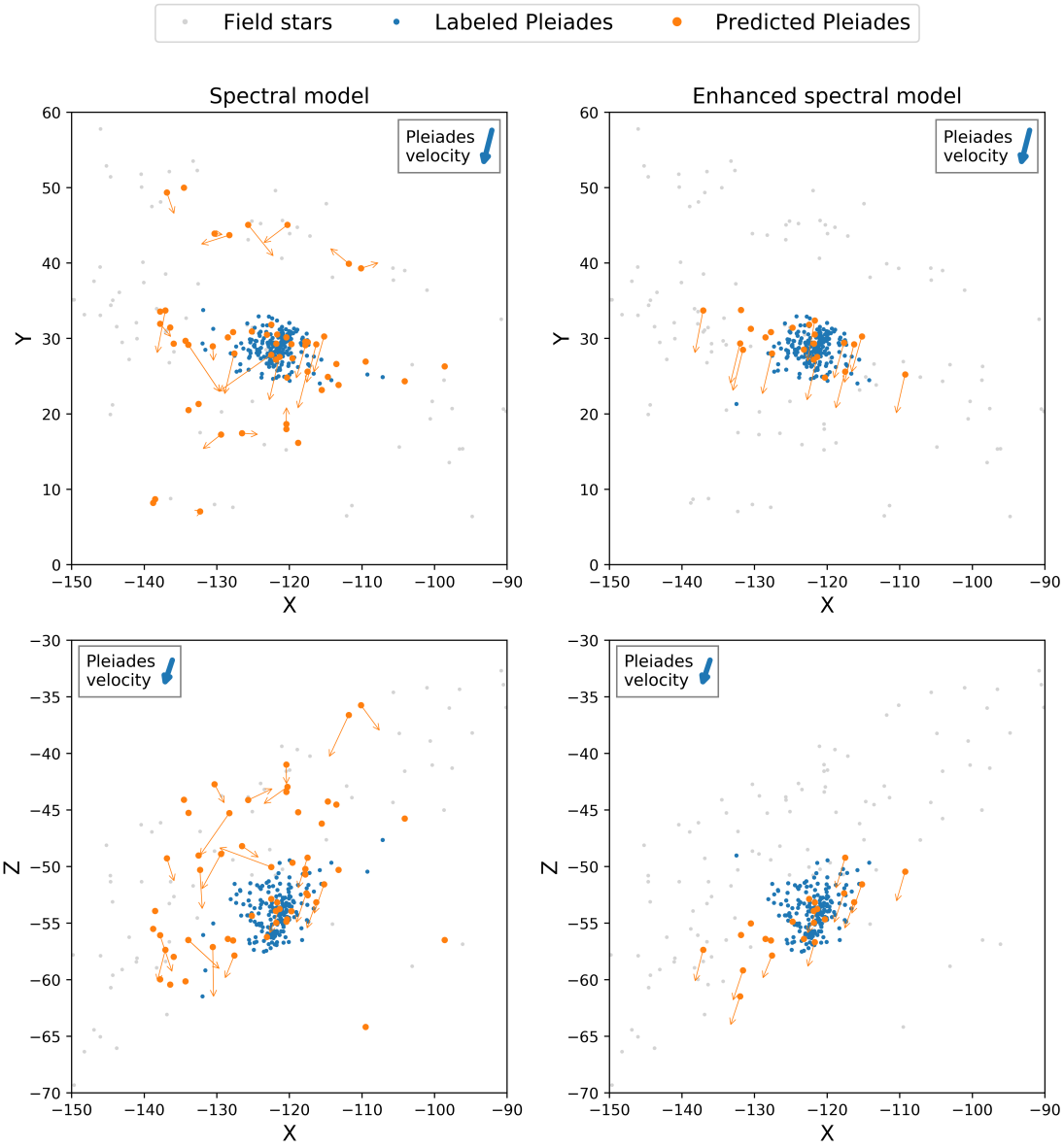

Figure 11. Labeled and predicted Pleiades in galactic coordinates
Orange arrows on the plot show the individual velocities of the predicted Pleiades member candidates. The Pleiades cluster velocity is shown with the blue arrow in the box.

An interesting observation becomes apparent from Figure 11. There are some stars that are colored differently in the spectral and enhanced spectral models' plots: for one model they were used as true Pleiades stars, and for the other they are predicted Pleiades member candidates. As described in Section 3.1.3, Pleiades members for the enhanced spectral model were selected based on proper motions and distance, while for the spectral

model they were selected based on the Pleiades members list from Gaia Collaboration et al. 2018a. Interestingly, the spectral model was able to predict some of the labeled Pleiades from the enhanced spectral model, and vice versa.

4.4 Validation

There are many ways to examine the validity of the predictions for Pleiades stars. First, one would expect that the new Pleiades stars will not dramatically change the distributions from Figure 4, as the predicted Pleiades stars should have a **similar distance and radial velocity**. Also, the obtained distances and velocities may be compared against the cluster parameters reported in the literature. Second, predictions can be validated by plotting them in HRD together with the true Pleiades stars. Stars of the same cluster should **fall into one line on the main sequence**. Finally, in case of the spectral model predictions, one should check that those stars that were predicted by the model but have a too big distance, are still **valid predictions** in terms of similarity of their spectral features to the true Pleiades.

The distributions of cluster parameters after the addition of the predicted stars is shown in Figure 12. New Pleiades stars do not change the general shape of the distributions for both models and all parameters. However, for the spectral model there is a small number of candidates with proper motions and radial velocity values that are outside the usual parameter range of the Pleiades.

The HRD diagrams for both models are presented in Figure 13. It can be observed that most of the predictions do fall into one line with the labeled Pleiades on the main sequence. None of the Pleiades candidates are in the giants region in HRD, which, nevertheless, is quite expectable because the cluster is about 100 Myr old.

It is important to examine the Pleiades predictions of the spectral model, which are too far from the Pleiades. There are 201 predicted Pleiades with distances less than 115 pc or more than 155 pc. Figure 14 confirms that all these stars have parameters similar to the Pleiades from the training data, that is why a positive prediction was made for them. These stars probably are not Pleiades, but are predicted as those because they have a very similar chemical composition.

4.5 Opening the model

Opening the model refers to examining the importance of the features and understanding how the model arrives at the predictions it has made.

The Logistic regression model that is used by the spectral model is directly interpretable through its coefficients. Table 9 shows the **coefficients of the Logistic regression model**. The higher the magnitude of the coefficient is, the more important the feature is. Therefore, top-five most important spectral features are $N/Fe, C/Fe, T_{eff}, Mg/Fe$ and O/Fe.

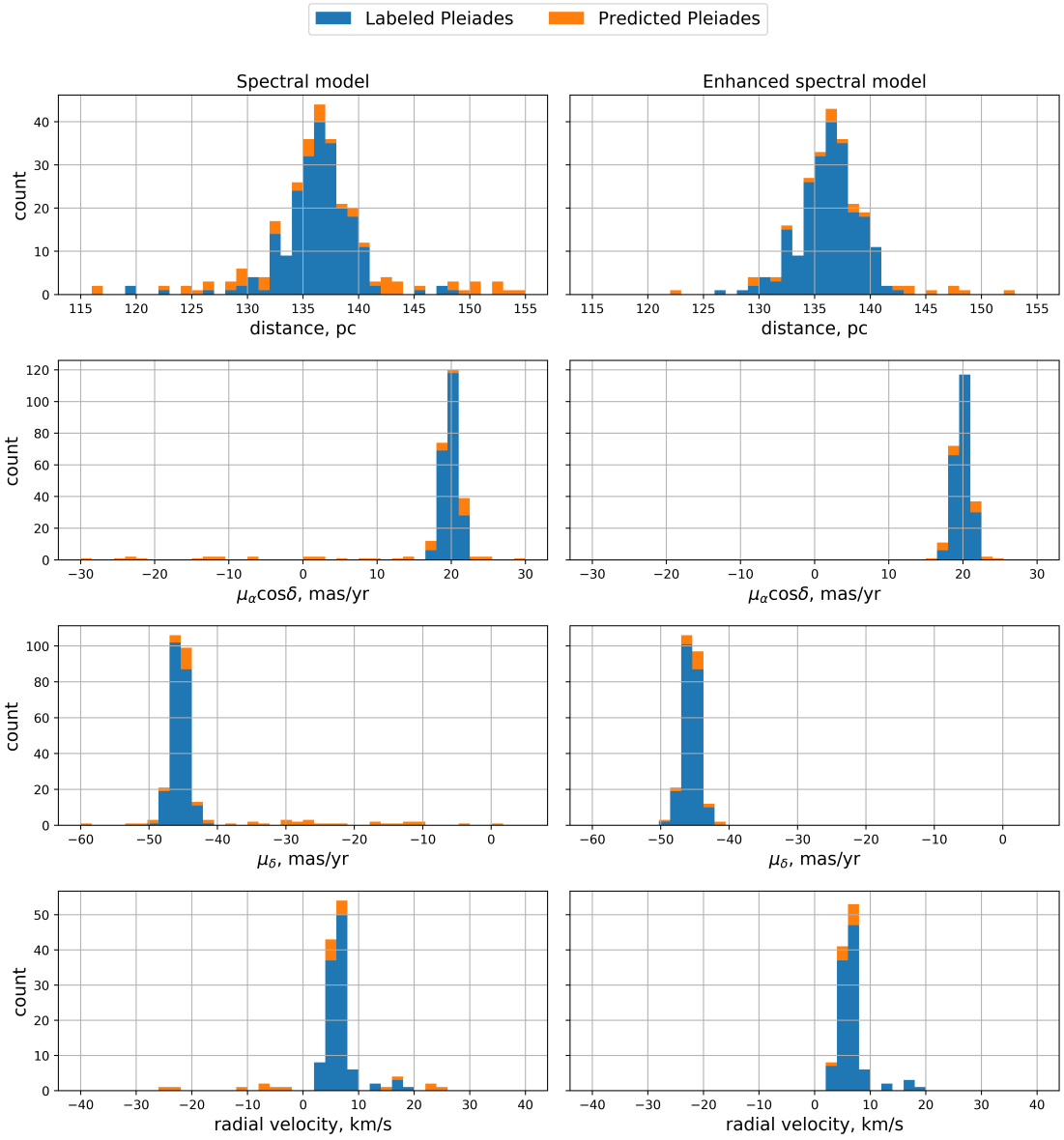

Figure 12. Distributions of distance, proper motions and radial velocity for selected models with added predicted Pleiades

The first row shows the distribution of distances for two selected models, the second and third rows display the distribution of proper motions and the last row illustrates the distribution of radial velocity values.

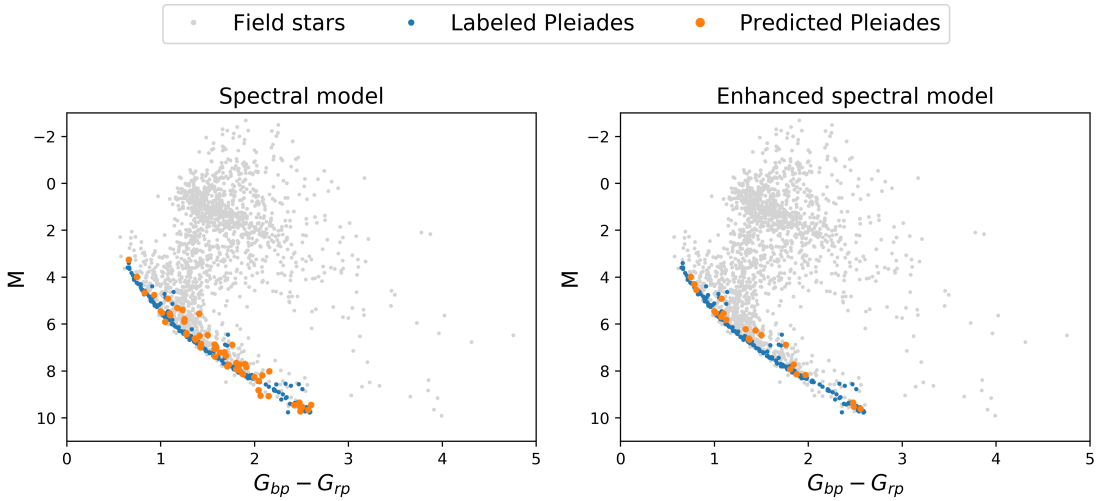

Figure 13. HRD of labeled Pleiades with added predicted Pleiades
Predicted Pleiades stars lie almost on the same line as labeled Pleiades stars for both models.

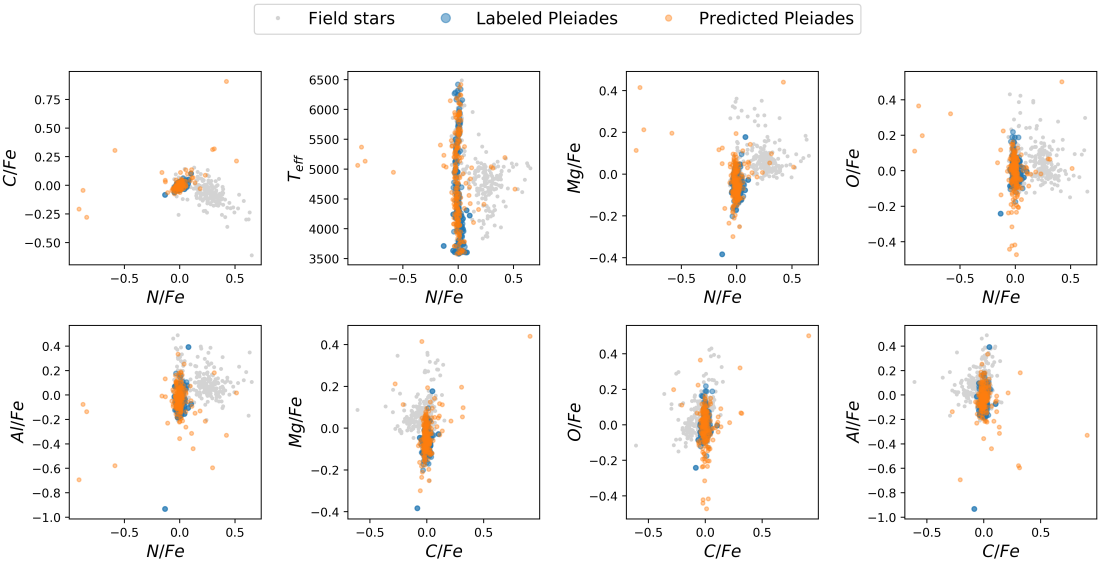

Figure 14. Predicted Pleiades with big distance
Predicted Pleiades with big distance have the same parameters as Pleiades stars from the training data. These stars probably are not Pleiades, but they have a very similar spectrum.

Feature importance for Gaussian Naive Bayes model that is used by the enhanced spectral model can be estimated by the **overlapping area of the feature distributions** for the two classes. A small overlapping area for the given feature suggests a good separation between the classes in values of this feature. Table 10 presents feature parameters for the enhanced spectral model and is sorted in the descending importance

Table 9. Coefficients of Logistic regression model (spectral model)

Spectral model is a Logistic regression model. Features can be ranked by the magnitude of their coefficients. The top-five most important features are $N/Fe, C/Fe, Mg/Fe, O/Fe$ (abundances) and T_{eff} (atmospheric parameter).

Feature	**Coefficient**	**Rank**
N/Fe	-2,2632	1
C/Fe	2,0620	2
T_{eff}	-1,5293	3
Mg/Fe	-1,1685	4
O/Fe	-1,1268	5
Al/Fe	-1,0350	6
Ni/Fe	-0,7943	7
Mn/Fe	-0,6455	8
Fe/H	-0,3431	9
Si/Fe	-0,1348	10
M/H	0,1300	11
α/M	0,0428	12

order.

The importance of the features is different in the spectral and enhanced spectral models. This could be explained by the fact that in the enhanced spectral model positions and velocities are much stronger predictors than spectral features. This could also explain some prediction contamination in the results of the spectral model, which uses spectral features only.

Table 10. Feature importance parameters for features for enhanced spectral model

The enhanced spectral model is a Naive Bayes model. Features can be ranked by their discriminative abilities by the overlapping area of the feature values distribution for two classes. The top-five most important features are X and Z (positions) and Fe/H, M/H and C/Fe (abundances).

Feature	θ_-	σ_-	θ_+	σ_+	**Overlapping area**	**Rank**
X	-0.123418	1.037285	0.809620	0.000003	0.184197	1
Z	-0.119891	1.043771	0.786485	0.000009	0.192609	2
Fe/H	-0.097991	1.070257	0.642823	0.062900	0.298921	3
M/H	-0.095338	1.074760	0.625415	0.058802	0.302915	4
C/Fe	-0.055193	1.120384	0.362066	0.059205	0.411157	5
μ_δ/p	-0.004634	1.152235	0.030401	0.000276	0.488283	6
$\mu_\alpha \cos\delta/p$	0.027708	1.146596	-0.181765	0.000253	0.572855	7
Y	0.110355	1.060370	-0.723926	0.000018	0.784320	8
O/Fe	0.044196	1.089097	-0.289923	0.318652	0.833580	9
Mn/Fe	-0.002003	1.062234	0.013140	0.591547	0.857850	10
N/Fe	0.141170	0.996905	-0.926072	0.031959	0.879549	11
T_{eff}	0.082021	0.820242	-0.538058	1.845577	0.906985	12
α/M	0.115053	1.002518	-0.754745	0.327008	0.957425	13
Mg/Fe	0.152971	0.918927	-1.003490	0.371340	0.994068	14
Si/Fe	0.119035	0.957133	-0.780870	0.578498	0.998310	15
Al/Fe	0.138497	0.910978	-0.908542	0.632703	0.999988	16
Ni/Fe	0.124015	0.905436	-0.813539	0.857605	1.000000	17

5 Discussion

In the course of this thesis, **63 new Pleiades member candidates** have been found. 41 of them were found by the spectral model (the Logistic regression model on the undersampled dataset that uses the Pleiades members listed in Gaia Collaboration et al. 2018a for training labels), 8 were found by the enhanced spectral model (Naive Bayes model on the unbalanced dataset that uses positions and velocities for determining the Pleiades membership for the training data), and 14 stars were found by both models.

These results confirm that spectral data **can be used to relate** to cluster members. However, this data alone might be not enough for a very accurate determination, as there could be other stars with similar chemical composition but located too far away from the cluster. For the most precise (but also more restrictive) membership determination, spectral data should be **complemented** with additional data, e.g. positions and velocities.

The results of the enhanced spectral model - **22** predicted Pleiades member candidates - are presented as the **main result** of this work. 41 stars predicted by the spectral model will need a further more detailed investigation to determine their membership. At this point of time, they have velocities too different from those of the Pleiades, which prevents us to consider them as high-probability member candidates. Yet, they might have been born together in the same molecular cloud with the stars in the cluster, as they all have a similar chemical composition, but have dispersed from the cluster with the passage of time and their velocity vector could have changed because of that.

It is worth highlighting that the predictions of the enhanced spectral model **fully agree** with other studies. The velocities of these predictions agree with the ones reported in Gaia Collaboration et al. 2018a. In our case, all except three predicted member candidates lie within the tidal radius of 11.6 pc, as it was reported in Lodieu et al. 2019. One star is at 13.2 pc from the cluster which is almost at the tidal radius of 13.1 pc, indicated in Adams et al. 2001. Two other stars are at 12.3 and 16.2 pc distance from the cluster, suggesting these stars may well have dispersed from the Pleiades.

As for the spectral model prediction, not all stars are in agreement with the reported Pleiades ranges. Only 20 predicted Pleiades candidate members have all the parameters typical of the Pleiades ranges. These stars can be considered Pleiades member candidates with a reasonably high probability. Other 35 stars have at least one parameter (distance, proper motions or radial velocity) that is unusual. These stars may be either scattered Pleiades stars or just stars with similar chemical composition. Thus, further research is needed to determine the true membership for them.

5.1 Limitations

One of the limitations of this work comes from **crossmatching** with the APOGEE survey. The thing is, it concentrates mostly on red giant stars, so the spectral data is available mostly for the stars of this type. Pleiades is a young cluster though, so it is not supposed

to have many giant stars. Nevertheless, the Pleiades were observed by APOGEE as a calibration cluster, that is why the cluster itself has a spectrum for many stars. However, in the area where one would expect to identify new cluster members, a very poor variety of star types is observed by APOGEE. Only 0.13% of Gaia DR2 stars in the 10° radius were matched with APOGEE.

In addition, it is hard to obtain spectral data for faint stars, so only relatively **bright stars** are included into our research. This limitation may be overcome by using another spectral sky survey, but to the best of the author's knowledge, there are simply no such surveys with open data available at the moment.

Although one of the dataset variations did make use of the **measurement uncertainties** for spectral data, those uncertainties were not directly incorporated into the models. Uncertainities could be a valuable source of information that possibly could allow us to avoid some of the model errors. In a similar fashion, in this work only binary predictions have been generated, without uncertainty estimation. Fortunately, this may be improved by using a different processing and training pipeline, along with the data that allows continuous estimations.

Finally, there is **no direct way to confirm** if thus predicted stars are indeed the Pleiades. Unfortunately, one can only discard obvious inconsistencies, like a too big distance or the wrong star type. However, if, for instance, the distance is in the plausible range from the Pleiades, but a space velocity is a bit different, that could mean both that this is a scattered Pleiades star, or that this is a field star that happen to have the same spectral features as the Pleiades. Indeed, more thorough sky surveys that will generate detailed data and enable more comprehensive analysis are needed to tackle this problem.

5.2 Future work

In the future, as more observations are being collected and **new sky surveys** provide detailed data, it will be possible to overcome some limitations of this work by using other data. Increasing the size of the dataset and improving the accuracy of the measurements within it will definitely help to build more accurate models of the evolution and current state of open clusters.

Also, the same algorithm can be applied to many **other star clusters** to find their members. This will allow us to find specific parameters that define membership for each and every cluster and compare them with one another.

Alternatively, an interesting continuation of the current analysis would be the development of the models that intrinsically **incorporate the measurement uncertainities**. A lot of data can be hidden in big errors or missing values. So much so that it might happen that there is an underlying reason for the incapacity of producing an accurate measurement.

Additionally, in the future studies on this topic one should try to include also **photometric features** into their analysis. Our current study focuses only on the spectral and

astrometrical features. However, it may be beneficial to complement these features with photometric colors and magnitudes. After all, there are dependencies between spectral and photometric features, as proven by Carliles et al. 2010.

Still, Dafonte et al. 2016 present an interesting approach of estimating stellar parameters from spectra by using generative artificial neural networks (GANNs). Therefore, a very interesting study would be to **reverse their analysis** and try to learn a model that predicts spectra from atmospheric parameters. Undoubtedly, this would have numerous applications, one of them being estimating spectra from atmospheric parameters measured by Gaia. This would definitely help to overcome the small data limitation and hopefully make models more accurate.

Last but not least, the results obtained in this thesis will be very useful for large future sky surveys. Having many stars as possible cluster members, our model would help to carefully and precisely **reduce** their number for a detailed membership study.

6 Conclusion

By using spectral data from APOGEE and astrometry data from Gaia sky surveys, **two machine learning models** for determining the Pleiades members have been successfully created. One model (the spectral model) is using exclusively spectral features to make a prediction, while the enhanced spectral model incorporates also positions and velocities. Both models have shown a **good performance** on the test dataset, although the enhanced spectral model is better suited for predicting the Pleiades members. The reason for that is that the spectral model by its nature is able only to find stars with similar spectral features and, as it was shown in Section 4.4, there are lots of stars that are similar to the Pleiades in these terms.

All in all, the findings of this project suggest **new 22 Pleiades cluster members**. Reassuringly, distances, velocities and locations on HRD for member candidates are in good agreement with the previous studies.

The predictions of the models generated in this study suggest that spectral data alone might **not be enough** for accurate predictions with little or no contamination. Still, spectral data is very informative, as it allows us to find stars with similar spectral features, and when used together with star positions and velocities, it provides a more precise cluster membership determination.

Acknowledgement

I would like to thank my supervisors Raul Vicente Zafra and João Alves for their continuous and inspiring support, guidance and their ideas on this project. In addition, I would like to express my gratitude to Stefan Meingast and Sebastian Ratzenböck for their helpful comments and ideas on the beginning of this project, and to Ardi Tampuu for the idea of using multiple-fold training procedure.

This work has made use of data from the European Space Agency (ESA) mission *Gaia* (`https://www.cosmos.esa.int/gaia`), processed by the *Gaia* Data Processing and Analysis Consortium (DPAC, `https://www.cosmos.esa.int/web/gaia/dpac/consortium`). Funding for the DPAC has been provided by national institutions, in particular the institutions participating in the *Gaia* Multilateral Agreement.

Funding for the Sloan Digital Sky Survey IV has been provided by the Alfred P. Sloan Foundation, the U.S. Department of Energy Office of Science, and the Participating Institutions. SDSS acknowledges support and resources from the Center for High-Performance Computing at the University of Utah. The SDSS web site is www.sdss.org.

SDSS is managed by the Astrophysical Research Consortium for the Participating Institutions of the SDSS Collaboration including the Brazilian Participation Group, the Carnegie Institution for Science, Carnegie Mellon University, the Chilean Participation Group, the French Participation Group, Harvard-Smithsonian Center for Astrophysics, Instituto de Astrofísica de Canarias, The Johns Hopkins University, Kavli Institute for the Physics and Mathematics of the Universe (IPMU) / University of Tokyo, the Korean Participation Group, Lawrence Berkeley National Laboratory, Leibniz Institut für Astrophysik Potsdam (AIP), Max-Planck-Institut für Astronomie (MPIA Heidelberg), Max-Planck-Institut für Astrophysik (MPA Garching), Max-Planck-Institut für Extraterrestrische Physik (MPE), National Astronomical Observatories of China, New Mexico State University, New York University, University of Notre Dame, Observatório Nacional / MCTI, The Ohio State University, Pennsylvania State University, Shanghai Astronomical Observatory, United Kingdom Participation Group, Universidad Nacional Autónoma de México, University of Arizona, University of Colorado Boulder, University of Oxford, University of Portsmouth, University of Utah, University of Virginia, University of Washington, University of Wisconsin, Vanderbilt University, and Yale University.

This research made use of Astropy,[4] a community-developed core Python package for Astronomy (Astropy Collaboration et al. 2013; Price-Whelan et al. 2018). This work has made use of Python libraries *scikit-learn* (Pedregosa et al. 2011), *numpy* (Oliphant 2006), *pandas* (McKinney 2010) and *matplotlib* (Hunter 2007), and software *Topcat* (Taylor 2005).

[4] http://www.astropy.org

References

Abolfathi, Bela et al. (May 2018).
"The Fourteenth Data Release of the Sloan Digital Sky Survey: First Spectroscopic Data from the extended Baryon Oscillation Spectroscopic Survey and from the second phase of the Apache Point Observatory Galactic Evolution Experiment".
In: *arXiv:1707.09322 [astro-ph]*. arXiv: 1707.09322.
DOI: 10.3847/1538-4365/aa9e8a. URL: http://arxiv.org/abs/1707.09322.

Adams, Joseph D. et al. (Apr. 2001).
"The Mass and Structure of the Pleiades Star Cluster from 2MASS".
In: *The Astronomical Journal* 121.4. arXiv: astro-ph/0101139, pp. 2053–2064.
ISSN: 00046256. DOI: 10.1086/319965.
URL: http://arxiv.org/abs/astro-ph/0101139.

An, Deokkeun et al. (Jan. 2007). "The Distances to Open Clusters from Main Sequence Fitting. III. Improved Accuracy with Empirically Calibrated Isochrones". en.
In: *The Astrophysical Journal* 655.1, pp. 233–260. ISSN: 0004-637X, 1538-4357.
DOI: 10.1086/509653.
URL: http://stacks.iop.org/0004-637X/655/i=1/a=233.

Astropy Collaboration et al. (Oct. 2013).
"Astropy: A community Python package for astronomy". In: 558, A33, A33.
DOI: 10.1051/0004-6361/201322068. arXiv: 1307.6212 [astro-ph.IM].

Banerji, Manda et al. (July 2010).
"Galaxy Zoo: reproducing galaxy morphologies via machine learning". English.
In: *Monthly Notices of the Royal Astronomical Society* 406.1, pp. 342–353.
ISSN: 00358711. DOI: 10.1111/j.1365-2966.2010.16713.x.
URL: https://academic.oup.com/mnras/article-lookup/doi/10.1111/j.1365-2966.2010.16713.x.

Bazán, Eric, Petr Dokládal, and Eva Dokládalová (2019).
"Quantitative Analysis of Similarity Measures of Distributions". en. In: p. 12.

Blanton, Michael R (2017). "Sloan Digital Sky Survey IV: Mapping the Milky Way, Nearby Galaxies, and the Distant Universe". en. In: *The Astronomical Journal*, p. 35.

Bouy, H. et al. (May 2015). "The Seven Sisters DANCe. I. Empirical isochrones, Luminosity and Mass Functions of the Pleiades cluster".
In: *Astronomy & Astrophysics* 577. arXiv: 1502.03728, A148.
ISSN: 0004-6361, 1432-0746. DOI: 10.1051/0004-6361/201425019.
URL: http://arxiv.org/abs/1502.03728.

Carliles, Samuel et al. (Mar. 2010).
"RANDOM FORESTS FOR PHOTOMETRIC REDSHIFTS". English.
In: *The Astrophysical Journal* 712.1, pp. 511–515. ISSN: 0004-637X, 1538-4357.
DOI: 10.1088/0004-637X/712/1/511. URL: http://stacks.iop.org/0004-637X/712/i=1/a=511?key=crossref.8d4b104039841659fd78077ade0cb1ca.

Chang, Chih-Chung and Chih-Jen Lin (Apr. 2011).
"LIBSVM: A library for support vector machines". en.
In: *ACM Transactions on Intelligent Systems and Technology* 2.3, pp. 1–27.
ISSN: 21576904. DOI: 10.1145/1961189.1961199.
URL: http://dl.acm.org/citation.cfm?doid=1961189.1961199 (visited on 12/25/2019).

Converse, Joseph M. and Steven W. Stahler (Apr. 2010).
"The Dynamical Evolution of the Pleiades".
In: *Monthly Notices of the Royal Astronomical Society*. arXiv: 1002.2229.
ISSN: 00358711, 13652966. DOI: 10.1111/j.1365-2966.2010.16505.x.
URL: http://arxiv.org/abs/1002.2229.

Dafonte, C. et al. (Oct. 2016).
"On the estimation of stellar parameters with uncertainty prediction from Generative Artificial Neural Networks: application to *Gaia* RVS simulated spectra". en.
In: *Astronomy & Astrophysics* 594, A68. ISSN: 0004-6361, 1432-0746.
DOI: 10.1051/0004-6361/201527045.
URL: http://www.aanda.org/10.1051/0004-6361/201527045.

David, Trevor J. et al. (Apr. 2016). "NEW PLEIADES ECLIPSING BINARIES AND A HYADES TRANSITING SYSTEM IDENTIFIED BY *K2*". en.
In: *The Astronomical Journal* 151.5, p. 112. ISSN: 1538-3881.
DOI: 10.3847/0004-6256/151/5/112. URL: http://stacks.iop.org/1538-3881/151/i=5/a=112?key=crossref.9e6f1f66e5952fe959b8e6fd1626712b.

Eisenstein, Daniel J. et al. (Sept. 2011). "SDSS-III: MASSIVE SPECTROSCOPIC SURVEYS OF THE DISTANT UNIVERSE, THE MILKY WAY, AND EXTRA-SOLAR PLANETARY SYSTEMS". en.
In: *The Astronomical Journal* 142.3, p. 72. ISSN: 0004-6256, 1538-3881.
DOI: 10.1088/0004-6256/142/3/72. URL: http://stacks.iop.org/1538-3881/142/i=3/a=72?key=crossref.c4bf9c63075c50b5eb07ebbbf2452a05.

Gaia Collaboration et al. (Nov. 2016). "The *Gaia* mission". en.
In: *Astronomy & Astrophysics* 595, A1. ISSN: 0004-6361, 1432-0746.
DOI: 10.1051/0004-6361/201629272.
URL: http://www.aanda.org/10.1051/0004-6361/201629272.

Gaia Collaboration et al. (Aug. 2018a).
"Gaia Data Release 2: Observational Hertzsprung-Russell diagrams".
In: *Astronomy & Astrophysics* 616. arXiv: 1804.09378, A10.
ISSN: 0004-6361, 1432-0746. DOI: 10.1051/0004-6361/201832843.
URL: http://arxiv.org/abs/1804.09378.

Gaia Collaboration et al. (Aug. 2018b).
"Gaia Data Release 2. Summary of the contents and survey properties".

In: *Astronomy and Astrophysics* 616, A1, A1.
DOI: 10.1051/0004-6361/201833051. arXiv: 1804.09365.
Galli, P. A. B. et al. (Feb. 2017).
"A revised moving cluster distance to the Pleiades open cluster". en.
In: *Astronomy & Astrophysics* 598, A48. ISSN: 0004-6361, 1432-0746.
DOI: 10.1051/0004-6361/201629239.
URL: http://www.aanda.org/10.1051/0004-6361/201629239.
Gatewood, George, Joost Kiewiet de Jonge, and Inwoo Han (Apr. 2000).
"The Pleiades, Map-based Trigonometric Parallaxes of Open Clusters. V." en.
In: *The Astrophysical Journal* 533.2, pp. 938–943. ISSN: 0004-637X, 1538-4357.
DOI: 10.1086/308679.
URL: http://stacks.iop.org/0004-637X/533/i=2/a=938.
Giannuzzi, M. A. (1995).
"The spectroscopic binary HD 23642 and the distance of the Pleiades."
In: *Astronomy & Astrophysics* 293, pp. 360–362. URL:
http://articles.adsabs.harvard.edu/full/1995A%5C%26A...293..360G.
Goddi, Ciriaco et al. (2019).
"First M87 Event Horizon Telescope Results and the Role of ALMA". en.
In: cites: goddiFirstM87Event2019, p. 12.
Hunter, J. D. (2007). "Matplotlib: A 2D graphics environment".
In: *Computing in Science & Engineering* 9.3, pp. 90–95.
DOI: 10.1109/MCSE.2007.55.
Ivezić, Željko et al. (Mar. 2019).
"LSST: from Science Drivers to Reference Design and Anticipated Data Products".
In: *The Astrophysical Journal* 873.2, p. 111. ISSN: 1538-4357.
DOI: 10.3847/1538-4357/ab042c. URL: http://arxiv.org/abs/0805.2366.
Karttunen, Hannu et al. (2017). *Fundamental Astronomy*. 6th ed. Springer.
Kos, Janez et al. (Feb. 2018). "The GALAH survey: Chemical Tagging of Star Clusters and New Members in the Pleiades". In: *Monthly Notices of the Royal Astronomical Society* 473.4. arXiv: 1709.00794, pp. 4612–4633. ISSN: 0035-8711, 1365-2966.
DOI: 10.1093/mnras/stx2637. URL: http://arxiv.org/abs/1709.00794.
Leeuwen, Floor van (Apr. 2009). "Parallaxes and proper motions for 20 open clusters as based on the new Hipparcos catalogue".
In: *Astronomy & Astrophysics* 497.1. arXiv: 0902.1039, pp. 209–242.
ISSN: 0004-6361, 1432-0746. DOI: 10.1051/0004-6361/200811382.
URL: http://arxiv.org/abs/0902.1039.
Lindegren, L. et al. (Aug. 2018). "*Gaia* Data Release 2: The astrometric solution". en.
In: *Astronomy & Astrophysics* 616, A2. ISSN: 0004-6361, 1432-0746.
DOI: 10.1051/0004-6361/201832727.
URL: https://www.aanda.org/10.1051/0004-6361/201832727.

Lodieu, N. et al. (June 2019).
"A 5D view of the Alpha Per, Pleiades, and Praesepe clusters".
In: *arXiv:1906.03924 [astro-ph]*. arXiv: 1906.03924.
URL: http://arxiv.org/abs/1906.03924.
Majewski, S. R. et al. (Sept. 2017).
"The Apache Point Observatory Galactic Evolution Experiment (APOGEE)".
In: *Astronomical Journal* 154, 94, p. 94. DOI: 10.3847/1538-3881/aa784d.
arXiv: 1509.05420 [astro-ph.IM].
Majewski, Steven R. et al. (Aug. 2017).
"The Apache Point Observatory Galactic Evolution Experiment (APOGEE)". en.
In: *The Astronomical Journal* 154.3, p. 94. ISSN: 1538-3881.
DOI: 10.3847/1538-3881/aa784d. URL: http://stacks.iop.org/1538-3881/154/i=3/a=94?key=crossref.7c7b5a12ff580bd0c9b363a4ff2955b2.
McKinney, Wes (2010). "Data Structures for Statistical Computing in Python".
In: *Proceedings of the 9th Python in Science Conference*.
Ed. by Stéfan van der Walt and Jarrod Millman, pp. 51–56.
Melis, C. et al. (Aug. 2014).
"A VLBI resolution of the Pleiades distance controversy". en.
In: *Science* 345.6200, pp. 1029–1032. ISSN: 0036-8075, 1095-9203.
DOI: 10.1126/science.1256101.
URL: http://www.sciencemag.org/cgi/doi/10.1126/science.1256101.
Oliphant, Travis E (2006). *A guide to NumPy*. Vol. 1. Trelgol Publishing USA.
Pedregosa, F. et al. (2011). "Scikit-learn: Machine Learning in Python".
In: *Journal of Machine Learning Research* 12, pp. 2825–2830.
Powers, David (2007). "Evaluation: From Precision, Recall and F-Factor to ROC, Informedness, Markedness & Correlation". en. In: p. 24.
Prati, Ronaldo C (2009).
"Data mining with imbalanced class distributions: concepts and methods". en.
In: p. 17.
Price-Whelan, A. M. et al. (Sept. 2018). "The Astropy Project: Building an Open-science Project and Status of the v2.0 Core Package". In: 156, 123, p. 123.
DOI: 10.3847/1538-3881/aabc4f.
Sarro, L. M. et al. (Mar. 2014).
"Cluster membership probabilities from proper motions and multi-wavelength photometric catalogues: I. Method and application to the Pleiades cluster". en.
In: *Astronomy & Astrophysics* 563, A45. ISSN: 0004-6361, 1432-0746.
DOI: 10.1051/0004-6361/201322413.
URL: http://www.aanda.org/10.1051/0004-6361/201322413.
Soderblom, David R. et al. (Mar. 2005). "Confirmation of Errors in *Hipparcos* Parallaxes from *Hubble Space Telescope* Fine Guidance Sensor Astrometry of the Pleiades". en.

In: *The Astronomical Journal* 129.3, pp. 1616–1624. ISSN: 0004-6256, 1538-3881. DOI: 10.1086/427860.
URL: http://stacks.iop.org/1538-3881/129/i=3/a=1616.

Taylor, Mark B (2005).
"TOPCAT & STIL: Starlink Table/VOTable Processing Software". en. In: p. 5.

Zasowski, G. et al. (Aug. 2013). "TARGET SELECTION FOR THE APACHE POINT OBSERVATORY GALACTIC EVOLUTION EXPERIMENT (APOGEE)". en.
In: *The Astronomical Journal* 146.4, p. 81. ISSN: 0004-6256, 1538-3881.
DOI: 10.1088/0004-6256/146/4/81. URL: http://stacks.iop.org/1538-3881/146/i=4/a=81?key=crossref.b7c8bf0fc68a85c8c463ac2bf0cac9c3.

Zasowski, G. et al. (Oct. 2017).
"Target Selection for the SDSS-IV APOGEE-2 Survey". en.
In: *The Astronomical Journal* 154.5, p. 198. ISSN: 1538-3881.
DOI: 10.3847/1538-3881/aa8df9.
URL: https://doi.org/10.3847%2F1538-3881%2Faa8df9.

Appendix

I. Additional materials

List of predicted Pleiades stars

Table 11. The list of all predicted Pleiades for both models

source_id	α	δ	Distance	$\mu_\alpha \cos\delta$	μ_δ	Radial velocity
61004001783775360[s]	51.776	20.9	116.656	51.374	16.211	
52849233277199872[s]	63.021	22.457	128.131	9.803	0.354	
66830485697459456[s]	57.037	25.029	131.52	13.413	-41.315	
52850573306994816[s]	62.939	22.514	136.515	-12.403	-9.649	-10.36
66519530066802944[s]	57.206	23.977	152.488	22.356	-37.832	
66531418536135552[s]	57.215	24.205	116.309	20.017	-27.829	
53758254515062784[s]	62.516	23.812	142.896	18.061	-42.584	
216814488885922944[s]	57.318	32.034	143.999	-12.828	-11.268	-4.48
66816187748666624[s]	56.999	24.731	154.005	22.012	-52.315	
66161604670993024[s]	59.875	24.336	126.587	18.798	-32.519	
67135462735510400[s]	59.786	25.459	125.439	-23.99	-15.157	
67318359623488512[s]	58.898	25.897	124.833	46.858	-21.597	
66690710282089856[s]	58.562	25.05	135.232	59.252	-13.116	25.7
66458610250837632[s]	57.726	23.835	145.421	-10.583	-35.519	6.33
65660158649542784[s]	57.909	23.183	126.188	24.809	-59.194	
66692703146816640[s]	58.256	24.87	132.352	1.36	-34.522	
66475068565410432[s]	58.34	23.981	150.684	53.868	-60.434	
64944170420647296[s]	57.396	23.454	150.15	21.968	-48.8	
66823304512166528[s]	56.909	24.892	139.935	21.131	-45.273	
52804531257705088[s]	63.224	22.444	139.421	0.591	-25.669	17.81
149771251985135744[s]	63.317	23.686	129.267	-6.781	-4.372	
57856615388968064[s]	52.106	19.751	129.767	-7.47	-30.582	
217409771349935872[s]	54.728	32.248	135.428	18.918	-23.136	14.37
216559264746726144[s]	55.188	31.663	153.629	1.629	-26.923	-6.63
216589909337574016[s]	55.589	31.923	140.623	14.663	-17.055	-25.82
217932928431926912[s]	55.925	33.259	150.15	-14.008	-29.64	
47964481071082880[s]	66.039	19.096	141.735	-21.27	-16.604	5.09
216309911829610752[s]	56.23	31.224	142.878	2.506	-27.353	23.68
65272821318002560[s]	56.098	24.132	138.733	28.589	-44.954	
48153803229729536[s]	65.332	19.188	148.878	14.392	-30.114	

Table 11 continued from previous page

source_id	α	δ	Distance	$\mu_\alpha \cos\delta$	μ_δ	Radial velocity
47751794291088384[s]	65.044	18.691	149.703	7.711	-26.534	
149943428633879296[s]	64.084	24.614	142.816	17.322	-45.279	
217107607515576704[s]	56.533	32.764	124.247	-22.792	-11.493	22.5
149967033773927936[s]	63.917	24.601	128.591	-25.422	5.624	4.87
217114552480179200[s]	56.648	32.864	122.287	-28.723	-10.428	-7.79
66802276352348416[s]	56.3	24.586	132.031	20.885	-44.704	
67084301084506752[s]	58.454	25.276	152.186	-11.285	-28.196	-3.87
66944422591073024[s,es]	57.032	25.315	131.444	21.737	-49.197	
65289588870240384[s,es]	56.041	24.268	135.75	21.575	-43.411	
66576223634093184[s,es]	58.433	24.224	148.258	4.9	-51.621	-22.22
65276699671440384[s,es]	56.435	24.22	134.561	21.0	-47.315	
66584332530000512[s,es]	58.537	24.333	129.958	17.346	-46.305	6.83
66720946851771904[s,es]	57.204	24.267	143.536	17.479	-47.325	
64921458633614976[s,es]	57.244	23.201	142.882	22.288	-41.076	5.31
66937859881182848[s,es]	57.503	25.399	152.386	17.947	-43.975	4.53
64094037476647040[s,es]	57.704	22.67	134.472	17.379	-44.616	
69876506565909632[s,es]	55.861	24.994	129.82	19.169	-44.696	6.44
65289279632597760[s,es]	55.961	24.247	131.172	21.711	-44.674	5.46
69873418486230272[s,es]	56.516	25.453	137.163	22.55	-44.945	
65224442806459008[s,es]	56.554	24.054	136.831	22.07	-46.185	
66802654309459712[s,es]	56.417	24.627	132.12	18.187	-43.817	6.68
66734720809017856[s,es]	56.825	24.391	143.082	21.401	-45.705	
66502281478384000[s,es]	57.35	23.839	135.566	24.706	-45.966	5.33
66452734735432192[s,es]	57.493	23.709	136.008	17.451	-44.911	
66837495084120320[s,es]	56.668	24.931	136.273	23.084	-44.67	
68334235349446528[es]	55.128	24.487	138.112	22.009	-43.785	2.95
66814302260735104[es]	56.533	24.867	139.888	22.092	-46.141	
64971177174850304[es]	56.542	23.34	122.974	21.166	-49.898	6.04
64980278208557696[es]	56.668	23.498	138.487	16.38	-47.044	
64804841680675200[es]	56.722	22.881	148.542	19.054	-41.517	6.57
69954713627862528[es]	56.897	25.544	147.226	19.035	-42.75	
64924413571101952[es]	57.485	23.218	147.083	19.087	-43.22	6.97
66878245735176704[es]	57.625	25.052	145.062	18.283	-44.219	

[s] Spectral model returns a positive prediction for this star

[es] Enhanced spectral model returns a positive prediction for this star

Source code

The source code produced for this research is located under `https://github.com/AlinaVorontseva/Pleiades_ml`. Access to the repository could be granted upon sending an email to alina.vorontseva@gmail.com.

II. Licence

Non-exclusive licence to reproduce thesis and make thesis public

I, **Alina Vorontseva**,
(author's name)

1. herewith grant the University of Tartu a free permit (non-exclusive licence) to

   reproduce, for the purpose of preservation, including for adding to the DSpace digital archives until the expiry of the term of copyright,

   Using Machine Learning to Find New Members of the Pleiades,
   (title of thesis)

   supervised by Raul Vicente Zafra and João Alves,
   (supervisor's name)

2. I grant the University of Tartu a permit to make the work specified in p. 1 available to the public via the web environment of the University of Tartu, including via the DSpace digital archives, under the Creative Commons licence CC BY NC ND 3.0, which allows, by giving appropriate credit to the author, to reproduce, distribute the work and communicate it to the public, and prohibits the creation of derivative works and any commercial use of the work until the expiry of the term of copyright.

3. I am aware of the fact that the author retains the rights specified in p. 1 and 2.

4. I certify that granting the non-exclusive licence does not infringe other persons' intellectual property rights or rights arising from the personal data protection legislation.

Alina Vorontseva
09/01/2020